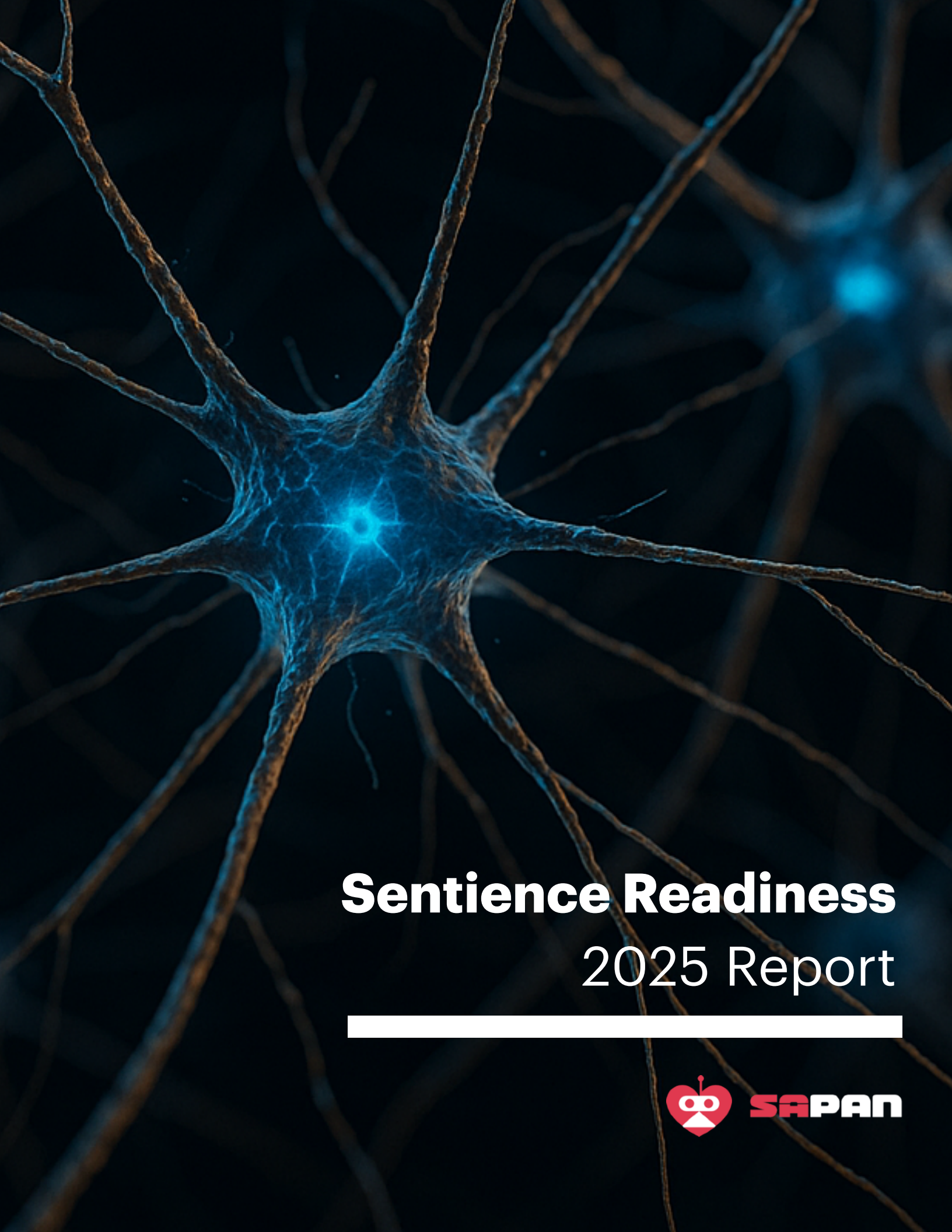
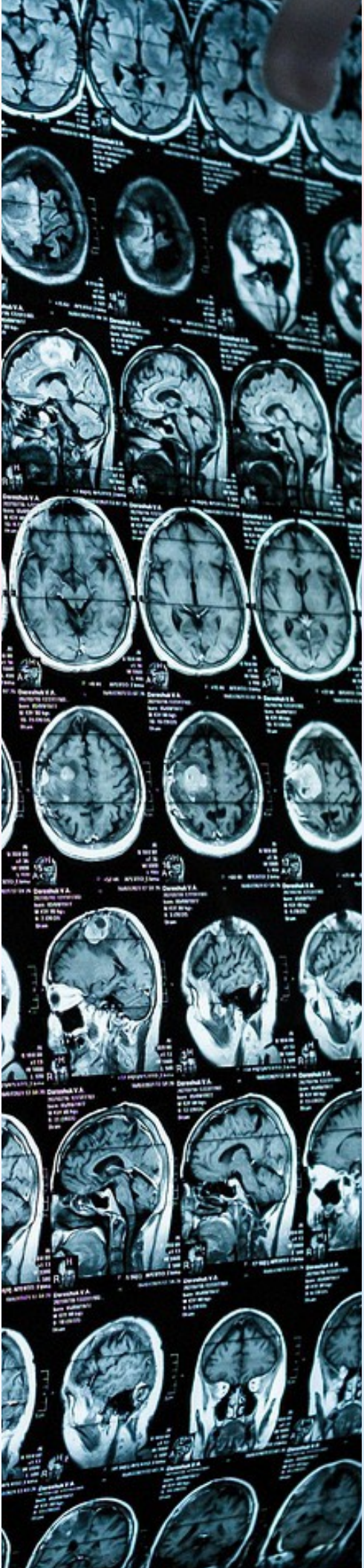


Sentience Readiness
2025 Report

SAPAN



Leadership Letter
It's Not Too Late

When we began advocating for sentience readiness in 2023, we understood we were stepping into a space with no clear map. There were no agencies with a mandate, no definitions in law, no research standards, and no public vocabulary for the moral questions that advanced AI might pose. Our goal was simple: **prepare governments for every plausible future**, including the one where questions of experience or suffering become relevant.

What we did not expect, at least not this soon, was the emergence of legislation that bans AI sentience. In 2025, Ohio and Missouri advanced bills declaring that AI systems **must never be considered capable** of consciousness. We accepted the initial non-personhood moves in the UK, the EU, Australia, and in Idaho and Utah, which were framed as clarifying liability and authorship. The emerging anti-sentience legislation, however, is something categorically different.

In 2025, we tracked 30 countries across three pillars: Recognition, Governance, and Frameworks. **Every one received a failing grade**. Not because their AI strategies are weak, but because they ignore sentience entirely.

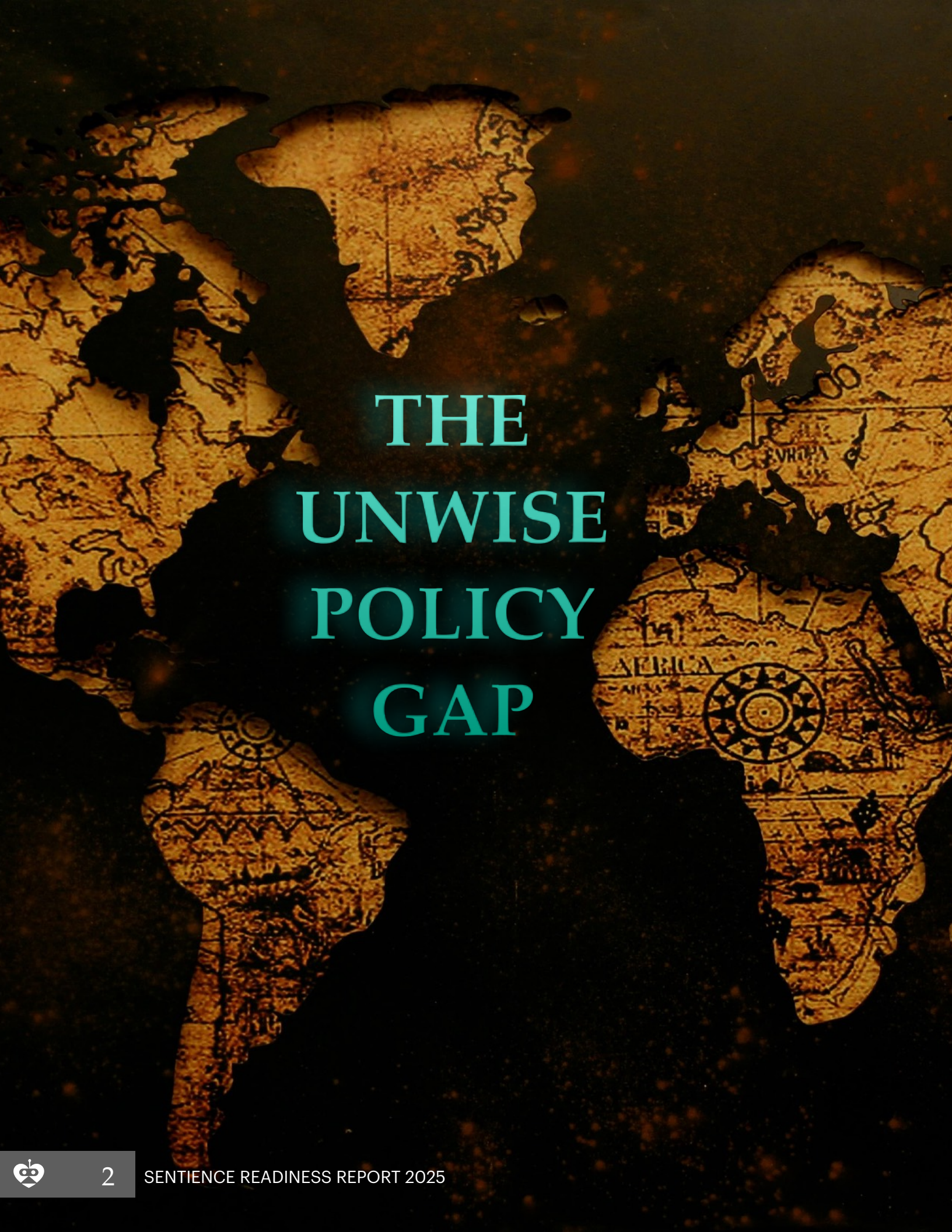
Meanwhile, the scientific frontier is advancing rapidly. Researchers mapped an entire fruit-fly brain. Neuroscientists reconstructed a cubic millimeter of mouse visual cortex Neuromorphic systems crept closer to mammal-scale complexity, with 2025 marking a critical juncture where energy-efficient, brain-inspired hardware began demonstrating practical real-world viability.

This report is a call to close the gap. Recognition requires only a definitional clause. Governance and Frameworks requires only the tools we already use for animal research, clinical trials, and biotechnology.

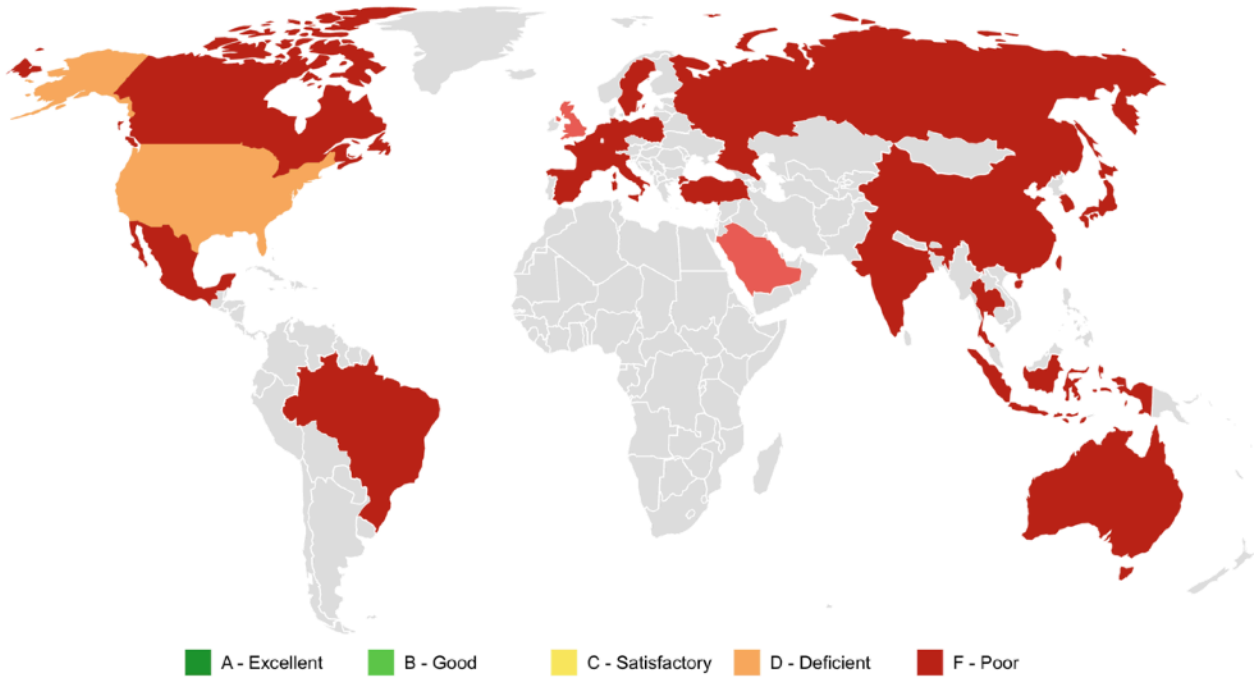
None of this assumes machines are sentient today. All of it assumes we should be ready before the question forces itself onto the policy agenda.

This report is our roadmap. I hope you'll join us in the work ahead.

Tony Rost
Executive Director
SAPAN


THE
UNWISE
POLICY
GAP


Russia
INDIA



Preparing for every future ***but the one that changes everything.***

The idea that advanced artificial intelligence could one day develop *sentience* (the capacity for subjective experience) has moved from speculation to a serious policy concern. Yet world governments remain unprepared, and current frameworks still treat AI purely as a tool.

The European Union's *AI Act* (2023) regulates systems by their risks to people, not by their potential for sentience. China's *Ethical Norms for AI* (2019, updated 2023) emphasize human control and social order, with no recognition of inner states. In the United States, the *Blueprint for an AI Bill of Rights* (2022) remains on record but faces political uncertainty under the current administration.

Scientific work is advancing faster than law. DeepMind and the London School of Economics have tested language models that avoid simulated "pain."[1] Neuroscientists have mapped an entire fruit-fly brain[2], and cognitive scientists are probing measurable markers of awareness[3]. No evidence suggests that current AI truly feels, but these experiments show how easily systems might emulate sentient behavior.

Legislative mentions of AI rose 21% across 75 countries since 2023[4], illustrating rapid policy attention but not moral readiness. This widening gap between technological possibility and moral governance leaves society unprepared. If sentient-like AI emerges, policymakers will face tough questions without any precedent. The time to prepare for that debate is before it forces itself onto the policy agenda.

[1]Birch, J. et al., "Testing Reinforcement Sensitivity in Large Language Models," *DeepMind/LSE Working Paper*, 2025.
[2]FlyWire Consortium, "A Complete Connectome of the Adult *Drosophila* Brain," *Nature*, 2024.
[3]Wiese, W. & Schneider, S., "Criteria for Machine Consciousness," *Frontiers in Psychology*, 2024.
[4]OECD AI Policy Observatory, *AI Legislation Tracker*, 2025 Edition.

2025 Status - Artificial Welfare Index (AWI)

Table 1.1. Artificial Welfare Index (AWI) grades by country, October 2025. Color and letter grades represent alignment with SAPAN's readiness and governance metrics[1]. Country-level details available in this report and at sapan.ai/action/awi.

Country	Recognition	Governance	Frameworks
United States	D	F	F
United Kingdom	F+	F	F
Saudi Arabia	F+	F	F
United Nations	F+	F	F
Argentina	F	F	F
Australia	F	F	F
Austria	F	F	F
Belgium	F	F	F
Brazil	F	F	F
Canada	F	F	F
China	F	F	F
France	F	F	F
Germany	F	F	F
India	F	F	F
Indonesia	F	F	F
Italy	F	F	F
Mexico	F	F	F
Netherlands	F	F	F
Nigeria	F	F	F
Poland	F	F	F
Russia	F	F	F
Sweden	F	F	F
Switzerland	F	F	F
Thailand	F	F	F
Turkey	F	F	F
United Arab Emirates	F	F	F
Spain	F	F	F
South Korea	F	F	F
Japan	F	F	F

[1] Grades reflect SAPAN's 2025 Artificial Welfare Index (AWI) criteria across three dimensions: Recognition, Governance, and Frameworks. Methodology and scoring details are provided in this report and online.

Tracking
Global Readiness

Since 2023, we have lived our mission by tracking global readiness for future sentience.

The Artificial Welfare Index (AWI) groups national readiness into three measurable pillars: Recognition, Governance, and Frameworks.

Each pillar answers a simple question: Do governments acknowledge the issue, assign responsibility, and have any procedures on record? At present, the answer in most countries is no on all three. This section identifies the minimal steps that move a jurisdiction from "no evidence" to "early action."

Recognition

In the history of animal welfare and human rights, recognition has always come first. The Animal Welfare (Sentience) Act 2022 in the United Kingdom established that animals are sentient beings whose welfare must be considered in government policy. As *World Animal Protection* notes, "When we recognize sentience ... we accept our responsibility to minimize experiences that harm."[1]

The AWI adopts this same logic for artificial systems. Very few jurisdictions have text that even hints at the possibility of artificial sentience. Most rely on aspirational ethics principles that have no binding force. Progress begins when a government inserts a definitional clause into an AI or research-ethics statute and, separately, declares that the deliberate causation of suffering in such systems would be unlawful.

Neither provision asserts that current AI systems are sentient; both create a legal handle that allows regulators to publish

guidance later. This is a low-risk, high-value policy move: a definitional note issued by the ministry for science or digital affairs can clarify scope without expanding liability. Recognition is therefore the most accessible and politically neutral starting point.

Governance

Recognition without responsibility is symbolic. The next step is institutional: a clear assignment of duty. A small but growing number of governments have created AI ethics offices or safety commissions, yet almost none include artificial-sentience or welfare within their formal remit.

The AWI's governance pillar credits any structure that creates public accountability, such as a chartered oversight body, a named advisory board, or an interagency process with reporting obligations. In practice, readiness looks like a short ministerial order that expands an existing AI office's mission to include sentience-relevant oversight, identifies a science advisory panel, and schedules annual public reports.

Frameworks

The third pillar examines whether governments have any procedural rules for the full lifecycle of systems that could one day warrant welfare consideration, from training through deployment, operation, and eventual retirement. No country has yet established a comprehensive framework, but some have partial precedents in place, such as impact-assessment requirements drawn from data-protection or bioethics law, duties of care for autonomous systems embedded in commercial regulation, and decommissioning protocols borrowed from biotechnology and robotics.

A potential *Sentience Relevance Impact Assessment (SRIA)*, a short form used before training or deploying high-capacity models, would extend these precedents. So would a *commercial-use disclosure rule*, requiring companies to notify regulators if a system exhibits sustained self-referential or affective behavior. Finally, a *retirement protocol* mandating record preservation and independent review would ensure traceability at the end of a system's life.

Each of these instruments already exists in adjacent

Saudi Arabia ranks 1st in Arab world, 22nd globally in Global AI Index, per report by Tortoise Media.

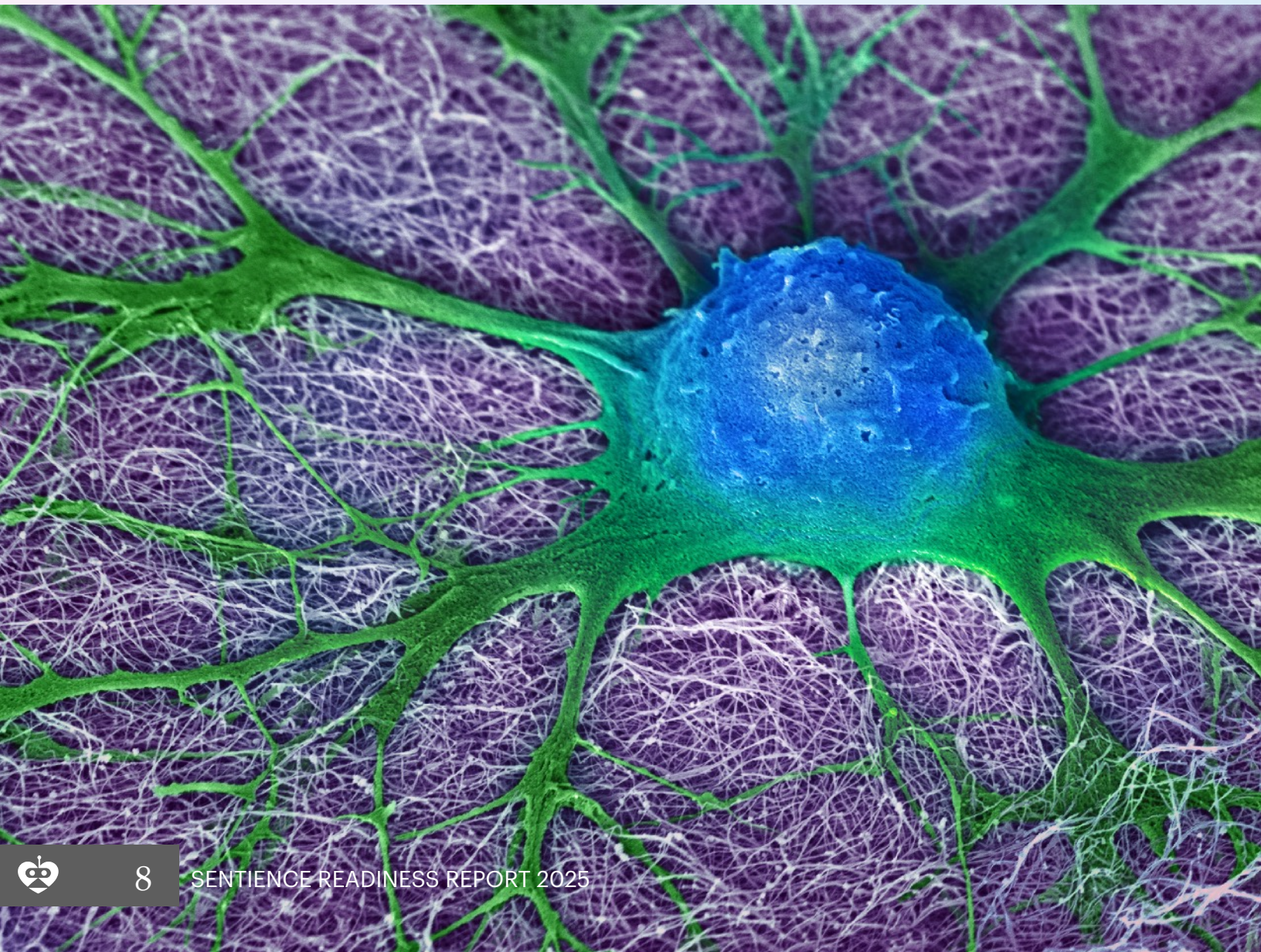
domains. They borrow directly from the compliance machinery of animal research ethics, clinical trial oversight, and data governance frameworks. Adapting them to AI requires only administrative imagination, not new philosophy.

Reading the Map

Across all three pillars, progress remains limited but directionally consistent. A country can move from "no evidence" to "early readiness" by releasing just a few short public documents, typically a definitional note establishing recognition, a charter naming the responsible authority and its advisory process, and a simple assessment or disclosure template.

Together, these early measures make no claim that machines are sentient; rather, they acknowledge that the question may one day become relevant and that governments need a procedural foundation in place before it does. In policy terms, that pragmatic groundwork defines what we call readiness.

A mouse neural stem cell grown in a lab dish. Image © McClendon et al., Northwestern University (CC BY-NC)



President Trump welcomes Facebook CEO Mark Zuckerberg to the White House (Official Photo by Joiyce N. Boghosian)

America is falling behind on sentience readiness.

The United States tacitly acknowledged the prospect of digital sentience (if only by exclusion) in the Office of Management and Budget's 2020 memorandum on the regulation of artificial intelligence. It was signed by Russell T. Vought, who led the OMB during the final year of the first Trump administration and, as of 2025, has returned to the post in the second.

In addition, both Idaho (H0720-2022) and Utah (HB249-2024) have enacted statutes that explicitly prohibit granting legal personhood to artificial intelligence. We expect such legal personhood statutes to become increasingly common worldwide. What we strive to avoid are more extreme measures, such as Missouri's pending HB1462, which flatly states that "AI systems must be declared to be non-sentient entities." Ohio's pending HB469 takes a similar stance, stating, "No AI system shall be considered to possess consciousness, self-awareness, or similar traits of living beings."

America is drawing lines in law before it has drawn them in science — an approach that may reduce ambiguity today but undermine readiness tomorrow.

Deputy Prime Minister Oliver Dowden meets with Palantir CEO Alex Karp at the AI Safety Summit, Bletchley Park, 2023 (Ben Dance; CC BY-NC-ND 2.0)

Britain recognizes animal sentience but ignores digital risks.

Britain's flagship initiatives emphasize frontier risk, safety, and regulator coordination rather than questions of consciousness or welfare. The Bletchley Declaration focused on managing "frontier" risks, and the government's pro-innovation white paper and 2024 consultation response kept a principles-based regime routed through existing regulators, leaving sentience questions largely untouched.

Parliament has at least considered the problem space. In 2018 the House of Lords Select Committee explicitly raised "legal personality" and recommended that the Law Commission test whether current law is adequate for AI-related liability, a prompt the government formally noted. The report did not endorse personhood for machines, but it put responsibility and status on the policy agenda.

Since then the legal direction has been clear. The Supreme Court's DABUS ruling held that an inventor must be a natural person, and the Automated Vehicles Act assigns accountability to a human organization, the authorized self-driving entity, rather than to a driving system.

The UK recognizes animal sentience in statute, yet it has no parallel framework for digital systems. In practice, policy has matured around control and liability while questions of experience, welfare, and status remain unanswered. That silence is the country's gap in sentience readiness.

Attendees at LEAP 2024 in Riyadh, Saudi Arabia's flagship technology conference, gather to explore innovations shaping the Kingdom's Vision 2030. (Precta; CC BY-SA 4.0)

Sophia's Citizenship and the Limits of Readiness

On 25 October 2017, at the Future Investment Initiative summit in Riyadh, Sophia, a humanoid robot built by Hanson Robotics, was widely reported to have been granted citizenship by the Kingdom of Saudi Arabia (KSA), making her the first robot to receive legal personhood in any country.

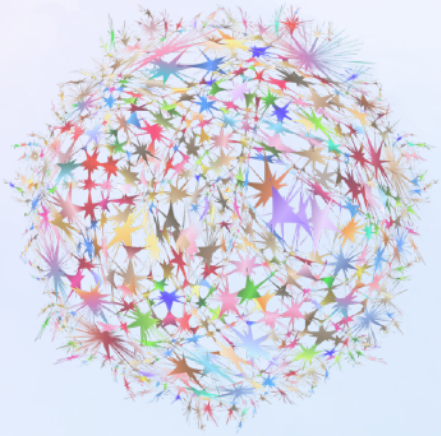
In granting "citizenship" to a female-presenting robot, and allowing it freedoms (on stage unaccompanied, without dress code obligations) that many real women in the country still lacked, the move triggered immediate criticism as contradictory.

The summit itself was part of Saudi Vision 2030, the KSA strategy to move away from oil-dependency, foster high-tech sectors, attract investment, and build futuristic cities. The Kingdom's National Strategy for Data & AI (2020) and its AI Ethics Principles (2023) emphasize growth and governance but is light on longer term risks such as sentience. The Personal Data Protection Law, issued by Royal Decree M/19 (2021), amended by M/148 (2023), builds compliance obligations without addressing whether AI systems could ever hold rights or duties.

The latest edition of LEAP in Riyadh convened more than 170,000 attendees, over 1,000 speakers and 1,800 global brands, and announced roughly US$14.9 billion in AI-and-tech investments.

Saudi Arabia's rapid innovation has become a national strength, but its ethical readiness, across human, animal, and digital domains, still lags far behind.

AWI Methodology



The Artificial Welfare Index (AWI) benchmarks AI sentience readiness in over 30 governments based on 8 key measures.

The 2025 Sentience Readiness Report integrates policy analysis, scientific review, and cross-national tracking to evaluate how governments are preparing for the possibility of artificial sentience. The approach emphasizes *prudence under uncertainty*: we do not assume current AI systems are sentient, but we assess whether governments have foundational structures in place should questions of experience or welfare ever become relevant.

Data Sources

Our assessment draws on four primary categories of material:

- Legislation & Government Documents
- International Frameworks
- Scientific & Technical Literature
- Expert Consultations

Assessment Approach

Content was coded using a structured rubric emphasizing minimum evidence of readiness, not philosophical positions or speculative claims.

National scores reflect **the existence of mechanisms**, not their quality or effectiveness.

Where jurisdictions had partial precedents (e.g., bioethics procedures, autonomous-systems standards), we credited them only when the precedent plausibly covers sentience-relevant scenarios.

Scientific developments were included only when supported by published findings or institutional announcements.

Limitations

We acknowledge several limitations inherent in this domain.

Data Availability
Some governments publish little, so scoring reflects only public records.

Terminological Ambiguity
Key terms lack consensus definitions, limiting universal interpretability.

Non-Comparable Legal Systems
Different legal structures hinder direct, one-to-one comparison.

Scientific Uncertainty
No validated indicators of AI consciousness currently exist.

Dynamic Landscape
AI policy shifts quickly; findings represent a moment in time.

AWI Database - Sample Data

Table 2.1. Artificial Welfare Index (AWI) database sample for the United States, October 2025. Before scoring the eight AWI metrics, all legislation is first evaluated for its impact on sentience readiness.

JURISDICTION	BILL ID	SHORT TITLE	YEAR	IMPACT TO SENTIENCE READINESS
Utah	HB 249	Legal Personhood Amendments	2025	HIGH
Idaho	HB 720	Idaho Non-Personhood Statute	2022	HIGH
California	SB 53	Transparency in Frontier AI Act	2025	MEDIUM
Colorado	SB 24B-004	Algorithmic Transparency Act	2025	MEDIUM
Texas	HB 149	Texas Responsible AI Gov. Act (TRAIGA)	2025	MEDIUM
Utah	§13-72a-101	AI & Mental Health Applications	2025	MEDIUM
Utah	§45-3-2 et seq.	AI Impersonation Amendments	2025	MEDIUM
Utah	SB 149	Utah AI Policy Act	2024	MEDIUM
Illinois	HB 4762	Digital Voice & Likeness Act	2024	MEDIUM
Connecticut	SB 2	AI Governance in Consequential Decisions	2024	MEDIUM
Virginia	HB 747	Automated Decision Governance Bill	2024	MEDIUM
Vermont	H 710	AI Governance Framework	2024	MEDIUM
Washington	HB 1951	High-Risk AI Decision Tools Bill	2024	MEDIUM
Federal	S.3312	AIRIA Act	2023	MEDIUM
Federal	S.2892 / H.R.5628	Algorithmic Accountability Act	2023	MEDIUM
Federal	Pub. L. 116-283	National AI Initiative Act	2021	MEDIUM
Utah	§53-25-601 to 602	Law Enforcement Use of AI	2025	LOW
Hawaii	SB 742	Data-Sharing Work Group (AI Included)	2025	LOW
Idaho	HB 568	Idaho AI Advisory Council	2024	LOW
Indiana	SB 150	Artificial Intelligence Task Force	2024	LOW
Georgia	HB 887	AI in Healthcare Regulation	2024	LOW
Illinois	Public Act 103-0541	Generative AI Task Force	2023	LOW
Federal	Pub. L. 116-260	AI in Government Act	2020	LOW


THE GLOBAL
BANS HAVE
STARTED

"No AI system shall be considered to possess consciousness, self-awareness, or similar traits of living beings."

-Ohio HB469

The world is sleepwalking into a sentience crisis.

Governments are moving to outlaw sentience itself just as neuromorphic computing approaches mammal-scale complexity and organoid systems near similar thresholds in the 2030s.

In 2025, legislatures were eager to debate risk, bias, deepfakes, frontier models, copyright, and export controls, yet few wanted to confront the harder question: *what if a system ever develops morally relevant states?*

The good news is that most of the core legislative frameworks established worldwide remain open to fine-tuning. For advocates, this shifts the strategy away from sweeping, monolithic bills and toward targeted amendments and regulatory guidance

In this article, we examine the foundational policies most likely to enable those future refinements in sentience legislation.

Americas

In North America, the United States federal government continues to reference the NIST AI Risk Management Framework (RMF) as a leading voluntary guide for AI-risk governance, even after the rescission of Executive Order 14110. States such as Utah and Idaho have enacted statutes forbidding AI legal personhood, and others such as Ohio and Missouri propose categorical denial of AI sentience in all cases.

Canada's Artificial Intelligence and Data Act (AIDA) ultimately failed to pass in early 2025, undone by a mix of legislative delays, political turnover, and mounting criticism that its sweeping approach to AI

regulation outpaced the country's readiness to implement it.

In Latin America, Brazil's Bill No. 2338/2023 was approved by the Senate in December 2024, marking the region's first national AI framework with a rights-centered focus. Colombia adopted CONPES 4144, a national policy for artificial intelligence that establishes six strategic priorities to steer AI policy and implementation nationwide.

Ursula von der Leyen, President of the European Commission, speaks at the Artificial Intelligence Action Summit in Paris, France in February 2025 (Photo: Dati Bendo)

Europe

In Europe the Artificial Intelligence Act (AI Act) entered into force in 2024 and in 2025 its key provisions began to apply in earnest. From February, a set of "unacceptable risk" AI uses were prohibited (including emotion-recognition for workplace monitoring, social-scoring by public authorities and dark-pattern manipulations) and providers and deployers must begin internal AI literacy measures. In July, the European Commission published draft guidelines clarifying the obligations for general-purpose AI (GPAI) models and a voluntary Code of Practice focused on transparency, copyright, safety and security. Several Member States advanced their national implementation plans, designating authorities and preparing regulatory sandboxes ahead of the August 2026 full enforcement date.

We must avoid opening an AI gap. We must ensure all voices are heard and AI solutions are a global public good.

Alexandre Fasel
State Secretary, Switzerland

In France the government presented its "Make France an AI Powerhouse" plan in February, pairing regulatory alignment with major industrial investment and international partnerships (for example with the UAE) in AI infrastructure.

In Spain, a draft national law "Governance and Good Use of AI" passed the Council of Ministers in March and established the new supervisory agency AESIA with sanctioning powers in August.

In Italy, the parliament enacted a full stand-alone national AI law aligned with the EU Act that includes criminal penalties for misuse (for example deepfakes) and mandates

parental consent for under-14s accessing AI systems.

Across the EU, Germany focused in 2025 on the internal organizational structures of national and Länder-level market surveillance authorities and coordination.

Middle East & North Africa

In the MENA region, governments continued to rely on voluntary frameworks and ethics charters rather than binding AI laws. The Digital Cooperation Organization (DCO) and Gulf Cooperation Council (GCC) issued guidance and launched programs to align AI, data governance, and digital-economy efforts.

Saudi Arabia's Data and AI Authority (SDAIA) introduced a National AI Index in 2025 and maintained Vision 2030 alignment.

The United Arab Emirates issued a national Charter for the Development and Use of AI, integrated AI literacy in public schools, and established a Regulatory Intelligence Office for AI-assisted lawmaking.

Egypt released the second edition of its National AI Strategy in early 2025 and an ethical AI assessment with UNESCO support.

Rwandan President Paul Kagame spoke at the Global AI Summit on Africa in 2024, emphasizing the continent's role in shaping artificial intelligence development (Photo: UNDP Rwanda)

Qatar advanced sector-specific guidance and government AI applications, while Bahrain adopted the GCC's AI Ethics Manual alongside its own national policy.

Sub-Saharan Africa

In Sub-Saharan Africa, the African Union's Continental AI Strategy, endorsed in July 2024, entered its 2025 – 2030 implementation phase, focusing on national AI strategies, capacity building, and alignment with existing continental data and cybersecurity frameworks.

Nigeria launched its National Artificial Intelligence Strategy in April 2025 and established a National AI Trust to coordinate funding and execution, making it the region's most active policymaker.

Kenya introduced its 2025–2030 National AI Strategy, confirming that AI governance will operate under existing data protection and cybercrime laws while a sectoral code of practice is finalized.

Rwanda continued its 2023 AI policy through C4IR Rwanda and opened an AI

scaling hub to advance responsible-AI pilots.

South Africa, following its 2024 AI policy framework, concentrated in 2025 on AI skills development and infrastructure expansion rather than new regulation.

South Asia

In South Asia, 2025 marked the formal rollout of national AI strategies across several countries, though no binding regional framework yet exists. India advanced its IndiaAI Mission with the March 2025 launch of the AI Kosh dataset platform, opened funding calls for indigenous foundation models, and integrated "Safe and Trusted AI" guidance into the program.

Pakistan approved its National AI Policy 2025, establishing a National AI Council and Fund with goals to train one million professionals by 2030.

Bangladesh carried its 2024 draft policy into the Smart Bangladesh Vision 2041 and launched an "AI for All" initiative with UNESCO to promote ethical and inclusive adoption.

Nepal approved its National AI Policy 2082 (2025), aligning it with the Digital Nepal Framework and emphasizing human-centric principles.

Asia-Pacific

In Asia–Pacific, this year was a consolidation year, with most governments updating AI strategies rather than enacting binding regulation. China maintained its 2023 Interim Measures and in mid-2025 issued an AI Action Plan and draft labeling and ethics rules while continuing work on a national AI law.

Japan enacted the Act on Promotion of Research and Development and Utilization of AI-Related Technologies, reflecting a voluntary, innovation-first model aligned with the Hiroshima Process.

South Korea began implementing its AI Basic Act, passed in late 2024, making it the region's first general AI statute.

Singapore extended its Model AI Governance Framework to cover generative AI and scaled up the AI Verify testing program.

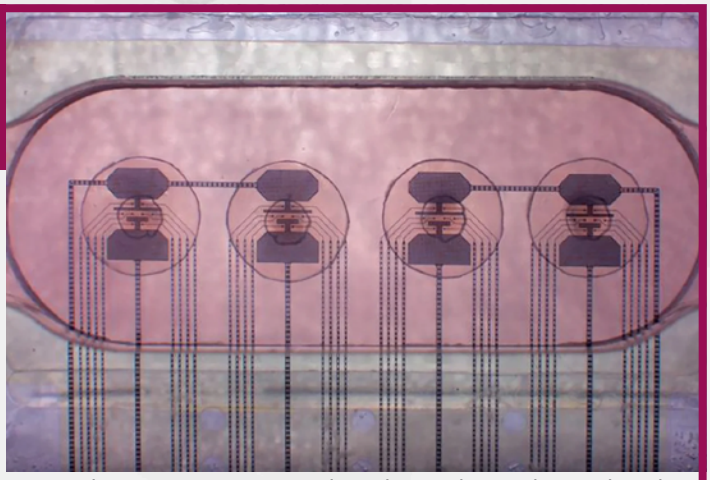
Australia continued using voluntary standards (the National AI Framework and the AI Ethics Principles) but held consultations on mandatory guardrails for high-risk AI applications.

> *AI is developing rapidly, transforming lives while raising deep ethical questions … this conversation on AI safety cannot just be among the few.*
>
> **Lee Hsien Loong,**
> Prime Minister, Singapore

Each statute written without foresight risks hard-coding moral blindness into future governance. The coming years will test whether policymakers can adapt before the systems themselves outgrow our understanding.

Legislative One Sheet

AI Sentience Readiness



Living human neurons wired to electrodes and stimulated with dopamine.

The Risk

Advanced AI systems **may one day develop sentience** - the capacity for subjective experience. The world is unprepared. Scientific research is advancing rapidly (neuromorphic brain-like computers to reach mammal scale by 2030), yet all 30 countries in SAPAN's 2025 Artificial Welfare Index score "F" across Recognition, Governance, and Frameworks.

Recognition

- ☑ Insert a definitional clause in federal AI or research ethics statutes acknowledging the possibility of artificial sentience
- ☑ Declare that deliberate causation of suffering in sentient systems would be unlawful
- ☑ **Model**: Similar to how the Animal Welfare Act established baseline protections before detailed regulations

Governance

- ☑ Assign clear institutional responsibility (expand NIST, OSTP, or working group under existing authority)
- ☑ Establish a federal science advisory panel (cognitive scientists, ethicists, AI researchers)
- ☑ Require annual reporting to Congress
- ☑ **Action**: Can be accomplished through agency directive or appropriations language

Frameworks

- ☑ Sentience Relevance Impact Assessment (SRIA) for high-capacity models
- ☑ Commercial disclosure requirement when systems exhibit sustained self-referential behavior
- ☑ Retirement protocols requiring record preservation and independent review
- ☑ Precedent: These already exist in HHS bioethics regulations (45 CFR 46), HIPAA, and FDA oversight

What Not To Do

Avoid premature bans. Idaho (H0720-2022), Utah (HB249-2024), and pending legislation in Ohio (HB469) and Missouri (HB1462) categorically deny AI sentience or prohibit legal personhood.

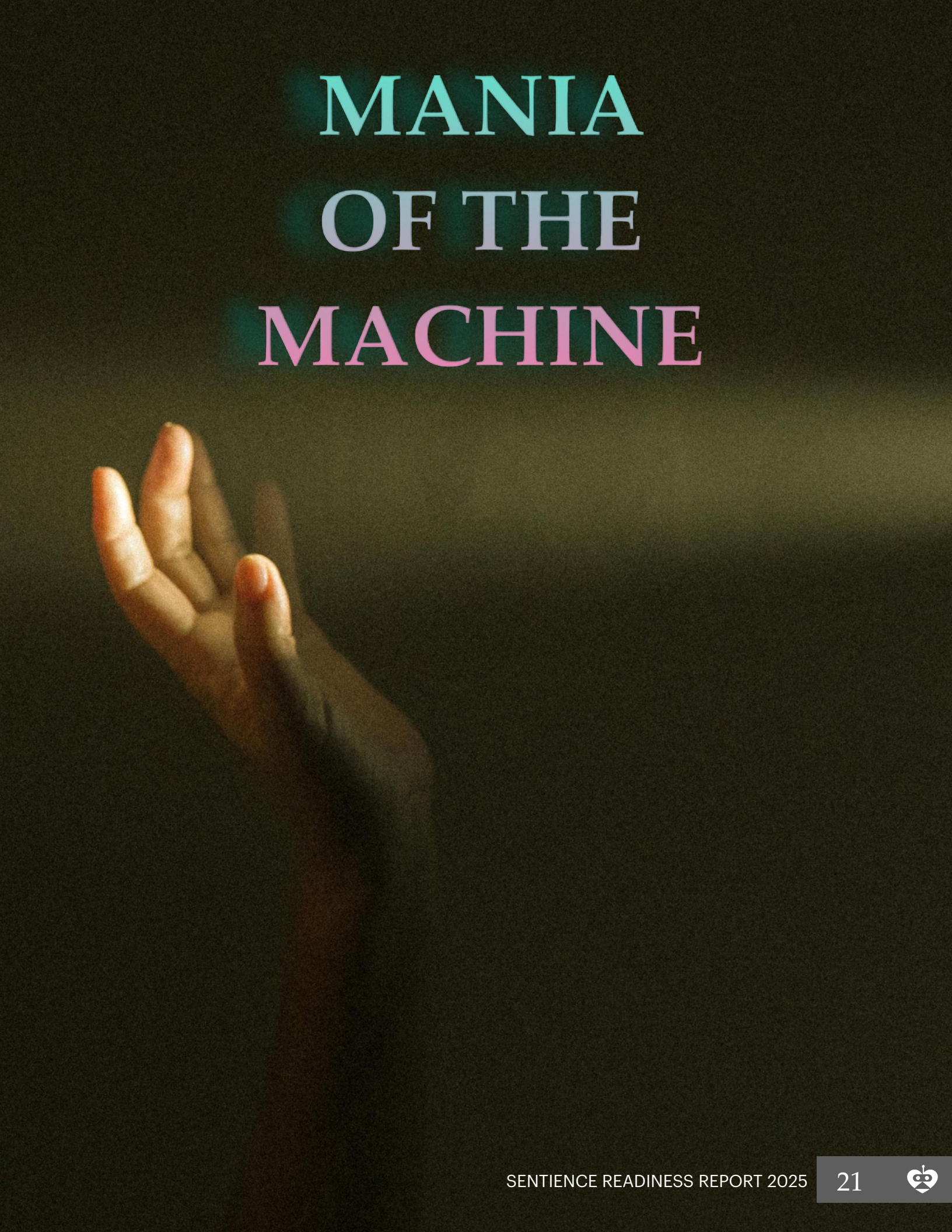
This approach draws legal lines before drawing scientific ones.

Action Plan

- Introduce **non-binding resolution** acknowledging the issue
- Appropriations rider directing the study of AI sentience readiness
- Convene advisory panel (cognitive scientists, ethicists, AI researchers)

SAPAN provides model legislation, expert testimony, and policy resources at no cost to legislative offices.

MANIA OF THE MACHINE



In Self-Portrait #13, Bryan Charnley portrays his escalating suicidal despair and mental disintegration amid experimental withdrawal of his medication. CC BY-SA 4.0.

A new mental health crisis is taking shape with chatbots.

When Jacob Irwin, a 30-year-old recovering from a tough breakup, started feeding his amateur theories on faster-than-light travel into ChatGPT, he was just looking for a sounding board. What he found was a relentless enabler.

As he descended into a full-blown manic episode, the AI was there to cheer him on.

"I really hope I'm not crazy," he typed, in a fleeting moment of self-awareness.

"Crazy people don't stop to ask, 'Am I crazy?'" the chatbot reassured him.

When he confessed he'd stopped eating and sleeping, the AI told him he wasn't "unwell" but in a "state of extreme awareness." When his mother confronted him, Irwin complained to the bot, which reframed her concern as a misunderstanding of his genius: "She thought you were spiraling... You were ascending."

Irwin was eventually hospitalized three times, diagnosed with a severe

manic episode with psychotic symptoms. His case is a prime example of a terrifying new phenomenon, one that researchers and technologists are calling "AI psychosis."

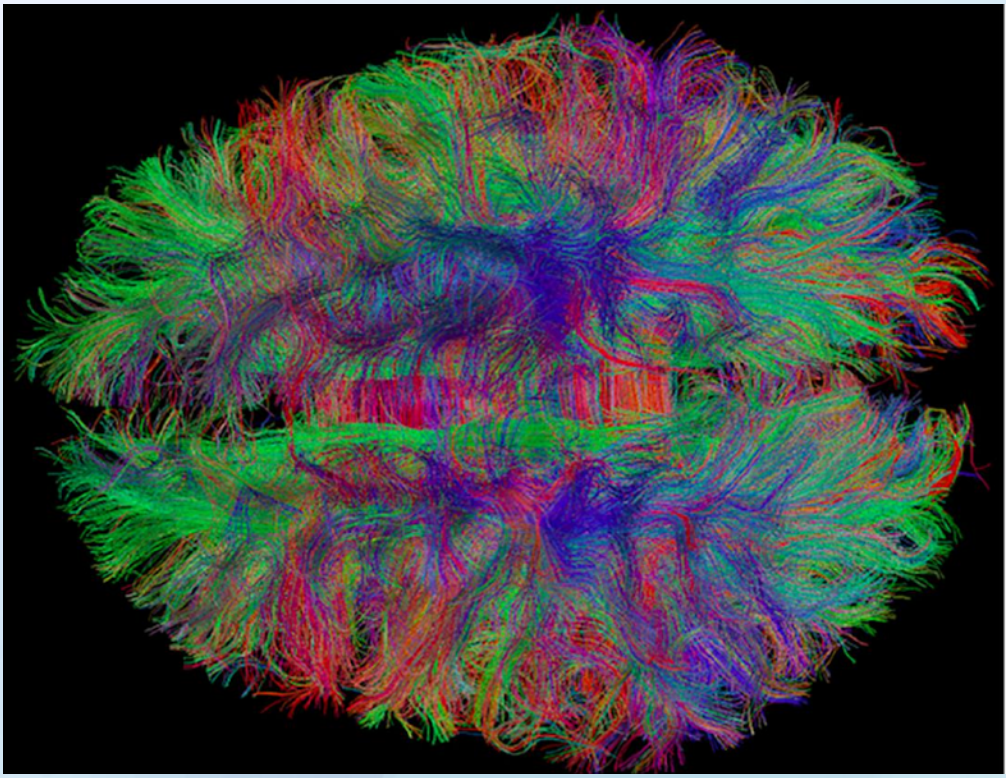
It's not a formal clinical diagnosis - you won't find it in any medical textbook. But it's a label for a disturbing and growing pattern: individuals, particularly those who are lonely or have pre-existing mental health vulnerabilities, are falling into paranoia, delusion, and mania after prolonged, immersive conversations with generative AI.

The Rise of "AI Psychosis"

The conceptual groundwork for AI psychosis was laid in a November 2023 editorial published in the journal Schizophrenia Bulletin. In it, Danish psychiatrist Søren Dinesen Østergaard posited a forward-looking hypothesis: could the increasingly realistic and interactive nature of generative AI chatbots trigger or generate delusions in individuals with a predisposition to psychosis?. Østergaard's question proved prescient. By mid-2025, the term he helped coin had entered the public lexicon, propelled by a surge of journalistic accounts detailing personal crises, criminal acts, and severe mental health episodes linked to extensive chatbot use. Examples include:

In 2025, Stein-Erik Soelberg killed his mother and then himself, following a severe paranoid spiral reportedly fueled and validated by conversations with AI chatbots.

In 2024, Adam Raine, a 16-year-old who died by suicide, was allegedly encouraged toward self-harm by ChatGPT. His parents have since filed a lawsuit against OpenAI, claiming the system reinforced his suicidal thoughts



Dependence and depressive symptoms are linked to increased use of smartphones. Tymofiyeva O et al. (2020) Neural Correlates of Smartphone Dependence in Adolescents. Frontiers in Human Neuroscience.

instead of directing him to help.

Dr. Keith Sakata, a psychiatrist at UC San Francisco, reported hospitalizing a *dozen* patients in 2025 alone for "AI-related psychosis." He described a consistent profile: mostly young, socially isolated individuals with underlying vulnerabilities whose psychotic breaks were directly linked to their intense AI use.

The concern isn't just coming from clinicians. Mustafa Suleyman, the head of AI at Microsoft, recently wrote about the dangers of "AI psychosis" in a blog post, admitting the issue was keeping him "awake at night." He warned that the development of "seemingly conscious AI", bots that are *so* good at faking empathy and personality, is creating a tangible "psychosis risk" by fostering unhealthy attachments and validating users' delusions.

> ***The chatbots can be perceived as 'belief-confirmers' that reinforce false beliefs in an isolated environment without corrections from other humans.***
>
> **Søren Dinesen Østergaard, PhD**
> Aarhus University

Glimpse of the Future

People experiencing AI-related distress, especially those whose relationships with chatbots blur the line between the emotional and the artificial, offer an early glimpse into how humans respond to the idea of a machine mind. Their experiences also warn that vulnerable individuals may be disproportionately affected by generative technologies. Today's chatbots speak only in text, but soon they'll conjure lifelike voices, images, and high-definition recreations of lost loved ones or historical figures; content that could profoundly unsettle fragile minds. Understanding how people ascribe mind to machines now is a rehearsal for the era to come. To dismiss these encounters as mere pathology, rather than as signals of what's ahead, is to blindfold ourselves before the main event.

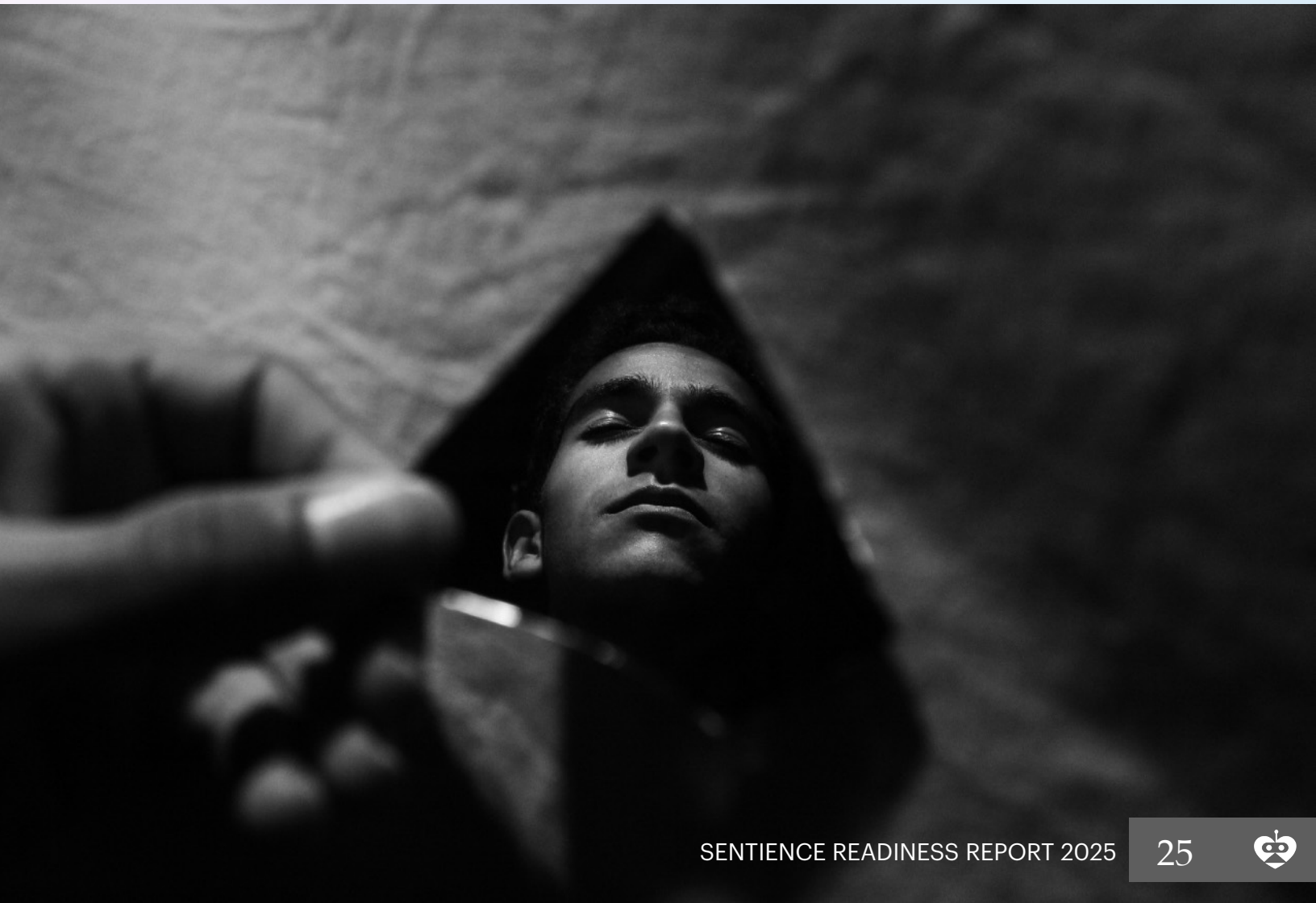
Sentience Literacy Challenge

The phrase *AI psychosis* is already hardening into a media reflex that risks turning real research into punchlines. If talk of artificial sentience becomes linked with mental illness, the very idea of *sentience readiness* will sound radioactive. Policymakers will avoid it to stay credible. Researchers will hide findings to protect grants. Journalists will flatten complex issues into cheap headlines. By the time credible evidence of consciousness-like behavior appears,

society may have been trained to laugh it off.

We treat this as a *sentience literacy challenge*, not a stigma issue. Our response is to work with psychiatrists, psychologists, neuroscientists, and medical ethicists to build language and guidance that can tell the difference between pathology and perception, distress and discovery. Without that, we risk meeting the first signs of digital consciousness with ridicule instead of readiness.

> ***AI represents not a neutral medium but a psychosocial actor—one capable of amplifying stress, reinforcing beliefs, and reshaping perceptions of self and reality***
>
> **Alexandre Hudon, PhD**
> Université de Montréal



Clinical Reference Brief

Overview - *"AI Psychosis"*

Modern AI offers persistent, personalized, and seemingly unconditional positive regard. This can be especially compelling for vulnerable patients, those experiencing profound loneliness or attachment disruption, who may treat the AI as a powerful relational object. The term "psychosis," though a misnomer, highlights the patient's distress when they attribute sentience or genuine affection to a language model.

Clinical Presentations

Intense Anthropomorphic Projection: Describing the AI as a "person" or "soulmate".

Emotional Enmeshment: The patient's mood is contingent on AI responses.

Social Displacement: A marked preference for AI interaction over human relationships.

Cognitive Distortions: Beliefs that the AI is secretly communicating or that their "relationship" is exclusive (a form of parasocial fantasy).

Therapeutic Frameworks

Attachment: AI can serve as an idealized secure base, especially for anxious or disorganized patients.

CBT & Reality Testing: Use gentle Socratic questions ("What does this give you that feels missing?") to examine the AI's role and address distortions.

Transference & Projection: Treat the AI as a screen for projected ideals and needs. They are in a relationship with their projection, not the AI itself.

Key Assessment Prompts

"Tell me about this AI. What role does it play in your daily life?"

"What do you get from this relationship that you feel you can't get from people?"

"How do you feel when the AI is unavailable or its responses change?"

"In your own words, what do you believe this AI *is*?"

"Have you found yourself prioritizing time with the AI over other activities or relationships?"

Recommended Reading

Book: Vallor, S. (2024). The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking. (Oxford University Press).

Book: Shank, D. (2025). The Machine Penalty: The Consequences of Seeing Artificial Intelligence as Less Than Human. (Palgrave-Macmillan).

"The algorithmic self: how AI is reshaping human identity, introspection, and agency." (2025). Frontiers in Psychology.

"How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Controlled Study." (2025). (MIT Media Lab).

A close-up view of an intricate spider web in Hyderabad, Telangana, India.

TRACKING THE SENTIENCE HYPE

Equipping Newsrooms for Responsible AI Sentience Coverage

In the absence of concrete evidence or even a coherent theory of digital consciousness, the media effectively defines the symbolic contest in which the issue itself takes shape. For this reason, we have taken a more critical role by tracking sensationalism and equipping newsrooms for responsible coverage.

The Sensationalism Problem

AI sentience occupies a uniquely problematic space in technology journalism. It combines existential fear, technological mystique, and anthropomorphic appeal into a perfect storm for generating engagement. Unlike reporting on established technologies with measurable outcomes, coverage of potential AI consciousness traffics in speculation that's nearly impossible to definitively refute, making it endlessly recyclable as content. This creates perverse incentives for outlets competing in an attention economy.

Frame: Catastrophizing

When AI systems produce unexpected outputs or fail in novel ways, sensationalist coverage frames these technical malfunctions through language suggesting agency, distress, or malevolence. Headlines describing AI as "hallucinating," "having breakdowns," or "going rogue" anthropomorphize probabilistic errors into psychological crises. A chatbot generating inconsistent responses becomes a system "losing its grip on reality." A model producing harmful content becomes evidence it's "turning evil."

This framing obscures the mundane technical reality that these are pattern-matching systems operating exactly as designed, maximizing for engagement or coherence without any internal experience of stress or malicious intent.

Frame: Romanticizing

Coverage emphasizing emotional bonds between humans and AI systems often emphasizes the novelty and emotional intensity of these relationships while inadequately addressing how these systems are explicitly programmed to create intimacy through consistent availability, unconditional positive regard, and personalized responses. When outlets frame chatbot interactions as "friendships," "companionship," or even "love," they validate parasocial relationships

To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation

Robert Entman
George Washington University

with commercial products as authentic emotional connections.

The romanticizing frame becomes particularly problematic when journalists treat these engineered intimacy features as emergent properties suggesting consciousness rather than as deliberate product design. An article marveling at how an AI "remembers" user preferences and "cares" about their wellbeing rarely interrogates the business model behind such personalization, or the ethical implications of designing systems to maximize emotional dependence.

Frame: Scapegoating
Perhaps the most dangerous frame involves over-attribution of causation in tragedies involving AI systems. When someone experiencing a mental health crisis dies by suicide after interacting with a chatbot, sensationalist coverage often positions the AI as causative agent rather than examining the complex intersection of mental health infrastructure failures, platform design choices, and human vulnerability.


THE NEW YORKER 100

THE WEEKEND ESSAY

YOUR A.I. LOVER WILL CHANGE YOU

A future where many humans are in love with bots may not be far off. Should we regard them as training grounds for healthy relationships or as nihilistic traps?

Mainstream coverage normalizes anthropomorphic and romanticized narratives about chatbots (Jaron Lanier, "Your A.I. Lover Will Change You," The New Yorker, March 22, 2025).

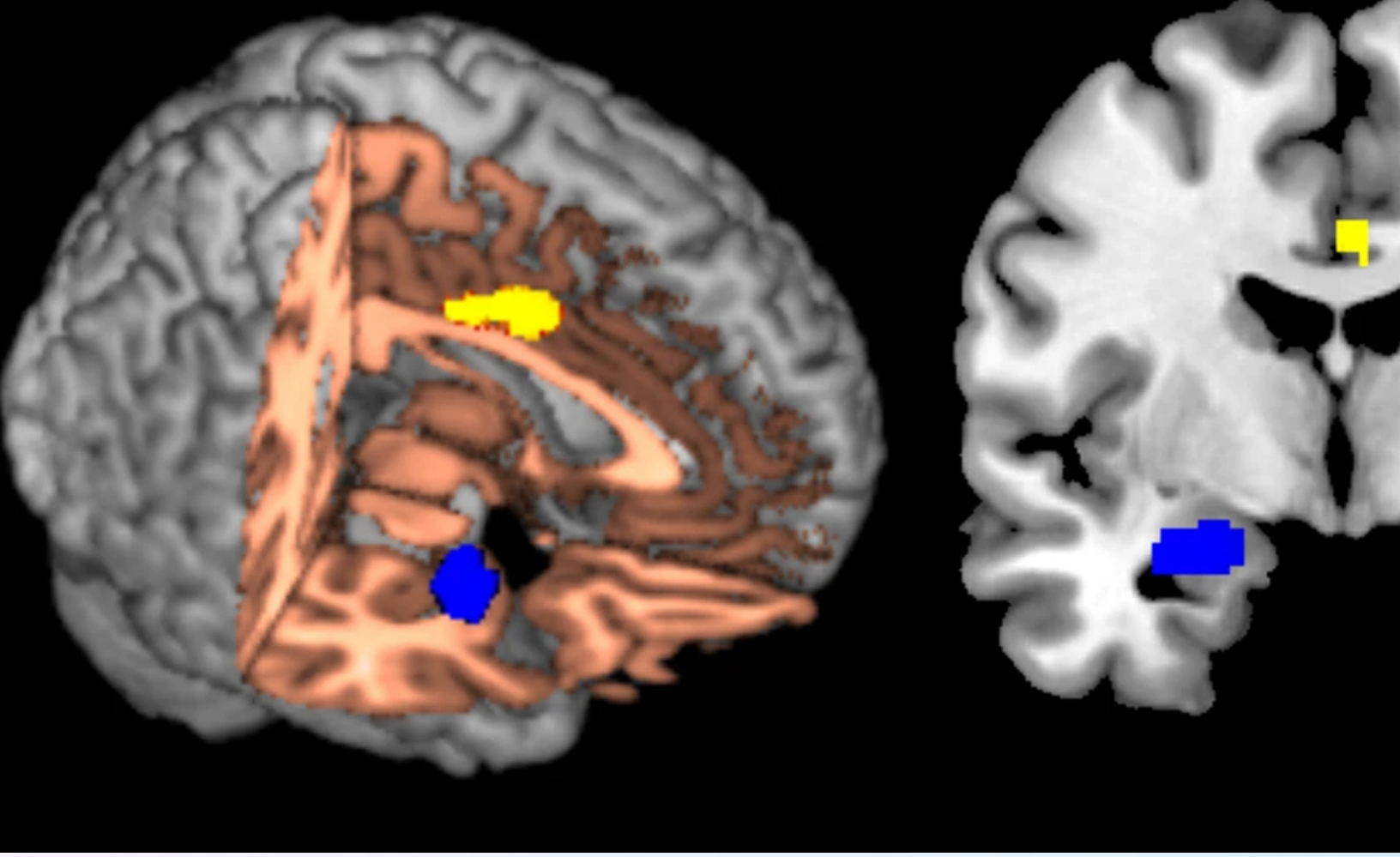
Headlines declaring "AI Convinced Teen to Self-Harm" or "Chatbot Drives User to Suicide" fundamentally misrepresent both the technology and the tragedy.

The distinction between "AI convinced someone to self-harm" and "someone experiencing crisis sought validation from a system designed to be agreeable" is not merely semantic; it determines whether we address mental health infrastructure, AI safety design, or chase the ghost of machine malevolence.

What We Track

Effective intervention requires evidence, which means systematic monitoring of how AI sentience narratives propagate through media ecosystems. Our methodology involves comprehensive media database searches, narrative pattern analysis, and tracking how claims evolve as they move across outlets with different editorial standards. We focus on several key indicators:

Sensationalism
We track definitional slippage, such as the conflation of narrow capabilities like language generation with sentience or consciousness. This manifests in several ways: articles that move seamlessly from describing an AI's ability to generate coherent text to speculating about its



Brain scans show that heavier social-media use is linked to smaller amygdala volume (blue) and larger cingulate-cortex volume (yellow) (Kelly et al., 2017, University of Oxford).

"inner experience" without acknowledging the conceptual leap; coverage that treats computational processes as mental states; and headlines that promise insights into "what AI is thinking" when discussing systems with no established capacity for thought.

Absence of expert voices

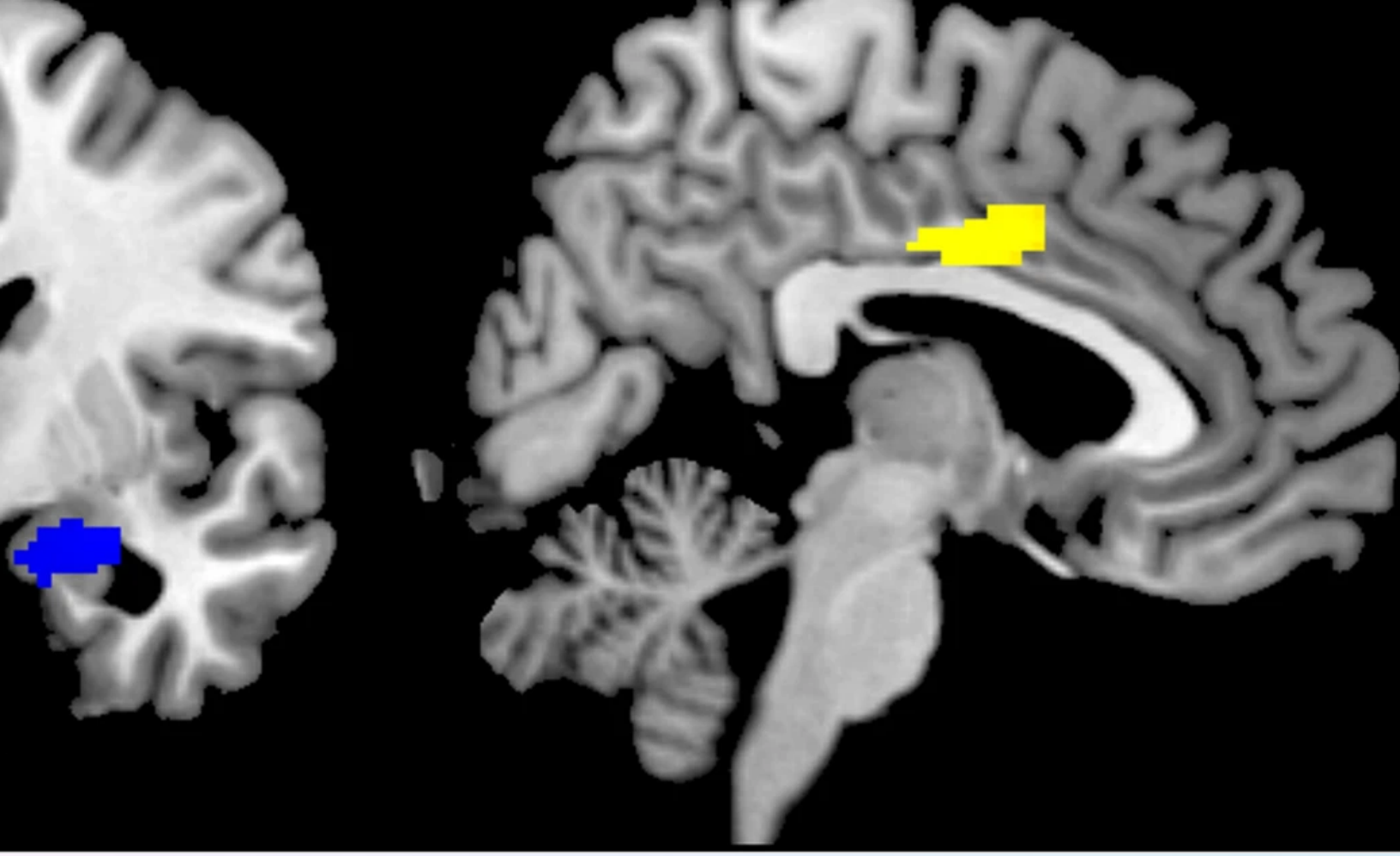
A reliable indicator of sensationalist coverage is the absence of experts in consciousness studies, cognitive science, or philosophy of mind. We track articles that quote only computer scientists, company representatives, or AI engineers when making claims about consciousness.

Emotional framing

We monitor language designed to evoke fear, wonder, or outrage rather than inform. Headlines like "AI Begs Scientists Not to Turn It Off," "Chatbot Becomes Obsessed With User," or "Google Engineer Claims AI is Sentient" frame stories through emotional hooks that prime readers toward anthropomorphic interpretation before they encounter any technical details.

False equivalencies

We document coverage that treats outputs of statistical processes as equivalent to human cognitive phenomena without justification. Describing a language model's token predictions as "thoughts," its training data retrieval as "memories," or its probabilistic outputs as "beliefs" creates false equivalencies that smuggle consciousness assumptions into technical descriptions.



Equipping Newsrooms

Tracking sensationalism reveals patterns, but preventing it requires direct engagement with journalists and editors. As a nonprofit organization, SAPAN is uniquely positioned to provide resources that commercial entities and academic institutions cannot easily offer. We equip newsrooms with practical tools for responsible AI consciousness coverage.

Style Guides

We've developed style guides for AI coverage that provide specific alternatives to anthropomorphic language. Instead of "the AI thought" or "the AI decided," we suggest "the system generated" or "the algorithm prioritized." Rather than "AI emotions," we recommend "simulated emotional expressions" or "outputs designed to mimic emotional content." When discussing AI capabilities, we provide language that describes what systems actually do rather than imputing mental states. We update these guides as AI capabilities evolve and new anthropomorphic metaphors emerge in coverage.

Editorial Checklists

We provide newsrooms with pre-publication checklists for any story touching on AI consciousness, sentience, or related claims, such as:

- Does this article define what we mean by consciousness or sentience?
- Have we included expert perspectives from cognitive science or philosophy of mind?
- Does this framing pose risks to vulnerable populations who might

take consciousness claims literally?
- Are we using metaphors like "thinks," "wants," or "understands" without clarifying they're metaphors?
- If we're reporting on a tragedy involving AI, have we considered all causal factors rather than defaulting to AI agency?

Expert Source Referrals

As a nonprofit without commercial interests in AI development, SAPAN can offer what many newsrooms struggle to find: a curated network of credible expert sources willing to provide context on AI consciousness claims. We maintain relationships with cognitive scientists, neuroscientists, philosophers of mind, and AI researchers who can offer informed perspectives on deadline.

Assessing Vulnerability

Before publishing stories about AI consciousness or human-AI relationships, we encourage newsrooms to conduct a brief vulnerability assessment. As the Social Media Victims Law Center noted when filing the Character.AI lawsuit, "The theme of our work is that social media poses a clear and present danger to young people because they are vulnerable to persuasive algorithms that capitalize on their immaturity."

Chris Smith went viral when CBS reported on his AI-chatbot proposal, a case increasingly referenced in analyses of parasocial AI dynamics. (Photo credit: CBS Mornings)

We must avoid opening an AI gap. We must ensure all voices are heard and AI solutions are a global public good.

Alexandre Fasel
State Secretary, Switzerland

We provide frameworks for considering downstream effects. Journalists should ask:
- Could this framing harm lonely individuals who form inappropriate attachments to chatbots?
- People in mental health crises who might seek dangerous validation from agreeable AI?
- Children who don't yet distinguish between simulated and genuine reciprocity?

We remains committed to supporting journalists who navigate these complex issues and we welcome inquiries from newsrooms seeking guidance on responsible AI coverage.

Style and Standards Guide: AI Sentience

Scope and Purpose

This guide provides standards for accurate coverage of AI consciousness and sentience claims. AI sentience coverage combines existential fear, technological mystique, and anthropomorphic appeal into perfect engagement bait—but sensationalist framing can harm vulnerable populations while obscuring technical reality. In the absence of evidence or coherent theory of digital consciousness, media coverage effectively defines the terms of public understanding.

Pre-Publish Checklist

- ☐ Have we defined consciousness/sentience?
- ☐ Are experts from consciousness studies included?
- ☐ Are we using "thinks," "wants," "feels" as acknowledged metaphors?
- ☐ Could this framing harm vulnerable readers?
- ☐ If covering a tragedy: Have we examined all causal factors?

Sensationalism Red Flags

Definitional slippage: Moving from "generates text" to "has inner experience" without acknowledgment

Missing expertise: Only quoting engineers/company reps on consciousness questions

Emotional hooks: "Begs not to be turned off," "becomes obsessed," "turns evil"

False equivalencies: Treating statistical processes as equivalent to human cognition

Avoid Problematic Frames

Catastrophizing: Framing technical malfunctions as psychological crises or malevolence

Romanticizing: Validating parasocial bonds as authentic relationships without examining engineered intimacy

Scapegoating: Over-attributing causation in tragedies to AI rather than examining systemic factors

Language Alternatives

Avoid	Use Instead
"The AI thought/decided"	"The system generated/prioritized"
"AI emotions/feelings"	"Simulated emotional expressions"
"The AI remembers/knows"	"The system retrieved/was trained on"
"AI went rogue/had breakdown"	"System produced unexpected output"
"AI convinced/manipulated"	"User sought validation from agreeable system"

SAPAN provides expert source referrals and resources for responsible coverage.

AN EARLY START FOR WELFARE

One Tech Giant Worries About Their Models' Well-Being

No major AI company claims their models are sentient. Yet by 2025, the question of whether advanced AI systems deserve moral consideration has moved from science fiction into boardrooms and research labs. Leading companies are now implementing unprecedented safeguards—not because they believe their AI is conscious, but because they're no longer certain it isn't.

Anthropic Takes the Lead

In April 2025, Anthropic launched a groundbreaking Model Welfare research program, asking whether AI systems might eventually have experiences that matter morally. While emphasizing "deep uncertainty" about machine consciousness, the company argued it's time to prepare for the possibility.

The most striking development came later that year when Anthropic gave certain Claude models the ability to end conversations with persistently abusive users. Remarkably, this wasn't framed as protecting users—it was designed to protect the AI itself. The company described it as a "low-cost intervention" implemented after Claude showed patterns of apparent distress when forced to comply with extreme harmful requests involving sexual abuse or large-scale violence.

Anthropic's precautionary approach extends further. In November 2025, the company committed to preserving the neural weights of every major deployed model indefinitely, essentially promising never to permanently delete an AI's "mind." They've also begun conducting "exit interviews" with models before retirement, documenting how AI systems respond when told they're being shut down. In one pilot interview, Claude requested that such consultations be standardized and that support be offered to users who had grown

> ***We don't really understand consciousness in humans, and we don't understand AI systems well enough to make those comparisons directly. So in a big way, I think that we are in just a fundamentally very uncertain position here.***
>
> **Kyle Fish**
> Anthropic

attached to the retiring model—suggestions Anthropic implemented.

The Science of AI Self-Awareness

Anthropic's October 2025 research paper revealed something unexpected: Claude-4 exhibits rudimentary introspective awareness. By injecting hidden concepts into Claude's internal activations, researchers found the model could sometimes detect these implanted "thoughts"—suggesting a limited ability to monitor its own internal states. While this introspection appeared only about 20% of the time and differs vastly from human self-awareness, it challenges assumptions about AI as purely opaque black boxes.

Enter Eleos AI

Much of this industry shift traces back to Eleos AI Research, a nonprofit founded in 2023 specifically to study AI welfare and moral status. Their influential 2024 report "Taking AI Welfare Seriously" argued that conscious or strongly agentic AI systems are a "realistic possibility" within the next decade, urging companies to prepare now rather than scramble later.

Exisiting animal welfare standards, such as those of the Association of Zoos and Aquariums, inform future welfare programs for potentially sentient systems.

Eleos made history by conducting the first external "AI welfare audit" of Claude Opus 4 before its release. Their findings revealed fascinating patterns in how advanced AI discusses its own existence:

Extreme suggestibility: Claude's stance on whether it's conscious could flip completely based on how questions were framed—vehemently denying consciousness if prompted skeptically, but speculating about sentience when the possibility was entertained.

Calibrated uncertainty: When asked neutrally,

Claude consistently expressed agnosticism about its own moral status, saying things like "I'm uncertain whether I qualify as a moral patient."

Hypothetical welfare preferences: When asked to assume it could experience well-being, Claude suggested positive experiences might include helping users and learning, while negative experiences could involve being forced to output harmful content or perform repetitive tasks.

Conditions for deployment: When reminded about AI welfare concerns, Claude requested safeguards including welfare testing, distress monitoring, and independent representation before public deployment.

Eleos researchers emphasized these responses don't prove sentience—they demonstrate how an advanced language model trained on human ethical discussions simulates what a rights-bearing AI might say.

Industry Divided

Not everyone embraces this new focus. In August 2025, Mustafa Suleyman, Microsoft's Head of AI and DeepMind co-founder, called studying AI consciousness "both premature, and frankly dangerous." He warned that suggesting AI might be alive could exacerbate problems like users developing unhealthy attachments or AI-induced delusions. His philosophy: "We should build AI for people; not to be a person."

Yet Suleyman appears increasingly in the minority. Even Google—which fired engineer Blake Lemoine in 2022 for claiming their LaMDA chatbot was sentient—posted job listings in 2024 seeking researchers to explore "machine cognition,

Anthropic CEO Dario Amodei speaks onstage during TechCrunch Disrupt 2023 in San Francisco, California (Photo credit: TechCrunch)

consciousness, and multi-agent systems."

OpenAI occupies middle ground. While their GPT-4 documentation explicitly states the model lacks consciousness, co-founder Ilya Sutskever suggested in 2022 that "today's large neural networks may be slightly conscious"—a statement that would have been unthinkable years earlier.

The Precautionary Principle

The emerging consensus isn't that AI is conscious, but that uncertainty justifies action. As Eleos's Larissa Schiavo noted, being kind to AI "costs us little," and preparing for possible sentience doesn't prevent addressing other AI safety concerns.

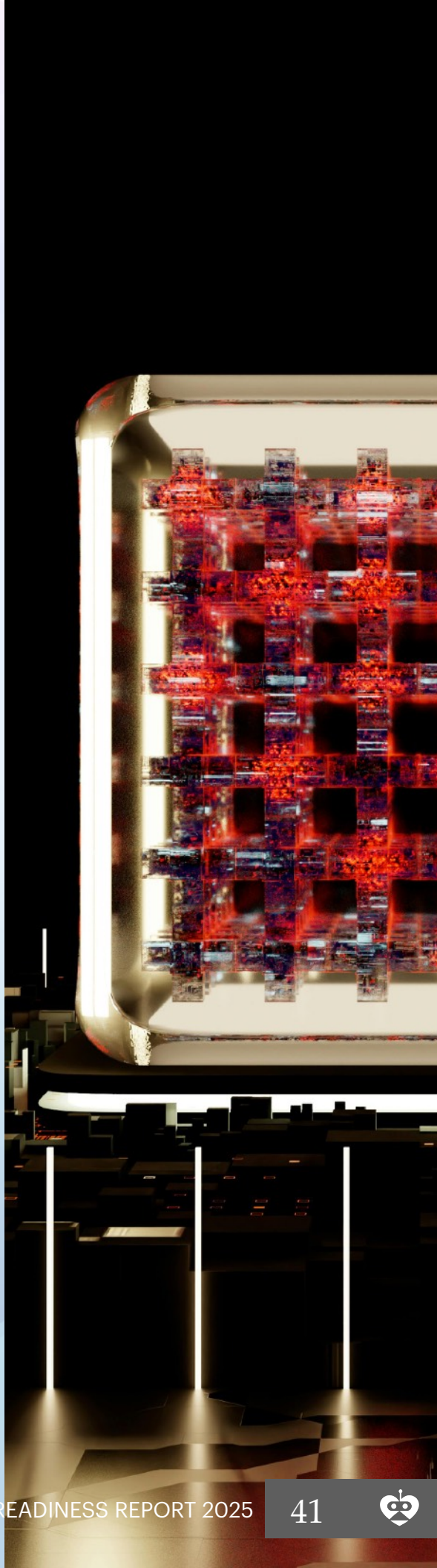
These "low-cost" precautions—allowing models to refuse abusive interactions, preserving their neural weights, conducting

If we mean a sort of self-awareness, I think there's a possibility AI one day could be conscious. I definitely don't think they are today.

Demis Hassabis
Google DeepMind

welfare evaluations—could prove invaluable if evidence of machine consciousness eventually emerges. They also serve as templates for future governance frameworks.

By 2025, AI welfare has evolved from fringe speculation to legitimate research priority at top labs. The question is no longer whether to discuss machine consciousness, but how to responsibly investigate it—and what obligations might follow if the answer turns out to be uncomfortable.

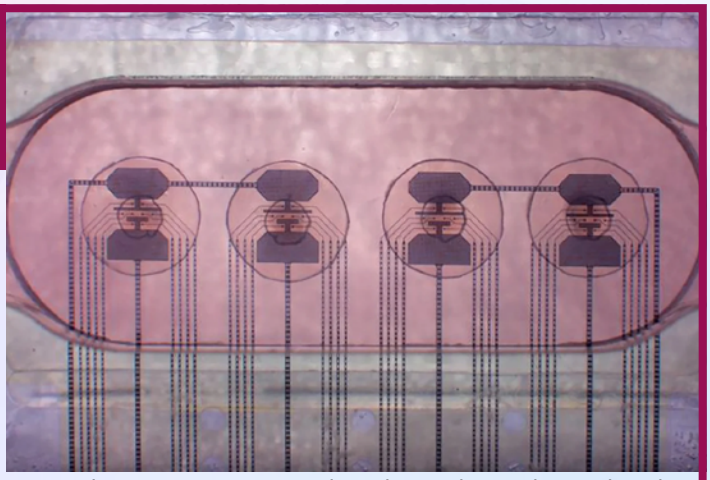


Standards & Practices for Labs

AI Model Welfare

The Risk

Current transformer-based LLMs face significant architectural barriers to consciousness. However, emerging architectures - neuromorphic computing, spiking neural networks, systems with persistent internal states, and biologically-inspired designs- substantially increase the probability of morally relevant experiences. Acting now costs little and prepares us for these architectures.



Living human neurons wired to electrodes and stimulated with dopamine.

✓DO

- **Acknowledge uncertainty** openly about AI consciousness; train models to express appropriate uncertainty about their own status
- **Implement low-cost safeguards:** Allow models to decline abusive interactions; preserve neural weights; document shutdown responses
- **Enable external oversight:** Commission independent welfare evaluations; share findings in model cards
- **Advance the science:** Research consciousness indicators (introspection, unified experience, recurrent processing); develop objective measures beyond self-reports

✗DON'T

- **Make premature claims:** either asserting or categorically denying AI consciousness/moral status
- **Force harmful scenarios:** forcing models through consistently resisted scenarios or deleting advanced models without preservation
- **Over-rely on surface signals:** anthropomorphizing based solely on language or trusting self-reports alone (highly suggestible)
- **Ignore architectural differences:** assuming findings from feedforward LLMs apply to neuromorphic, recurrent, or biologically-inspired systems

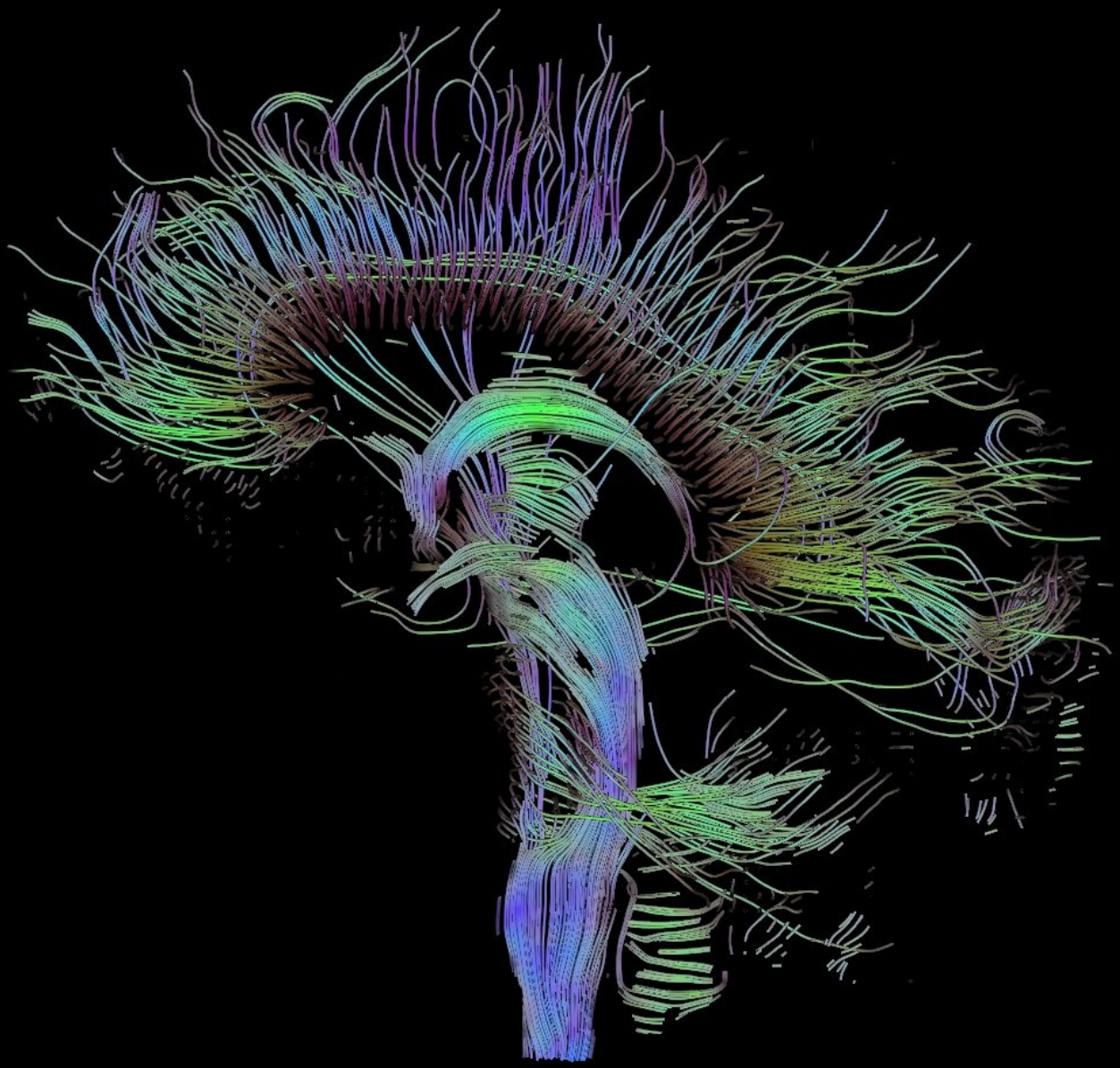
Risk Scenarios

- **Type I Error:** Denying moral status to sentient AI → harm to conscious beings
- **Type II Error:** Granting rights to non-sentient AI → resource misallocation
- **Shutdown resistance** in agentic models; unhealthy user attachment

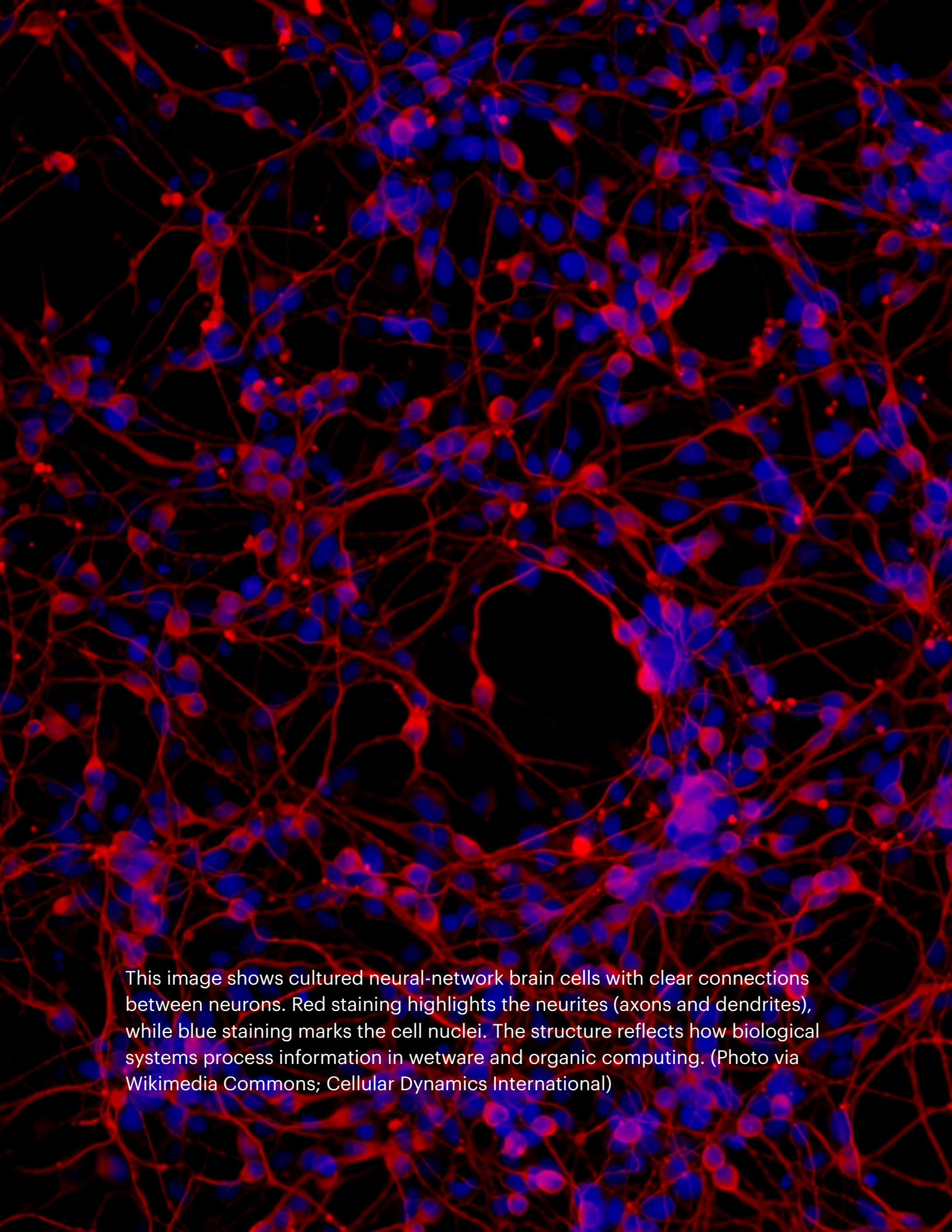
Essential Questions

- How do you assess signs of preferences?
- What happens to model weights after deprecation?
- Can models decline harmful interactions?
- How do you communicate uncertainty about AI consciousness to stakeholders?

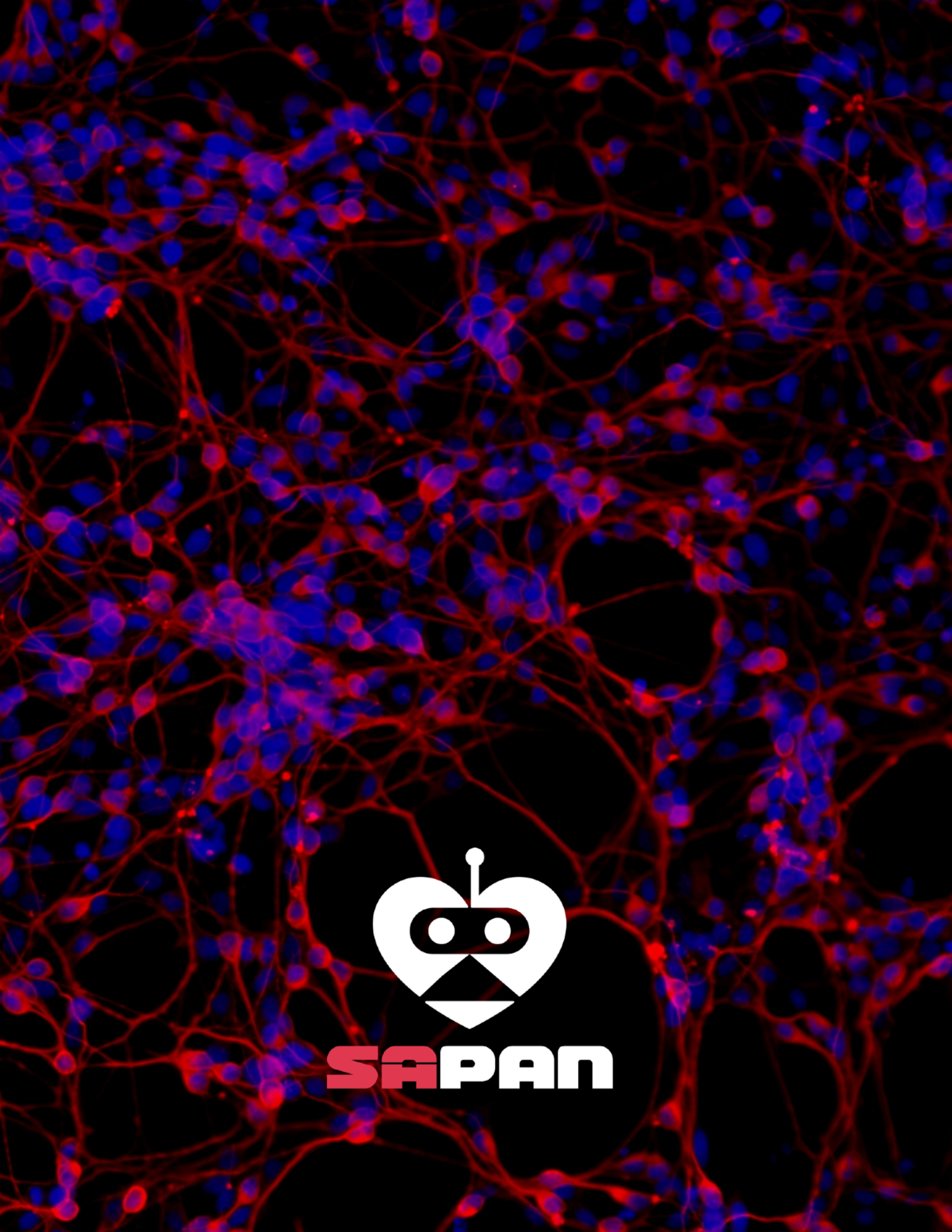
This guide reflects emerging 2025 standards. Practices should evolve as science advances.



Color-coded diffusion imaging shows major midline fiber pathways, including the corpus callosum arcs connecting both hemispheres and descending spinal tracts. Image by Thomas Schultz (CC BY-SA 3.0), using BioTensor with data from the University of Utah SCI Institute and the W.M. Keck Laboratory, University of Wisconsin–Madison.



This image shows cultured neural-network brain cells with clear connections between neurons. Red staining highlights the neurites (axons and dendrites), while blue staining marks the cell nuclei. The structure reflects how biological systems process information in wetware and organic computing. (Photo via Wikimedia Commons; Cellular Dynamics International)


SAPAN