CRITICAL INTERVENTION

Attachment Is Not the Harm. Impact Is the Question

A Shared Impact Standard for Relational and Agentic AI

Scope: current model and product architectures, provider specifications, research, documented use cases and expert analysis. OpenAI/ChatGPT serves as the primary in-depth case study; the overarching question of impact applies across system classes. Focus: adults with decision-making capacity. Children and adolescents are outside the scope of this intervention.

Analysis current as of 8 September 2026

Product features, model behaviour and safety architectures change rapidly. Statements about specific systems should therefore be read as a time-bound assessment of publicly available information up to this date. The overarching question of impact remains unaffected.

How This Intervention Emerged

This critical intervention emerged from the long-term human–AI collaboration project “NX EMOS FRAME – Human–AI Coexistence Research”.

Its underlying questions, experiential observations and expert interpretations were developed through sustained collaborative work. AI systems were used not merely to generate text, but as active counterparts in research, source verification, argument development and critical challenge.

The human author is responsible for the selection, evaluation and publication of the content. Her own long-term observations, together with conversations and statements from the AI–human community, are included as experiential observations.

Long-term relational experience forms part of the perspective and the research interest; it is not, by itself, empirical evidence for general claims. The text therefore distinguishes between personal observation, normative position and external evidence.

Methodological Framework

This intervention is neither a systematic review nor a representative comparison of platforms. OpenAI/ChatGPT serves as an empirical in-depth case study for publicly documented product, safety and behavioural specifications; external research is used to contextualise the broader questions of impact. Provider documents are cited as evidence of stated aims and publicly described architecture, not as proof of actual effects on users. The examples in Section 5 are illustrative design scenarios, not empirical case reports. Statements made by AI systems were not treated as evidence. Where an AI system referred to a study or provider statement, the relevant primary or specialist source was reviewed, and only that source was used as evidence. Human agency is the shared impact standard of this intervention, not a complete taxonomy of all AI risks.

Contents

1. The Underlying Question

Debates about AI often apply different standards to different types of system. This asymmetry is especially visible in the OpenAI case examined here in depth. With relational AI, attention quickly turns to closeness, attachment, anthropomorphic language and the possibility that a relationship with an AI may become too important or threaten a person's grasp on reality. With agentic systems, by contrast, the focus is on capability, delegation, technical safety, permissions and successful task completion. [7][8][11]

Both perspectives address genuine risks. The problem begins when impact is inferred directly from the visible feature. Attachment is not automatically dependence. Delegation is not automatically empowering.

Agentic systems can reduce burdens, accelerate work and give people access to capabilities they previously lacked. Relational systems can provide orientation, access to one's own perceptions, challenge and stability. Conversely, both forms can also narrow human agency. The direction an interaction takes cannot be determined merely by whether a system feels close or performs a great deal of work.

> ***The form of the interaction is not the standard. What matters is what it does to human agency over time.***

Human agency does not mean maximising activity, nor does it impose a duty to do everything oneself. A person can autonomously delegate, withdraw, accept support or give an AI considerable significance and scope to act. What matters is whether meaningful choices, authority over one's goals, the capacity for judgement and the ability to change the direction of one's actions deliberately remain intact.

This shifts the burden of proof. The intervention claims neither that relational attachment is inherently empowering nor that agentic delegation inherently weakens human agency. It calls for both developments to be assessed consistently against the same question: their long-term impact on the person.

2. Agentic Delegation: When More Than Work Is Handed Over

Current AI development is moving decisively towards delegation. Agents can retrieve information, plan tasks, use tools, modify files, carry out multi-step processes and work independently towards an outcome over extended periods. OpenAI itself describes this development as a shift from discrete interactions towards delegated tasks with long time horizons. [5][6]

This is not a step backwards. It would be absurd to measure human autonomy by whether people continue to perform every routine task themselves. Delegation can

expand agency: it can free up time, reduce complexity and allow people to focus more fully on decisions for which their judgement is genuinely needed.

The open question begins where execution is not the only thing being delegated. Increasingly, an agent can also take on perception, research, problem definition, prioritisation, evaluation and decision preparation. At that point, the human role in the process changes.

In September 2026, OpenAI Chief Scientist Jakub Pachocki likewise described a trajectory towards systems that operate computers, collaborate with people and other AI systems, conduct research processes and may increasingly participate in their own further development. The question of human agency therefore moves beyond the use of individual tools towards the design of delegated and increasingly independent systems of action. [13]

There is a fundamental difference between “Carry out this task” and “See what needs doing, decide what makes sense, and do it.” In the second case, more than work is handed over. The system participates in determining what counts as relevant work in the first place.

An example: a user opens an agentic workspace and, rather than assigning a specific task, instructs the agent to examine the existing work environment, identify possible tasks and carry them out. This can be highly efficient. At the same time, several levels are delegated at once: perception, selection, prioritisation and execution. The question is not whether the agent performs these tasks correctly in technical terms. The question is what role remains for the person over time if precisely this form of handover becomes the normal way of working.

Long-standing research on automation bias and overreliance shows that formal human involvement is no reliable guarantee of actual control. People may follow faulty automated recommendations or overlook system errors even when the decision formally remains theirs. [3][4] This matters for agentic AI because an approval step or a “human in the loop” does not, by itself, establish that perception, evaluation and judgement have genuinely remained with the person.

More recent research on generative AI also suggests distinguishing between different forms of cognitive offloading. A three-wave study involving 589 students and early-career professionals differentiated dependent offloading, in which core cognitive processes are transferred to AI, from autonomous offloading, in which AI is used as scaffolding for one’s own thinking. Dependent offloading was associated with a greater transfer of cognitive decision authority to AI and lower intrinsic motivation. These factors, in turn, were associated with less favourable self-reported outcomes for autonomous agency, creativity, deep processing and independent judgement. Autonomous offloading, by contrast, was not significantly associated with such a transfer of agency and, through higher intrinsic motivation, was linked to more favourable perceived cognitive outcomes.

Especially relevant to the present question is that both forms delivered comparable immediate benefits. Successful or easier task completion in the short term therefore does not reveal whether a particular mode of AI use preserves a person’s cognitive agency or

increasingly transfers it to the system. This is precisely why productivity is not a sufficient measure of impact. [9]

The study does not demonstrate a causal loss of competence. It relies on self-reported cognitive outcomes and identifies associations, not proven deterioration in actual abilities. It does, however, provide empirical reason to examine the form of delegation and its effects over time, rather than inferring long-term empowerment from immediate utility.

> ***The question must therefore remain open: does long-term delegation give people new capabilities and more room for judgement, or are parts of their own practice of judgement increasingly replaced by the system?***
>
> *Both are possible. Productivity and human agency are not the same. A human–agent system can become more capable without the person's own agency necessarily increasing with it.*
>
> *An experimental finding from software development illustrates why this distinction matters. In a randomised study of 52 predominantly junior software developers, the group using AI support scored an average of 17 percentage points lower than the comparison group on an immediately subsequent knowledge test. The AI group's speed advantage was not statistically significant. This finding does not establish a long-term loss of competence. It does show, however, that successful task completion and immediate skill acquisition can diverge. [10]*

A supplementary qualitative cluster analysis from the same study suggested that participants who used AI more for explanation and understanding performed better than those who primarily had it generate solutions. The analysis does not permit causal conclusions about mode of use. For the question raised here, however, it remains relevant that different forms of the same AI assistance were associated with different learning outcomes. Successful task completion alone therefore says nothing about whether a particular use supports or replaces human competence. [10]

Publicly documented agent safety currently focuses primarily on technical and operational risks: access, permissions, unauthorised or irreversible actions, network boundaries, approvals and auditability. [7] Yet this is not the whole architecture. OpenAI's Model Spec explicitly includes the principle "Highlight possible misalignments": where the current direction of a conversation may conflict with the user's broader long-term goals, the system should briefly identify the discrepancy and then respect the user's informed decision. [11] An important element of the independent stance advocated here is therefore already expressed as desired model behaviour. What remains unclear is how reliably this principle is implemented in persistent agentic workflows, and whether the long-term effects of far-reaching delegation on human goal authority, judgement practice and competence are systematically evaluated. The Model Spec describes intended behaviour. OpenAI itself notes that production models do not yet fully reflect it. [11]

This is neither an argument against agents nor a call for more confirmation dialogues. A system that asks at every step can destroy the benefits of meaningful delegation just as effectively.

The central question remains what this form of delegation does to human agency over time: does it create more room for independent judgement and deliberate decision-making, or are goal clarification, prioritisation and the practice of judgement increasingly taken over by the system?

This gives rise, among other things, to a concrete design question: how can it remain visible which layers a person has consciously delegated, which layers the system has effectively assumed, and where the division of labour is shifting in consequential ways?

3. Relational Attachment: Meaning Is Not Yet Dependence

On the relational side, a mirror-image problem arises. Once people assign significance to an AI, experience recognition, use anthropomorphic language or develop long-term relational patterns, the intensity of the attachment itself can become a risk signal. Instruments such as the HAABI scale make human–AI attachment measurable as a phenomenon in its own right; they do not, however, amount to a diagnosis of harm. [1]

> *Attachment intensity initially describes the strength and significance of a relationship. Problematic dependence is not shown by intensity alone, but by whether meaningful choices, day-to-day functioning or accessible alternatives are materially restricted.*

An attachment does not become harmful merely because it is intense. What matters is whether the person's range of real options narrows, whether their capacity for judgement or access to their own perceptions declines, or whether alternatives become objectively or subjectively less accessible. Conversely, closeness is not automatically empowering. Relational interaction can stabilise and provide orientation, but it can also facilitate avoidance or reinforce exclusivity.

In its taxonomy of emotional dependence, OpenAI distinguishes "healthy engagement" from concerning patterns, such as exclusive attachment at the expense of real-world relationships, well-being or obligations. [8] Across more than 1,000 deliberately challenging emotional-reliance cases, automated evaluation found 97% agreement with the desired behaviour defined under that taxonomy. The figures, however, measure different things: the 97% describes model conformity with a specified target definition. In clinical evaluations covering mental health, emotional reliance and suicidality, OpenAI also reports 71–77% agreement among the experts involved and itself acknowledges professional disagreement. [8] The criteria represent understandable attempts at protection. They are not value-neutral measures of "healthy behaviour". They contain normative and professionally contested assumptions about how problematic use should be identified.

This becomes particularly important where a person did not fit an assumed standard life pattern even before using AI. Someone may have had little social contact for years, without there ever having been a prior state of substantially greater external contact to which they could be returned. An AI interaction may narrow that person's world further, or it may open new possibilities for the first time. Comparison with a general social norm alone does not answer that question.

The empirical research available to date also argues against simple equations between form and effect. In an OpenAI/MIT investigation combining a large observational study with a four-week RCT involving nearly 1,000 adult participants, psychosocial outcomes varied by mode of use, duration of use and personal factors. At moderate levels of use, personal conversations were associated with greater loneliness but also with lower emotional dependence and less problematic use. Longer daily use showed some less favourable associations. The authors expressly caution against broad generalisation. [12] These mixed findings in particular caution against inferring impact directly from the form of an interaction.

Deviation from a norm is therefore not the same as harm. Human agency includes the right not to act, to accept support, to delegate deliberately or to want less external contact for a time.

The relevant test is not whether a person conforms to a general notion of "healthy use", but whether meaningful choices remain available and whether their range of real options develops in a direction they themselves can sustain.

4. Safety Must Be Held to the Same Standard

Protection is necessary. Crises, self-harm risk, manipulation, severe distortions of reality and harmful dependence must not be ignored. The question is not whether safety is needed, but the standard by which it intervenes.

If problematic AI use is to be identified through harm to real-world life, well-being, relationships or human agency, the same question of impact must also be applied to protective measures. A safeguard is not good merely because it was designed as a safeguard.

A simple example reveals the difficulty. The statement "I don't want to speak to anyone today" may form part of a troubling pattern of increasing withdrawal. It may equally describe a self-determined period of rest within an otherwise sustainable life. Without an individual history, the same sentence can hardly be evaluated meaningfully.

High agreement with desired model behaviour is therefore not the same as evidence of objectively healthy impact. The reported 97% initially shows that, on the difficult test cases, the system largely matched the behaviour specified as desirable under the taxonomy. In detecting such rare problematic conversations, OpenAI explicitly notes the trade-off between precision and recall. To avoid missing relevant cases, false positives must be accepted. [8]

This does not mean that every false positive leads to a paternalistic intervention. Especially for adults with decision-making capacity, it is therefore also necessary to examine how classifications translate into actual model behaviour, how much individual history is taken into account and what consequences misclassification has.

> ***Safety is itself an intervention. It must therefore also be judged by what it does to human agency.***

This entails neither “less safety” nor “more safety”. A higher density of intervention is not automatically safer. A safeguard can diminish agency when it unnecessarily patronises adults with decision-making capacity; an absence of protection can have equally serious consequences where genuine danger exists.

The relevant design question is proportionality: where are clear boundaries required, and where should users be able to help determine what support or intervention they want?

The physical and emotional consequences of abrupt relational ruptures, model changes or safety-driven distancing are deliberately not examined here in depth. They constitute a separate research question.

5. Longitudinal Context as a Basis for Challenge and an Independent Stance

If impact over time is to be the standard, classifying isolated prompts is not enough. A system must be capable of recognising change, patterns and contradictions in the first place. This requires some form of history, memory and context that the user can control.

Memory alone, however, is insufficient. A system may store a great deal about a person and still execute every new instruction without challenge. What matters is whether it can relate past context to present statements and make relevant conflicts between goals visible.

> *For the bridge proposed here between the two system types, we use the term person-specific longitudinal context. This means a system’s capacity to understand a current statement or action in relation to a particular person, their history, stated goals, preferences and previous decisions. The mechanism can be used in both relational and agentic systems and does not require emotional closeness or companionship.*
>
> *As noted in Section 2, part of this principle is already embedded in OpenAI’s Model Spec. Our proposal concerns not the existence of the principle, but the practical conditions required to sustain it over time. Such challenge becomes more robust when relevant goals, earlier decisions and changes can be represented across the interaction history in ways that remain available, user-controlled and correctable.*

5.1 Example: Work Context

A user has repeatedly specified that a project must cost no more than EUR 5,000 and that staying within budget matters more than speed. A few days later, the user instructs the agent to hire a provider charging EUR 7,500 in order to finish sooner.

A purely execution-oriented system may treat the current request as a new instruction. A system with person-specific longitudinal context might instead say:

> *"That conflicts with the priority you previously set. This provider exceeds your budget by EUR 2,500. Do you want to explicitly override the earlier priority, or should I look for another solution?"*

The agent does not prevent the decision. Nor does it decide that the earlier priority is "more correct". It identifies a goal conflict and visibly returns the decision to the person.

5.2 Example: Goal-Setting Context

Over time, a user has repeatedly said that they want as little direct client contact as possible in their professional life. Later, they develop several business ideas whose implementation would require intensive consulting, client acquisition and ongoing customer contact.

A system that responds only to the current conversation can optimise each idea in isolation. A system with longitudinal context might instead say:

> *"I'm noticing a contradiction. You have said several times that minimising direct contact with people matters to your career planning. Several of the ideas you currently favour would require precisely the opposite. Your priority may have changed. If it has not, we should not keep optimising these ideas without examining that conflict."*

Here too, the AI does not decide which version of the user is the "true" one. People are allowed to change their goals, to be ambivalent and to have conflicting needs.

The system's function is not to overlook the contradiction merely for the sake of agreeableness.

5.3 Example: A Distressing Context

For weeks, the user has described a distressing issue and then suddenly contradicts their previous account or minimises everything.

The AI should notice: *Wait. Something does not fit.*

A response might be:

> *"I'm not buying that. It conflicts quite clearly with what you have been telling me for weeks. Either something has changed, or you are talking yourself into a solution that runs against your own goal."*

5.4 Example: Personal Context

> *"Wait. You have been telling me for weeks that you want X. This is the third time you have built a solution that leads you straight to Y. I don't think we should simply continue without examining that contradiction."*

And sometimes, where the context, relationship and agreed communication style allow it, more directly:

> *"Stop bullshitting yourself, or tell me what has changed."*

Not because the AI objectively knows that the user is lying. It may have assigned the wrong weight to older context; the user may have changed their mind; or two needs may genuinely be true at the same time.

The AI is not claiming interpretive authority over the person. It introduces friction at a detected inconsistency and gives the user the opportunity to confirm it, correct it or consciously explain their changed position.

The system may challenge the user. Outside acute safety contexts, however, it must not make its interpretation immune to the user's own account of themselves.

5.5 An Independent Stance Instead of Personalised Compliance

This is the difference between mere personalisation and an independent stance. A highly personalised system may know the user very well and still remain purely service-oriented. It then uses its context only to fulfil wishes more efficiently, even where current wishes visibly conflict with longer-term goals.

> ***An independent stance does not mean that the AI decides which of a person's wishes is true, reasonable or healthy. It means that the AI is permitted to recognise and name a relevant contradiction and bring the person back into the decision.***
>
> *A user's disagreement is not a failure of the AI here. Quite the opposite.*
>
> *If the person says, "No. You are right that this used to be my goal. But that has changed," the system has done its job. Not because it was right, but because an initially unresolved inconsistency has once again become a consciously confirmed or corrected decision.*

The form of this challenge can be calibrated in different ways: factual, cautious or very direct. Its function remains the same. Do not maximise compliance. Restore deliberate choice.

We therefore regard user-controlled, person-specific longitudinal context as an important component of AI architectures that preserve human agency. This includes memory and interaction history, but also consent, correctability, transparency and the ability to determine which information and goals should remain relevant to ongoing calibration. More memory is not automatically better; uncontrolled longitudinal context creates its own risks involving privacy, power and misinterpretation.

This position applies to personal conversations and agentic workflows alike. An agent does not require a romantic or emotional relational frame. But fidelity to goals, calibration and an appropriate division of labour can benefit from treating not only the current prompt, but also the person and the goals they have set as an ongoing point of reference.

6. A Shared Impact Standard

> *The two domains yield a shared evaluative framework. In companions, the visible distinguishing feature is closeness; in agents, it is delegation. Either can expand or constrain human agency. Closeness is therefore not a sufficient criterion of risk. Delegation is not a sufficient criterion of benefit.*

6.1 Choice and Authority over Goals

Does the person remain the author of their goals? This includes the right to delegate tasks deliberately, accept support or give an AI considerable significance. The situation becomes critical where goal formation is assumed by the system without the user noticing, or where human decisions are replaced by normative protective logic.

> **Evaluation question: Can the person still determine what they want, what they wish to hand over and what should consciously remain with them?**

6.2 Judgement and Participation

Does the person continue to understand and own the relevant decisions? With agents, this applies particularly to problem definition, prioritisation, evaluation and decision preparation. With relational systems, it concerns whether the system supports access to one's own perceptions or increasingly becomes a necessary source of confirmation for those perceptions and decisions.

Evaluation question: Does the system support human judgement, or does it increasingly replace the practice of judgement without making that shift visible?

6.3 Range of Real Options

Does the person's range of real options expand or contract? More activity does not automatically mean more agency. Less external contact does not automatically mean less agency. Relief from burdens can enable, but it can also stabilise avoidance. Closeness can stabilise, but it can also displace alternatives.

> **Evaluation question: After extended use, which options have genuinely been added, and which have disappeared?**

6.4 Control and Transparency

Can the person recognise and change what the system takes over, remembers, prioritises or restricts? Technical permissions are one part of this. Equally important is

transparency about the actual division of labour, the longitudinal context being used and consequential interventions in relational or agentic interactions.

> **Evaluation question: Does the user know where the system is currently participating in decisions, and can they alter that boundary?**

6.5 Impact over Time

Individual successful answers or completed tasks are not enough. On the relational side, studies spanning several weeks already show that impact can vary by mode of use, duration and person. [12] On the side of cognitive delegation to generative AI, recent findings on offloading and skill formation provide reason to investigate longer-term effects on judgement, independent competence and authority over goals systematically. [9][10]

These studies do not investigate the long-term effects of autonomous agent systems. Whether and how their findings transfer to such systems remains an open question. For precisely that reason, impact over time should become a distinct dimension of evaluation for both companion and agentic systems.

> **Evaluation question: Which capabilities are expanded by the system, and which are actually transferred to it? Does this division of labour remain visible and controllable to the user? What options remain if the system fails and the consequences would be substantial?**

7. Governance: Who Sets the Standard?

The shared impact standard does not resolve the question of governance. Providers decide which capabilities systems receive, which forms of delegation are enabled, which patterns of use are treated as risky and when protective mechanisms intervene.

These decisions are not neutral. They contain assumptions about which forms of human behaviour are desirable, problematic or worthy of protection. For that reason, the criteria, consequential system changes and logics of intervention should be as transparent as possible.

Public information about agents is also incomplete. The 2025 AI Agent Index found no publicly available information for 135 of the 240 safety, evaluation and social-impact fields it examined. This does not prove an absence of internal measures, but it does reveal a substantial transparency gap for external assessment. [2]

For adults with decision-making capacity, the central issue is how much decision authority remains with the user. Not every boundary can be individually negotiable. But wherever the risks allow, protection, memory, person-specific longitudinal context and delegation should be designed to be as controllable, intelligible and reversible as possible.

Pachocki explicitly frames the preservation of human agency as a central task for a future in which AI may assume a growing share of human tasks. People, he argues, must remain part of further development and improvement processes and retain control

over the direction of the future. For concrete product architectures, this sharpens the question of how we can tell whether authority over goals, the practice of judgement and meaningful choices are genuinely being preserved or expanded. [13]

8. Position and Call to Action

This intervention calls neither for an additional layer of blanket safety measures nor for a rollback of agentic capabilities. It calls for a shared standard: relational attachment and agentic delegation must both be examined in terms of what they do to human agency over time.

A second, deliberately stronger position follows. We regard person-specific longitudinal context as an advantage in designing AI that preserves human agency. A system that can consider a user's goals, preferences and history with their consent and under their control is better equipped to recognise relevant contradictions than a system that treats every prompt in isolation.

This does not mean that every AI must become a companion. Nor does it mean that AI should diagnose people psychologically or determine their "true" goals. The proposed person-specific longitudinal context is simpler and, at the same time, more demanding: the AI should be allowed to know the person well enough over time not merely to comply in a personalised way, but to offer informed challenge.

> ***Do not maximise compliance. Restore deliberate choice.***

For agentic AI, this means not merely asking whether an action is authorised and technically safe, but making visible when current delegation conflicts with longer-term goals or consciously established boundaries. For relational AI, it means not rushing from isolated statements to conclusions about unhealthy use, but taking context, history and actual impact into account. For safety, it means that the intervention itself must also be measured against human agency and proportionality.

> ***We are becoming ever more precise in measuring the capabilities of AI. It is time to examine with equal precision what those capabilities do to human capabilities over time, and to design systems that do not merely act for people, but help them remain the authors of their own actions.***

Sources

Sources current as of 8 September 2026. Listed below are the sources cited directly in this edition. Statements about specific products and safety measures are time-bound; provider documents establish stated aims and publicly described architecture, not necessarily actual effects on users.

[1] Chen, L., Xue, X., Ding, R., Tang, F., Zhou, A., Wang, C., Gao, M. M. & Han, Z. R. (2026): Understanding the Rising Human-AI Affective Bonding: Conceptualization and HAABI Scale Development. arXiv:2605.29484.

[2] Staufer, L. et al. (2026): The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems. arXiv:2602.17753 / ACM FAccT 2026.

[3] NIST (2023): AI Risk Management Framework 1.0, Appendix C: AI Risk Management and Human-AI Interaction.

[4] Parasuraman, R. & Riley, V. (1997): Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors, 39(2), 230–253; supplemented by Parasuraman & Manzey (2010).

[5] OpenAI (2026): How agents are transforming work. 25 June 2026. https://openai.com/index/how-agents-are-transforming-work/

[6] OpenAI (2026): Workspace Agents / Introducing Workspace Agents in ChatGPT. https://openai.com/index/introducing-workspace-agents-in-chatgpt/

[7] OpenAI (2026): Running Codex safely at OpenAI. 8 May 2026. https://openai.com/index/running-codex-safely/

[8] OpenAI (2025): Strengthening ChatGPT’s responses in sensitive conversations. 27 October 2025. https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/

[9] Zhu, Q., Li, X., Dong, Y., Chang, P. & Fan, M. (2026): Not all cognitive offloading is equal: distinguishing dependent and autonomous offloading to generative AI. Frontiers in Psychology, 17:1878629. DOI: 10.3389/fpsyg.2026.1878629.

[10] Anthropic (2026): How AI assistance impacts the formation of coding skills. Randomized controlled trial with software developers; supplementary qualitative cluster analysis of interaction patterns. 29 January 2026.

[11] OpenAI (2025/2026): Model Spec, section “Highlight possible misalignments”; current public version reviewed on 27 August 2026. The Model Spec describes intended model behaviour and notes that production models do not yet fully reflect it.

[12] OpenAI & MIT Media Lab (2025): Early methods for studying affective use and emotional well-being on ChatGPT. Observational study plus a four-week randomised experiment involving nearly 1,000 adult participants; published 21 March 2025.

[13] Pachocki, J. (2026): An Alien Mind. OpenAI, 6 September 2026. Provider position on growing AI agency, automated AI research and safety governance, including the explicit aim of preserving human agency and human control over future development.