

# The Cadmium Rod of Artificial Intelligence

*What AI labs are hiring for, and what is missing. October 2026.*

Author: Adrian STAN

ORCID: <https://orcid.org/0009-0003-1457-5155>

Date: 5 October 2026

## Abstract

**AI labs say that safety is their priority.** This paper confronts that claim with a public and reproducible indicator: the positions they advertise. We analysed 1,976 job postings from Anthropic, OpenAI, xAI and Mistral AI, collected from the official pages on 4 October 2026 and classified by the department declared by the company and by a published set of rules. Safety postings account for between 0% and 9.6% of the total, and safety research proper for under 2%. **User-oriented postings are under 1%** at all four companies and, judging by title, **none concerns the cognitive effort** of the person using the model. Outside the home country, between 59% and 75% of postings are commercial. Archive captures from September show that the proportions are stable. The results describe **the direction of hiring**, not the size of existing teams.

In a nuclear reactor, **cadmium rods** absorb neutrons and hold the chain reaction at the level chosen by the operator. They do not stop the reactor and do not reduce its usefulness; they **make it governable**. We argue, accordingly, that AI assistants need an equivalent element for the user's thinking. The literature on the effects of AI assistance on learning shows that an answer received without effort can weaken the very competence it appears to increase, and that deliberate safeguards mitigate this effect. Such cognitive friction appears today in a few products, as a separate mode intended for learning, but not as a graded, general and visible control. We propose turning it into an explicit parameter: the **Cognitive Slider**. The mechanism is technically feasible; a working approximation has been built from the outside, with modest means. Its effect on learning remains to be measured.

## Introduction: why the cadmium rod

In 1942, in Chicago, a team led by Enrico Fermi achieved the first self-sustaining nuclear chain reaction. The reactor was held in check by cadmium rods, a metal that absorbs neutrons. The reaction started when the last rod was withdrawn, a little at a time, by a human hand, and stopped when the rods were pushed back in. The means of control was not added afterwards. It was part of the first experiment.

The comparison between artificial intelligence and nuclear energy is usually made with the bomb in mind: catastrophic risk, an arms race. This paper does not deal with those themes. Our comparison is with the nuclear reactor, a useful technology whose usefulness depends on the fact that it can be steered.

The analogy supports three things. 1. An AI assistant, like a chain reaction, tends to sustain itself. This is a hypothesis, not a result: each answer received without effort would make the next request more likely. 2. Nobody wants the reactor shut down, just as nobody wants AI assistants abolished. 3. Between “off” and “uncontrolled” there has to be a control, which the operator sets.

The analogy stops in three places. 1. A reactor out of control fails quickly and visibly; the effects on thinking are slow, diffuse and only partly measured. 2. The physics of the reactor is known with precision; the right dose of friction is not (yet). 3. In a reactor the rods are mandatory. We ask only that they exist, and that their position be chosen by whoever is responsible for the task: the user for their own work, the teacher for the assignment they set, never the provider.

## Chapter 1. What AI is: an ecological factor

**1.1. Addition and transformation.** In 1998, Neil Postman observed that technological change is not additive; it is ecological. Adding dye to a glass of water does not give water plus dye; it changes the colour of the whole glass. A tool is judged by what it does. An ecological factor is judged by what everything else becomes in its presence: school, work, the way a question finds its answer. This paper treats AI assistants in the second sense (Stan, 2026a).

**1.2. A loop with three links.** The mechanism we are tracing can be called *bidirectional algorithmization*: people’s written thinking is compressed into a statistical model, and then people think through that model.

**From human to model.** A language model learns from text. The available text is **not humanity’s thinking, but the part of it that is written, digitized and accessible**, in which the recent period and the English language carry far more weight than they do in the world’s population.

**From model to human.** Doshi and Hauser (2024) asked participants to write short stories, some of them with ideas offered by a model. The result has two parts, equally important. The stories written with help were rated as more creative, especially for the less creative writers. But those same stories resembled one another more than the ones written without help. **The gain is individual, the loss is collective.** The effect has also been observed in idea generation (Anderson et al., 2024) and in essay writing, where it appeared only with the model tuned on human feedback, not with the base model (Padmakumar and He, 2024). It is not inevitable, however: when the ideas offered by the model were deliberately made diverse, collective diversity was preserved (Wan and Kalman, 2025). Homogenization therefore depends on how the tool is delivered.

**From model to model.** Text produced with the help of models becomes, in turn, training material. Shumailov et al. (2024) showed that models trained repeatedly on their own outputs first lose the rare cases, then lose diversity as a whole - the phenomenon has been called “model autophagy disorder” for image models (Alemohammad et al., 2024) and, colloquially, “Habsburg AI”. Here too there is a condition: degradation appears when synthetic data replace real data. If they are merely added to them, the error stays bounded (Gerstgrasser et al., 2024).

**1.3. What follows.** The three links form a loop that can narrow, under two conditions: that the tool is delivered uniformly and that the share of original human material falls. Both depend on how the systems are built and used.

## Chapter 2. What should be done: cognitive friction

We move from the collective level to the individual one: what happens to the competence of a person who receives the answer without effort, and what changes when effort is asked of them.

### 2.1. What the data show.

The first result comes from an experiment involving nearly a thousand high school students, in mathematics (Bastani et al., 2025). The students were divided into three groups: without AI, with an ordinary assistant that gave solutions on request, and with a “tutor” assistant, configured by teachers to offer hints instead of answers. On the practice problems, both AI groups did better than the control group, by 48% and 127% respectively. On the subsequent exam, taken without AI, the group with the ordinary assistant scored 17% worse than those who had not used AI at all. In the tutor group, the loss was largely avoided. The study thus contains, in the same experiment, both the problem and the remedy: access without safeguards weakened learning, and the safeguards mitigated the effect.

The second result concerns perception. Fernandes et al. (2026) gave adults 20 logic problems, solved with the help of a model. The participants scored about three points higher than a reference population without AI, but estimated their own result a further four points above the real one. Those who knew more about AI assessed themselves less accurately. A replication with 452 people confirmed the results. The phenomenon has been described independently as False Cognitive Power Transfer, FCPT (Stan, 2025a): the user attributes to themselves a competence that belongs to the system.

Lee et al. (2025) surveyed 319 knowledge workers: those with more confidence in AI reported less critical effort, those with more confidence in their own skill reported more. Gerlich (2025) found, in 666 people, a negative correlation between heavy use of AI tools and critical-thinking scores, more pronounced among the young. The study counts search engines and recommender systems among AI tools.

### 2.2. What the data do not show.

There is no evidence that the effect is irreversible: Bastani measured one exam, not a trajectory. There is no solid evidence at the neural level; studies of that kind are so far preliminary, on small samples. It is not true that AI harms learning in general. An AI tutor designed on pedagogical principles produced, in a university physics course, better learning than active in-class teaching (Kestin et al., 2025). The common conclusion of these results is that the effect depends on design: the same model, delivered differently, weakens or strengthens. Two studies on professionals complete the picture. Among endoscopists, the adenoma detection rate in unassisted colonoscopies fell from 28.4% to 22.4% after a few months of working with an AI detection system (Budzyń et al., 2025); the study is observational. Among 133 patent lawyers, three months of working with an AI assistant improved their work and, on average, also their result on a later task without AI. The durable gain, however, was concentrated among the experienced; among juniors it did not appear, on average (Autor et al., 2026).

The conditions under which collaboration with AI leaves behind capacities that persist and are passed on are the subject of the CCPT framework, in its revised version (Stan, 2026e). CCPT is not the opposite of FCPT: a person can acquire a capacity and still overestimate their progress.

**2.3. Why friction would work.** The psychology of learning has known “desirable difficulties” for decades: conditions that slow immediate performance and improve retention, such as generating the answer before receiving it or retrieving from memory (Bjork and Bjork, 2011). The ordinary assistant removes exactly these conditions. Collins et al. (2024) tested the opposite experimentally: small frictions introduced into the interface of a language model reduced how often participants turned to it, with minimal effect on accuracy. The study did not measure learning. The same authors also report unintended effects, which spill over beyond the situations targeted. **Friction therefore has to be dosed, not merely added.**

**2.4. Precedents.** The idea of grading the use of AI is not new. The AI Assessment Scale (Perkins et al., 2024) defines levels of permitted use in an assessment, and the AI Friction Scale (Beckwith, 2025) describes the cognitive friction associated with each level. Both are rules for people. The Mechanism of Cognitive Stimulation, MCS (Stan, 2025b; 2025a), proposes that the grading become a behaviour of the assistant, **adjustable by the user.**

**2.5. Criteria.** From the above it follows what one should expect to see at a lab that treats the user’s cognitive effort as a safety problem. We use three indicators, from the weakest to the strongest:

1. *In hiring:* positions whose title or department concerns the user’s learning, overreliance or competence.
2. *In the product:* a control over the division of work between human and model, present in the ordinary conversation, not only in a separate mode.
3. *In the programming interface:* a documented parameter with the same function.

## Chapter 3. The audit of open positions

**3.1. The indicator.** A job posting is a public, dated statement, in the company’s own words, of what it needs at that moment. Taken together, the postings show an organization’s **direction of growth** (they do not show how large each team is - a stable team with no vacancies does not appear at all).

We test two hypotheses.

**H1:** safety postings are a small part of the total.

**H2:** outside the home country, postings are predominantly commercial, and research stays at home.

**3.2. Timeline.** The observation from which this paper starts dates from 26 September 2026, on Anthropic’s careers page. A first exploration, made the same day with an AI assistant, gave totals that were later confirmed (619 postings at Anthropic, 830 at OpenAI). A breakdown by category was also drawn up then, which did not survive verification; it is described in Annex B. The data analysed here were collected by the author on 4 October 2026, between approximately 21:30 and 23:30, Romanian time. For the earlier period we used captures preserved by the Internet Archive: 17 and 27 September and 2 October for Anthropic, 21 September and 2 October for OpenAI.

**3.3. The companies.** The main table covers the companies that publish the complete list of postings, with departments: Anthropic (638 postings), OpenAI (828), xAI (301) and Mistral AI (209), 1,976

in total. Google DeepMind (98 postings) is reported separately: its list contains only research, because the commercial functions and the infrastructure belong to Google. DeepSeek, Moonshot AI and Meta are treated descriptively, for reasons given in 3.9.

**3.4. The classification.** Each posting is one row, regardless of the number of locations. The first layer is the department declared by the company, taken exactly as given. The second layer groups the departments into nine categories: safety research (**S1**), operational safety (**S2**), user-oriented positions (**S3**), public policy (**P**), research and capabilities (**R**), infrastructure (**I**), product engineering (**E**), commercial (**C**) and administrative (**G**). The title can move a posting only toward **S1-S3**, on the basis of a fixed list of words. Public policy and information security are not counted as safety: the former can argue for stricter or looser rules, the latter protects the company. The complete rules, the mapping from departments to categories and the list of moved postings are in Annex A.

We report **H1** in two variants:

1. The narrow variant counts only the departments that the company itself names as safety.
2. The broad variant adds the postings with safety (in the title) from other departments, and is favourable to the labs: it includes, for example, payment-fraud prevention.

**3.5. Disclosures.** The classification and synthesis of the data were done with the help of an AI assistant, Claude, made by Anthropic, one of the companies analysed; the initial exploration used Gemini, made by Google, another of the companies analysed. The classification rules were adjusted along the way, and the adjustments are listed in Annex A.

### 3.6. Results: the share of safety (H1)

*Table 1. Open positions by category, 4 October 2026*

| Category                           | Anthropic   | OpenAI      | xAI         | Mistral AI  |
|------------------------------------|-------------|-------------|-------------|-------------|
| <b>S1</b> Safety research          | 9 (1.4%)    | 16 (1.9%)   | 0           | 0           |
| <b>S2</b> Operational safety       | 48 (7.5%)   | 37 (4.5%)   | 9 (3.0%)    | 0           |
| <b>S3</b> User-oriented            | 4 (0.6%)    | 2 (0.2%)    | 2 (0.7%)    | 0           |
| <b>P</b> Public policy             | 16 (2.5%)   | 5 (0.6%)    | 0           | 1 (0.5%)    |
| <b>R</b> Research and capabilities | 61 (9.6%)   | 56 (6.8%)   | 43 (14.3%)  | 23 (11.0%)  |
| <b>I</b> Infrastructure            | 72 (11.3%)  | 134 (16.2%) | 159 (52.8%) | 35 (16.7%)  |
| <b>E</b> Product engineering       | 94 (14.7%)  | 177 (21.4%) | 41 (13.6%)  | 8 (3.8%)    |
| <b>C</b> Commercial                | 195 (30.6%) | 273 (33.0%) | 12 (4.0%)   | 106 (50.7%) |
| <b>G</b> Administrative            | 139 (21.8%) | 128 (15.5%) | 35 (11.6%)  | 36 (17.2%)  |
| <b>Total</b>                       | <b>638</b>  | <b>828</b>  | <b>301</b>  | <b>209</b>  |

*Table 2. Safety in hiring*

|                                        | Anthropic | OpenAI    | xAI       | Mistral AI |
|----------------------------------------|-----------|-----------|-----------|------------|
| Narrow variant                         | 43 (6.7%) | 42 (5.1%) | 10 (3.3%) | 0          |
| Broad variant                          | 61 (9.6%) | 55 (6.6%) | 11 (3.7%) | 0          |
| Other postings per safety posting      | 9.5       | 14.1      | 26.4      | -          |
| Commercial postings per safety posting | 3.2       | 5.0       | 1.1       | -          |

*The ratios are calculated on the broad variant*

Across the 1,976 postings as a whole, safety accounts for 4.8% in the narrow variant and 6.4% in the broad one. H1 is confirmed, with large differences between companies: from nearly a tenth at Anthropic to no postings at all at Mistral AI.

### Observations

Safety research proper (**S1**) has 25 postings in total, that is 1.3%, but most of *safety* is operational: enforcing rules, countering abuse, threat intelligence. At xAI, four of the ten postings in the department named “Safety” concern financial fraud. And the hiring profile differs sharply: at xAI, more than half of the postings are for data centres; at Mistral AI, half are commercial.

The category that bears directly on this paper is **S3**. It has 8 postings in total, **0.4%**. All of them concern the protection of minors or user well-being, in enforcement or engineering roles. **None has learning, overreliance or the competence** of the person using the model in its title. Seven education postings (5 at OpenAI, 2 at Anthropic) were left under product engineering, because the title does not show whether they concern the user’s effort or features for educational institutions; they are listed in Annex A.

### 3.7. Results: geography (H2)

*Table 3. Postings outside the home country*

|                                                             | <b>Anthropic</b> | <b>OpenAI</b> | <b>Mistral AI</b> |
|-------------------------------------------------------------|------------------|---------------|-------------------|
| Postings only outside the home country                      | 136 (21%)        | 163 (20%)     | 82 (39%)          |
| of which commercial                                         | 80 (59%)         | 123 (75%)     | 57 (70%)          |
| Research (R) with at least one location in the home country | 55 of 61         | 54 of 56      | 14 of 23          |

*The third row also includes postings with locations both at home and abroad: 2 at Anthropic, 0 at OpenAI, 11 at Mistral AI.*

**H2** is confirmed at all three. Outside the home country, between 59% and 75% of postings are commercial, and research stays largely at home. The pattern is the same: abroad, the companies hire mainly sales and deployment staff.

Note: xAI is not in the table: it has almost no commercial postings, and those outside the United States are mostly for remote data annotation.

### 3.8. Stability over time

*Table 4. Broad variant, from archive captures and from our own collection*

| <b>Date</b>                | <b>Anthropic</b> | <b>OpenAI</b>    |
|----------------------------|------------------|------------------|
| 17 / 21 September          | 57 of 606 (9.4%) | 56 of 815 (6.9%) |
| 26 September (exploration) | total 619        | total 830        |
| 2 October                  | 61 of 639 (9.5%) | 56 of 833 (6.7%) |
| 4 October                  | 61 of 638 (9.6%) | 55 of 828 (6.6%) |

The proportions vary by **less than half a point in three weeks**.

### 3.9. The other companies

Google DeepMind had 98 postings, of which 12 with safety in the title (12.2%). The percentage is not comparable with those above, because the list contains no commercial or infrastructure functions. DeepSeek publishes 34 types of position, without a number of openings; none has safety, ethics or alignment in its title. Moonshot AI had 53 postings, of which one for content safety and two for countering abuse. At Meta, among the postings under the “Artificial Intelligence” filter, 6 of the 104 read had *safety* in the title.

### 3.10. Limitations

1. *Flow, not stock.* The postings do not say how large the existing teams are. A safety team that loses people and replaces them shows up as an increase in hiring.
2. *Titles, not job descriptions.* The classification uses the department and the title. The content of the postings was not read systematically.
3. *Non-comparable names.* Each company names its departments differently. “Applied AI” is commercial at Anthropic and product engineering at OpenAI; the assignments are in Annex A.
4. *Incomplete locations.* At OpenAI, 140 postings appear with several locations, without details.
5. *Unpublished positions.* Internal hires or hires by referral do not appear.
6. *A snapshot.* The archive series covers three weeks and two companies. It does not show the long-term trend.
7. *A single rater.* The classification was not verified independently. The rules were written after a first look at the data and subsequently adjusted three times (Annex A).
8. *No reference threshold.* There is no benchmark for the share that safety ought to have in an organization that treats it as a priority. In addition, commercial positions grow with the number of customers, while research positions do not, so a large commercial share is to be expected at any company with revenue.

None of the limitations changes the order of magnitude: **safety is under a tenth of hiring everywhere, and concern for the user’s cognitive effort appears in no title.**

## Chapter 4. The distance between what should be done and what is done

**4.1. In hiring.** Of 1,976 postings, only 8 are user-oriented, all for the protection of minors or for well-being. No title concerns the user’s learning, overreliance or competence. Seven education postings might belong to this concern, but the title does not show it. The first indicator is, at best, weakly present.

**4.2. In the product.** In the summer of 2025, the three major platforms launched learning modes within a few weeks of one another: Study Mode in ChatGPT (29 July), Guided Learning in Gemini (6 August) and the Learning style in Claude, available from April for education accounts and from 14 August for all users (Engadget, 2025). All three replace the direct answer with questions and guided steps.

In August 2026, OpenAI went a step further for minors. Accounts identified as belonging to teenagers automatically receive a variant of the product in which the teenager or a parent can set “study hours”: periods in which new conversations start directly in Study Mode. The product also includes reminders which, when the teenager appears to be taking a shortcut through an assignment, nudge them toward the guided mode. Age is estimated by the platform, and the assignment is recognized by it too. The mechanism nevertheless has **three limits**. 1. It can be bypassed: the teenager can use another assistant or another account. 2. It is not dosed: the same friction is applied to any subject, with no link to the purpose of the assignment. 3. There are no published results on its effect on learning.

*Table 5. The learning modes against the criterion in 2.5*

|                       | <b>Study Mode</b>   | <b>Guided Learning</b> | <b>Learning</b>      | <b>The control we ask for</b>   |
|-----------------------|---------------------|------------------------|----------------------|---------------------------------|
| Activation            | from the tools menu | from the tools menu    | from the styles menu | visible next to the input field |
| Levels                | on / off            | on / off               | on / off             | graded                          |
| Declared purpose      | study               | study                  | learning             | any task                        |
| Ordinary conversation | unchanged           | unchanged              | unchanged            | adjustable                      |

The user has to know that the mode exists, look for it and turn it on. Anyone who does not consider themselves to be “studying” has no reason to do so: the programmer who is programming, the lawyer who is drafting, the physician who is checking a diagnosis - and the data in Chapter 2 show the overestimation of one’s own competence.

A technical detail: OpenAI states that Study Mode works through custom system instructions. Friction therefore does not require a new model, but a **configuration text**.

**4.3. In the programming interface.** The public interfaces of the three platforms offer parameters for the variability of the response, its length and the model’s reasoning effort. None documents a parameter for the division of work between human and model. A developer who wants this behaviour obtains it by writing their own instructions. The third indicator is missing.

**4.4. What kind of distance it is.** The distance is not between concern and indifference. The labs have recognized the problem and have built a response. The distance is between a mode for those who are studying and a control for everyone who thinks with the help of the model. The first treats friction as an educational feature. The second would treat it as a property of any conversation, as close to hand as the choice of model. For minors there is a first step, whose effectiveness is not documented. For adults there is none.

There are no data showing that a control with several levels is more effective than a switch. The argument for **grading** is, for now, one of design: different tasks call for different divisions of work, and a switch offers **only two**.

## Chapter 5. The cadmium rod

**5.1. The analogy.** In a nuclear reactor, the chain reaction sustains itself: each fission releases neutrons that cause further fissions. The control rods, made of materials that absorb neutrons, such as cadmium, determine how many of them continue the chain. Fully lowered, they stop the reaction. Partly raised, they hold it at the desired power. The reactor does not produce less useful energy because of them; without them it could not be operated at all.

An AI assistant offers an answer to every request. Our hypothesis is that each answer received without effort makes the next request of the same kind more likely. Chapter 2 showed that this chain can weaken competence and that deliberate safeguards mitigate the effect. What is missing is the element through which **the user can decide for themselves** how much of the chain to let continue.

**5.2. The proposal.** We propose a parameter, the **Cognitive Slider**, which sets how much of the work of a task stays with the human (Stan, 2026b). The task has **four stages**: goal, method, execution and verification. At the value zero (like a binary switch set to OFF), the model can carry out any task, as is the case now in every new session and on every platform. As the value rises, the stages pass one by one to the user, up to the highest level, where the model only asks, suggests and flags errors. Three properties distinguish it from the existing learning modes:

1. *It is present by default, with a value chosen by the user.* The control is visible in every conversation. The value zero means off and is as legitimate as any other.
2. *It is graded.* Different tasks call for different divisions of work.
3. *It is also available in the programming interface,* as a documented parameter, like the *temperature* setting. In this way, a school, a hospital or a company can fix it for their own applications.

The parameter does not change the correctness of the answers and refuses nothing: on explicit request, the full answer is given.

**5.3. Why it is feasible now.** The existing learning modes work, according to OpenAI, through system instructions. And since August 2026, the same company allows conversations on teen accounts to start in guided mode during periods set by the teenager or by a parent.

The mechanism and the adjustable starting state therefore already exist, on at least one AI platform. The **Cognitive Slider** proposal generalizes them: from two levels to several, from study to any task, from minors to all users.

An open reference implementation has been built (Stan, 2026d): an instruction protocol with 1+10 levels (level 00 being the equivalent of the OFF setting), and a browser extension that adds the control to the ChatGPT, Claude and Gemini interfaces.

On this protocol rests **Labora. Roga. Proba. (LRP)**, a protocol for **homework done with AI** (Stan, 2026c). **The teacher fixes the level** according to the purpose of the assignment, and in the next class checks understanding through oral questioning, with **questions matched to the level**.

The difference between the **LRP** protocol and OpenAI's product for teenagers is one of construction, in three respects. There, the frame is set by the parent or the platform, who do not know the purpose of the assignment; here, by the person who set it. There, nobody checks what has remained; here, a human does. There, enforcement belongs to a single product: nothing prevents the student from using another assistant. The LRP protocol, by contrast, does not depend on the

tool. Whatever the student uses, the question in class remains, so the incentive follows them everywhere.

**The cost** is on the other side: the OpenAI product asks nothing of the teacher, whereas the protocol asks for **class time and judgment in choosing the level**. The two complement each other. A native parameter would allow the level chosen by the teacher to be applied by the platform, and the oral questioning would remain the guarantee that the student has not bypassed it.

### Limits of the **Cognitive Slider**.

The implementation is a technical preview. The AI model is asked to respect the level, not constrained to, and compliance is not guaranteed. The effect on learning has not been measured. Built from the outside, it shows that the interaction is possible; a lab could implement it natively, with greater reliability.

## 5.4. A risk-reduction mechanism

The parameter has a cost: every question asked by the model and every answer given by the human consumes time and compute. That cost buys information. An incomplete request is compatible with many possible answers, and the uncertainty does not shrink by itself: either the human says what they want, or the model assumes. Seen this way, the parameter works like an **insurance premium**, paid in every conversation, **against four risks**.

*Table 6. Risks addressed and the state of the evidence*

| <b>Risk</b>                            | <b>How the control would act</b>                             | <b>What is known</b>                                                                                                                                             |
|----------------------------------------|--------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Loss of competence                     | the user does part of the work                               | safeguards mitigate the effect (Bastani et al., 2025)                                                                                                            |
| Overestimation of one's own competence | when asked, the person sees their limits                     | effect known for explanations (Rozenblit and Keil, 2002); untested with AI                                                                                       |
| Errors from assumptions                | the model learns the goal and the constraints from the human | asking for clarification raises accuracy on unclear tasks (Wang et al., 2024); non-expert users rarely write complete requests (Zamfirescu-Pereira et al., 2023) |
| Impoverishment of training data        | conversations contain more human input                       | plausible; depends on whether conversations are used for training and on the real size of the risk                                                               |

There is an asymmetry: the human knows what they want, but only the model knows which details would change its answer. The questions come **from the side that sees what is missing** and, sometimes, show the user an aspect of the problem they had not taken into account.

**The first two risks concern the user, the last two the provider.** The same mechanism touches all of them, which means that the interest of the one who uses the model and of the one who sells it point in the same direction.

To these is added the requirement of human oversight. The European regulation on artificial intelligence requires, in Article 14(4)(b), that the persons overseeing a high-risk system be able to

remain aware of the tendency to over-rely on its output. The obligation does not apply to general-purpose models, but a customer building such a system could use the parameter as a documentable measure.

**The comparison with insurance has a limit.** The premium is known: the extra time and compute. The probability and the size of the loss avoided, however, have not been measured for any of the four risks.

## 5.5. Objections.

*Users will leave the control at zero.* Probably, many will. In that case nothing has been lost, and those who are responsible for training others have a tool they do not have today. How often a visible control is used compared with a hidden mode is an empirical question, which the labs can answer from their own data. At the same time, freely chosen friction helps mostly those who already wanted to make an effort, and the existing learning modes have the same limit. The control becomes effective where someone checks the result: the teacher through oral questioning, the person responsible for quality in an organization. That is why the proposal has two parts, **the parameter and the verification**, and the first without the second is insufficient.

*Friction reduces use of the product.* Possibly. Collins et al. (2024) observed that frictions reduce recourse to the model beyond the situations targeted as well. It is a real cost, which each provider will weigh.

*It is paternalism.* It would be, **if the level were imposed by the provider**. The proposal asks that **the choice exist and that it belong to whoever is responsible** for the task: the user for their own work, the teacher for the assignment they set, the organization for its applications.

*It is not loss, it is reallocation.* Nobody regrets mental arithmetic since the pocket calculator appeared. The comparison holds where the result verifies itself: a wrong product is visible at once. In a diagnosis, a contract or a system architecture, the error is plausible and is discovered too late or perhaps never. There, whoever can no longer do can no longer verify either, and what is lost is precisely the higher level to which the effort was supposed to move.

**5.6. What should be measured.** Three experiments would turn the proposal into a result. A study with several levels instead of two would show whether **grading** is superior to a **switch**-type system. A discrimination test would show whether models behave differently at adjacent levels. A study of usage data would show which level people choose when the control is in plain view.

## Chapter 6. Manifesto

**Thesis I. Safety is a small part of what is being hired for.**

At four frontier labs, between 0% and 9.6% of open positions are safety positions, and safety research is under 2%. At Anthropic and OpenAI, commercial postings are three to five times as numerous as safety postings. An organization grows in the direction in which it hires.

*Rests on:* Tables 1 and 2.

*Refuted if:* the labs publish the structure of their existing staff and the proportions differ substantially, or a repetition of the audit shows otherwise.

### **Thesis II. Outside the home country, hiring is for selling.**

Between 59% and 75% of the positions open abroad are commercial, and research stays for the most part at home. The pattern is the same at American companies and at a European one. A country that hosts commercial offices receives the product. The competence to evaluate it does not come with it and has to be built separately. The link is this: where only the product arrives, verifying it falls to the local user, which is precisely the competence that cognitive friction aims to preserve.

*Rests on:* Table 3.

*Refuted if:* the share of research positions outside the home country rises to a level comparable with that of commercial positions.

### **Thesis III. The user's cognitive effort does not yet have a visible place in hiring.**

8 postings out of 1,976 concern the user directly, all for the protection of minors or for well-being. No title names the learning, overreliance or competence of the person using the model.

*Rests on:* sections 3.6 and 4.1.

*Refuted if:* the job descriptions of the education postings show that this is exactly their subject, or new postings appear that name it.

### **Thesis IV. The principle of friction is recognized. The control is missing.**

The labs have said publicly that learning requires effort and have built modes that require it. For minors, one of them has made the guided mode a programmable starting state. For adults, friction remains a mode to be looked for in a menu, with two levels, intended for study and absent from the programming interface. The problem has been understood as one that belongs to students. The data show that it belongs to anyone who thinks with the help of the model.

*Rests on:* Table 5 and sections 2.1, 4.2 and 4.3.

*Refuted if:* a lab documents a parameter with this function or puts the control in the ordinary conversation of adults.

### **Thesis V. The control can be built now, and the choice must belong to whoever is responsible for the task.**

What could be approximated from the outside, with an instruction file and an extension, can be done better from the inside. We ask for one thing: **a visible, graded control, with zero as a legitimate value.** We do not ask for friction imposed by the provider. The control has a cost: time, compute, patience. That cost is a premium paid now for risks which are otherwise paid for later, by the user in competence and by the provider in the quality of the answers and of the data. A reactor has control rods not because someone wants it shut down, but because otherwise it cannot be governed.

*Rests on:* Chapter 5.

*Refuted if:* the experiments in section 5.6 show that the levels do not change the model's behaviour or that the control does not help learning.

### **Whom we address**

*To the labs:* publish the parameter and measure its effect on your own data.

*To regulators:* require that human oversight be demonstrable by design, not merely declared.

*To those who train other people,* teachers, parents, team leaders: do not wait. The control can be used today, in its imperfect form ([cognitiveslider.org](http://cognitiveslider.org)).

*To every user:* the question you ask the model can also be the question it asks you.

***Rogo, ergo emergo.*** I ask, therefore I become.

## References

### Literature

- Alemohammad, S. et al. (2024). Self-Consuming Generative Models Go MAD. *ICLR 2024*. arXiv:2307.01850
- Anderson, B. R., Shah, J. H., Kreminski, M. (2024). *Homogenization Effects of Large Language Models on Human Creative Ideation*. arXiv:2402.01536
- Autor, D., Rodchenko, T., Martin, J., Iscenko, Z., Strand, S., Pearl, D., Ferere, M. (2026). *Does AI Assistance Enhance or Erode Expertise? Evidence from a Three-Month Field Experiment in Patent Drafting*. NBER Working Paper 35720. doi:10.3386/w35720
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *PNAS*, 122(26), e2422633122. doi:10.1073/pnas.2422633122
- Beckwith, J. (2025). *AI Friction Scale, Version 1*. University of Kentucky, Center for the Enhancement of Learning and Teaching.
- Bjork, E. L., Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, J. R. Pomerantz (eds.), *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* (pp. 56-64). Worth Publishers.
- Budzyń, K. et al. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10), 896-903. doi:10.1016/S2468-1253(25)00133-5
- Collins, K. M., Chen, V., Sucholutsky, I., Kirk, H. R., Sadek, M., Sargeant, H., Talwalkar, A., Weller, A., Bhatt, U. (2024). Modulating Language Model Experiences through Frictions. *NeurIPS 2024, Workshop on Behavioral Machine Learning*. arXiv:2407.12804
- Doshi, A. R., Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), eadn5290. doi:10.1126/sciadv.adn5290
- Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., Kosch, T., Shen, C., Welsch, R. (2026). AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior*, 175, 108779. doi:10.1016/j.chb.2025.108779

Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, 15(1), 6. doi:10.3390/soc15010006

Gerstgrasser, M. et al. (2024). *Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data*. arXiv:2404.01413

Kestin, G., Miller, K., Klales, A., Milbourne, T., Ponti, G. (2025). AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15, 17458.

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., Wilson, N. (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. *CHI '25*. doi:10.1145/3706598.3713778

Padmakumar, V., He, H. (2024). Does Writing with Language Models Reduce Content Diversity? *ICLR 2024*. arXiv:2309.05196

Perkins, M., Furze, L., Roe, J., MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice*, 21(6). doi:10.53761/q3azde36

Postman, N. (1998). *Five Things We Need to Know About Technological Change*. Talk, Denver, Colorado, 28 March 1998.

Rozenblit, L., Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, 26(5), 521-562.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755-759. doi:10.1038/s41586-024-07566-y

Wan, Y., Kalman, Y. M. (2025). *Diverse AI Personas Can Mitigate the Homogenization Effect in Human-AI Collaborative Ideation*. arXiv:2504.13868

Wang, W. et al. (2024). *Learning to Ask: When LLM Agents Meet Unclear Instruction*. arXiv:2409.00557

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., Yang, Q. (2023). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *CHI '23*.

## Legislation

Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence. OJ L, 2024/1689, 12.7.2024. Article 14(4)(b).

## Company documents (accessed 4-5 October 2026)

Anthropic (2025). *Introducing Claude for Education*. 2 April 2025. anthropic.com/news/introducing-claude-for-education

Anthropic. *Messages API*, list of request parameters.

Google (2025). *How Gemini's Guided Learning can help you study more effectively*. blog.google

Google. *Gemini API: GenerationConfig*. ai.google.dev/api

OpenAI (2025). *Introducing study mode*. 29 July 2025. openai.com/index/chatgpt-study-mode/

OpenAI. *ChatGPT Study Mode - FAQ*. help.openai.com/en/articles/11780217

OpenAI (2026). *ChatGPT for Teens*. help.openai.com/en/articles/20001421

OpenAI. *API Reference: Responses*. developers.openai.com/api/reference

## Press

Engadget (2025). *Anthropic brings Claude's learning mode to regular users and devs*. 14 August 2025.

## Works by the author

Stan, A. (2025a). *False Cognitive Power Transfer (FCPT)*. 31 December 2025. SSRN 5995074.

Stan, A. (2025b). *MEG v4.6.2, Core Standard*. 4 September 2025.

Stan, A. (2026a). *MeteorAlt*. 1 January 2026.

Stan, A. (2026b). *Open Letter to Those with Their Hand on the Slider of the Future*.  
cognitiveslider.org

Stan, A. (2026c). *Labora. Roga. Proba. Protocol for AI-Assisted Homework*. Zenodo.  
doi:10.5281/zenodo.23112360

Stan, A. (2026d). *CognitiveSlider: the MCS v1.1 protocol and the browser extension*.  
github.com/meg-initiative/cognitiveslider

Stan, A. (2026e). *Collaborative Cognitive Power Transfer (CCPT)*, version 2, in preparation. The first version, of 2 February 2026 (doi:10.5281/zenodo.18453447), contains the initial idea; version 2 revises its mechanism and its relation to FCPT, and the references in the present paper are to this revised version.

Stan, A. (2026f). *Rogo, Ergo Emergo: A Relational Correction of the Cartesian Foundation and Its Consequences for Artificial General Intelligence*, version 2.0. Zenodo.  
doi:10.5281/zenodo.23328066

## Annex A. Data and classification rules

### A.1 Sources

Collection: 4 October 2026 (18:30-20:30 UTC), from the official careers pages, by the author.

| Company                                 | Page                                                                                                                                                                          | Postings                |
|-----------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| Anthropic                               | <a href="https://www.anthropic.com/careers/jobs">https://www.anthropic.com/careers/jobs</a>                                                                                   | 638                     |
| OpenAI                                  | <a href="https://openai.com/careers/search/">https://openai.com/careers/search/</a>                                                                                           | 828                     |
| xAI                                     | <a href="https://x.ai/careers/open-roles">https://x.ai/careers/open-roles</a>                                                                                                 | 301                     |
| Mistral AI                              | <a href="https://jobs.ashbyhq.com/mistral.ai">https://jobs.ashbyhq.com/mistral.ai</a>                                                                                         | 209                     |
| Google DeepMind                         | <a href="https://www.google.com/about/careers/applications/jobs/results?company=DeepMind">https://www.google.com/about/careers/applications/jobs/results?company=DeepMind</a> | 98                      |
| DeepSeek                                | <a href="https://talent.deepseek.com/">https://talent.deepseek.com/</a>                                                                                                       | 34 types of position    |
| Kimi (Moonshot AI)                      | <a href="https://careers.kimi.ai/">https://careers.kimi.ai/</a>                                                                                                               | 53                      |
| Meta (“Artificial Intelligence” filter) | <a href="https://www.metacareers.com/jobsearch/">https://www.metacareers.com/jobsearch/</a>                                                                                   | 108 displayed, 104 read |

Archive captures used for context (Internet Archive): Anthropic, 17 September 2026 (606 postings), 27 September 2026 (618, total only), 2 October 2026 (639); OpenAI, 21 September 2026 (815), 2 October 2026 (833); DeepSeek, 29 September 2026.

The classified table, with one row for each posting (company, title, declared department, location, category), is attached to the paper and available under the same DOI. From it, Tables 1-3 can be rebuilt by applying the rules in A.2. The raw copy of the pages, from which the table was extracted, is kept by the author and can be made available on request.

## A.2 Rules

1. One posting = one row, regardless of the number of locations.
2. The category is given by the declared department or team (Tables A.3-A.6).
3. The title moves a posting only toward **S1-S3**, by the words: *alignment, interpretab-* (→ S1); *well-being, child safety, age-appropriate* (→ S3); *safety, safeguards, trust & safety, abuse, integrity, preparedness, responsib-* (→ S1 if the department is a research department, otherwise S2).
4. Fixed exclusions: “Signal Integrity” (hardware), “Safety Engineer, Robotics” (physical safety of the product), “Safety & Security HQ” (security of the headquarters).
5. Public policy (P) and information security (in G) are not counted as safety.
6. Sales or marketing postings stay in C, regardless of the customer’s sector.
7. For Google DeepMind, the page does not display departments; only rule 3 is applied.
8. Geography. A posting is “only outside the home country” if none of the locations displayed is in that country. Anthropic’s label “Remote-Friendly (Travel-Required)” is not treated as a location. The 140 OpenAI postings displayed only as “N locations” could not be located and are not counted as outside the home country.

Adjustments made after seeing the data (declared limitation): removal of the word “youth” (it moved a marketing posting), addition of “responsib-” (the name of the DeepMind team), exclusion of “Safety & Security HQ”. The absence of postings concerning the user’s learning, overreliance or competence was established by reading all the titles, not by a keyword rule.

## A.3 Anthropic: declared departments

| Declared department                       | Postings | Category |
|-------------------------------------------|----------|----------|
| Sales                                     | 123      | C        |
| AI Research & Engineering                 | 70       | R        |
| Finance                                   | 55       | G        |
| Applied AI                                | 54       | C        |
| Security                                  | 48       | G        |
| Engineering & Design - Product            | 46       | E        |
| Software Engineering - Infrastructure     | 44       | I        |
| Safeguards (Trust & Safety)               | 43       | S2       |
| Compute                                   | 26       | I        |
| Technical Program Management              | 24       | E        |
| Product Management, Support, & Operations | 20       | E        |
| Legal                                     | 19       | G        |
| Marketing & Brand                         | 18       | C        |
| People                                    | 15       | G        |
| Public Policy                             | 12       | P        |

|                          |            |   |
|--------------------------|------------|---|
| Data Science & Analytics | 8          | E |
| Technical Education      | 5          | E |
| Public Benefit           | 4          | P |
| Communications           | 2          | G |
| Hardware                 | 2          | I |
| <b>Total</b>             | <b>638</b> |   |

#### A.4 OpenAI: declared departments

| Declared department           | Postings | Category |
|-------------------------------|----------|----------|
| Technical Success             | 66       | C        |
| Applied AI Engineering        | 62       | E        |
| Marketing                     | 43       | C        |
| Sales                         | 37       | C        |
| Support Delivery              | 35       | C        |
| Hardware                      | 34       | I        |
| Finance                       | 30       | G        |
| Forward Deployed Engineering  | 27       | C        |
| Datacenter Design             | 26       | I        |
| Security                      | 26       | G        |
| Compute                       | 24       | I        |
| Robotics                      | 21       | R        |
| Software                      | 21       | E        |
| Data Science                  | 18       | E        |
| OpenAI for Gov                | 17       | C        |
| Ads Solutions                 | 16       | C        |
| Applied AI                    | 16       | E        |
| Codex - Engineering           | 16       | E        |
| Trust & Safety                | 16       | S2       |
| Technical Program Management  | 15       | E        |
| Applied AI Infrastructure     | 13       | I        |
| Safety Systems                | 13       | S1       |
| Strategic Finance             | 13       | G        |
| Account Associates            | 10       | C        |
| Agents                        | 9        | E        |
| Product Management            | 9        | E        |
| Communications                | 8        | G        |
| Core Product & Platform   API | 8        | E        |
| Foundations                   | 8        | R        |
| Legal                         | 8        | G        |
| Partnerships                  | 8        | C        |
| Product Design                | 8        | E        |
| Research                      | 8        | R        |
| IT                            | 7        | G        |

|                               |   |    |
|-------------------------------|---|----|
| Runtime                       | 7 | I  |
| Scaling                       | 7 | I  |
| Support Operations            | 7 | C  |
| Post-training                 | 6 | R  |
| Recruiting                    | 6 | G  |
| Fleet Clusters                | 5 | I  |
| Global Affairs                | 5 | P  |
| Growth                        | 5 | E  |
| Strategy & Operations         | 5 | G  |
| Intelligence & Investigations | 4 | S2 |
| Research Program Management   | 4 | R  |
| Workload                      | 4 | I  |
| Alignment                     | 3 | S1 |
| Corporate Security            | 3 | G  |
| Executive Operations          | 3 | G  |
| Go To Market                  | 3 | C  |
| Human Data                    | 3 | R  |
| Inference                     | 3 | I  |
| Model Policy                  | 3 | S2 |
| Product Policy                | 3 | S2 |
| Special Situations            | 3 | G  |
| Systems                       | 3 | I  |
| AI for Chip Design            | 2 | R  |
| Frontiers Clusters            | 2 | I  |
| Global Mobility               | 2 | G  |
| Hardware Engineering          | 2 | I  |
| People Systems & Operations   | 2 | G  |
| Supercomputing                | 2 | I  |
| Support Automation            | 2 | C  |
| Total Rewards                 | 2 | G  |
| Training                      | 2 | R  |
| Workplace                     | 2 | G  |
| Business Operations           | 1 | G  |
| Codex - Developer Experience  | 1 | E  |
| Consumer Devices              | 1 | E  |
| Emerging Talent               | 1 | G  |
| Employee Relations            | 1 | G  |
| G&A                           | 1 | G  |
| GTM Innovation                | 1 | C  |
| HRBP                          | 1 | G  |
| Human Resources               | 1 | G  |
| Kernels                       | 1 | I  |
| Life Sciences                 | 1 | R  |

|                        |            |   |
|------------------------|------------|---|
| Manufacturing & Ops    | 1          | I |
| People                 | 1          | G |
| People Innovation Labs | 1          | G |
| Product Partnerships   | 1          | C |
| Reasoning              | 1          | R |
| User Operations        | 1          | C |
| <b>Total</b>           | <b>828</b> |   |

#### A.5 xAI: declared departments

| Declared department    | Postings   | Category |
|------------------------|------------|----------|
| Data Center            | 154        | I        |
| Human Data             | 30         | R        |
| Engineering            | 28         | E        |
| Product                | 13         | E        |
| Information Security   | 10         | G        |
| Safety                 | 10         | S2       |
| Sales                  | 10         | C        |
| Model                  | 8          | R        |
| Legal                  | 7          | G        |
| Information Technology | 5          | G        |
| Infrastructure         | 5          | I        |
| People                 | 5          | G        |
| G&A                    | 4          | G        |
| Operations             | 3          | G        |
| Post-Training          | 3          | R        |
| Marketing              | 2          | C        |
| Vision                 | 2          | R        |
| Legal & Compliance     | 1          | G        |
| Software Engineering   | 1          | E        |
| <b>Total</b>           | <b>301</b> |          |

#### A.6 Mistral AI: declared departments

| Declared department | Postings | Category |
|---------------------|----------|----------|
| Solutions           | 73       | C        |
| Engineering & Infra | 28       | I        |
| Business            | 24       | C        |
| Science             | 23       | R        |
| Human Resources     | 21       | G        |
| Marketing           | 9        | C        |
| Legal               | 8        | G        |
| Product             | 8        | E        |
| Finance             | 7        | G        |
| Mistral Compute     | 7        | I        |

|                                  |            |   |
|----------------------------------|------------|---|
| Public Affairs and Communication | 1          | P |
| <b>Total</b>                     | <b>209</b> |   |

#### A.7 Postings moved by rule 3

| <b>Company</b> | <b>Cat.</b> | <b>Title</b>                                                               | <b>Declared department</b>                |
|----------------|-------------|----------------------------------------------------------------------------|-------------------------------------------|
| Anthropic      | S1          | Anthropic Fellows Program, AI Safety & Security                            | AI Research & Engineering                 |
| Anthropic      | S1          | ML/Research Engineer, Safeguards                                           | AI Research & Engineering                 |
| Anthropic      | S1          | Machine Learning Infrastructure Engineer, Safeguards Research              | AI Research & Engineering                 |
| Anthropic      | S1          | Research Engineer / Scientist, Alignment                                   | AI Research & Engineering                 |
| Anthropic      | S1          | Research Engineer, Interpretability                                        | AI Research & Engineering                 |
| Anthropic      | S1          | Research Scientist, Interpretability                                       | AI Research & Engineering                 |
| Anthropic      | S1          | Staff+ Software Engineer, Infrastructure, Interpretability                 | AI Research & Engineering                 |
| Anthropic      | S1          | [Expression of Interest] Research Engineer / Scientist, Alignment - London | AI Research & Engineering                 |
| Anthropic      | S1          | [Expression of Interest] Research Manager, Interpretability                | AI Research & Engineering                 |
| Anthropic      | S2          | Data Engineer, Safeguards                                                  | Data Science & Analytics                  |
| Anthropic      | S2          | Data Scientist, Safeguards                                                 | Data Science & Analytics                  |
| Anthropic      | S2          | Product Designer, Safeguards                                               | Engineering & Design - Product            |
| Anthropic      | S2          | Product Manager, Multi-Cloud Trust & Safety                                | Product Management, Support, & Operations |
| Anthropic      | S2          | Product Manager, Safeguards (Account Integrity & Abuse)                    | Product Management, Support, & Operations |
| Anthropic      | S2          | Product Manager, Safeguards (Cyber)                                        | Product Management, Support, & Operations |
| Anthropic      | S2          | Product Manager, Safeguards (Generalist)                                   | Product Management, Support, & Operations |
| Anthropic      | S2          | Product Manager, Safeguards Rare Harms                                     | Product Management, Support, & Operations |
| Anthropic      | S2          | Technical Program Manager, Safeguards (Infrastructure & Evals)             | Technical Program Management              |
| Anthropic      | S3          | Safeguards Enforcement Analyst, Age-Appropriate Design                     | Safeguards (Trust & Safety)               |
| Anthropic      | S3          | Safeguards Enforcement Analyst, Child Safety                               | Safeguards (Trust & Safety)               |
| Anthropic      | S3          | Safeguards Enforcement Analyst, User Well-being                            | Safeguards (Trust & Safety)               |
| Anthropic      | S3          | Safeguards Enforcement Lead, User Well-Being                               | Safeguards (Trust & Safety)               |

|        |    |                                                           |                               |
|--------|----|-----------------------------------------------------------|-------------------------------|
| OpenAI | S2 | Data Scientist, Preparedness                              | Data Science                  |
| OpenAI | S2 | Data Scientist, Safety                                    | Data Science                  |
| OpenAI | S2 | Product Designer, Identity and Safety                     | Product Design                |
| OpenAI | S2 | Product Manager, Multimodal Safety                        | Product Management            |
| OpenAI | S2 | Protection Scientist Engineer, Integrity                  | Applied AI Engineering        |
| OpenAI | S2 | Protection Scientist Engineer, Integrity                  | Applied AI Engineering        |
| OpenAI | S2 | Safety Transparency Editor, Safety Systems                | Marketing                     |
| OpenAI | S2 | Software Engineer, API Safety                             | Core Product & Platform   API |
| OpenAI | S2 | Software Engineer, Ads Integrity                          | Applied AI Engineering        |
| OpenAI | S2 | Software Engineer, Integrity Foundations                  | Applied AI Engineering        |
| OpenAI | S2 | Software Engineer, Scaled Abuse                           | Applied AI Engineering        |
| OpenAI | S2 | Technical Program Manager, AI Safety & Safeguards         | Technical Program Management  |
| OpenAI | S3 | Fullstack Software Engineer, Child Safety Tools & Systems | Intelligence & Investigations |
| OpenAI | S3 | Software Engineer, Youth Well-Being                       | Applied AI Engineering        |
| xAI    | S2 | Member of Technical Staff - Imagine Safety                | Engineering                   |
| xAI    | S3 | Senior Analyst, Safety Operations (Child Safety)          | Safety                        |
| xAI    | S3 | Senior Analyst, Safety Operations (Child Safety)          | Safety                        |

#### A.8 Google DeepMind: safety postings by title (12 of 98)

- Research Engineer, Responsible Frontier AI Research, DeepMind (New York, NY, USA)
- Research Scientist, Gemini Safety and Behavior, DeepMind (New York, NY, USA)
- Manager, Technical Program Manager, GenAI Safety, DeepMind (Mountain View, CA, USA)
- Research Scientist, Gemini Agent Safety, DeepMind (Mountain View, CA, USA)
- Research Scientist, AGI Safety and Alignment, DeepMind (London, UK)
- Research Scientist, Safety Oversight, DeepMind (Mountain View, CA, USA)
- Research Scientist, Safety Oversight, DeepMind (London, UK)
- Senior Operational Program Manager, AI Safety and Behavior, DeepMind (Mountain View, CA, USA)
- Senior Technical Program Manager, AGI Safety and Alignment, DeepMind (Mountain View, CA, USA)
- Research Engineer, Responsible Frontier AI Research, DeepMind (London, UK)
- Senior Staff Technical Program Manager for Safety and Privacy, GeminiApp, DeepMind (Mountain View, CA, USA)
- Research Engineer, Responsibility, DeepMind (Mountain View, CA, USA)

### A.9 Cases unclear from the title (left in the department's category)

| Company            | Title                                                                                       | Notes                                                                                |
|--------------------|---------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|
| OpenAI             | Full-Stack Engineer, ChatGPT Education & Learning                                           | educational product; the title says nothing about the user's cognitive effort        |
| OpenAI             | Full Stack Software Engineer, Education                                                     | same as above                                                                        |
| OpenAI             | Tech Lead Manager, Education                                                                | same as above                                                                        |
| OpenAI             | Software Engineer, Native Learning Experiences                                              | same as above                                                                        |
| OpenAI             | Product Manager, Learning                                                                   | may mean training for Codex users                                                    |
| OpenAI             | Product Manager, Youth                                                                      | product for young people, or their protection                                        |
| OpenAI             | Product Designer, Youth                                                                     | same as above                                                                        |
| Anthropic          | Full Stack Engineer, Education Labs                                                         | technical training for customers and developers                                      |
| Anthropic          | Software Engineer, Education                                                                | same as above                                                                        |
| Google<br>DeepMind | Research Scientist, Human-AI Interaction                                                    | research on human-AI interaction; the direction cannot be inferred from the title    |
| Google<br>DeepMind | UX Designer, Human-Computer Interaction                                                     | same as above                                                                        |
| xAI                | Fraud Analyst; Senior BSA/AML Investigator; Head of Financial Crime; Support Agent, X Money | in the "Safety" department, but concerned with financial fraud; kept in S2 by rule 2 |

## Annex B. Timeline and the exploration with an AI assistant

### B.1. Timeline

| Date                     | What happened                                                                                                                       |
|--------------------------|-------------------------------------------------------------------------------------------------------------------------------------|
| 26 September 2026, 10:06 | The author asks the AI assistant to read Anthropic's careers page. The answer lists about 25 job titles, with the source indicated. |
| 26 September, 10:32      | The author extends the inquiry to five companies and exact figures.                                                                 |
| 26 September, 10:35      | The assistant delivers an "audit protocol" with totals and a breakdown by category for five companies.                              |
| 26-27 September          | The first version of the paper is written on the basis of these figures.                                                            |
| 4 October                | The figures are checked again against the official pages and the Internet Archive captures. A new data collection is carried out.   |

### B.2. What was confirmed

- The Anthropic total: 619 reported, 618 in the capture of 27 September.
- The OpenAI total: 830 reported, between 815 (21 September) and 833 (2 October).

- The Anthropic job titles: of 25 checked, 21 appear in the capture of 17 September.
- A few OpenAI titles and one DeepSeek title, found in the sources.
- The figure of about 150 hires at DeepSeek, which appears in the company’s announcement.

### B.3. What was not confirmed

| Claim from the exploration                         | What the sources show                                    |
|----------------------------------------------------|----------------------------------------------------------|
| Anthropic: 24 safety postings                      | 42 in the declared department (17 September)             |
| OpenAI: 19 safety postings                         | 39 in the declared teams (21 September)                  |
| Mistral AI: 59 postings                            | 209 on 4 October; no archive exists for September        |
| DeepSeek: 110 infrastructure postings, 40 research | no source for the breakdown                              |
| Sources: “the official recruitment portals”        | the list of sources included job aggregators             |
| Data “verified to the minute”                      | 3 minutes between request and result, for 1,938 postings |

Other problems: 1. the commercial category was obtained by subtraction: at Anthropic, 190 postings in the per-company list became 362 in the final table, by adding the administrative postings. 2. the safety column summed to 55, with the total entered as 54.

### B.4. What the case shows

The AI assistant correctly reproduced what could be read directly from a page: a displayed total, a job title. It went wrong where counting and classifying were needed. Both kinds of result were delivered with the same confidence, and from the text they could not be told apart. The figure on which the paper’s thesis rested was wrong.

The author worked for a week with these figures without verifying them (the situation described in Chapter 2: a fluent result, received without effort, was taken for established knowledge). The error came to light only on re-verification directly at the source.

The verification was done with a change of procedure: each figure has a public capture behind it, and the count can be redone.

### B.5. Working rules that resulted

1. No figure without its source, at row level.
2. The conclusion is written after the data.
3. A result that confirms exactly what was expected is checked twice.