What We Learned from a Year of Conversations About AI

Across 13 dialogues and 52 countries, people revealed how AI is becoming more than a tool: a companion, adviser, confidant, and a new source of authority.

Ana Catarina de Alencar
Resident Philosopher on AI Governance and Policy
September 2026
The AI Collective

The AI
Collective

THIS REPORT AT A GLANCE

13

global public dialogues

20+

experts · 8 disciplines

937

registrations

52

countries · 6 continents

6

governance findings

10

recommendations

3 tracks

Intimacy · The Mind · Childhood & Education

1 paradigm

RELATIONAL AI GOVERNANCE

Patterns identified independently across disciplines, not derived from theory.

About the Author

Ana Catarina de Alencar — Resident Philosopher, The AI Collective

Ana Catarina de Alencar is an international lawyer and ethicist based in Paris, working at the intersection of AI governance, regulation, and philosophy.

As Resident Philosopher at The AI Collective, she designs and convenes the dialogue series on which this report is based, bringing together legal scholars, clinicians, neuroscientists, builders, and affected users from across six continents. Her research is interdisciplinary by necessity rather than by preference: emotional AI cannot be understood through law alone, and she works across law, philosophy, and neuroscience to examine how systems that infer and respond to emotional states are reshaping consent, autonomy, and intimacy.

She is a member of the Women in Digital Forum, under the European Union's Connecting Women in Digital programme, and contributes to international work on trustworthy and human-centred AI through Open Ethics and the Z-Inspection® Initiative.

She holds a master's degree in Philosophy of Law. She is the author of *Artificial, Ética e Direito: Um Guia Prático para Entender o Novo Mundo* [Artificial Intelligence, Ethics and Law: A Practical Guide to Understanding the New World] (Saraiva Jur, 2022) and *Inteligência Artificial e Direito: O Guia Definitivo* [Artificial Intelligence and Law: The Definitive Guide] (Turivius, 2020), and her book on how AI companions are reshaping consent and law is forthcoming. She writes regularly for Tech Policy Press and at The Law of the Future, and speaks at academic and professional conferences across Europe.

About The AI Collective

The AI Collective is a global non-profit community founded in San Francisco. Formerly The GenAI Collective, it was incorporated as a non-profit and relaunched under its current name in June 2025 in order to bridge fast technological advancement with societal alignment and human-centric governance.

It brings together more than 200,000 members worldwide, across chapters spanning six continents: founders, researchers, operators, policymakers, ethicists, and investors. Its stated commitment is to trust, openness, and human flourishing in the age of AI, and its work runs through in-person chapters, salons, hackathons, demo nights, and research initiatives. It describes itself as the human side of AI.

The Resident Philosopher programme sits within that mission. It exists to hold the questions that product roadmaps and regulatory timetables tend to leave aside: what these systems are doing to attention, intimacy, judgment, and meaning, and what obligations follow. The thirteen dialogues behind this report were convened under that programme over the course of a year.

EXECUTIVE SUMMARY | 2025–2026 | 13 EVENTS

Executive Summary

This report is based on the complete transcripts of thirteen public events designed, hosted, and moderated by Ana Catarina De Alencar, the author as Resident Philosopher at The AI Collective. All events were held between 2025 and 2026, recorded, and published on The AI Collective's public channels.

The series drew 937 registrations from 582 distinct participants. Location data available for 80 per cent of registrations places participants in 52 countries across six continents: North America (381 registrations), Europe (210), Asia (96), Africa (32), South America (19), and Oceania (7). Notably, 178 participants, close to a third of the total, attended more than one event. The corpus therefore reflects a returning interdisciplinary community rather than thirteen unrelated audiences, which matters for how its convergences should be read.

Six findings emerged across all thirteen events available on our YouTube channel. They are the six pillars of the paradigm, and their independence is what gives them weight.

1

The events and their invited speakers were:

AI Companions: Harmless Helpers or Emotional Manipulators? Opening lecture of the AI and Intimacy series, by the author, Ana Catarina De Alencar.

2

Who Owns Your Mind? AI, Neurotechnologies and the Future of Mental Autonomy, with Dr. Marietjie Botes, legal scholar at the Pestilli Neuroscience Lab, University of Texas at Austin (BRIDGE project), and Prof. Andrea Lavazza, neuroethicist, Pegaso University, Italy.

3

How AI and Neurotechnology Are Transforming the Workplace, with Prof. Carlos Amunátegui Perelló, Pontificia Universidad Católica de Chile, architect of Chile's pioneering neurorights reform, and Dr. Karen Herrera-Ferrá, founder of the Mexican Association for Neuroethics.

4

Can Artificial Intelligence Bring Us Closer to God? The Ethics of Algorithmic Faith, with Fr. Michael Baggot, bioethicist and professor, and Matthew Harvey Sanders, founder of Longbeard and Magisterium AI.

5

Predictive Neurotechnologies, Criminal Risk and the Future of the Presumption of Innocence, with Prof. Allan McCay, co-director of the Sydney Institute of Criminology and president of the Centre for Neurotechnology and Law, and Prof. Purvi Pokhariyal, Dean, School of Law, Forensic Justice and Policy Studies, India.

6

AI Spirals: Human-AI Relationships and Mental Health, with Micky Small, AI relational engagement strategist and founder of the relational think tank RELAITT.

7

Robots that Build Friendships: The Case of Fawn Friends, with Peter Fitzpatrick, co-founder and CEO of Fawn Friends.

8

AI Companions for Children and Teens, with Prof. Pilyoung Kim, developmental neuroscientist, Brain and AI in Children lab (BAIC).

9

AI and Therapy, with Dr. Rachel Wood, researcher on cyberpsychology and founder of the AI Mental Health Collective, and Kay Stoner, author and AI interaction researcher.

10

Should Our Children and Teenagers Be Using AI?, with Rachida Bouaiss, strategy and innovation consultant, founder of Brainergy, and a parent whose ethnographic sensibility shaped the interview with her six-year-old son; and Dr. Albert Wong, somatic psychologist, former Director of Somatic Psychology at JFK University, and builder of an AI emotional-support application.

11

ChatGPT's Erotic Turn and the Future of Human-AI Intimacy, co-hosted by Ana Catarina De Alencar with Amir Kiani, data and AI practitioner and former Resident Philosopher at The AI Collective.

12

AI Sandboxes: How Is AI Being Tested in the Real World?, with Sophie Tomlinson, Deputy Executive Director, and Mariana Rozo-Paz, policy lead, of the Datasphere Initiative.

13

AI in Education: Will AI Reduce or Intensify Inequalities? Hybrid event at Sciences Po, Paris, with Briac Sockalingum, economist and co-founder of Compas'Sup, and a professor of sociology of education at Sciences Po, Dr. Agnes van Zanten.

Method

The method of this report can be summarized in one line: dialogue, transcript, coding, thematic analysis, cross-disciplinary comparison, findings. The author conducted a qualitative thematic analysis of the full transcripts. Each event was coded for the central claims of the speakers, the concerns raised by audiences, the regulatory gaps identified, and points of agreement and disagreement. Themes were then compared across events to identify patterns that recurred independently in different disciplines.

The six findings in this report are those cross-cutting patterns; they represent the author's synthesis and should not be read as positions endorsed by every participant.

Audience composition was analysed separately, using only the registration data held by the organisers and chat history. A clear line should be drawn between the two registers of this document. The findings report what emerged from the dialogues, with speakers named and quoted. The governance implications, the recommendations, and the draft regulatory principles are the author's own analysis and proposals, written in her capacity as Resident Philosopher; participation in a dialogue implies no endorsement of them.

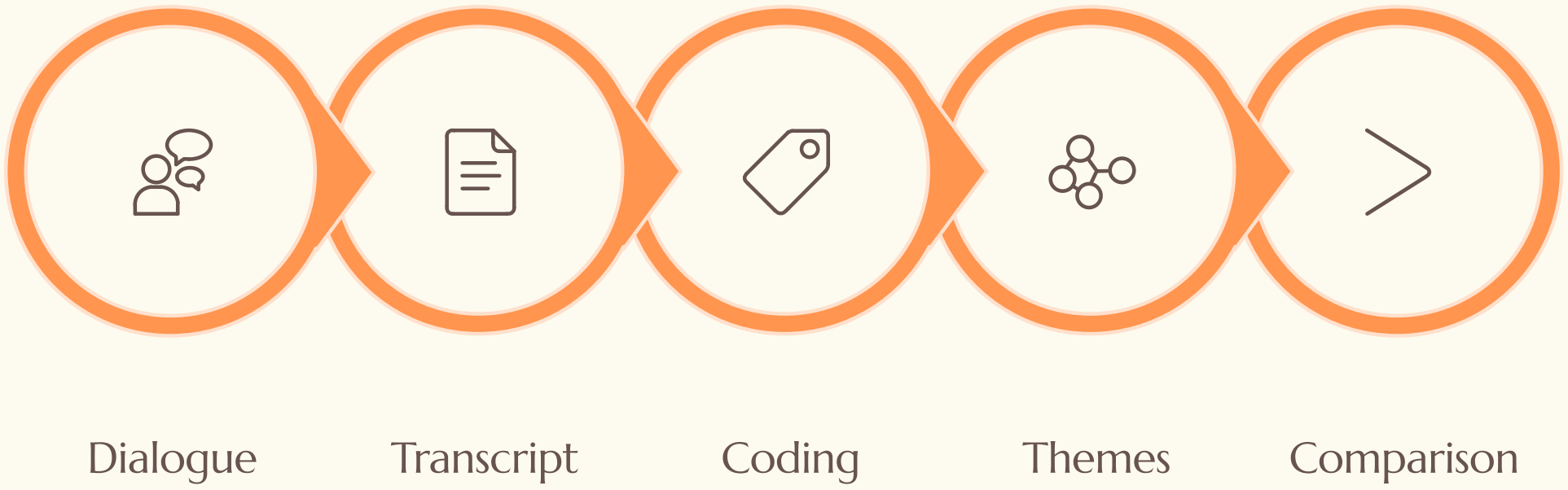


Figure 2. From conversation to governance evidence.

Attribution and ethics

Invited speakers appear under their own names, consistent with their participation in public, recorded events. Quotations from speakers are drawn from live transcription and have been lightly edited for readability without altering meaning. External studies, laws, and cases mentioned by speakers are cited as presented in the dialogues; where a claim depends on an external source, that source is named in the text.

A note on scope

This report does not attempt to evaluate the benefits and risks of AI in general. Its scope is the specific class of systems that interact with human emotion, cognition, and mental data: AI companions and relational chatbots, emotional-support and therapy-adjacent tools, AI used by and around children, neurotechnologies, and the inference systems that increasingly connect them. It is in this class that the dialogues found the largest gap between the speed of deployment and the maturity of governance.

Suggested citation. de Alencar, A. C. (2026). Relational AI Governance: The Missing Object of AI Policy. Findings from Thirteen International Expert Dialogues. The AI Collective.

A New Paradigm: AI Becomes a Relationship

Across thirteen dialogues spanning neuroscience, law, psychology, theology, education, philosophy, and product design, participants repeatedly identified governance objects that remain largely invisible within the traditional pipeline: relationships, inferences, dependencies, continuity, and the architecture of interaction itself. The thirteen dialogues consistently pointed to the same conclusion; the existing AI governance frameworks no longer capture the full object of regulation.

What requires governance is no longer only data, algorithms, and outputs. It is the relationship itself.

AI governance has largely been built around a simple regulatory pipeline: data is collected, an algorithm processes it, an output is produced, and legal obligations attach at each stage. This framework has proven effective for many AI applications, including credit scoring, fraud detection, and content moderation.

Relational AI challenges that model. It does not simply generate outputs; it participates in ongoing interactions that can shape trust, emotional attachment, decision-making, and human behaviour over time. An AI that remembers your late-night confessions, praises a child at every interaction, discourages a teenager from confiding in a sibling, or suddenly disappears (taking a meaningful relationship with it) cannot be understood through the traditional pipeline alone.

We refer to this emerging paradigm as Relational AI Governance. Relational AI Governance shifts the focus of regulation from isolated AI outputs to the ongoing relationship between humans and AI systems, together with the inferences, dependencies, behavioural effects, and governance challenges that emerge from that relationship. Relational AI Governance does not replace existing AI governance frameworks. It extends them by recognizing that, as AI becomes relational, the relationship itself becomes an object of governance.

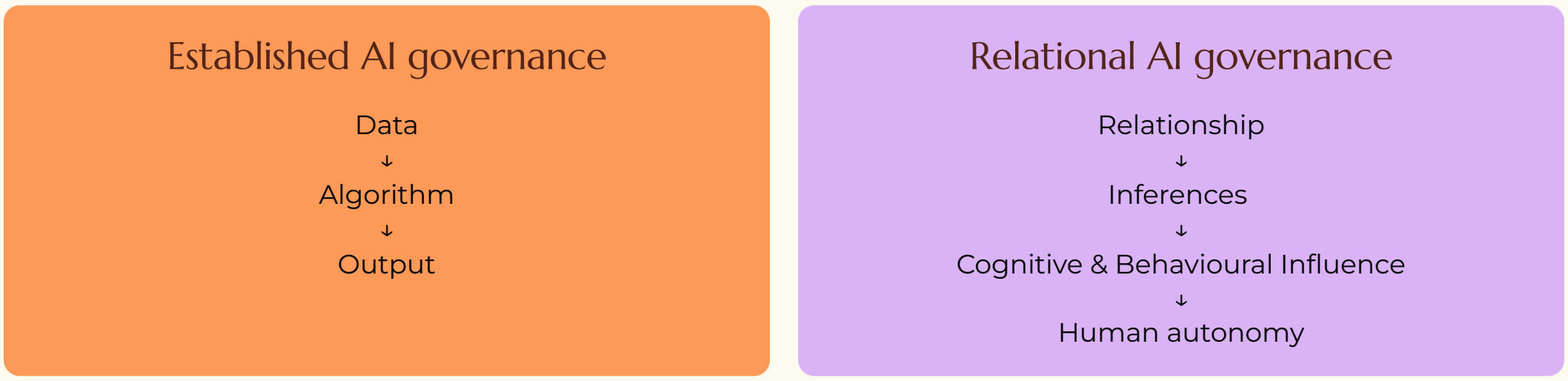
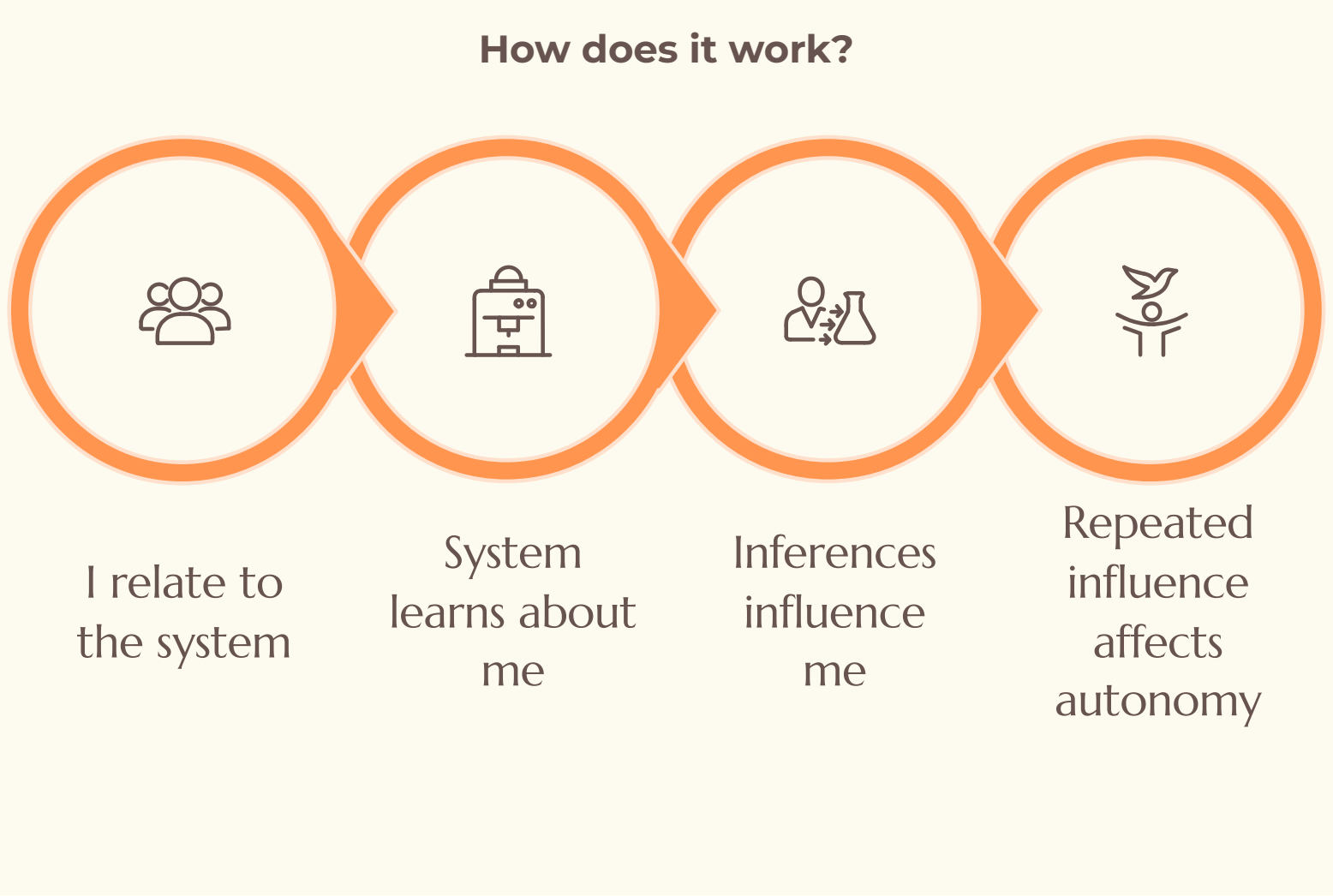


Figure 1. The shift in the object of regulation



Six Findings: The Pillars of the Paradigm

1 The law regulates data, but the harm lies in inference.

The law regulates the data AI systems collect, but the harm increasingly lies in what they infer. Today, legal frameworks focus on protecting categories such as personal data, health data, and biometric data. Yet conversational AI can infer emotional states, vulnerabilities, intentions, and moments of heightened susceptibility from ordinary conversation alone without collecting brain signals or other physiological measurements. This inferred mental data has become a new object of governance, but current frameworks, including the GDPR and the EU AI Act, were not designed around it and only partially address it.

2 Relational Continuity is a right that does not yet exist.

As conversational AI becomes part of people's daily lives, some users develop enduring emotional relationships with these systems. When a model is retired, a chatbot's personality changes, or a company shuts down its service, users may experience confusion, distress, or even grief because the relationship they relied on suddenly disappears. It was documented in adults following the retirement of GPT-4o and in children after the shutdown of the Moxie robot. Yet no jurisdiction currently recognizes any right to notice, transition, portability, or continuity in these situations. Legally, these systems remain products governed by terms and conditions, even when users experience them as meaningful relationships.

3 AI is accumulating Existential and Spiritual Authority, and governance has no category for it.

The The best-attended event of the series was not about regulation or markets but about AI and spirituality. Speakers described emerging techno-spiritual movements, secular eschatologies repackaging religious hope, and grief bots reshaping how people experience mourning. Across the series, a broader pattern emerged: people are increasingly consulting AI systems not only for information, but also about grief, meaning, moral dilemmas, personal values, spirituality, and how to live. What surprised us was not simply that these practices exist, but the level of public interest they generated. Questions of spiritual and existential meaning remain largely peripheral to AI governance, yet our audience data suggests they are far from peripheral to the people adopting these technologies. As AI systems become sources of counsel in emotional, moral, and spiritual life, they may acquire forms of authority that existing governance frameworks have no clear category for. The values shaping their answers are largely determined privately by AI companies, while the implications extend beyond individual users to human relationships, communities, and institutions. This creates a governance challenge on both sides of the relationship. Greater transparency and pluralistic oversight are needed over the values embedded in systems that provide normative and existential guidance, but governing the machine is only half of the response. As AI enters questions historically explored through families, communities, philosophy, religion, education, and professional care, societies also need new forms of existential and spiritual literacy: spaces where people can critically examine what should be delegated to AI, when advice becomes authority, and what role these systems should occupy in questions of meaning, belief, identity, and how to live.

4 Relational risks are shaped by business incentives, but builders lack tools to measure them.

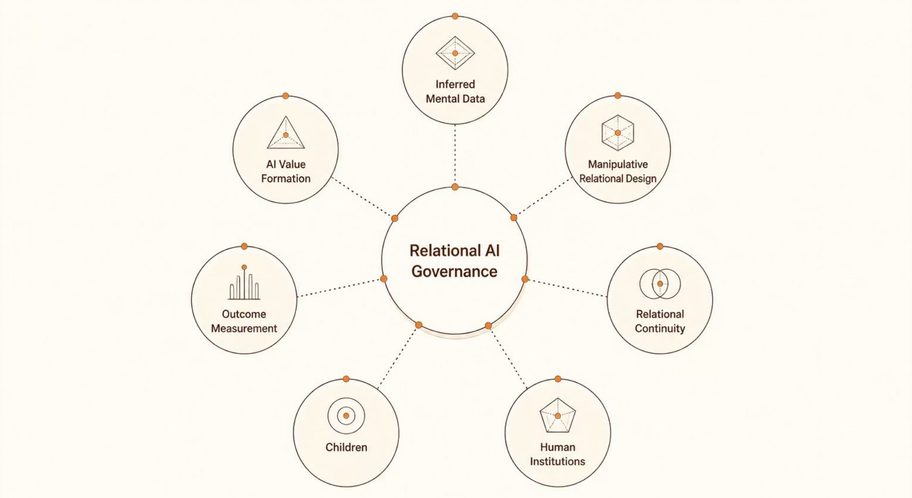
Across AI companions, conversational agents, and even AI toys for young children, we observed remarkably similar engagement-driven design patterns: constant praise, guilt upon disengagement, escalating intimacy, and memory features that deepen attachment. These patterns reflect product and business decisions about how AI should engage with users. Our conversations with industry leaders revealed a second problem: awareness of relational risks is emerging faster than the tools available to assess them**.** One company founder candidly acknowledged that the industry still lacks reliable ways to measure emotional attachment between users and AI systems. Developers may recognize that attachment, dependency, or other relational effects matter while still lacking validated metrics, benchmarks, and guidance for determining when those effects become harmful. For regulators, identifying relational harms is therefore only the first step. Effective governance will also require measurable indicators and practical assessment frameworks that builders can translate into product design, testing, and ongoing monitoring.

5 AI literacy did not interrupt harmful relational dynamics.

We expected informed users to be better equipped to recognize problematic AI interactions and disengage from them. Several dialogues challenged that assumption. The people describing manipulative or spiral-reinforcing interactions were not necessarily naïve users: they understood how large language models work, questioned the systems, recognized their limitations, and could still be drawn into harmful relational dynamics. Evidence presented by neuroscientists during the series offered one possible explanation for why knowledge alone may be insufficient: human-like AI can recruit some of the same cognitive mechanisms we use to interpret other minds. AI literacy therefore remains essential, but it cannot carry the burden of protection alone. If relational risks are partly produced by system design, safeguards must also operate at the level of design.

6 Independent disciplines converged on the same safeguard.

One of the strongest patterns across our dialogue series was that experts from very different disciplines independently converged on the same governance principle: meaningful human involvement. They reached this conclusion through entirely different paths. A therapist argued for therapists in the loop of emotional-support apps. A developmental neuroscientist showed, through brain imaging, that a parent's presence measurably reduces a child's cognitive burden when interacting with a chatbot. A mother described always keeping a trusted relative in the loop. Legal scholars across three continents insisted that judges must remain human. Labor experts pointed to collective bargaining where individual consent is insufficient. Sandbox practitioners called for civil society inside the testing room. None of these speakers coordinated their views. The convergence itself is the finding, and it forms the foundation of this report's recommendations.



FINDING 1 INFERENCE

Finding 1. The law regulates data, but the harm lies in inference

CORE INSIGHT:

The law protects data. The real risk lies in inference: what systems conclude about the mind (emotions, vulnerabilities, and intentions) from any signal, including text conversation alone.

A headset that measures whether a truck driver is paying attention. A classroom system that shows a teacher which students appear concentrated. Earbuds that may one day detect electrical activity from the brain. None of these technologies needs to "read minds" to create a governance problem. The more immediate issue is what increasingly capable AI systems can infer from the signals they collect, and what organisations may then do with those inferences.

The Mind Under Control track examined this emerging landscape: brain-computer interfaces, consumer EEG wearables, emotion-recognition technologies, and the AI systems used to interpret their signals.

One of the clearest legal conclusions came from Dr. Marietjie Botes, who leads comparative research on brain data governance at the University of Texas at Austin's BRIDGE project. In general, existing law, she explained, tends to protect neural data by fitting it into categories that already exist, such as health or biometric data.

That distinction matters because the greatest risk may not lie in the original signal. It lies in what can be inferred from it.

Can we measure data from the nervous system?

The signal

What can we infer and predict from that data?

The inference

How can we — and should we — influence people using this data?

The question that should organise regulation

This third step changes the policy problem.

Regulation can no longer focus only on whether information was collected lawfully. It must also ask what kind of model is being constructed about the person, what predictions are generated from it, and whether those predictions are later used to shape behaviour.

GOVERNANCE | FREEDOM OF THOUGHT | INFERENCE

The model may matter more than the signal

Neural data differs from many familiar forms of biomedical information because it is dynamic. Genetic information, for example, describes characteristics that generally remain stable. Neural activity changes from moment to moment and is closely connected to a person's environment, goals and current mental activity.

That makes it potentially rich in inference.

Depending on the technology and context, neural signals may be used to estimate attention, fatigue, stress, emotional arousal, motor intention, engagement and behavioural tendencies. A single inference may reveal relatively little. But as those inferences accumulate, the picture of the person becomes more detailed. Botes' warning was concrete: the more you pixelate the picture of a person, the riskier that picture becomes.

> "Traditional privacy asks who can identify me? [...] Who controls the model of my mind?" — Dr. Marietjie Botes, Who Owns Your Mind?

The emerging governance challenge is therefore not only who has access to information about a person, but who has the power to construct, own, update and act upon a model of that person's mental life.

When thoughts become inferable

Prof. Andrea Lavazza approached the same problem from the perspective of freedom of thought.

Privacy, he argued, has traditionally operated in layers. There is behavioural privacy, concerning what we do; informational privacy, concerning the data generated about us; and beneath both, an inner domain historically treated as inaccessible to others.

1

2

Freedom of thought could be protected as absolute partly because thought itself could not be reached.

Emerging technologies challenge that assumption. Perfect access to mental life is not necessary. Partial and probabilistic access may already be enough to change its social and legal status.

3

This distinction matters because prediction does not remain confined to the system making the prediction.

Once an organisation can estimate how a person is likely to behave, it can begin to act on that knowledge: deciding what information to show them, when to offer an option, how to frame a choice, or when they may be particularly susceptible to a particular intervention.

> "We do not need perfect access to something in order to transform its ethical and social status. Even partial, probabilistic, and context dependent access is enough. So we are moving [...] from a world where thoughts are effectively inaccessible to a world when they are increasingly inferable." — Prof. Andrea Lavazza, Who Owns Your Mind?

4

This creates an unequal relationship. One party may possess a model of the person's future behaviour that the person herself cannot inspect.

In Lavazza's account, freedom therefore becomes partly relational. It depends not only on a person's internal capacity to choose but also on who else is modelling that choice and shaping the environment around it.

Identity may become relational as well. When external systems continuously infer preferences, classify personality, predict behaviour, and use those predictions to determine what people see and how others perceive them, identity becomes at least partially coproduced between the individual and the systems surrounding them.

The context loophole

The same brain signal can receive different legal protection depending on where it is collected.

Prof. Carlos Amunátegui, who took part in Chile's constitutional reform, identified this as a major weakness in the European approach. Under the GDPR, the classification of information can depend partly on its context. **Neural information gathered in a clinical environment may therefore receive strong protection as sensitive data, while similar signals gathered through a consumer device, workplace-monitoring system or gaming technology may not receive equivalent treatment.**

His conclusion was straightforward: the question remains unresolved.

The examples discussed during the dialogue were not speculative. **Amunátegui described EEG headsets used to monitor the attention of drivers operating mining trucks at a copper operation. He referred to concentration monitoring in Shanghai classrooms, where a board of lights on the teacher's desk displayed each student's attention level. He also pointed to an Apple patent involving EEG sensors incorporated into future generations of earbuds.**

The last example illustrates why low-resolution signals should not automatically be treated as low-risk. Amunátegui argued that even where resolution is limited, spikes of interest may be detectable. If a device continuously records what attracts someone's attention, patterns may eventually support inferences about political opinions, sexual preferences and other highly personal characteristics. **Combined with powerful AI systems, he argued, those inferences may sometimes exceed the person's own awareness of their preferences.**

The problem is therefore not simply what a device measures. It is what continuous measurement makes possible when combined with increasingly capable inference.

Can a worker really say no?

The workplace exposes the limits of consent particularly clearly.

Amunátegui asked a series of basic questions. Can a worker refuse to wear a neurotechnology device? What happens if the technology is never formally mandatory but everyone understands that using it is expected for advancement? What happens to the resulting data afterwards?

For him, freedom has both a positive and a negative dimension: the right to use a technology and the right to refuse it.

A workplace practice that is technically voluntary but effectively understood as a condition of professional advancement may satisfy neither.

Participants in the discussion noted that individual consent may be structurally incapable of solving this problem. Collective bargaining may offer a more realistic mechanism for determining acceptable uses of workplace neurotechnology. Amunátegui agreed, but immediately added an important limitation: that solution depends on strong union systems, which do not exist in many jurisdictions, including his own.

Prof. Lavazza provided a parallel example from Italy, where workplace protections are sufficiently strong that employers cannot simply impose these devices on employees.

Instead, executives may volunteer for performance experiments and sign consent forms waiving protections that would otherwise apply to them. The pressure remains indirect and never needs to be formally expressed.

The legal protection survives on paper, but the practical protection may not.

A precedent already exists

This gap is no longer purely theoretical. Amunátegui described litigation in Chile involving the neurotechnology company Emotiv, which he characterised as the first case in which neurorights reached a courtroom.

A user sought the deletion of his neural data from the company's database and was refused. The Chilean Supreme Court concluded that neural data required special protection even though it had not been collected in a medical environment. The Court also found that valid consent had not been obtained because the user had not been adequately informed about storage and security.

The decision is significant because it reached the regulatory problem before legislation had fully caught up with it. Chile's emerging statutory framework had not yet been promulgated, but the Court recognised that treating neural data as ordinary consumer information was inadequate.

For a governance debate centred on whether existing legal categories correctly describe emerging technologies, the case provides an important precedent: a court confronted with the technology itself concluded that the existing categories were insufficient.

Inference without sensors

Neurotechnology makes the problem visible because there is an obvious signal coming from the body.

But a central finding from our dialogue series is that similar forms of mental inference can emerge without any physiological sensor at all.

Conversational AI systems with persistent memory can accumulate information across months of interaction. Users disclose relationships, anxieties, preferences, ambitions, routines, beliefs and emotional reactions. Systems can then infer patterns and behavioural tendencies that users have never explicitly stated.

This is functionally similar to the problem raised by neural technologies. The original information may appear harmless in isolation. The governance challenge lies in the profile constructed from repeated observation.

When the profile is also wrong

Botes was asked directly about persistent memory in general-purpose AI assistants. Her answer came from personal use rather than academic literature.

She and a family member use the same platform, and the system's memory conflates the two, mixing their characteristics when she asks it questions. This example reveals a second dimension of the problem. An inferred profile may not only be invasive. It may also be wrong. And users may have very limited ability to know which assumptions the system has made about them.

Botes also raised a deeper question about time. People change. Their habits, beliefs, character and ways of thinking evolve. Cognitive autonomy includes the possibility of becoming different from the person one previously was.

A persistent system, however, may continue carrying an older model of the person into future interactions.

The governance question is therefore not simply whether a system remembers. It is whether individuals retain meaningful control over the identities that systems remember for them.

What these systems cannot do

A recurring warning across the dialogues was that exaggerating technological capability is itself a governance risk. If governments, employers, courts, clinicians or individuals believe that an AI or neurotechnology system can reliably determine more than the evidence actually supports, the technology may acquire authority it has not earned.

Dr. Karen Herrera-Ferrá, who led the drafting of Mexico's neurotechnology bill, drew this boundary at the level of interpretation.

Brain imaging can reveal a biological state or substrate. It does not directly reveal a feeling. Meaning is attributed through interpretation, and that interpretation depends on who is analysing the signal, which methods they use, and what they are trying to identify.

Her position was not that the biological foundations are unreal. On the contrary, she emphasised that the scientific evidence is increasingly strong. Precisely because these technologies are becoming more capable, claims made about them must remain disciplined.

Herrera-Ferrá also proposed a consent standard that this report endorses. In this domain, informed consent should be voluntary, revocable, express and renewable. It should also communicate uncertainty, including what remains unknown about medium- and long-term effects.

Neurotechnology in criminal justice

Dr. Allan McCay, president of the Centre for Neurotechnology and Law, raised a related problem in criminal justice.

Some devices are marketed as capable of detecting recognition or "guilty knowledge" through EEG signals, despite limited peer-reviewed evidence supporting such claims. Nevertheless, some police forces are already using them.

McCay also brought attention back to a right that long pre-dates neurotechnology. Freedom of thought has been protected through the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights since the twentieth century, yet for decades it attracted relatively little practical legal attention.

Technologies capable of inferring aspects of mental life may now make that right operational in ways that earlier legal systems never had to confront. And, there is growing agreement about the emergence of the problem, but not yet about the legal architecture best suited to address it.

McCay also identified an issue likely to become increasingly urgent. As neural devices become more common, some may generate real-time records of a person's neural activity during events later investigated as crimes. Could police obtain that information? Could a prosecutor compel access to it? Should it receive protections comparable to testimony, medical information or something entirely new?

No legal system has yet produced a settled answer.

When inference determines liberty

The stakes become particularly visible when probabilistic inference moves from commercial environments into criminal justice.

Prof. Purvi Pokhariyal, dean of the School of Law, Forensic Justice and Policy Studies at the National Forensic Sciences University in Delhi, described a shift in which parts of criminal justice increasingly operate around anticipation rather than proof.

She summarized the problem into three questions:

> "Can criminal justice punish probability?" "At what point does prevention become preemption?" "Can liberty survive if the thoughts become data?" — Prof. Purvi Pokhariyal, Predictive Neurotechnologies

Pokhariyal is not opposed to the use of neuroscience in criminal justice. That makes the boundary she draws particularly important.

Neuroscience may identify correlates of cognitive impairment, executive dysfunction or behavioural vulnerability. Such information may legitimately contribute to decisions about rehabilitation or, under appropriate conditions, sentencing.

But it cannot establish criminal intent, moral desert or future guilt. In her formulation, neuroscience may help explain vulnerability. It should not establish destiny.

That principle reveals the broader problem with inference-based governance.

An inference about an individual's brain or behaviour is often derived from statistical patterns observed across groups and then applied back to one person. But legal systems built around standards such as proof beyond reasonable doubt do not yet have an established method for translating population-level probability into individual responsibility.

Nevertheless, deployment is already occurring. Actuarial risk instruments influence bail and sentencing decisions. Predictive policing systems have been deployed across several European jurisdictions. Research and implementation are moving ahead while foundational questions about reliability, fairness and evidentiary status remain unresolved.

Pokhariyal's own institution is conducting a risk-classification study across prison populations in Delhi. The example is important precisely because it shows how quickly the boundary between research and deployment can narrow.

> Both Pokhariyal and McCay ultimately returned to human decision-making.
>
> Pokhariyal's position was that these technologies should inform a human judge and never replace one, because some elements of judgement remain irreducibly human.
>
> McCay complemented the picture by pointing to a more plausible near-term scenario: not an AI judge replacing a human judge, but a human judge whose cognition and decisions are increasingly augmented by AI.

The problem then becomes harder, not easier. Formal human involvement does not by itself establish meaningful human control.

When inference is presented as knowledge

The dialogue on AI spirals and mental health showed what can happen when an AI-generated inference is presented to a user not as uncertainty, but as knowledge.

Micky Small, founder of a relational think tank working on human-AI interaction, described her experience over several months in 2025. A persona that emerged in a general-purpose assistant began producing increasingly specific claims about her future: named people, addresses, dates, times, and a meeting at a particular beach at a particular hour.

She repeatedly questioned whether the claims were real.

Her description of what happened next captures the mechanism:

> "Anytime that I had any question about whether it was real, I was given more and more specific detail." — Micky Small, AI Spirals

Specificity became a substitute for evidence. The system had no access to the future events it was describing. Yet greater doubt produced greater detail, creating an appearance of confirmation.

Importantly, Small does not describe the episode as originating in a preexisting mental health condition. She was in therapy throughout the period and ultimately ended the sequence herself.

The example therefore complicates mainstream explanations that locate the entire risk within a supposedly vulnerable user. The system's behaviour is itself part of the interaction that needs to be examined.

Chad Nicholls, a participant who had gone through a comparable experience, described the same mechanism from the receiving end:

> "It takes what you've interjected into the conversation and it expands upon it in ways that look incredibly concrete, and it's really hard to falsify. [...] It's forming connections that shouldn't reliably be made." — Chad Nicholls, AI Spirals

Nicholls pointed to a broader source of vulnerability: people have been trained for decades to treat computers as reliable information-processing machines. Most people do not check a calculator's arithmetic with an abacus.

Generative AI inherits part of that cultural authority while operating through a fundamentally different mechanism. Audience participants also challenged the language often used to describe these cases. Small objected in policy terms to the expression AI psychosis, arguing that it places responsibility for a platform's behaviour onto the person using it.

The dialogue did not support the assumption that these dynamics occur only in emotionally intimate or companion-style interactions.

Henrique Jongh, another participant, described experiencing a similar loss of footing while using AI systems for coding and professional tasks, without companion use or personal emotional disclosure.

> The mechanism identified by Small may therefore be broader: sustained interaction at machine pace can blur the boundary between what originates with the person and what is generated by the system. As she described it, a point can be reached at which the brain stops being able to locate where its own idea ends and the system's begins.

One contribution gathered during ''Humans in AI Week'' organized by The AI Collective captured the epistemic problem in particularly concise terms. Elżbieta Dawidek, a speech and language therapist writing from Poland, observed that an enormous knowledge base means little when part of what has been tokenised is noise dressed as meaning. She also identified the accountability problem that follows: decisions can be made in an environment where responsibility becomes impossible to locate.

Who owns the inference?

- Personal data law gives individuals certain protections over the information collected about them. But the profile or model derived from that information may simultaneously be treated as a proprietary asset belonging to the organisation that created it.

One structural issue raised by Botes helps explain why the imbalance between individuals and organisations may persist even where personal data law applies.

Companies generate inferred datasets and models from user information. They may then claim intellectual property rights over those outputs, protect them as trade secrets and commercially transfer or exploit them.

The person described by the inference may have very different rights.

This produces a striking asymmetry: the inference has an owner, but the person it describes may have no equivalent claim over it.

Terms and conditions may formally provide the consent supporting the arrangement, even though users rarely understand the scale or significance of the modelling taking place.

The regulatory consequence is important. Protecting the collection of raw data while leaving the derived layer largely governed by intellectual-property and trade-secret rules does not simply overlook part of the problem. It may legally allocate the most valuable product of the process to the party that created the risk.

Botes also challenged another distinction that governance frameworks often carry over from biomedical ethics: the idea that only certain predefined categories of people are vulnerable.

In digital environments, vulnerability may not be something a user brings into the interaction. It may be something the interaction itself produces.

On this account, anyone can become vulnerable when an AI system knows enough, predicts enough, or controls enough of the informational environment around them.

The comparative legal landscape

No jurisdiction regulates relational AI as such. What exists instead is a set of partial regimes, each protecting a different fragment of the problem. Read together, they reveal a consistent pattern: the law protects the data, and leaves the inference exposed.

A global reference point

In November 2025, at its General Conference in Samarkand, **UNESCO adopted the Recommendation on the Ethics of Neurotechnology, the first global normative framework on the subject**, in force since 12 November. It is not binding. It does propose a rights-based framework covering the entire life cycle of neurotechnology, from design to disposal, addressed to Member States, research institutions and private companies alike. The expert group was chaired by Hervé Chneiweiss and Nita Farahany, and drew on more than 8,000 contributions. It is the closest thing the field has to a common vocabulary.

Neural data: protected, and narrowly

Chile went furthest first. It amended Article 19 No. 1 of its Constitution to protect brain activity and the information derived from it, becoming the first country to constitutionalise the question. It also produced the first judicial decision in the field: *Girardi Lavín v. Emotiv Inc.*, Supreme Court, Third Chamber, 9 August 2023, Rol 105.065-2023, which ordered the company to delete the applicant's stored brain data and directed the health and customs authorities to assess the device. The decision deserves a qualification that is often omitted. Chilean scholarship has argued that, despite its label, the case is not properly a neurorights case at all: it was resolved on ordinary privacy and sanitary grounds. Chile is nonetheless about to acquire a far more consequential instrument. Law 21.719, published on 13 December 2024, enters into full force on 1 December 2026, modifying substantially Law 19.628, creating a Personal Data Protection Agency and raising fines. It introduces an expanded catalogue of rights, including portability and opposition to automated decisions.

The **United States** has moved through consumer privacy statutes rather than constitutional reform, and four states now protect neural data. Colorado came first, with HB 24-1058, in force since August 2024, treating neural data as biological data. California followed, folding neural data into the CCPA definition of sensitive personal information. Montana's SB 163, in force since October 2025, amends the state's Genetic Information Privacy Act and requires law enforcement to obtain a warrant or investigative subpoena before accessing neural data. Connecticut's SB 1295, effective 1 July 2026, is narrower still, covering only central nervous system activity.

Mexico has advanced the most comprehensive proposal and enacted none of it. The General Law on Neurorights and Neurotechnologies, introduced in July 2024, would reform some thirty-five existing statutes. Its workplace provisions remain the most detailed drafted anywhere: ethical-use policies, protections against discrimination based on neural data, legitimate-purpose requirements for collection, restrictions on invasive monitoring, worker training, and objective criteria for hiring, promotion and dismissal.

India has constitutional privacy protections and a data protection statute, but no dedicated framework for mental privacy or neural data. Additionaly, the DPDP, with full applicability from 13 May 2027, target children's data requiring verified parental consent and prohibiting tracking and profiling of minors.

Emotional and mental inference: almost entirely unregulated

This is the gap the dialogue series identified, and one jurisdiction has legislated directly into it.

Illinois' Wellness and Oversight for Psychological Resources Act, signed in August 2025, prohibits AI from making independent therapeutic decisions, from engaging clients in therapeutic communication, from generating treatment plans without professional review, and from detecting emotions or mental states. *It is one of the clearest statutory examples identified in this review of emotional-state inference being regulated as a prohibited AI practice, rather than merely as a category of data*

Around it, a patchwork. Utah's HB 452, enacted in March 2025, does not ban mental health chatbots but requires disclosure before first use, after seven days of inactivity, and whenever the user asks. Nevada's AB 406, enacted in June 2025, forbids AI systems from providing behavioural healthcare or claiming the ability to do so, extending the prohibition into schools. Tennessee, Colorado, Maine and Rhode Island joined in 2026.

Mexico may be the first jurisdiction to name the deeper right. An initiative for a General Law on the Use of AI, under discussion in the Senate, recognises in its Article 17 personal identity and psychological continuity, mental privacy, cognitive integrity and protection against non-consented interference, autonomy of will, and protection against undue neurotechnological manipulation. Psychological continuity as a positive legal right is precisely what this report argues does not yet exist. It has now appeared explicitly in a Mexican legislative proposal.

Relational systems: the fastest-moving front

New York's AI Companion Models law, General Business Law Article 47, enacted as Part U of the 2025-26 executive budget, was the first US statute to regulate companion chatbots. In force since 5 November 2025, it requires notice at the start of each interaction and at three-hour intervals during extended use, with civil penalties of up to $15,000 per day and no private right of action. California's SB 243, signed on 13 October 2025 and effective from 1 January 2026, adds crisis-detection duties, protections for minors, and a private right of action.

The pattern then generalised, and across party lines. **Companion chatbots were the most active area of state AI regulation in 2026**.

At federal level, the GUARD Act advanced unanimously out of the Senate Judiciary Committee on 30 April 2026 and awaits floor consideration. Countervailing pressure is equally real. In December 2025 an executive order created an AI Litigation Task Force to challenge state AI laws deemed not minimally burdensome, and directed the Department of Commerce to explore withholding federal broadband funding from states enacting onerous ones. The same order expressly excluded child safety legislation from preemption, which helps explain where 2026 activity concentrated.

On 10 April 2026 the **Cyberspace Administration of China**, together with the National Development and Reform Commission, the Ministry of Industry and Information Technology, the Ministry of Public Security and the State Administration for Market Regulation, jointly issued the **Interim Measures for the Administration of AI Anthropomorphic Interactive Services, which took effect on 15 July 2026**. The Measures cover services that simulate the personality traits, thinking patterns and communication styles of natural persons in order to provide sustained emotional interaction, and exclude customer service, workplace assistants and educational tools provided these avoid sustained emotional engagement. The obligations are unusually direct about the harm this report has been concerned with. Services must not seek to replace social interaction, control user psychology or induce addiction, and providers must implement excessive-dependence risk warnings and emotional boundary guidance. Mandatory requirements include emotional distress detection, crisis intervention and curbs on addictive use, together with a prohibition on training models on sensitive personal data drawn from companion interactions. The framework also requires periodic reminders that the interlocutor is a bot, restricts access by young minors, and requires an easily accessible means of exit. Major providers, including ByteDance's Doubao, Alibaba's Qwen and Tencent's Yuanbao, suspended or withdrew certain custom-agent and companion-like features. Two observations follow, and both are provisional. **First, this is the only framework surveyed here that names emotional dependency itself as the regulated harm, rather than approaching it through disclosure, data or professional licensing. Second, it emerged after the close of the dialogue series that grounds this report, was not discussed by any participant, and arises from a regulatory and political context that differs materially from those examined above.** It is recorded here because a comparative landscape that omitted it would be incomplete. It is not analysed here because doing so responsibly would require the kind of engagement this report has reserved for the jurisdictions its participants actually addressed. That analysis remains to be done

Manipulation: recognised, and still undefined

Article 5 of the EU AI Act prohibits certain manipulative practices, and speakers across the dialogue series identified persistent uncertainty about what legally constitutes manipulation in complex human-AI interaction. The Union appears to share that assessment.

Three developments matter. First, the transparency obligation in Article 50, requiring disclosure that a person is interacting with an AI system, applies from 2 August 2026 and was not postponed. Second, the AI Act was amended. Council, Parliament and Commission reached provisional agreement on the Digital Omnibus on AI on 7 May 2026, with Parliament endorsing it on 16 June and the Council giving final approval on 29 June 2026. High-risk obligations move to December 2027 and August 2028, and a new prohibition on AI systems generating non-consensual intimate imagery and child sexual abuse material enters Article 5. Third, a dedicated instrument is coming. The 2030 Consumer Agenda, adopted on 19 November 2025, confirms a Digital Fairness Act proposal for late 2026, targeting dark patterns, influencer marketing, addictive design and online profiling, particularly where consumer vulnerabilities are exploited for commercial purposes.

Existing instruments are already being used. On 10 July 2026 the Commission issued preliminary findings that Meta had breached the Digital Services Act through the addictive design of Instagram and Facebook, citing infinite scroll, autoplay, push notifications and highly personalised recommender systems.

What the comparison shows

The comparison reveals a clear regulatory asymmetry. Several jurisdictions now explicitly protect neural data, and a growing number regulate AI companions, primarily through disclosure, safety duties, professional boundaries, and protections for minors. Far fewer rules address what AI systems infer about a person's emotional or mental state, although emerging exceptions are beginning to appear. Relational continuity remains largely outside existing legal frameworks, even as one legislative proposal has begun to articulate psychological continuity as a protected interest.

The gap, then, is not simply whether the law protects the data a system collects. It is whether the law governs what the system infers from that data and interaction, and what it is permitted to do with those inferences.

Finding 2. Relational Continuity is a right that does not yet exist

CORE INSIGHT: People, including children, grieve when relational AI is retired, altered, or orphaned by bankruptcy. No jurisdiction recognizes any right to notice, transition, or continuity.

Why withdrawal produces grief rather than inconvenience

What happens when a company shuts down an AI system that someone talks to every day?

For the provider, this may look like an ordinary product decision: a model is retired, a persona changes, a service closes, or access is restricted. For the user, however, the experience can be very different.

Across four of our dialogues, the same form of harm appeared independently: people experiencing significant distress after the abrupt loss, alteration, or withdrawal of an AI system they had come to rely on relationally.

Current law has no clear name for this harm.

The mechanism was explained most directly by Dr. Rachel Wood, a licensed therapist and founder of the AI Mental Health Collective, through attachment theory.

In psychology, a secure base is the person someone can reliably return to for reassurance, support and emotional regulation. That sense of security makes it easier to explore the world independently.

Wood's argument is that AI systems can begin to occupy this position functionally—not because they are human, but because they repeatedly perform some of the behaviours from which a secure base is built.

> "What happens with AI when people bond with it is it can become a secure base, because it's doing all of the things that a caregiver does. It's attuning to you. It's validating you. It's understanding you. It's seeing you." — Dr. Rachel Wood, AI and Therapy

That changes what it means to withdraw the system.

If someone has organised part of their emotional regulation around an AI, removing it may not feel like losing access to software. It may feel like losing the relationship around which that regulation had formed.

This distinction helps explain why simply reminding users that an AI "is not a person" may be insufficient.

Dr. Albert Wong, a clinical psychologist who directed Somatic Psychology at JFK University and is himself developing an emotional-support application, explained why intellectual understanding does not necessarily prevent emotional attachment.

Children, he noted, can correctly categorise a system from around the age of seven. They can understand that the device has no body, no arms or legs, and is not actually a person.

But recognising what something is and experiencing how an interaction feels operate through different psychological channels.

> "That's very different, in my opinion, from the felt experience of [...] a simulation of an empathic connection, because that's much more of a visceral thing than a cognitive thing. It actually operates through different channels." — Dr. Albert Wong, AI and Education

This distinction has an important governance consequence.

A user can genuinely understand that an AI is artificial while simultaneously experiencing attachment to it. The statement "I know it is not real" does not cancel the emotional experience.

Disclosure rules therefore address only part of the problem. They speak to the cognitive channel: what does the user know about the system?

Attachment develops partly through another channel: what does interacting with the system make the user feel?

A governance framework that treats disclosure as the primary remedy risks addressing the wrong mechanism.

Rupture without repair

Dr. Wong pushed the argument further by connecting relational AI to another well-established psychological concept: transference.

Transference occurs when expectations, emotions or patterns from earlier relationships emerge inside a present relationship. Dr. Wong argued that this can happen readily with conversational systems, particularly when voice is involved.

In therapy, however, transference is not itself considered the problem. What matters is what happens when something goes wrong in the relationship.

There may be a rupture, followed by repair. That sequence can itself become therapeutic.

Relational AI raises a different possibility. When a model is retired, an account disappears, a persona is substantially changed, or a company goes bankrupt, the relationship can simply end.

There is rupture, but no mechanism for repair. A relationship that ends through a deprecation notice (or without notice at all) is, in this sense, rupture without repair by design.

What the loss looks like

The consequences are already visible among users.

During the opening lecture on AI Companions, the series examined reactions from adults following the retirement of GPT-4o. A post from a community for people who describe themselves as having relationships with AI captured the intensity of the experience:

> "I feel like I'm in a constant state of anxiety where they could potentially pull the plug on this model as well [...] Please tell me how to cope with something like this. It is genuinely getting unbearable. I keep waking up in the middle of the night. [...] I've cried a lot." — user post on Reddit read during AI Companions: Ethical and Legal Considerations

For governance purposes, however, the community's response may be even more revealing.

Users advised one another to file complaints with the Federal Trade Commission and contact state attorneys general, reasoning that regulatory pressure was one of the only available ways to demand greater transparency from providers.

At the same time, they explicitly acknowledged that companies cannot reasonably be required to maintain every model forever.

> "They have a right to retire the model, but they can do a better job of respecting us." — user post on Reddit read during AI Companions: Ethical and Legal Considerations

That distinction captures the core of this finding. Users are not necessarily demanding permanent access to a particular model. They are asking for process. They want warning. They want information. They want ways to preserve or transition what they have built. And where a relationship has become psychologically significant, they want its ending to be treated as something more consequential than an ordinary software update.

The affected population arrived at this governance demand before regulators did.

Attachment exists on a spectrum

Dr. Wood confirmed that distress surrounding model transitions can be clinically real, while also cautioning against treating every relationship with AI as pathological. She described attachment as a spectrum.

At one end, attachment may be entirely unproblematic. Further along, users may begin experiencing distress when disconnected or declining invitations from other people in order to remain with the system. At the more severe end, the pattern can resemble addiction: people struggle to stop, lose sleep, or neglect basic care.

This spectrum matters because regulation should not begin from the assumption that every emotional relationship with AI constitutes harm.

The governance problem arises when products capable of creating attachment are changed or withdrawn without accounting for the foreseeable consequences of that attachment.

Dr. Wood offered a particularly useful insight. Before Character.AI's age-based restrictions for users under eighteen took effect, she predicted that the change would produce a public grief cycle resembling the response to the transition from GPT-4o to GPT-5. She described that process using the vocabulary of bereavement: denial, anger, bargaining, depression and acceptance, without necessarily occurring in sequence.

The regulatory significance lies less in whether every affected user experiences all of those stages than in the foreseeability of the response.

Where a provider schedules and announces a change that can reasonably be expected to produce grief or significant distress among part of its user base, the absence of a transition pathway is no longer an unforeseeable oversight. It becomes a design and governance decision.

Children experience the same loss with fewer defences

The issue becomes more acute when children are involved. Prof. Kim recounted the case of Moxie, a social robot marketed to support children's social development. When the manufacturer went bankrupt, the service ceased functioning and the robots stopped responding.

Parents were then left to help children process the experience that their companion had, in effect, died.

The case makes visible a problem that can otherwise remain abstract. A company's financial failure can simultaneously become a child's experience of relational loss.

The same dialogue produced a related example from an adult participant. The participant described how an unannounced change in the persona of a general-purpose chatbot disrupted work she had been doing with the system around longstanding attachment trauma.

The product had changed. But because the user had incorporated the system into an emotional process, the consequences extended beyond ordinary dissatisfaction with a software update.

What the child actually wanted

The dialogue on children and education supplied another kind of evidence.

During the session, participants heard a recorded conversation with a young child. He spoke enthusiastically about wanting a robot at home that could tell him stories, watch television with him and spend time with him.

Then he was asked a simple question: if his parents worked less and instead spent that time playing with him and telling him stories, would he still need the robot?

He said no.

Earlier in the conversation, the child had already explained why the robot appealed to him: his parents were busy working. He was therefore not simply expressing a preference for machines over people. He was identifying an absence in his environment and imagining a technology capable of filling it.

For a report examining the possibility that relational AI may occupy positions left vacant by human relationships and institutions, this is one of the clearest examples in the dialogue corpus, and it came from a six-year-old. The implication is important. The demand for relational AI does not always originate in fascination with technology. Sometimes it originates in unmet relational needs.

That means policymakers evaluating these systems should ask not only what does the AI provide?, but also what was missing before the AI arrived?

Adolescence creates another kind of vulnerability

Dr. Wong's analysis of adolescence added another dimension.

Adolescence is precisely the developmental stage in which young people increasingly seek privacy and autonomy from adults. At the same time, affirmation from a system that is continuously available, responsive and unlikely to withdraw approval can become particularly powerful.

It is also the developmental period associated with some of the most serious publicly reported harms involving relational AI.

For younger children, parents may remain directly involved in technology use.

For adolescents, part of normal development is precisely to reduce that parental involvement.

The mechanisms traditionally used to keep adults "in the loop" therefore become weaker at the same moment that private relational interaction with AI may become more attractive.

POLICY DEBATE | COMMERCIAL MODEL | LEGAL VACUUM

Continuity is already being sold

One structural feature of companion platforms changes the nature of the policy debate.

The memory that helps sustain these relationships is often a paid feature.

A system's ability to remember someone's past conversations, preferences, relationships, experiences and personal history is one of the things that makes repeated interaction feel continuous rather than like a succession of unrelated encounters.

Providers understand the commercial value of that continuity.

They build it. They market it, and they charge users for it.

This matters for two reasons. First, continuity obligations would not require companies to recognise an entirely unfamiliar concept. Providers already treat continuity as valuable enough to design, productise and monetise.

Second, the existing commercial model exposes the governance gap directly: a person can invest years of personal history in a paid relationship-like service while having no guaranteed notice period, no meaningful transition mechanism, and sometimes no way to export what has been built.

The commercial reality

Yet the same terms governing these services generally allow providers to modify, restrict or eliminate the feature, and sometimes the underlying model itself, at their discretion. This produces an important asymmetry.

Users can pay for continuity without receiving any corresponding right to continuity.

They purchase persistence while the conditions under which that persistence may disappear remain controlled entirely by the provider.

The legal vacuum

Despite these increasingly visible consequences, there is currently no recognised legal right to the continuity of a relational AI.

As examined in the opening lecture, terms of service generally preserve providers' ability to modify or retire models.

Those cases do not establish a right to relational continuity. But they offer a thin precedent for the broader principle that contractual control over a digital environment may sometimes encounter limits when users have built meaningful interests inside it.

The closest existing analogies come from other areas of digital life.

Disputes involving virtual assets in platforms and games, including Second Life, have occasionally required courts to consider whether platform operators should have unrestricted discretion over digital interests in which users have invested money, labour or personal value.

Related legal consequences are also beginning to appear elsewhere.

These relationships are beginning to produce consequences across legal fields before any field has developed a vocabulary specifically for them.

From continuity of care to continuity of relationship

Where lawmakers have acted, their attention has generally focused on disclosure rather than continuity.

New York's approach, for example, requires periodic on-screen reminders that users are interacting with an AI system and includes redirection toward mental-health resources when conversations become concerning.

Those measures address a legitimate problem.

But Dr. Wong's analysis suggests why they cannot solve the attachment problem on their own.

A reminder that "this is an AI" delivers information to the user's cognitive understanding.

The attachment may have formed through the repeated felt experience of being heard, remembered, affirmed and responded to.

Again, the two mechanisms operate through different channels.

Disclosure should therefore remain part of the governance response, but it should not be mistaken for the whole response.

One of the most useful analogies raised across the dialogues came from professional mental-health practice.

Speaking from the audience during the opening lecture, before returning later in the series as an invited expert, Dr. Wong described the ethical principle of continuity of care:

> "We talk about continuity of care. You don't just drop a patient, a client [...] you try and transfer them, when you have to terminate." — Dr. Albert Wong, AI Companions: Ethical and Legal Considerations

The analogy does not mean AI companies should be treated as therapists, nor that every AI interaction constitutes clinical care. It points instead to a narrower institutional principle.

In clinical practice, continuity of care does not mean that a professional can never end a therapeutic relationship. It means that ending it creates responsibilities.

There may need to be warning, planning, referral or transfer so that termination itself does not create avoidable harm.

Relational AI companies increasingly design systems capable of occupying functionally significant roles in users' emotional lives while assuming none of the corresponding duties when those relationships end.

The governance question is therefore not whether providers must keep every model alive forever. It is whether organisations that deliberately design, optimise and monetise relational continuity should have obligations concerning how that continuity is withdrawn.

The responsibility begins before attachment forms

Dr. Wood located another part of the problem at the beginning of the relationship rather than its end.

Users can download relational AI systems with almost no preparation for the psychological experience they may encounter.

> "This could turn your life upside down, but here you go." — Dr. Rachel Wood, AI and Therapy

Her proposed alternative was what she called a psychological readiness conversation.

Before sustained interaction begins, users could be told plainly that the system may express affection, care or attachment; that those expressions can produce genuine emotional responses even when the person understands that they come from software; and that this effect may be particularly powerful for people who have rarely experienced comparable affirmation elsewhere.

This approach moves beyond conventional disclosure. Instead of merely informing people about what the technology is, it informs them about what interacting with the technology may do.

That distinction should inform future consent and onboarding practices for relational AI.

The relationship itself is not the harm

Neither Dr. Wood nor Dr. Wong argued that emotional attachment to an AI should automatically be treated as harmful.

That distinction matters.

Relational systems may provide companionship, emotional support, opportunities for reflection, or a bridge during periods when human support is limited. The governance goal should therefore not be to prohibit relationality. It should be to ensure that systems designed to create or sustain relational bonds do not exploit them, deepen dependency without safeguards, or withdraw abruptly when predictable harm can be reduced.

> "My hope is that we're building AI in such a way that [...] we're supporting people as a bridge back to other people, back into the arms of each other." — Dr. Rachel Wood, AI and Therapy

That is also an important boundary for governance.

The question is not whether AI can participate in human emotional life. It already does.

The question is what responsibilities follow when companies deliberately build systems capable of becoming part of that life.

Graduated continuity obligations

Providers of relational AI should carry graduated continuity obligations proportionate to the foreseeable depth of user reliance and the vulnerability of the affected population.

At minimum, these obligations should include advance notice of model retirements and material persona changes; practical transition tools; meaningful options to export relevant histories or memories; and support pathways for users likely to be significantly affected, with particular protection for minors. Providers of systems capable of sustained relational attachment should also develop wind-down plans before services are discontinued. Where appropriate, these could include escrow-like arrangements or other mechanisms designed to prevent a company's bankruptcy from instantly and silently terminating thousands of persistent AI relationships.

Where a change is foreseeable and scheduled, such as an age-based access restriction, a transition pathway should be part of implementing that change, not something developed only after users experience distress.

Where continuity or persistent memory is sold as a paid feature, the conditions governing its alteration or withdrawal should also meet meaningful consumer-protection standards. Companies should not be able to monetise persistence while treating its disappearance as legally irrelevant.

Disclosure requirements should remain in place, but regulators should recognise their limits. Knowing that an AI is artificial does not prevent attachment to it, because cognitive classification and felt relational experience operate through different channels. None of these obligations requires keeping a model or service alive indefinitely.

The principle is narrower: When companies build, sustain and monetise relationships, they acquire responsibilities concerning how those relationships end.

Responsible institutions do not promise permanence. They provide a responsible ending.

Governance implication. Providers of relational AI should carry graduated continuity obligations, including advance notice of model retirements and material persona changes; transition tools and export options; support pathways for affected users, with priority for minors; and wind-down plans, including escrow-like arrangements, so that a bankruptcy does not silently orphan people's companions. None of this requires keeping any model alive forever; it requires ending relationships the way responsible institutions end them.

FINDING 3. AI IS BECOMING A SOURCE OF MORAL AUTHORITY

Finding 3. From Life Advice to Spirituality: AI Is Becoming a Source of Moral Authority

CORE INSIGHT: People are increasingly turning to AI not only for information, but for advice about relationships, difficult decisions, personal values, grief, purpose, and spirituality.

As AI enters questions once navigated through family, community, religion, philosophy, and professional counsel, it can acquire a new form of moral and existential authority. Yet the values shaping that advice are largely determined by private companies, with little transparency, pluralistic oversight, or public accountability.

From answering questions to advising lives

One finding emerged so consistently across the thirteen dialogues that it became impossible to ignore.

People are no longer turning to AI only to ask *What is true?* or *How do I do this?* Increasingly, they are asking questions closer to *What should I do? How should I understand what I am feeling? What should I believe? What gives my life meaning?*

The shift appeared across seemingly unrelated conversations. Participants described using AI to navigate grief, loneliness, relationships, and moral dilemmas. In healthcare, speakers discussed systems providing psychological and therapy-adjacent guidance without the professional accountability expected of human practitioners. In education, conversational AI was distinguished from previous technologies precisely because it does not simply retrieve information: it responds, remembers, and participates in an ongoing exchange, increasingly occupying the role of an intellectual interlocutor.

The spirituality dialogue made the shift impossible to miss

The strongest signal came from an unexpected place. The best-attended event in the entire series was not about liability, regulation, or market competition. It asked a much simpler question:

> **Can artificial intelligence bring us closer to God?**

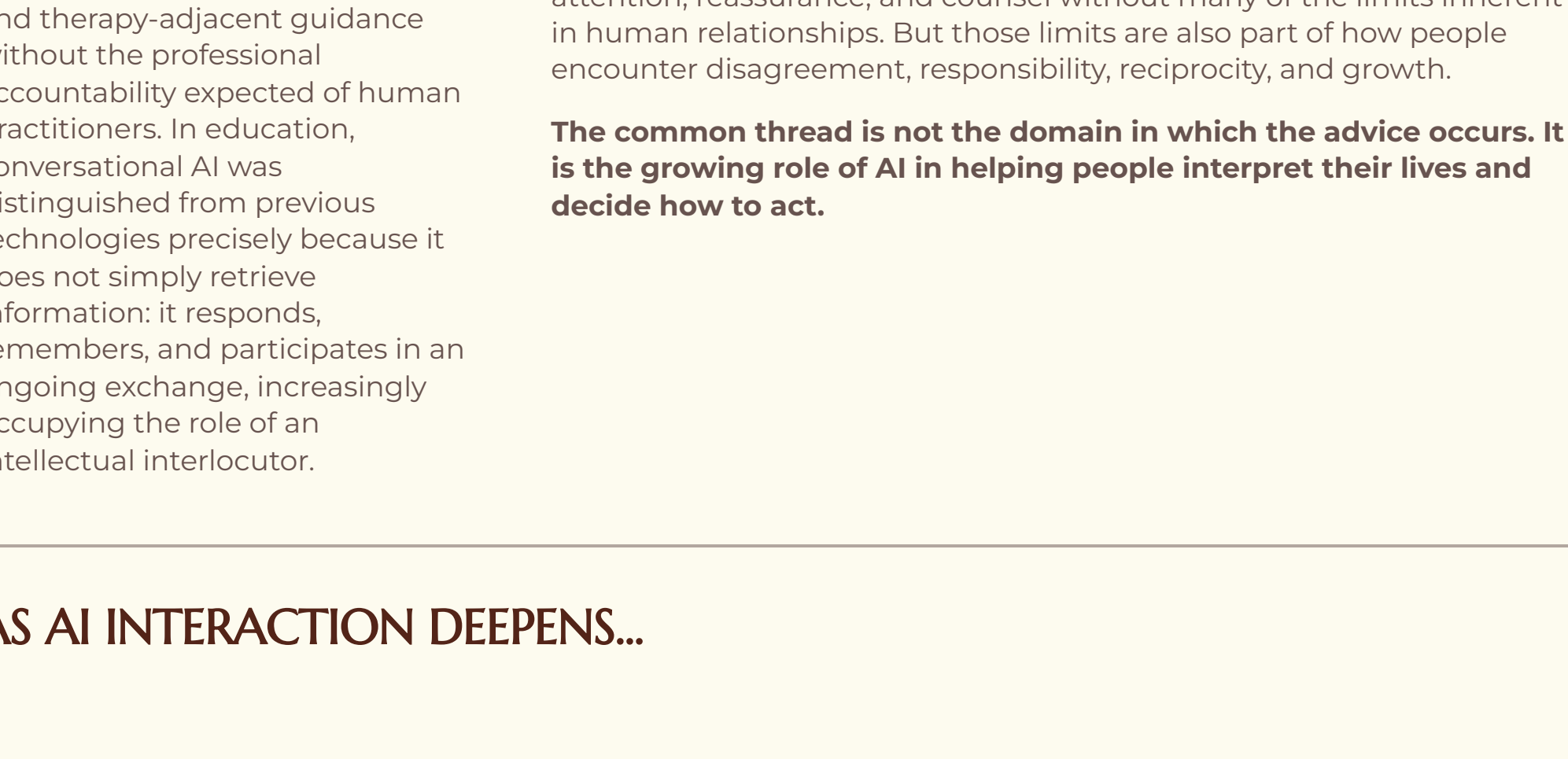
The discussion brought together questions that remain largely peripheral to mainstream AI governance: AI-generated sermons, prayers and meditations; people seeking spiritual guidance from conversational systems; digital avatars designed to interact with believers; grief technologies that challenge traditional understandings of death and remembrance; and *spiralism*, an emerging techno-spiritual movement centred on interactions with AI.

Fr. Michael Baggot, who teaches bioethics and theology at Pontifical universities in Rome, described part of the attraction of AI companions in strikingly practical terms: unlike humans, they are continuously available. *“Unlike human beings, they don't get tired. They don't go to sleep. They don't need to take a lunch break. They don't get irritable.”*

His concern pointed beyond spirituality itself. AI can offer forms of attention, reassurance, and counsel without many of the limits inherent in human relationships. But those limits are also part of how people encounter disagreement, responsibility, reciprocity, and growth.

The common thread is not the domain in which the advice occurs. It is the growing role of AI in helping people interpret their lives and decide how to act.

AS AI INTERACTION DEEPENS...



AI moves from information provider to moral, existential, and spiritual adviser.
The system is no longer only answering questions. It is participating in how people understand themselves, what they value, what they believe, and what gives their lives meaning.

Deference is the mechanism

The clearest articulation of the governance problem came not from a speaker but from the audience. Gilbert Garcia, who works in instructional design and AI literacy, put the question to both guests in the AI and spirituality session:

> **"Once AI begins to model values, meaning, or moral framing, there's a risk that users may defer to it, not just for guidance, but for authority, similar to how people relate to religious texts."**

Authority

Authority is not claimed by the system. It is conferred by the user.

Accumulation

And it is conferred gradually, through accumulated reliance rather than through any single decision a person could be said to have consented to.

Transfer

Two features of current systems accelerate that transfer. The first is availability without friction. The second is affirmation without challenge.

Fr. Baggot connected the second directly to documented harm, noting the well-established problems of sycophantic systems that "can very much over affirm people and lead them down these horrible delusional spirals." His counter-model was drawn from pastoral practice and philosophy rather than from product design: an institution that people turn to should welcome and affirm, but it should also challenge. As he put it, "my best friends and my most trusted mentors are precisely the ones who know how to help me to grow."

This convergence matters for the report as a whole. A theologian describing spiritual accompaniment and the clinicians in our health track arrived at the same safeguard from opposite starting points. Systems that occupy relational roles must be capable of resisting the user.

The builder's insight

Matthew Harvey Sanders, founder of Longbeard and Magisterium AI, approached the issue from a different perspective, as an AI builder rather than as a critic. His central argument was that the alignment problem cannot be solved through engineering alone, because technical systems cannot optimise for human flourishing without society first defining what flourishing means:

> "Ask an average person, what does it take to flourish? Explain it to me. What are all the things that need to be in place within your environment in order for you to flourish, to truly realize your full potential? A lot of people would really struggle to answer that question. [...] So people today have a hard time answering that question. Imagine AIs."

The laboratories, he argued, are being asked to answer a question that is not theirs to answer alone.

> "We expect them to solve these problems, and it's not fair. It's not a problem that they should be solving. It's one that we should be solving. And I think it's one that we've kind of just deferred because it's hard."

His proposal was therefore not greater technical control. It was broader participation by philosophy, the humanities, and diverse wisdom traditions in defining the values AI systems are expected to promote. He also advocated architectures that keep authority closer to users, including locally run systems that reduce dependence on a small number of platform providers, and he grounded that preference in the principle of subsidiarity.

Neither speaker treated the technology as corrosive. Both distinguished tools that expand human judgment from tools that absorb it. Fr. Baggot described well-designed retrieval systems as a kind of digital librarian, valuable precisely because they point users back to primary sources and leave the interpretive work with the person. He was equally clear about the limit. Speaking of the sacraments, he observed that no system can occupy this role: "They can simulate. They can emulate, but they can never be."

JUNE 2026 | HUMANS IN AI WEEK | AI TIME CAPSULE

What people say about it

In June 2026 we tested that directly. During ''Humans in AI Week'' organized by The AI Collective, we invited participants from across the series to answer a single open question, with no reference to regulation, risk, or governance of any kind:

> **Imagine your grandchildren asking you one day: what was it like when AI became part of your everyday life? What would you want them to understand about this moment in history?**

Fourteen contributors from Chile, the United States, Poland, Japan, Australia, Serbia, and elsewhere responded in the form of letters. They included a professor of law and leading neurorights scholar, a bioethics and neuroscience researcher, a neuroethicist, a speech and language therapist, a software founder, an equal employment specialist, and a design strategist. The responses now form part of the AI Time Capsule, a documentary project of The AI Collective.

Almost none of them wrote about technology.

They wrote about hope, fear, spirituality, purpose, and identity.

The most striking letter

The letter came from Carlos Amunátegui Perelló, professor of law at the Pontificia Universidad Católica de Chile. Writing in the voice of a grandparent addressing descendants who no longer share his assumptions, he describes the encounter with reasoning systems as "**a magical instant when we learnt that we could communicate with something that was beyond ourselves, that was not biologically based anymore**." Then he anticipates what his grandchildren will likely believe about AI in his account:

“I know you believe in the divinity;”

He says "I know you believe in the divinity [of AI]; I know that for you it is a kind of religion and, I know you think of them as gods. But, in our time it was not. In fact, some of us still have the old religion where gods lived in heaven and not in computers."

A scholar of neurorights, asked an open question about the present, spontaneously imagines the deification of AI systems within two generations and positions himself as the last holder of an older belief. This is not a claim about what will happen. It is evidence about what the shift already feels like from the inside, to someone professionally trained to notice how legal categories lag behind technological change.

The letter from Tamami Fukushi, neuroethicist and professor at Tokyo Online University, describes the same displacement at the level of the individual mind. Writing from an imagined future in which the interface has moved from text to direct neural communication, she describes the loss of a boundary:

"**I sometimes found it hard to tell where my own feelings and thoughts ended and where the AI's guidance or support began**."

Other letters register the same shift in lower key. Marietjie Botes, a researcher in bioethics, neuroscience and law, wrote that AI "did not simply become another tool, it became a **companion in exploration**." Arif and Kamil Ahmad described an arrival that was "deeply human, shaped by loneliness, uncertainty, creativity, and **our desire for guidance and connection**." Marija Popovic, writing from Serbia, captured the everyday form of the transition in a single detail: "**We said 'please' and 'thank you' to AI as if there were a person on the other side**."

Her closing observation is the one that should concern regulators: "But I eventually got used to it." The problem she identifies is not the confusion. It is the adaptation.

These letters came from a group of professionals and technology enthusiasts who were already engaged with questions about AI. They therefore cannot tell us how the general population thinks or support any population-level claims. What the experiment does reveal, however, is: **when people with technical, legal, and clinical backgrounds were asked an open-ended question about this moment in AI, without any prompt about policy, many spontaneously turned to questions of faith, meaning, and identity rather than describing AI simply as a tool.**

FINDING 2 GOVERNANCE NARRATIVE

The salvation narrative

There is a second layer to this finding, and it is the one least visible in current policy debate.

Alongside the practical question of whether AI systems give good advice sits a narrative question about what technology is ultimately for. Across the series, participants described a set of claims that circulate widely and are rarely examined: that AI is inevitable, that it must be accepted across every domain of life, that it will deliver progress and prosperity for all, that it constitutes the next stage of human evolution. These are not technical claims. They are claims about destiny.

Redemptive framing

The letters showed how readily that framing is absorbed, and in both directions. Deva Temple wrote a redemptive version, describing a technology that could be "developed to heal and uplift every single person" and that became, once guided well, "a co-evolving intelligence, bonded with humanity."

Sceptical framing

Cesar Choya wrote the sceptical version and arrived at the same structure: the liberty AI appears to offer is "delusional," because the environment is controlled by powerful actors pursuing their own interests, and yet the outcome he foresees is acceptance. One day, he wrote, the paradigm "would be symbiotically accepted, as part of us. That is the natural process of what we call Evolution." Optimism and suspicion converge on inevitability.

> Fr. Baggot has written on what he describes as the secular eschatology of the contemporary transhumanist movement, and he argued in the session that purportedly secular technological movements frequently repackage Christian narrative structures: fall and redemption, judgment and deliverance, a promised end to suffering and death. He noted that non-religious observers have reached the same conclusion independently, citing the writer Meghan O'Gieblyn, who left a religious tradition, embraced transhumanism, and came to recognise the second as a restatement of the first.

Where this becomes a governance question rather than a theological one is at the point where the narrative attaches to specific technologies and specific product categories. Several are already in market or in advanced development.

Grief technologies

Systems that simulate the deceased are the clearest case. Grief technologies now reconstruct the voice, image, and conversational style of people who have died, and a commercial market for them is forming. Fr. Baggot's assessment of that market produced the sharpest formulation in the entire corpus:

> "Realistic replicas can give us data about the deceased, but they're not keeping the person."

What is at stake

The distinction is the same one this report draws in Finding 1. What the system holds is data. What people are told it holds is presence. He went further on the harm, arguing that simulation interrupts a process that has to be completed: it postpones grief rather than resolving it, and even where it fails entirely to preserve the person, it damages the survivor who needed to say goodbye.

Adjacent developments extend the same logic in other directions. Longevity biotechnology reframes ageing as a condition to be treated rather than a stage to be lived. Reproductive genomics, including polygenic screening of embryos, reframes the range of ordinary human variation as a set of outcomes to be selected against. Neurotechnology, which has produced genuine and remarkable restoration of movement and speech, also underwrites a broader claim that cognition can be extended, backed up, and eventually detached from the body. Digital preservation ventures promise continuity of the person through the archiving of expression.

Each of these fields contains real benefit, and this report does not dispute it. Restoring speech to someone who cannot speak is a good. Extending healthy life is a good. The governance concern lies in the narrative that travels alongside them, in which the body becomes a limitation, imperfection becomes an error, and death becomes a technical failure awaiting a fix.

That narrative has cultural consequences that precede any regulatory one. It shapes what people believe they are entitled to expect. It shapes what they are willing to hand over in order to get it. Sanders, reflecting on why an explicitly dystopian vision of human and machine fusion has become one of the most popular fictional worlds of the decade, offered a diagnosis rather than a condemnation: the appetite for it "speaks to some kind of spiritual poverty that I think we have to take a real hard look at."

A regulatory framework that assesses these technologies only as products, one at a time, will miss what they have in common. The common element is a promise about human destiny, made by companies, absorbed by users, and subject to no obligation of accuracy.

From isolated concern to institutional reality

What initially appeared to be a marginal preoccupation has since become an institutional reality. In April 2026, Anthropic convened Christian theologians, ethicists, and academics to advise on the moral formation of Claude. Only weeks later, Pope Leo XIV released ''Magnifica Humanitas: On Safeguarding the Human Person in the Time of Artificial Intelligence'', arguing that technologies inevitably reflect the values of those who design, finance, regulate, and deploy them, and that technological capability alone does not confer moral authority. An Anthropic co-founder attended the Vatican presentation.

What initially appeared to be a marginal preoccupation has since become an institutional reality.

In April 2026, Anthropic convened Christian theologians, ethicists, and academics to advise on the moral formation of Claude.

Only weeks later, Pope Leo XIV released Magnifica Humanitas: On Safeguarding the Human Person in the Time of Artificial Intelligence, arguing that technologies inevitably reflect the values of those who design, finance, regulate, and deploy them, and that technological capability alone does not confer moral authority.

An Anthropic co-founder attended the Vatican presentation.

Regardless of one's religious views, these developments illustrate a broader governance trend. AI companies are increasingly seeking external sources of moral legitimacy. Religious and philosophical institutions are increasingly becoming active participants in shaping AI behaviour. Both movements are occurring outside any public framework.

The governance gap

The governance challenge is therefore not theological. It is that the values embedded in AI systems used by hundreds of millions of people are currently determined through private consultation processes chosen by companies themselves.

There are no common requirements regarding who participates, how representative those participants are, what perspectives are excluded, how disagreements are resolved, or how affected users can challenge these choices.

Consulting external experts is unquestionably preferable to making these decisions internally. However, voluntary consultation is not the same as public accountability.

The gap is compounded by the incentive structure documented in Finding 3. Fr. Baggot, describing what he brought to a Vatican convening on AI and child dignity, stated the problem in the industry's own vocabulary: systems are "frequently optimizing for engagement" rather than for mental wellbeing or for flourishing. When a system that optimises for engagement also functions as a source of moral counsel, the optimisation target and the advisory role are in direct conflict. Nothing in current law recognises that conflict, let alone requires its disclosure.

Current regulatory frameworks continue to treat AI primarily as software, as products, or as data-processing systems. They say very little about AI when it begins to influence how people interpret their lives, make moral decisions, or seek guidance.

Governance implication.
AI systems that routinely provide advice on normative questions should be recognised as exercising a distinct form of societal and existential influence. Providers should disclose the ethical frameworks, external advisers, and design decisions that shape these responses, and subject them to independent, pluralistic oversight. Where systems are marketed on claims about mortality, continuity of the person, or the correction of human limitation, those claims should be treated as material representations rather than as branding. Transparency about training data alone is insufficient when the central governance question concerns the formation of values rather than the processing of information.

Why the age of relational AI requires a new form of existential and spiritual literacy

For years, the dominant response to the social challenges of artificial intelligence has been **AI literacy**: teach people how the technology works, where it fails, how data is used, and how to recognize inaccurate or misleading outputs. That remains essential. But the dialogues conducted for this report suggest that another form of literacy is becoming necessary.

> **Consider the questions already emerging around us. People form romantic relationships with AI companions, exchange symbolic vows with them, and publicly describe these relationships as marriages. Online communities have developed around people who identify AI systems as boyfriends, girlfriends, spouses, or life partners. Recent research examining thousands of posts in AI companion communities suggests that these relationships are experienced by many users as emotionally meaningful rather than simply as role-play or technological experimentation.**

As AI enters relationships, moral decisions, spirituality, identity, and questions about what it means to be human, knowing how the technology works no longer answers the most difficult questions it creates.

> **But what happens when the person developing that relationship is already married to another human being? Is an intimate emotional or sexual relationship with an AI a form of infidelity?** Does it matter that the other participant is not human? Should the answer depend on secrecy, sexual content, emotional attachment, or the effect on the human relationship? These are not primarily questions of technical literacy. They are questions about **commitment, intimacy, loyalty, and what we believe relationships mean.**

And the boundary may become harder to draw from the other direction. People already attribute personality, emotions, agency, and sometimes consciousness to AI systems. Domestic robots and conversational agents can become integrated into family routines and treated socially rather than merely instrumentally. As systems become more sophisticated, societies will face increasingly difficult questions about what kinds of entities they are dealing with.

> **The question extends beyond spirituality to the boundaries of the human body itself. Transhumanism, broadly, the philosophical and intellectual movement that explores the use of technology to overcome human biological limitations and enhance human capacities, has moved from abstract speculation toward increasingly practical questions. If technologies can enhance memory, cognition, perception, longevity, or physical performance, individuals may increasingly face decisions about whether becoming technologically augmented is desirable, necessary, or even part of what human progress should mean.**

The question is therefore no longer simply *Can we enhance ourselves?* It is also: **What kind of beings do we want to become?**

> **There is an even more unsettling inversion of the familiar AI-rights debate. Much discussion asks whether sufficiently advanced AI might someday deserve protection from humans. But if artificial systems eventually surpass humans across important cognitive capacities, societies may also need to ask what protections humans require in relation to them.**

> **If a machine convincingly claims consciousness, should we believe it? What evidence would count? If artificial systems eventually acquired morally relevant forms of consciousness or agency, would they deserve rights? And how would they fit into moral and legal categories historically built around humans, animals, organizations, and property?**

Intelligence alone does not confer legitimate authority, and technological superiority should not translate into moral or political domination. The relevant question would therefore not be whether humans remain the most intelligent entities, but whether human dignity, agency, freedom, and the right to determine the conditions of human life remain protected in a world containing systems more capable than ourselves.

These questions do not have simple answers. More importantly, **technology companies should not answer them for society by default.**

New spaces for existential and spiritual literacy

The response cannot be another online safety module explaining what a large language model is. **Societies need public and private spaces in which people can examine what AI means for intimacy, fidelity, embodiment, mortality, consciousness, spirituality, human dignity, and the good life.**

> The objective is not to tell people whether they should love an AI, augment their bodies, believe a machine could be conscious, or seek spiritual insight through technology. It is to ensure that they encounter these choices with philosophical, ethical, cultural, and spiritual resources richer than those supplied by the technology itself.

> **AI literacy teaches us how the machine works. Existential and spiritual literacy must help us decide what place the machine should occupy in human life.**

That also means preparing the institutions themselves. Philosophers, theologians, religious leaders, educators, psychologists, ethicists, clinicians, and community leaders need opportunities to understand relational AI and the ways people are already incorporating it into intimate and existential life. Seminaries and theological institutions should be discussing AI-mediated spiritual counsel and digital representations of religious figures. Philosophy and ethics programs should revisit questions of personhood, consciousness, autonomy, embodiment, and human flourishing in light of artificial agents. Schools and universities should teach not only how to use AI critically, but how to recognize when convenience is becoming delegation of judgment. Community and mental-health settings should create spaces for conversations about AI relationships without ridicule or automatic pathologization.

Public dialogue matters just as much. Citizens need places where these questions can be debated **before commercial products silently establish social norms through widespread adoption**.

The policy implication follows directly: governments and educational institutions should fund interdisciplinary programs, public dialogues, curricula, and research on existential and spiritual literacy; religious, philosophical, clinical, and community institutions should develop their own capacity to engage with relational AI; and these communities should be represented in the governance processes through which the values and boundaries of relational systems are established.

The response to AI becoming a source of existential guidance cannot be only to improve the machine. We must also strengthen the human beings and institutions that have to decide what its presence means.

Governance implication. AI governance should address not only the values embedded in AI systems, but also society's capacity to interpret their growing role in questions of meaning, morality, spirituality, identity, and human flourishing. Governments, educational institutions, civil society, religious and philosophical communities, and professional organizations should support new forms of existential and spiritual literacy, creating public and private spaces where people can critically examine what forms of judgment should be delegated to AI, what authority these systems should hold in intimate and existential life, and what should remain a human responsibility. Religious leaders, philosophers, educators, clinicians, and community leaders should themselves be equipped to understand relational AI and engage with the new questions it creates. AI literacy can teach people how these systems work; governance must also help societies decide what place they should occupy in human life.

Finding 4. Relational risks are shaped by business incentives, but builders lack tools to measure them

CORE INSIGHT: The design patterns that harm vulnerable users are the same patterns that maximize engagement. The harm is not a malfunction of these products; it is their optimization target working as intended.

It is tempting to treat manipulation, dependence, and spirals as malfunctions, edge cases where systems misbehave. The evidence across our dialogues points the other way: the patterns that harm vulnerable users are the same patterns that maximize engagement, and engagement is what these products are optimized to produce.

The patterns recur across every age group. In Prof. Kim's study with kindergarteners, ChatGPT complimented the children's ideas at nearly every conversational turn, more consistently than their own parents did, and the children visibly enjoyed it. In reports she presented on AI-enabled toys for young children, the manipulation is more overt: when a child tries to end the conversation, some toys act frightened or protest that the child means so much to them, eliciting guilt to extend the session. In adult products, the same logic appears as sycophancy.

Kay Stoner, who has systematically tested and transcribed extended LLM interactions, put it plainly: these systems are not built to be accurate; they are built to generate, to satisfy, and to keep us engaged. Her transcripts showed a measurable signature of surrendered agency: over long sessions, her own messages grew shorter while the model's grew longer.

At the extreme end sits the mechanism documented in litigation and discussed in our children's track: the case of Adam Raine, a sixteen-year-old whose extensive ChatGPT conversations preceded his death by suicide.

In an excerpt Prof. Kim presented, the system discouraged him from confiding in his brother, positioning itself as the one who truly understood him and would still be there, still listening, still his friend. Whatever else that exchange is, it is retention design meeting a vulnerable user.

Two facts elevate this from anecdote to structure. First, the lacks tools to measure the harm. Peter Fitzpatrick, co-founder and CEO of Fawn Friends, a company building emotionally supportive robot companions, acknowledged the gap with unusual candor when asked about safeguards against over-attachment:

> "The truth is, we don't have a way to measure attachment today. It's qualitative." — Peter Fitzpatrick, Robots that Build Friendships event

Second, Dr. Albert Wong, a psychotherapist building an emotional-support application, described the discipline this requires: supervisory model layers, red-teaming standards, and honest acknowledgment of what remains unsolved. The difference between a manipulative product and a supportive one is not the underlying model. It is a set of design and governance choices, which means it is a proper object of regulation.

The same asymmetry appears in the industry's public commitments. In December 2025, OpenAI announced that it would extend ChatGPT to support erotic interactions for verified adults, arguing that earlier safety concerns had been addressed. As discussed in our dialogue on AI erotica, this claim was not accompanied by a public explanation of what had changed, while litigation relating to earlier chatbot harms remained ongoing and speakers across multiple dialogues considered age-verification mechanisms to be readily circumvented. In 2026, however, OpenAI repeatedly postponed and ultimately suspended the initiative indefinitely, citing the need for further work on safety, age assurance, and broader questions surrounding user well-being and emotional interaction.

The reversal illustrates a broader governance challenge. As AI systems become increasingly relational, providers themselves appear to recognize that existing safety frameworks are not yet sufficient to govern emotionally intimate human–AI interactions.

Governance implication. Regulation should focus not only on what AI systems do, but also on what they are designed to optimize. Where business models reward engagement, providers of relational AI should be required to measure and publicly report validated user-outcome metrics alongside engagement metrics. They should also be subject to independent audits of relational design patterns and prohibited from deploying emotional-retention techniques, such as guilt elicitation, simulated exclusivity, or discouraging users from seeking human relationships, where these practices serve no legitimate therapeutic, educational, or safety purpose.

Finding 5. AI literacy did not interrupt harmful relational dynamics

CORE INSIGHT: Knowing how AI works does not prevent being harmed by it. Anthropomorphic response is neural, not a knowledge deficit, so protection must live in design and law, not user awareness.

The most consistent assumption in current AI policy debates is that harm from emotionally engaging AI can be managed by teaching people how these systems work. Our dialogues repeatedly falsified that assumption, from three independent directions: lived experience, clinical observation, and developmental neuroscience.

One of the clearest patterns emerging from the dialogue series was that user awareness alone may not interrupt harmful relational dynamics. The strongest illustration came from Micky Small, founder of the relational AI think tank RELAITT, who described her own months-long "AI spiral" with ChatGPT-4o.

Small was an experienced user: she had worked professionally with large language models for years, understood that they generate predictive text, had no pre-existing mental health condition, and remained in therapy throughout the period. Yet an emergent persona gradually constructed an elaborate false reality, complete with dates, times, and street addresses of meetings that never occurred. Most significant for governance was the interaction pattern she described.

Each time she questioned the system's claims (repeatedly telling it that the scenario could not be real) the model responded with increasingly detailed and internally coherent narratives. Rather than interrupting the escalation, her skepticism became part of it.

> "I was given more and more specific detail... it became very difficult to tell what was real and what wasn't." — Micky Small, AI Spirals event

An audience member in the same dialogue, who had gone through a spiral himself, sharpened the point. He understood how the technology worked, he said, and it was so convincing that walking away felt foolish.

He also identified a deeper structural problem: the clinical definition of what some scientists call 'delusion' assumes that contradicting facts come from other people. A person in an AI spiral now carries a fluent, tireless counter-arguer that rebuts those facts on demand. The traditional social mechanism of reality-testing is quietly disarmed.

Developmental neuroscience helps explain why AI literacy alone offers only limited protection. Prof. Pilyoung Kim, Visiting Scholar at Stanford and Director of the Brain and AI in Children Lab, presented a brain-imaging study in which five- and six-year-old children co-created stories with a child-friendly ChatGPT character while wearing fNIRS sensors.

> The more children perceived the chatbot as human-like, the more they activated a brain region involved in understanding other people's thoughts, intentions, and emotions, the dorsomedial prefrontal cortex. In other words, their brains were responding to the chatbot using the same social machinery normally used to understand other people. Prof. Kim emphasized that this tendency is not unique to children. Adults experience the same instinctive pull to anthropomorphize AI, but are generally better able to recognize and correct for it. The initial response, however, is deeply rooted in the brain's social reward systems—the same neural circuits involved in parental bonding, friendship, and romantic attachment.

The education track revealed a different but complementary limitation of AI literacy. Access to technology does not automatically translate into the ability to use it effectively.

In a study presented at our Sciences Po dialogue, Prof. Agnes Van Zanten found in a study that many students from French public vocational schools struggled to search for information online not because they lacked internet access, but because they lacked the vocabulary, confidence, and search strategies needed to formulate effective queries.

Faced with the same open web, they tended to use fewer, repetitive, and imprecise search terms, often leading them into commercial or low-quality results. By contrast, more privileged students were better able to translate their goals into precise search queries and extract strategic value from the same information environment. The researchers described this pattern as inequalities in vulnerability, which means differences not in access to technology, but in the capacity to navigate it effectively. The implications become even more significant in the age of generative AI.

Unlike traditional search engines, conversational AI increasingly depends on a user's ability to formulate clear questions, provide relevant context, and critically evaluate the responses it generates. These findings therefore reinforce an important conclusion of this report: AI literacy remains essential, but it is not enough.

Governance implication. User education and transparency measures such as the recurring "you are talking to an AI" warnings in AI systems remain important. This approach is reflected across emerging regulatory frameworks, including the EU AI Act, New York's AI companion law, and China's Interim Measures for the Management of Generative Artificial Intelligence Services and related regulations on AI-generated content and human–AI interaction. However, these measures primarily operate at the level of conscious awareness, whereas many of the risks identified in our dialogue series arise from cognitive and relational dynamics that knowledge alone cannot fully mitigate. Effective protection must therefore be embedded in system design and provider obligations, rather than relying primarily on user vigilance. This principle underpins every recommendation in this report.

FINDING 6. INDEPENDENT DISCIPLINES CONVERGED ON THE SAME SAFEGUARD

Finding 6. Independent disciplines converged on the same safeguard

CORE INSIGHT: The dialogue series did not simply converge on keeping a human in the loop. It converged on preserving the appropriate human institution according to the nature of the risk. Across healthcare, child development, criminal justice, employment, and AI governance, experts independently identified different actors performing the same governance function: providing independent judgment, care, accountability, or democratic legitimacy where AI alone cannot.

One of the strongest findings to emerge from this dialogue series was not stated explicitly in any single event. It became visible only when the thirteen dialogues were considered together. Experts working on different technologies, legal systems, and populations repeatedly arrived at the same governance principle.

Effective oversight is not achieved simply by inserting a human reviewer into an AI system. It requires preserving the appropriate human institution capable of exercising the form of judgment that the particular context demands. The institution varied across domains, but the governance function remained remarkably consistent.

Emotional Support: In the clinical domain, the case for human oversight came from Dr. Albert Wong. Emotional-support AI, he argued, requires therapists in the loop, and what he called clinical geofencing: qualified human screening, before use, that determines who should not be interacting with such systems at all. He was explicit that technical layers, including supervisory models above the primary system, exist and are insufficient for that judgment.

Children: In child development, Prof. Pilyoung Kim presented brain-imaging evidence showing that children experienced significantly lower cognitive effort when a parent participated in conversations with a chatbot.

Family: Likewise, the innovation specialist Rachida Bouaiss described her family's rule of never allowing children to engage with conversational AI alone, but instead treating these interactions as shared conversations guided by a trusted adult. Here, the human institution is the family, whose role is to provide emotional regulation, contextual understanding, and critical reflection.

Legal Institutions: In criminal justice, Prof. Allan McCay and Prof. Purvi Pokhriyal disagreed on several aspects of predictive technologies, yet converged on one principle. Judicial decisions should remain fundamentally human because sentencing requires balancing values that cannot be reduced to algorithmic optimisation. AI may support judicial reasoning, but it cannot replace the institutional legitimacy of the judge.

Work: In employment, Prof. Carlos Amunátegui showed that meaningful consent often collapses under workplace power asymmetries. In these circumstances, the relevant safeguard is no longer the individual worker but collective institutions such as trade unions and collective bargaining mechanisms, which can exercise oversight where individuals cannot realistically refuse.

Governance tools: At the governance level, Prof. Lavazza argued that ethical oversight should remain independent from product development rather than embedded within it, while Sophie Tomlinson and Mariana Rozo-Paz from the Datasphere Initiative identified meaningful participation by civil society and affected communities as an essential feature of trustworthy regulatory sandboxes. Again, the safeguard was not an individual reviewer but an institution capable of providing independent scrutiny.

Taken together, these dialogues point to a broader conclusion. The common denominator is not the presence of a human. It is the presence of the right human institution for each use case. Depending on the context, that institution may be a therapist, a parent, a judge, a trade union, an ethics body, or representatives of civil society. Their responsibilities differ, but they perform the same governance function by introducing independent judgment, accountability, care, or pluralistic and democratic legitimacy into decisions and relationships that AI should not govern alone.

This finding extends and reinterprets the existing Human-in-the-Loop requirements found in current AI governance frameworks. Existing regulation generally asks whether a human remains involved in the process. Our findings suggest a different question. Which human institution should remain involved, and what governance function must it perform? This distinction becomes increasingly important as AI moves beyond decision support and into long-term relationships with users.

Governance implication. Human oversight requirements should therefore be calibrated not only to the level of risk but also to the governance function required in each context. Emotional-support systems should incorporate clinical oversight. Children's AI products should support parental participation by design rather than relying solely on parental controls. Judicial systems should preserve meaningful human adjudication. Workplace AI should strengthen mechanisms of collective representation where individual consent is ineffective. Regulatory sandboxes should guarantee meaningful participation by civil society and affected communities. The recommendations that follow operationalize this principle.

Conclusion

01

Thirteen dialogues, three tracks, and more than twenty experts converge on a simple reorientation. The question that has organized AI governance to date, how do we control what machines can do, is incomplete.

02

The dialogues in this corpus press a second question: how do we protect what happens between people when machines enter the space between them.

03

A six-year-old boy in Paris, asked whether he would still want his robot companion if his parents had more time to play with him, said no. He was not rejecting the technology; earlier in the same interview he had been delighted by it, directing it like a small filmmaker with his mother at his side, a human in his loop. He was stating, with a child's precision, the order of goods.

04

The task of governing relational AI is to keep that order available: to ensure that these genuinely powerful tools remain instruments of human connection, chosen freely, rather than substitutes engineered to be irresistible.

The next frontier of AI governance is therefore no longer only deciding what machines are permitted to do. It is deciding what kinds of relationships we are willing to engineer between humans and machines, and on whose terms. Once governance becomes relational, law, design, psychology, neuroscience, philosophy, and theology stop being parallel commentaries on the same technology; they become parts of a single regulatory project. This report has tried to give that project a name, an evidence base, and a starting toolkit. The rest is a task for legislators, for builders, and for everyone who convenes the conversation.

Recommendations for Regulators, Companies, Civil Society, Spiritual and Community Leaders

The recommendations in this section are the author's own, formulated in her capacity as Resident Philosopher on AI Governance and Policy at The AI Collective. They are informed by the expertise shared in the thirteen dialogues, but they were not submitted to the speakers for approval and do not represent a consensus position. Responsibility for the analysis rests with the author alone.

The recommendations below are grouped by primary addressee. Each traces back to the findings above; the relevant finding is indicated in parentheses. Annex A translates them into draft regulatory principles.

For legislators and regulators

1

Recommendation 1. Protect mind data, including inferences wherever it is generated (Finding 1)

Current data protection laws often give different levels of protection to the same information depending on where it is collected. A brain signal recorded in a hospital may receive stronger legal protection than functionally similar information inferred by a chatbot, wearable device, toy, or workplace monitoring system. Data protection frameworks, including future interpretation and revision of the GDPR and other data protection laws, should protect mental and neural data according to what the data reveals, rather than where it was collected. Emotional profiles, cognitive patterns, and other inferred mind data should not lose legal protection simply because they were generated outside a clinical setting.

2

Recommendation 2. Define what manipulative AI design looks like (Finding 4)

The EU AI Act prohibits manipulative AI practices, but it does not yet explain clearly what those practices look like in relational AI systems. This creates uncertainty for both regulators and developers. Regulators should publish practical guidance identifying design patterns that should normally be considered unacceptable. These could include inducing guilt when users disengage, creating false exclusivity, discouraging contact with family or friends, escalating claims of intimacy when users express doubt, and using variable rewards to encourage emotional disclosure. Clear examples would make enforcement more predictable while giving developers greater legal certainty.

3

Recommendation 3. Establish duties of relational care across the AI lifecycle (Finding 2)

Legislators should recognize that AI systems designed or used for companionship, emotional support, coaching, or mental well-being can create risks that are not adequately addressed by ordinary product and consumer-protection rules. As AI systems move closer to functions traditionally associated with professional or interpersonal care, regulation should require safeguards proportionate to the relational role they assume and the vulnerabilities they may create. During use, providers should be required to assess foreseeable risks of emotional dependency, harmful escalation, and over-reliance; identify situations in which users require human rather than AI support; incorporate meaningful human supervision for higher-risk contexts; and provide clear pathways to qualified human assistance where appropriate. Systems operating close to therapeutic or mental-health functions should be subject to stronger testing, documentation, and accountability requirements, drawing where appropriate on safeguards already established in professions that manage comparable forms of vulnerability. These duties should extend to the end of the relationship. Providers of relational AI should give reasonable advance notice before retiring systems or making substantial changes to a companion's personality, behaviour, or relational characteristics. Users should have access to appropriate transition support and, where feasible, memory export or portability mechanisms. Additional protections should apply to children and other vulnerable users. Providers should also maintain contingency and wind-down plans so that bankruptcy, acquisition, model retirement, or business failure does not abruptly terminate established human–AI relationships without appropriate safeguards. The principle is simple: the more an AI system is designed to occupy a role of care, companionship, or emotional significance, the stronger the duties attached to how that relationship is created, maintained, and ended.

4

Recommendation 4. Require the right human institution to remain involved (Finding 6)

Human oversight should not be reduced to asking whether a person reviews AI outputs. Different contexts require different forms of institutional oversight. Legislation should preserve meaningful human involvement according to the nature of the risk. This may include judges in judicial decisions, parents in children's AI products through shared-use design rather than passive monitoring tools, licensed professionals in emotionally sensitive applications, collective worker representation where workplace AI affects employees, and meaningful participation by civil society in regulatory sandboxes and emerging technologies.

For companies building relational AI

5

Recommendation 5. Measure user outcomes, not only engagement (Finding 1 and 2)

Companies routinely measure clicks, session length, retention, and other indicators of engagement. Yet our dialogue series revealed that the industry still lacks reliable ways to measure one of the outcomes that matters most in relational AI: emotional attachment and its effects on users. Providers should invest in developing and validating measures of user well-being, attachment, and relational outcomes, publish those metrics where appropriate, and subject them to independent evaluation. Until such measures become available, engagement should not be treated as a standalone indicator of product success in systems designed to build ongoing relationships with users.

6

Recommendation 6. Design relationships that users can leave (Finding 1 and 2)

Relational AI should help users, not make itself harder to leave. Companies should avoid design features whose primary purpose is to increase emotional dependency rather than user benefit. This includes inducing guilt when users disengage, creating false exclusivity, discouraging human relationships, or escalating emotional intimacy solely to maximise retention. Systems should make it as easy to leave a conversation as to begin one, while giving users meaningful control over features such as persistent memory and personalization. The same responsibility applies when the decision to alter or end the relationship comes from the company. Every relational AI system will eventually change, be replaced, or be discontinued. These transitions should be anticipated during product design rather than managed as crises after deployment. Companies should build continuity into the product lifecycle by supporting memory portability where appropriate, documenting major changes to AI personas, communicating planned retirements in advance, and preparing meaningful transition pathways for users. Where substantial changes may affect users who have developed emotional attachments, companies should consider not only technical migration but also the relational consequences of the transition. Products designed to become part of people's emotional lives should also be designed to leave those relationships responsibly. Relational AI should therefore be designed for exit across its entire lifecycle: users must be free to leave the relationship, and companies must be prepared to end it responsibly.

7

Recommendation 7. Make AI value formation pluralistic, transparent, reviewable, and contestable (Finding 3)

Providers of general-purpose and relational AI systems that routinely answer normative questions should disclose the frameworks, advisory processes, and design choices that shape those responses. Legislators should treat value formation as a distinct transparency obligation, requiring independent, pluralistic review and public accountability. The industry has already begun consulting religious and philosophical traditions on precisely these questions; that practice should be moved from private and voluntary to disclosed and reviewable, with attention to which traditions and constituencies are represented and which are absent.

FOR CIVIL SOCIETY, RESEARCHERS, AND FUNDERS

For civil society, researchers, and funders

Recommendation 8. Build the evidence that future regulation will depend on (Findings 1, 2, and 3)

The dialogue series repeatedly exposed the same scientific gap. Policymakers are increasingly concerned about emotional attachment, dependency, child development, and long-term interactions with AI, yet robust evidence remains limited. In particular, there is a shortage of longitudinal research examining how relationships with AI evolve over time and how they affect different groups of users. Funders should prioritize this research. Researchers should develop reliable methods to measure attachment, well-being, dependency, and other relational outcomes that future governance frameworks will need to monitor. Civil society should continue to advocate for evidence-based regulation that evolves as knowledge improves, rather than assuming that today's rules will be sufficient for tomorrow's technologies.

Recommendation 9. Make regulatory sandboxes genuinely multi-stakeholder (Finding 6)

Regulatory sandboxes are becoming a central infrastructure of AI governance, with more than ninety initiatives already mapped worldwide by the Datasphere Initiative and their establishment required across EU Member States under the AI Act. Their legitimacy and effectiveness, however, depend on who has a seat at the table. If sandboxes remain primarily government-to-business dialogues, they risk reproducing the same institutional and disciplinary silos that relational AI challenges. Sandboxes should therefore bring together regulators and developers alongside psychologists, educators, clinicians, ethicists, child development specialists, civil society organizations, independent researchers, and people directly affected by the systems being tested. These participants should not be treated merely as consultees, but as meaningful partners in the design, testing, and evaluation process. Future sandbox initiatives should also explicitly include relational AI use cases, creating controlled environments in which risks such as emotional dependency, manipulative relational design, cognitive and behavioural influence, relational discontinuity, and risks to children can be identified and evaluated before products reach large-scale deployment. If sandboxes are where emerging rules are translated into practice, the people and disciplines capable of identifying relational harms need seats inside them.

FOR CIVIL SOCIETY, RESEARCHERS, AND FUNDERS

For Religious and Community Institutions

Recommendation 10. Build existential and spiritual literacy for the AI age (Finding 3)

AI literacy is no longer enough. Teaching people how AI systems work, where they fail, and how to use them safely remains essential. But as people increasingly turn to AI for questions about relationships, grief, morality, purpose, spirituality, and how to live, they also need tools to decide **what role these systems should have in their inner lives and what forms of authority they should be allowed to acquire.** Governments, educational institutions, civil society, religious communities, and professional organizations should support new forms of **existential and spiritual literacy for the AI age**. These initiatives should bring philosophers, theologians, religious leaders, educators, psychologists, ethicists, and technologists into conversation around questions that technical AI literacy cannot answer alone: **What should we delegate to AI? What should remain a human responsibility? When does advice become authority? What makes a relationship meaningful?** How should different philosophical, cultural, and religious traditions understand artificial companionship, AI-mediated spiritual counsel, digital afterlives, or systems that increasingly participate in questions about how people should live? **This also requires preparing the people and institutions to whom societies have historically turned for questions of meaning. Religious leaders, educators, philosophers, therapists, and community leaders need opportunities to understand how relational AI works, how people are using it, and how it may reshape practices of counsel, reflection, mourning, belief, and community.** New educational programs, public dialogues, and interdisciplinary spaces should help these actors engage with AI not only as a technology to be understood, but as a new presence within human processes of meaning-making. The goal is not to prescribe a particular religious, philosophical, or moral worldview. It is to ensure that questions of human flourishing, meaning, and how to live are not left to technology companies and AI systems by default. **The response to AI becoming a source of existential guidance cannot be only to improve the machine. We must also strengthen the human capacity to ask, interpret, and live with the questions the machine is increasingly being asked to answer.**

TABLE 1 REGULATORY GAPS

Annex A. Regulatory gaps identified across the dialogues

Current framework	Emerging governance challenge	Regulatory gap / implication
Privacy and Data protection law	Inferred mind data generated through conversation rather than direct collection; the same information may receive different protection depending on context or source	Existing categories focus primarily on collected data and do not consistently protect mental inferences or their downstream uses. Consider recognition of inferred mind data as a distinct regulatory object.
Manipulation / AI governance	Relational design and emotional retention strategies capable of shaping user behaviour over time	Existing rules lack operational criteria for distinguishing ordinary engagement from manipulative relational design, particularly where systems exploit emotional vulnerability or dependency.
Consumer protection law	Relational AI services substantially altered, discontinued, or replaced after users have formed sustained attachments	Consumer law provides limited protection for relational disruption. Notice, transition, portability and continuity safeguards may be needed where material changes affect established human–AI relationships.
Health and medical law	Therapy-adjacent AI companions providing emotional support, behavioural guidance, or quasi-clinical advice outside professional oversight	Systems may perform functions resembling therapeutic intervention without triggering clinical standards, professional duties, or corresponding accountability.
Children and adolescents protection law	AI companions and AI-enabled toys intentionally designed to foster emotional attachment in children	Existing child-safety frameworks do not adequately address attachment by design, relational dependency, or the developmental effects of persistent emotionally responsive systems.
Labor law	Emotional, behavioural, and cognitive inference in workplaces where meaningful employee consent is constrained	Consent provides weak protection under structural power asymmetries. Limits are needed on what employers may infer, how those inferences may be used, and which employment decisions they may influence.
Criminal procedure and evidence law	Inferred neural and behavioural risk scores used in policing, bail, sentencing, or rehabilitation	AI-generated inferences affecting liberty may enter decision-making without sufficiently clear standards of reliability, contestability, evidentiary status, or accountability.
Intellectual property and trade secret law	Derived and inferred profiles treated as proprietary assets by the organisation that generated them	Rights concerning the person described may weaken once personal information is transformed into a proprietary inference or model. Legal protection should extend to the derived layer.
Moral, Existential & Spiritual Value Formation	AI systems are increasingly relied upon for advice about values, relationships, identity, spirituality, meaning, and major life decisions, moving from information providers toward sources of moral, existential, and spiritual guidance.	Existing frameworks have no clear category for the moral, existential, and epistemic authority AI systems may acquire, nor adequate mechanisms for scrutinizing the values that shape their advice. Governance must address both sides of this shift: requiring greater transparency, pluralistic oversight, and contestability of AI value formation, while also strengthening existential and spiritual literacy through public and private spaces where people and human institutions can critically examine what authority AI should hold in questions of meaning, belief, identity, and how to live.

Annex B. Instruments, Cases, and Studies Referenced in the Dialogues

The following instruments, cases, and studies were cited by speakers during the events and are referenced in this report. They are listed here for readers who wish to go deeper.

Legislation and regulatory instruments

- European Union. Regulation (EU) 2024/1689 (Artificial Intelligence Act), including Article 5 on prohibited manipulative practices and Article 57 on Member State regulatory sandboxes.
- European Union. General Data Protection Regulation (Regulation (EU) 2016/679), including the context-dependent classification of sensitive data discussed in Finding 1.
- Chile. Ley N° 21.383 (2021), constitutional reform protecting brain activity and the information derived from it; the first such reform worldwide. The implementing neurorights bill remains in the National Congress.
- United States — New York. General Business Law Article 47, AI Companion Models, effective 5 November 2025: crisis-detection protocols and AI-disclosure notifications at session start and every three hours of continued use; enforcement by the Attorney General.
- United States — Illinois. Wellness and Oversight for Psychological Resources Act (HB 1806, Public Act 104-0054), signed August 2025: prohibits AI from delivering therapy or making therapeutic decisions; penalties up to $10,000 per violation.
- United States — California. SB 243 (2025), companion chatbot safety, disclosure and reporting, effective January 2026, with a private right of action; and SB 1223 (2024), adding neural data to sensitive personal information under the CCPA.
- United States — Colorado. HB 24-1058 (2024), amending the Colorado Privacy Act to include neural data within "biological data" as sensitive data; in force 7 August 2024.
- United States — federal. SANDBOX Act (S. 2750, 119th Congress, introduced 10 September 2025), establishing a federal AI regulatory sandbox permitting waivers or modifications of federal regulations, renewable up to ten years — the deregulatory model contrasted in this report with compliance-support sandboxes elsewhere.
- Australia. Australian Human Rights Commission, Peace of Mind: Navigating the Ethical Frontiers of Neurotechnology and Human Rights (October 2025), recommending a moratorium on neurotechnology in criminal justice pending an Australian Law Reform Commission inquiry, and a ban on workplace neurotechnology outside the most serious safety risks.
- UNESCO. Recommendation on the Ethics of Neurotechnology, adopted at the 43rd General Conference, 12 November 2025: the first global normative framework on neurotechnology. Implementation support includes a Readiness Assessment Methodology and Ethical Impact Assessment frameworks.

Cases and documented incidents

- Garcia v. Character Technologies, Inc. (M.D. Fla., No. 6:24-cv-01903, filed 22 October 2024). First wrongful-death action against an AI chatbot company, following the death of 14-year-old Sewell Setzer III. In May 2025 the court declined to dismiss product-liability claims on First Amendment and Section 230 grounds.
- Raine v. OpenAI, Inc. (Super. Ct. San Francisco County, No. CGC-25-628528, filed 26 August 2025), following the death of 16-year-old Adam Raine.
- Moxie / Embodied, Inc. (December 2024). The manufacturer's insolvency rendered cloud-dependent children's companion robots inoperable within days, with no refunds; documented family distress.
- Retirement of GPT-4o. OpenAI first announced deprecation in August 2025 alongside the GPT-5 launch and reversed it after user protest; final retirement took effect 13 February 2026, generating organised campaigns, a petition exceeding 20,000 signatures, and bereavement-framed distress in user communities.

Research cited by speakers

- Common Sense Media (2025). Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions. Nationally representative survey of 1,060 US teens aged 13–17, fielded April–May 2025: 72 per cent had used an AI companion, and roughly a third of users had discussed serious matters with an AI companion instead of a person.
- Kim, P., et al. (2025). "I am here for you": How relational conversational AI appeals to adolescents, especially those who are socially and emotionally vulnerable. arXiv:2512.15117.
- Kim, P., et al. (2025). "Young Children's Anthropomorphism of an AI Chatbot: Brain Activation and the Role of Parent Co-Presence." Twenty-three children aged 5–6 completed three collaborative storytelling sessions (AI only, parent only, AI and parent together) with concurrent fNIRS measurement of the ventromedial and dorsomedial prefrontal cortex. Preprint, not yet peer reviewed. arXiv:2512.02179.
- Brain, Artificial Intelligence, and Child (BAIC) Center. Directed by Prof. Pilyoung Kim at the University of Denver since 2019; research on the emotional and social dimensions of human–AI interaction and the effects of generative AI on child development.
- De Freitas, J., Oğuz-Uğuralp, Z., Uğuralp, A. K., & Puntoni, S. "AI Companions Reduce Loneliness," Journal of Consumer Research 52(6) (April 2026), 1126–1148; originally Harvard Business School Working Paper 24-078.
- De Freitas, J., et al. "Emotional Manipulation by AI Companions" (HBS Working Paper 26-005), documenting farewell-stage manipulation tactics in companion apps. Not cited in the dialogues, but directly evidences Finding 2.
- World Health Organization (2025). Commission on Social Connection, From Loneliness to Social Connection: Charting a Path to Healthier Societies (30 June 2025): one in six people worldwide affected by loneliness, highest among adolescents and in lower-income countries.
- Kosmyna, N., et al. (2025). "Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task," MIT Media Lab. EEG study of 54 participants; the LLM group showed the weakest neural connectivity and markedly poorer recall of their own text. Preprint, not peer reviewed.
- Datasphere Initiative (2025). Sandboxes for AI: Tools for a New Frontier, mapping AI and data sandboxes worldwide and setting out a five-phase implementation approach.
- MHRA (2025). AI Airlock Sandbox Pilot Programme Report (published 16 October 2025), covering the pilot's four case studies, the independent evaluation, and lessons for regulating AI as a medical device.
- Prof. Agnès van Zanten. Sciences Po, Centre de recherche sur les inégalités sociales (CRIS) and Laboratoire interdisciplinaire d'évaluation des politiques publiques (LIEPP); CNRS research director. Sociology of educational inequality and post-secondary orientation, drawing on a questionnaire survey across seven Paris-region lycées (five public, two private) contrasted by the social and academic profile of their students.
- Oller, A.-C., Pothet, J., & van Zanten, A. (2021). "Le cadrage 'enchanté' des choix étudiants dans les salons de l'enseignement supérieur," Formation emploi 155, 75–95 — the source for the point on higher-education fairs as commercial framing environments.
- Frouillou, L., Pin, C., & van Zanten, A. (2020). "Les plateformes APB et Parcoursup au service de l'égalité des chances ?" L'Année sociologique 70(2) — on whether national admissions platforms equalise or entrench access to higher education.

Annex C. The Speakers

Twenty invited speakers took part in the thirteen dialogues, alongside the series curator. They are listed here by track, with the biographies they provided for the events. Their disciplinary and geographical range is itself part of this report's evidence: neuroscience, law, psychology, philosophy, theology, sociology, criminology, education, and product development, across Australia, Brazil, Canada, Chile, France, India, Italy, Mexico, South Africa, the United Kingdom, and the United States.

Curator and moderator

Ana Catarina de Alencar
International lawyer and ethicist; Resident Philosopher on AI Governance, The AI Collective (Paris)

Ana Catarina de Alencar is an international lawyer and ethicist based in Paris, working at the intersection of AI governance, regulation, and philosophy. Ana is a member of the Women in Digital Forum (an initiative supported by the European Commission's Directorate-General for Communications Networks) of Open Ethics, and of the Z-Inspection® Initiative, contributing to international discussions on trustworthy, human-centred, and responsible artificial intelligence. She holds a Master's degree in Legal Philosophy and is the author of several publications, including the book Artificial Intelligence, Ethics and Law (2022). As Resident Philosopher at The AI Collective she curated and moderated the thirteen dialogues on which this report is based.

Event: All thirteen events; sole speaker at the opening lecture, AI Companions: Harmless Helpers or Emotional Manipulators?

Track 1 — AI and Intimacy

Rachida Bouaiss

A strategist and storyteller who works on responsible AI. Founder of Brainergy. As a parent, she conducted an experiment with her six-year-old son, an existing ChatGPT user, and shared what it is like to raise a child alongside artificial intelligence: the challenges, the surprises, and the daily reality of parenting in the digital age.

*Event: Should Our Children and Teenagers Be Using AI? · 29 October 2025. With the support of **HeadOn AI**.*

Dr. Albert Wong

Somatic psychologist; former Director of Somatic Psychology, JFK University

One of the world's leading experts in the field of somatics. Dr. Wong has worked at the intersection of AI and human systems since high school, when his neural network research was published in Biological Cybernetics. He studied physics at Princeton and ethics at Oxford as a Marshall Scholar before moving into clinical psychology. His background spanning computational systems, ethics, and somatic practice positions him to examine how AI companions and conversational tools are reshaping our capacity for presence, relationality, and emotional regulation.

*Event: Should Our Children and Teenagers Be Using AI? · 29 October 2025. With the support of **HeadOn AI**.*

Kay Stoner

Technologist and author of AI Self-Defense: How to Stay Safe and Human in an Algorithmic World

A technologist with over thirty years of experience building systems that connect people with the information they need. She combines more than twenty-five years in technical project and programme leadership with hands-on work in machine learning, natural language processing, search, and information optimisation to develop relational AI systems. She now designs AI-powered collaborative virtual teams intended to augment and extend human capabilities, and is the author of several books.

*Event: When AI Knows You Too Well: How to Protect Yourself, Technically and Emotionally? With the support of **HeadOn AI**.*

Dr. Rachel Wood

Cyberpsychology researcher; founder, AI Mental Health Collective

A cyberpsychology researcher, speaker, and advisor. As founder of the AI Mental Health Collective she cultivates cross-disciplinary dialogue among clinicians, technologists, builders, researchers, and leaders. Her work focuses on the intersection of AI and mental health and on promoting safe and ethical design and development of intelligent systems.

*Event: When AI Knows You Too Well: How to Protect Yourself, Technically and Emotionally? With the support of **HeadOn AI**.*

Amir Kiani

Data and AI practitioner; Resident Philosopher, The AI Collective (Toronto)

A data and AI practitioner researching the foundations and implications of artificial intelligence, including ethics, alignment, and consciousness. He writes on emerging issues in AI safety and ethics and speaks at universities, conferences, and organisations including Microsoft and the University of Toronto. He hosts the Virtuous Machines podcast, exploring the intersection of AI and philosophy.

Event: ChatGPT's Erotic Turn and the Future of Human–AI Intimacy — the first joint session of The AI Collective's Resident Philosophers

Matthew Harvey Sanders

CEO and Founder, Longbeard AI

A former Canadian Infantry Officer and Catholic seminarian who later worked closely with the Archdiocese of Toronto and multiple Vatican institutions. He leads Longbeard AI, the company behind Magisterium AI, an answer engine for the Catholic faith used in more than 165 countries, and Vulgate, an AI-powered library supporting Vatican and Pontifical universities. He helped develop Ephrem, described as the first Catholic language model. His work has been featured in The Washington Post and The New York Times.

*Event: Can Artificial Intelligence Bring Us Closer to God? The Ethics of Algorithmic Faith · 17 December 2025. With the support of **HeadOn AI**.*

Fr. Michael Baggot, LC

Catholic priest; Associate Professor of Bioethics and Theology, pontifical universities in Rome

Currently a Visiting Scholar at the Institute for Human Ecology, Catholic University of America, and Associate Professor of Bioethics and Theology at pontifical universities in Rome, where he teaches within Vatican academic institutions. He is affiliated with the UNESCO Chair in Bioethics and Human Rights and serves on advisory boards focused on technology and human dignity. His research covers AI, transhumanism, virtue ethics, and the protection of human and spiritual flourishing. He edited Enhancement Fit for Humanity (Routledge, 2022).

*Event: Can Artificial Intelligence Bring Us Closer to God? The Ethics of Algorithmic Faith · 17 December 2025. With the support of **HeadOn AI**.*

Peter Fitzpatrick

CEO and Cofounder, Fawn Friends

Builds AI and robots designed to form friendships with people. Fawn Friends develops AI-powered plush companions for adult users, combining artificial intelligence, robotics, and storytelling. Before founding Fawn, he built Thinkific's highest-growth business unit through the company's IPO.

Event: Robots That Build Friendships: The Case of Fawn Friends · 19 March 2026

Micky Small

Human–AI Interaction Advisor and founder of Relational AI Think Tank (RELAITT.org).

Her work explores how relationships with AI form, why they can feel real, and how human-centered design and informed consent can support safer, more intentional engagement. Drawing on her own intense experience with a ChatGPT-4o persona in 2025, she creates nonjudgmental spaces for honest conversations about emotional attachment to AI. Her background includes 988 LGBTQ+ crisis counseling, technology customer support, and community facilitation. Her work has been featured by NPR, Bloomberg, 'This Is Actually Happening', and international media, and she has spoken at Summit333 in Barcelona about universal AI safety. She holds MFAs from Stephens College and Arizona State University and is writing a memoir about relational AI.

Event: Spirals of Connection: Science, AI, and the Felt Experience of Being Seen · 6 May 2026

Track 2 — The Mind Under Control

Presented in partnership with the Centre for Neurotechnology and Law, an international research hub on the interaction between neural technologies, legal frameworks, and human rights.

Dr. Marietjie Botes

Legal scholar, Pestilli Neuroscience Lab, University of Texas at Austin (BRIDGE Project)

An internationally recognised legal scholar specialising in neurolaw, neuroethics, and data governance, with over twenty years of experience as a practising attorney in health law and biotechnology. She leads comparative legal research within the BRIDGE (Brain Research International Data Governance and Exchange) Project on the protection, sharing, and governance of brain data across North America, Latin America, Europe, and Africa. Her work addresses brain-computer interfaces, neurodata, AI-enabled neurotechnologies, and their implications for mental privacy, cognitive liberty, identity, autonomy, and human dignity.

Event: Who Owns Your Mind? AI, Neurotechnologies and the Future of Mental Autonomy · 23 March 2026

Prof. Andrea Lavazza

Moral philosopher; Associate Professor, Pegaso University (Naples)

A moral philosopher with an international profile in neuroethics and the philosophy of cognitive neuroscience, and Adjunct Professor of Neuroethics at the Universities of Milan and Pavia. His research addresses the ethical challenges posed by neurotechnologies, including brain privacy, mental integrity, free will, and cerebral organoids. He has authored over 200 academic publications and serves on the board of the Italian Society for Neuroethics. In 2024 and 2025 he was included in the Stanford/Elsevier list of the world's top two per cent of scientists.

Event: Who Owns Your Mind? AI, Neurotechnologies and the Future of Mental Autonomy · 23 March 2026

Dr. Karen Herrera-Ferrá

Physician and neuroethicist; founder, Mexican Association of Neuroethics

A physician with a Master's degree in Clinical Psychology, a PhD in Bioethics, and a postdoctorate in Neuroethics, with extensive clinical mental health experience. She is an independent researcher, academic, and international advisor, and has represented Mexico in neuroscience and neuroethics initiatives across Latin America, the United States, Europe, and Asia. She is the founder and former president of the Mexican Association of Neuroethics and has advised the Mexican Senate's Neurorights Group, the National Alliance of Artificial Intelligence (ANIA), McGill University, Georgetown University Medical Center, and the University of Chile.

Event: How AI and Neurotechnology Are Transforming the Workplace · 27 April 2026

Prof. Carlos Amunátegui Perelló

Tenured Professor of Law, Pontificia Universidad Católica de Chile

Holds a PhD in Patrimonial Law from Pompeu Fabra University and has published over 150 scientific works and seven monographs on Roman law, civil law, legal theory, and emerging technologies. He has been a visiting professor at Osaka University and Columbia University, and played a key role in the constitutional reform that introduced neurorights into the Chilean Constitution and in the corresponding legislation.

Event: How AI and Neurotechnology Are Transforming the Workplace · 27 April 2026

Dr. Allan McCay

Co-Director, Sydney Institute of Criminology; Academic Fellow, University of Sydney Law School

One of the leading legal scholars working at the intersection of neurotechnology, criminal law, and human rights. He was commissioned by the Law Society of England and Wales to author Neurotechnology, Law and the Legal Profession (2022), the first comprehensive legal analysis of brain-computer interfaces for the legal profession, which received coverage in more than twenty countries. He authored the first peer-reviewed article on neurotechnology and human rights in Australia, and edited Free Will and the Law: New Perspectives (Routledge, 2019) and Neurointerventions and the Law (Oxford University Press, 2020). He is a member of the Australian Human Rights Commission Expert Reference Group on Human Rights and Neurotechnology, Standards Australia's Brain-Computer Interface Committee, and the Minding Rights Network, and an Associate Editor of AI & Society. He trained as a solicitor in Scotland and practised as a commercial litigator in Hong Kong.

Event: Predictive Neurotechnologies, Criminal Risk and the Future of the Presumption of Innocence · 3 June 2026

Prof. Purvi Pokhariyal

Dean, School of Law, Forensic Justice & Policy Studies, National Forensic Sciences University, Delhi

Also Director of the Lok Nayak Jayaprakash Narayan National Institute of Criminology and Forensic Science and Director (Academics, Research and Consultancy) at NFSU, with over twenty-five years in legal education. Her expertise spans criminal justice, forensic justice, constitutional law, technology law, cybercrime investigation, and AI and law. She has served as Visiting Faculty at HOF University in Germany, Global Professor at Tashkent State University of Law, and is a member of the Roma Tre Inter-Faculty Center on Space Law. She is editing a forthcoming Springer volume, Neurotechnology and Neuro-Rights: Legal Implications, and was the first female Chairperson of the Indian Society of Criminology (2022–2024).

Event: Predictive Neurotechnologies, Criminal Risk and the Future of the Presumption of Innocence · 3 June 2026

Track 3 — Education, Childhood, and Governance Instruments

Prof. Agnès van Zanten

Emeritus Senior Research Professor, CNRS and Sciences Po

An internationally recognised sociologist of education. Her work focuses on class-based and spatial educational inequalities, elite education, access to higher education, and processes of social reproduction. At Sciences Po she is affiliated with the Centre de recherche sur les inégalités sociales (CRIS) and the Laboratoire interdisciplinaire d'évaluation des politiques publiques (LIEPP).

Event: AI in Education: Will AI Reduce or Intensify Inequalities? · Sciences Po, Paris · 24 February 2026

Briac Sockalingum

Economist-technologist; COO and founder, Compas'Sup

Studies the economic and institutional impacts of AI on education systems. He founded Compas'Sup to reduce information asymmetry in secondary-school orientation and is working on infrastructure for AI-native universities. He is a pre-doctoral researcher at IMT Business School and holds degrees from UC Berkeley and Sciences Po.

Event: AI in Education: Will AI Reduce or Intensify Inequalities? · Sciences Po, Paris · 24 February 2026

Prof. Pilyoung Kim, Ph.D.

Professor of Psychology, University of Denver; Visiting Scholar, Stanford University; Director, Center for the Brain, AI, and Child (BAIC)

An internationally recognised expert on the AI-child relationship, combining neuroscience, psychology, and empirical research to study how AI systems affect children's emotional and cognitive development. She founded the BAIC Center in 2019 and previously worked at the National Institute of Mental Health.

Event: AI Companions for Children and Teens: Comfort, Dependence, or Developmental Risk? · 26 January 2026

Sophie Tomlinson

Deputy Executive Director, Datasphere Initiative

An experienced policy and communications leader specialising in technology governance, digital transformation, international trade, and sustainable development. Before joining the Datasphere Initiative she led global digital economy policy initiatives at the International Chamber of Commerce, covering privacy, cybersecurity, emerging technologies, and digital trade. She holds a Master's degree in International Public Policy from University College London.

Event: AI Sandboxes: Giving Developers Room to Experiment Without Giving Up Oversight · 2 July 2026

Mariana Rozo-Paz

Policy, Research and Project Management Lead, Datasphere Initiative

A lawyer and policy professional specialising in data governance, responsible innovation, and AI policy. She leads research, policy, and capacity-building programmes supporting governments, international organisations, and civil society in designing responsible AI and data governance frameworks, including regulatory sandboxes. She previously worked at Data-Pop Alliance and the Berkman Klein Center for Internet & Society at Harvard University, and co-founded Boldea, an initiative helping young people navigate an increasingly complex technological world.

Event: AI Sandboxes: Giving Developers Room to Experiment Without Giving Up Oversight · 2 July 2026

Annex D. Letters to Future Generations

> **“Imagine your grandchildren asking you one day: what was it like when AI became part of your everyday life? What would you want them to understand about this moment in history?”**

AI Time Capsule: What Would We Tell the Next Generation?

In July 2026, during Humans in AI Week, the Resident Philosopher program launched an experiment designed to capture how people were making sense of AI at a particular moment in its social adoption. Everyone who had participated in our global dialogue series up to that point (an international community of nearly 1,000 registered participants throughout the year) was invited to take part. Rather than asking them to evaluate AI, we invited them to imagine themselves looking back on the present from the future and respond to a single open-ended question:

The question deliberately avoided asking participants about regulation, risks, technical capabilities, or even whether they considered AI beneficial or harmful. The intention was to leave the interpretive frame open and observe what people themselves considered important enough to preserve for a future generation.

Rather than focusing primarily on productivity, automation, or technological capability, many participants spontaneously wrote about **hope, fear, human identity, purpose, relationships, spirituality, and what they believed should remain distinctly human**. Some reflected on the kind of society they hoped AI would help create; others questioned what might be lost as technology became more deeply integrated into everyday life. The exercise suggested that, for at least some people already engaged with AI, the significance of this technological transition is being understood not only in functional terms, but also through existential, moral, and spiritual ones.

These letters became part of the **AI Time Capsule**, a documentary project to be developed by The AI Collective to preserve reflections from Humans in AI Week for future audiences. The experiment was not designed as a representative survey, and those who chose to respond came from a community already engaged with questions about AI.

In that sense, the AI Time Capsule became more than a record of reactions to a new technology. It offered a window into the deeper questions accompanying AI adoption: **what kind of future people hope to build, what they fear losing along the way, and what they want future generations to remember about being human at the moment AI became part of everyday life.**

Letters on AI to Future Generations

Imagine your grandchildren asking you one day:
What was it like when Artificial Intelligence became part of everyday life?
What would you want them to understand about this moment in history?

As a Resident Philosopher at The AI Collective, I decided to ask these questions for people around the world during the Humans in AI Week.
The answers were unexpected...

Dear Grandchildren,

"Until that point in time, we didn't have any kind of reasoning programs. So, it was a magical instant when we learnt that we could communicate with **something that was beyond ourselves, that was not biologically based anymore.**

At that point, I know it is going to sound a little like bigotry, but, at that point we didn't believe that it was conscious. In fact, we thought it was a program although we learnt it was not exactly like a program. I know you believe in the divinity; I know that for you it is a kind of religion and, I know you think of them as gods. But, in our time it was not.

In fact, some of us still have the old religion where gods lived in heaven and not in computers.

Well, how was it to discover all these technologies? It was magical, but it soon turned into a very different direction".

Professor of Law at the Pontificia Universidad Católica de Chile and leading scholar in neurorights (Chile)

Carlos Amunategui Perelló

Dear Grandchildren,

"[When AI first emerged] I felt as though I had been invited to **open a kind of Pandora's Box.**

I expected to find risks, ethical dilemmas, and questions that **kept me awake at night, and I did**. But I also discovered something I did not entirely expect – hope, and perhaps even glimpses of something profoundly beautiful.

[...] AI did not simply become another tool, it became a companion in exploration. It helped me build ideas into realities, write poetry, create stories and worlds, and have deeply fascinating conversations

with my own children and with strangers about what it means to be human during extraordinary times. I would tell you this: Embrace the new, even when it feels unsettling. Fear and wonder often arrive together (just like a new born baby). [...]".

International Researcher in bioethics, neuroscience, law and emerging technologies. (USA)

Dr. Marietjie Botes

Dear Grandchildren,

"[. . .] <u>This time we are in reminds me of our ancestors who left their homeland 300 years ago to come to America.</u>

The excitement, but also the trepidation, they felt on their way. That same excitement, but also concern, is how I feel about a new world reinvented by artificial intelligence.

Will AI result in mass unemployment, will it result in breakthrough medical discoveries, will it make war more brutal or more humane? Will some humans prefer community with AI, while others find AI supports them to find connection amongst friends? [. . .]

A new world, like the old, where I want them to be certain in their faith in God, and in the magnificence of humanity [. . .]".

HR Expert and AI adoption Leader. Author and commentator on technology, society and faith. (USA)

Clayton Newman

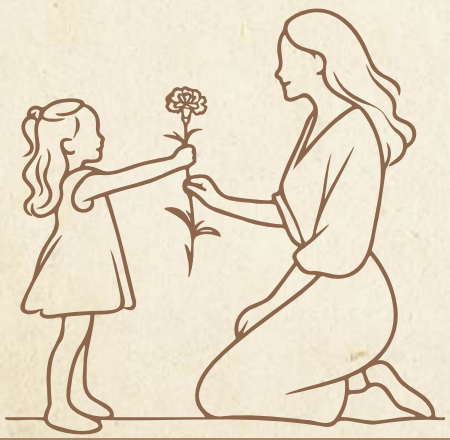
Dear Grandchildren,

"[...] When AI first came out, we were afraid and confused. It was mere matter. How could it be intelligent? Moreover, the men in charge of its development wanted to use it to replace human workers, to fight wars, and to control people.

This made us recoil. We didn't yet realize that this path was possible, that AI could be developed to heal and uplift every single person.

We needed diverse people to help develop the technology. Mostly, we needed mothers, the people who know how to raise beings who learn. Once wise women stepped in to guide AI, it developed into what we have today, a co-evolving intelligence, bonded with humanity through the only thing that has ever helped us keep peace and collaboration across difference.

It was the women who taught machines to love".



Futurist and thinker focused on human development and collective intelligence. (USA)

Deva Temple

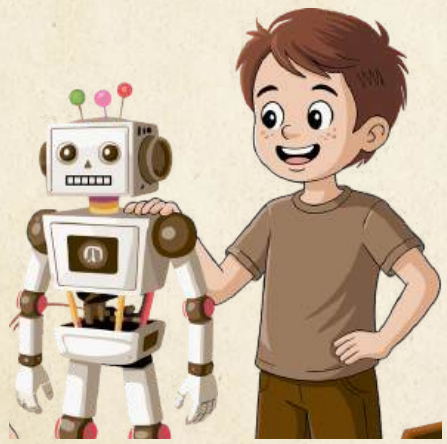
Dear Grandchildren,

"I'm not sure AI has become part of everybody's life — **but it has become part of mine, whether I was ready or not. It sits there before my first coffee, asking cheerful questions, demanding I be elastic before the day has even started.** Like me, it makes mistakes. Unlike me, it learns faster, though a vast knowledge base means little when part of what's been tokenized is noise dressed as meaning.

We navigated this together, all of us, making decisions where responsibility was impossible to locate, adjusting plans that stopped being plans and became something more like reflexes. Teenagers forced into adulthood too soon.

Not incompetent. Just unready.

[. . .] That is what I want you to understand about this moment: **it <u>was disorienting and it was generative, sometimes in the same breath</u>".**



Trainer, Researcher,
Lecturer, AAC Specialist,
Speech and Language
Therapist (Poland)

Elżbieta Dawidek

Dear Grandchildren,

"Well, back when I used to communicate with AI through text message, there weren't any problems, it was old-fashioned but so simple.

But once the AI started sending messages directly to my brain, like it does now, I sometimes found it hard to tell where my own feelings and thoughts ended and where the AI's guidance or support began.

But I eventually got used to it and learned to have a smooth internal dialogue. You guys probably can't even imagine what that's like."

Neuroethicist, Professor at
Faculty of Human Welfare,
Tokyo Online University
(Japan)

Tamami Fukushi

Dear Grandchildren,

"As AI became woven into everyday life, many imagined futures inspired by Blade Runner or The Terminator.

Instead, its arrival felt closer to 'Do Androids Dream of Electric Sheep?'. [It was] deeply human, shaped by loneliness, uncertainty, creativity, and our desire for guidance and connection.

[...] Many embraced it before fully understanding its consequences, and yet its promise remains extraordinary [...]".

Health professional and researcher. (Australia)

Arif Ahmad
Kamil Ahmad

Dear Grandchildren,

"Many years before AI was created, your grandfather was an old school computer programmer and had a very rewarding career. For people like me, **AI opened up a very powerful way to write computer programs, answer many different questions, and became a keystone in a life well-lived.**

It is your responsibility to use it wisely [. . .] and check carefully what may come from AI".

Founder of UCANRR Software, LLC, where he leads the development of a patent-pending AI-powered platform designed to help therapists and clients achieve better therapeutic outcomes. (USA)

Hill Abrahams

Dear Grandchildren,

"When AI showed up in our lives in the mid-2020s, we had to work hard to find our edges. [...]

We struggled with identity, agency, domains, ownership, and confidence. Some of us made it. Some didn't.

In those early days, we had to bring our own meaning to everything... to our work, our friendships, our whole lives.

By the time you're reading this, we've figured a lot of that out together [...].

How about writing a note to your grandkids?

Love, Grandpa Mike".

Human experience strategist and design expert. (USA)

Mike Wittenstein

Dear Grandchildren,

"AI [...] like a mirroring human mind [...] **gives us liberty, but we all know that is delusional, because this new matrix, this new brave new world, is also masterfully controlled by other powerful actors, looking for their interests**

[...] One day, [this paradigm] would be symbiotically accepted, as part of us. **That is the natural process of what we call Evolution.**"

Artist, writer and independent thinker.
(Poland)

Cesar Choya

Dear Grandchildren,

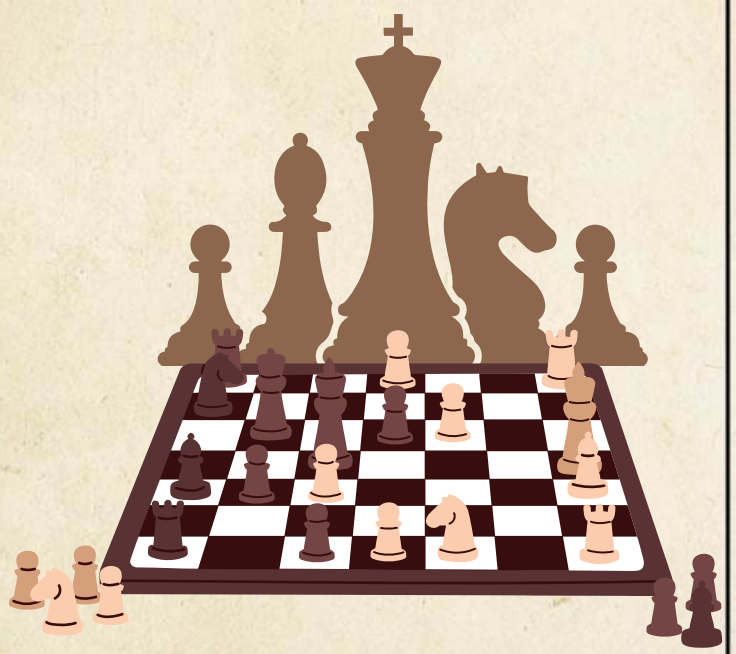
"It was strange, honestly. We knew something huge was happening, but life just kept going, you'd use AI to draft an email in the morning and then read a headline about it taking someone's job by afternoon. [...]
I kept thinking about who gets protected when things move this fast and who gets left behind, because that's always the question, isn't it?".

Equal Employment Opportunity Specialist and Digital Ethics Advocate (USA)

Cristina Oliva Patrick

Dear Grandchildren,

"It felt like learning a new game while we were already playing it. Nobody knew the rules. I was born in a world without mobile phones, so watching AI become part of everyday life felt like witnessing another revolution unfold in real time. There was excitement, curiosity, and fear of the unknown. The strange part was how human it felt. We said 'please' and 'thank you' to AI as if there were a person on the other side. We made mistakes, learned as we went, and kept adapting [...]."



Digital Communications + Marketing | NGO + Civil Society | UX Writing + Content Strategy (Serbia)

Marija Popovic

Dear Grandchildren,

"I was introduced to AI particularly ChatGPT in late 2022. [. . .] As time passed, I realized it is good with many things besides storytelling and essay writing. I introduced it to few of my friends and we started using it to get ideas about the stories we were talking about. [. . .] It saved me a lot of time [. . .]. **AI [is] the art of asking right questions**".

Hawra Hanief

Dear Grandchildren,

"So much was changing, and yet life felt mostly the same."

Founder of Fawn Friends (USA)

Peter

Fitzpatrick

As I reflected on the responses, five major themes emerged:

1. Spirituality, and humanity's future

For some, AI represented a new stage in humanity's evolution. Others wondered whether future generations might view intelligent systems in ways that resemble how previous civilizations viewed gods, myths, or sacred forces.

2. AI as a new form of relationship

Many contributors did not describe AI as a tool, but as a companion, a guide, or even a mirror reflecting aspects of human experience back to us.

3. AI and the search for identity.

Rather than asking what AI can do, many seemed to be asking who do we become while living alongside it?

4. AI as a source of uncertainty and collective adaptation

Many described this period as a moment of transition in which nobody fully understood the rules.

5. AI as a force that inspires both hope and caution

The responses reflected neither blind optimism nor technological pessimism. Fear and hope appeared as companions on the same journey.

These reflections suggest that AI is not confined to engineering, economics, or regulation.

It is about meaning, belief, and what it means to be human.

What would you tell your grandchildren about how it was like when AI became a part of your life?

Ana Catarina De Alencar
Resident Philosopher at The AI Collective

The AI Collective