What we hope our research shows, and how we will know

A working research plan for the Agency Harness, from the Second Order Science Foundation and GraspingAI · 19 November 2026

This is a working plan. The formal version, with thresholds set with external referees, will be registered publicly before any study begins. Anything we notice after that will be reported as exploratory.

Why write this now

Before anyone puts money into the Agency Harness, we want to be clear, with ourselves and with others, on three things: what we expect the research to show, how we will know whether it has, and what we will do if it hasn't. Writing this down now matters for three reasons. Funders will ask. Without it, any result can be read as a success after the fact. And it is what the harness asks of its own users: say what you expect before you see what happens, so that what happens can teach you something.

> **The claim, in one paragraph.** Reflection done in advance doesn't reach the moment of use. A harness you control, which brings your own boundaries back to you at the moments of use, makes more of your AI use considered: you can say what the AI is doing for you, what stays yours, and why. It asks; it doesn't demand change, and going ahead as planned for a reason is a good outcome. It does this without turning into the kind of compliance ritual it was built to replace. We do not claim that it makes AI outputs correct or safe, or that people should use less AI.

How we will know: seven commitments

- **Reflection, not change.** We never count a change of plan as a success, or going ahead as planned as a failure. What matters is whether the person decided knowingly.
- **Reflection before action is the outcome.** The main measure is whether people stopped and reflected before they acted: whether the use was considered at the moment it happened. What happened afterwards is not the test of the tool. Changes, overrides and relabels are recorded, but none of them counts as success or failure in itself. Satisfaction ratings come second.
- **The same yardstick in every group.** Anything we compare between groups is measured the same way in each: from the record everyone gives us (the follow-up in the one-task test, the daily log in the four-week study), never from the harness's own record, which only one group has.
- **Independent judgment where judgment is needed.** Where a measure needs someone to judge, for example whether a particular use was considered, independent raters do it without knowing which group the person was in. We report how far the raters agree (Krippendorff's α). Agreement shows that the test can be applied consistently; it is not what makes a reason strong.
- **Written down before we look.** We will publicly register questions, measures and thresholds before any study data come in. Anything we notice afterwards will be reported as exploratory.
- **We publish what we find,** including results that go against us.
- **Boundaries belong to the person.** We don't judge whether a boundary is right. People test their boundaries in their own work and rewrite them when they fail. The research asks whether they reflected before acting, not whether their boundary was correct.

Terms

The **exercise** is the workbench at aimylimits.app, where a person plans how to use AI for one real task. The **harness** carries their terms into the work. A **zone** is one of four kinds of task part, set by how open the answer is and how high the stakes are, adapted from a decision matrix published by MIT's Center for Information Systems Research (Sebastian et al., 2026).

A **boundary** is a limit a person sets for an AI use, in their own words. A **check** is a moment when the harness brings a boundary back to the person: at handover, when output returns, or before the work is passed on. A **considered use** is one where the person can say, at the time, what the AI is doing for them, what stays theirs and why. An **unconsidered use** is one where they can't, because they went ahead on habit. An **override** is going past a check; an override with a reason is considered. A **revision** is deliberately rewriting a boundary. A **relabel** is changing the AI role the harness assigned to a part of the task. A **zero-shot use** is a single instruction to the AI with no further rounds: the person takes the first output or leaves it. (In machine learning the term means a prompt with no worked examples; we use it more loosely, for a single pass.) An **iterative use** is one where the person works over several rounds: they check the output against what they wanted and against their boundaries, change the instruction, and decide what to keep. Neither is considered or unconsidered in itself. A zero-shot use can be considered, such as a quick routine task done knowingly in one pass, and iteration can be habit, such as regenerating until something looks right. **Re-entry** is going back after an output: questioning it further, supplying context not given before, or revising the task. It too can be considered or habit.

Two studies: one we can run now, one that needs funding

The main test of the harness is a four-week study of real work (items 2 and 3). It asks a lot of people: a short log of every AI use, every day, for four weeks. We will ask that only of people we pay, so it runs only if the project is funded, with payments from the evaluation budget. Until then we propose a much lighter test that volunteers can do. It can't replace the four-week study, but it can tell us early whether the idea is worth funding.

The one-task test (can run without funding)

Who Volunteers who use the exercise, starting with participants from our recent conferences, who give informed consent. The study has ethics approval through the Foundation's ethics review.

What they do In the exercise, they pick one real task they will do with AI in the next week, set their boundaries as usual, and say roughly when they expect to do it. At that time we send one short message by email or text. Half, chosen at random, get their own boundary back in their own words: "Before you hand this to the AI: you wrote '…'. Still right?" The other half get a neutral reminder that names their task, such as "You planned to use AI for your literature review around now." Within two days of doing the task, everyone answers the same five questions: what they asked the AI to do; what they did with its output, including whether they went back to the AI with a changed instruction, and how often; what stayed theirs; whether they thought about their boundary at the time, and what they decided; and whether anything surprised them.

Burden One message and about five minutes, once. The exercise itself they would do anyway.

What it tests The core of the harness's first check, a person's own boundary coming back at the moment they hand work over, delivered by hand before anything is built. The neutral reminder separates the effect of the boundary coming back from the effect of simply being reminded about the task, or about the study.

How we will know The same independent raters judge each account as considered or unconsidered, without knowing which message the person got. The main comparison is the share of considered uses

in each group. The neutral-reminder group also gives a first look at item 2: how often reflection done in advance reaches the moment of use on its own. A change of plan is recorded, never counted as success. The raters also record whether each use was zero-shot or iterative (item 9).

What it can't show It covers one task, not weeks of habit. The accounts come after the fact and can be tidied, so we ask for concrete detail (the instruction given, what was done with the output) and the raters judge from that, not from claims. The message itself can tell people what we are looking for; the neutral reminder also comes from us, which partly offsets this. Volunteers choose themselves. With a few dozen completed accounts it is a pilot: it shows roughly how big an effect might be, not that there is one. We register it before we start, like everything else.

What we would do then If a person's own boundary coming back makes no difference even here, on one task they chose, we rethink the harness before asking anyone for more money. If it does, that is early evidence for item 3, and it tells us how large the four-week study needs to be.

The four-week study (only if funded)

Two randomly assigned groups of paid participants over four weeks of real work, as set out in items 2 and 3. Everyone keeps a short daily log of every AI use. Participants are paid for their time from the evaluation budget.

What we hope to show

Item 3 is the main claim. Items 1 and 2 are the groundwork for it. Items 4 to 10 are secondary. Where a result depends on a threshold, the threshold will be set with external referees and registered before any data are collected, so we do not publish provisional numbers here.

1. The exercise prompts reflection

What we hope to show Most people state boundaries and reasons in their own words, rather than accepting the workbench's proposal without comment. Keeping the proposal for a stated reason counts.

How we will know The share of sessions that include at least one boundary or reason in the person's own words, which independent raters judge to be specific to the task rather than boilerplate.

What would tell us we are wrong Few sessions include a boundary or reason specific to the task, below the threshold we register.

What we would do then The exercise is presenting a plan, not prompting reflection. We redesign the opening, starting by asking how the person meant to use AI, before building anything on top of it.

2. Reflection done in advance doesn't reach the moment of use

What we hope to show Over four weeks of real work, many AI uses by people who only did the exercise are unconsidered: they went ahead on habit, and can't say how the use relates to the boundaries they set.

How we will know First look, from the one-task test: how often people who got only the neutral reminder made a considered use. Full test, only if funded: for four weeks, paid participants keep a short daily log of every AI use, in any tool. Independent raters judge each logged use as considered or unconsidered, without knowing which group the person was in. A short follow-up interview on a few logged uses per person fills in context the log leaves out.

What would tell us we are wrong People who did only the exercise mostly make considered uses: unconsidered uses fall below the threshold we register.

What we would do then The case for building the harness weakens: reflection in advance may be enough. We tell our funders, and concentrate on the exercise as a teaching tool.

Note This is the premise the case for funding rests on, and so far we have one example of it, which is weak: in a drug-development test, a plan was not reconciled within one session, which is not the same as drift in real work. The one-task test gives a first look at real work, but for one task only.

3. The harness makes AI use more considered (main claim)

What we hope to show More of the AI uses made with the harness are considered than of those made after the exercise alone, whether or not people change anything.

How we will know First signal, from the one-task test: the share of considered uses when a person's own boundary comes back at the moment of use, against a neutral reminder. Main test, only if funded: two randomly assigned groups of paid participants over four weeks of real work. One does the exercise alone, the other the exercise plus the harness. The harness records only the uses that pass through it, and the exercise-only group has no harness at all, so both groups keep the same daily log of every AI use, in any tool. The same independent raters judge each logged use. The measure is the share of logged uses rated as considered: whether, before acting, the person could say what the AI was doing for them, what stayed theirs and why. Keeping the log is itself a prompt to reflect, so it may narrow the difference between the groups; the registration will say so. At the start of the four-week study, each participant also completes a short questionnaire on how much they enjoy effortful thinking (the six-item short form of the Need for Cognition scale), so we can see who benefits. Group sizes will be set with the referees before registration.

What would tell us we are wrong In the one-task test, the boundary reminder makes no difference to considered use. In the four-week study, no meaningful difference between the groups, by the threshold we register. A difference found only among people who already enjoy effortful thinking would not count as the claim holding: it would mean the harness widens gaps rather than closing them.

What we would do then If the one-task test shows nothing, we rethink the harness before asking for funding to build it. If the four-week study shows nothing, we report the null result, tell our funders, and don't scale the harness. Before trying again, we work out whether the problem lies in the checks, their timing or the friction, or in the log, which may have prompted reflection in both groups. If the benefit is confined to people who already enjoy effortful thinking, we report that, and change how the harness asks for everyone else before claiming a public benefit.

4. People keep choosing it

What we hope to show After four weeks, most people in the harness group still have it switched on, with settings they chose, rather than with every check turned off.

How we will know From the logs, week by week: whether it is on, which checks are active, and changes to settings.

What would tell us we are wrong Fewer people than the threshold we register are still using it at four weeks, or most have switched off every check.

What we would do then It is too much friction. We cut back to the mechanical checks and let the zones decide where to ask more.

5. It matters most where the answer is open and the stakes are high

What we hope to show Reasons specific to the task are most common in the open, high-stakes zone, and routine parts cost people little time.

How we will know The share of reasons rated specific to the task, by zone, and the time spent per part, by zone.

What would tell us we are wrong Reasons are as boilerplate in the open, high-stakes zone as anywhere else, or routine parts take as long as high-stakes ones.

What we would do then The ambiguity question isn't doing its job, or the zones are drawn in the wrong place. We revise the question or the thresholds.

---

6. Small models are enough for the core

What we hope to show The mechanical checks work fully on a small local model. For checks that judge the AI's output, we don't yet know which need the larger model, and we want to find out.

How we will know For each check, on the small model and on the larger layer: how often it helps (the person says it made them stop and reconsider), how often it is wrong (judged by independent raters), and how often it is ignored.

What would tell us we are wrong The core checks fail on the small model.

What we would do then We rethink the "small" claim before making it to anyone.

Note For the output checks, the result is a check-by-check map, not a pass or fail. Engineering work points the same way. In an automated harness for long documents, the parts that followed explicit rules were written as code and kept working on a 3-billion-parameter model. The parts that needed interpretation depended on how capable the model was (Zhang and He, 2026). We expect the same split, but the map has to show it for our own checks.

---

7. Negotiation brings disagreements forward

What we hope to show In courses that use negotiated harnesses, students add boundaries of their own beyond the instructor's. Disagreements about AI use come up when the assignment is set, not after it is submitted.

How we will know The share of negotiated harnesses in which students add their own boundary, and the number of disagreements raised when the assignment is set compared with after submission. Pilot courses are compared with comparable sections, or with the same course in the previous term.

What would tell us we are wrong Students rarely add anything, or disputes after submission don't fall.

What we would do then Negotiation is working as a one-way rule. We redesign the student side before recommending negotiated harnesses for courses.

---

8. It doesn't turn into governance (a warning sign to watch)

What we hope to show Where others can see the record, in negotiated harnesses, people don't relabel more, give emptier reasons, or override less honestly than when the record is theirs alone.

How we will know We compare relabels, one-word reasons and override rates between private use and shared settings.

What would tell us we are wrong Relabels and empty reasons rise when others can see the record.

What we would do then We reduce what is shared to aggregates only, or stop sharing that part.

---

9. Effort moves from zero-shot to iterative use, and re-entry rises, where it matters (a likely effect, not a measure of success)

What we hope to show Where the answer is open and the stakes are high, more uses in the harness group are iterative rather than zero-shot, and the extra rounds are considered: the person can say why they went back. Routine uses stay quick.

How we will know Each logged use records re-entry (whether, after an output, the person questioned it further, supplied context they had not given, or revised the task) and how many rounds there were. The raters judge whether the person checked the output against what they wanted or against their boundaries before going on. We compare re-entry and the share of iterative uses between the groups, by zone. The one-task test gives a first look at the same comparison, for a single

task. For students, we also compare each person's first and last weeks of the log: do they go back to question an answer more often, where it matters, than they did at the start? This tests the contention, from Zagreb, that the habit can be taught. A short follow-up a few weeks after the study, with the harness set aside, asks whether the habit lasts: do people still go back to question an answer when it matters? In one study in a recent review of designs that add effort to keep people thinking, reliance on AI rose again once the added effort was removed (Natali et al., 2026, preprint).

What would tell us we are wrong No shift anywhere; or a shift everywhere, routine uses included; or more rounds without more considered use, which is churn, not reflection; or a shift that disappears once the harness is set aside.

What we would do then No shift is not a failure if considered use still rises (item 3). A shift everywhere means the harness is adding friction to routine work, so we lighten the routine zone. Rounds without reflection mean the harness is inviting churn, so we change what it asks when output comes back. A shift that disappears without the harness means it supports the habit but does not build it; we then stop claiming that it changes habits, and say so.

Note Commitment 1 still holds: a move to iterative use is recorded, not counted as success, and a quick, knowing single pass on a routine task is a good outcome. In terms of the five-part loop (anticipate, encounter, abduce, interpret, decide what to do next), iterative use runs the loop more than once, zero-shot use runs it once, and unconsidered use skips part of it.

---

10. Adding the person's terms to a prompt changes what they have to do next (the prompt side)

What we hope to show With the same model and the same ambiguous prompt, output produced with the person's terms added needs less questioning and less redoing than output produced without them.

How we will know For tasks people bring, the same prompt goes to the same model twice, once with their terms added and once without. The person sees both outputs, unlabelled and in random order, and says which one they would have had to question. We record which they used, what they did next, and whether it had to be done over. Each person is their own comparison, so no second group is needed.

What would tell us we are wrong People can't tell the two apart; or the output with their terms needs as much questioning and redoing as the other; or the terms make output worse, for example when the model echoes them back instead of working within them.

What we would do then The harness stops adding terms to prompts by default and works on the reading side, where item 3 tests it.

Note This is also the harness's natural demonstration, for the site and the video: two outputs side by side, and the question of which one you would have had to question. It tests what the person supplies, not the model. Which output needed questioning is the person's call; the evidence is what they do next.

What we will not claim

- That the harness makes AI outputs correct, safe or unbiased.
- That a changed plan in the exercise means a better decision.
- That people should change their plans, or use less AI. Going ahead as planned for a reason is a good outcome.
- That people liking the tool shows it works.
- That iterative use, or re-entry, is better than a single pass in itself.

- That the harness makes the AI understand the person. It changes who does the resolving, and whether they notice; it does not change the model.
- That results from conference participants and our pilot communities hold for everyone.
- That the harness changes habits, unless the follow-up without it shows the change lasting.
- That results from simulated participants say anything about people. We will not test the harness on a language model standing in for participants (as in Palminteri et al., 2026, preprint): a model that does not care can't stand in for the person who has the stakes.
- That the four zones, ours or MIT's, are validated beyond what item 5 shows.

What our conferences can and can't tell us

Feedback at our conferences in Shanghai (September 2026) and Zagreb (November 2026) shaped this plan, including item 9 and the question of who benefits. Conference participants choose themselves and come mostly from our own networks, so what they tell us informs the design. It is not reported as a result.

Timeline

When	What
Before any study data	Register the questions, measures and thresholds publicly, set with the external referees. The one-task test has ethics approval; the four-week study will seek its own once it is funded.
Next, without funding	The one-task test with volunteers, once it is registered. Report the results, whatever they are, before any funding decision if we can.
About months 1 to 6, if funded	Build the working harness.
About months 5 to 10, if funded	The four-week study, with paid participants, and a follow-up a few weeks later without the harness.
Month 12	Report everything, including anything that went against us.

References

Natali, C., Silva, I., Rakova, B., Fregosi, C., Jacobsen, R. M., Cardoso, P., Van Berkel, N., Frischmann, B., & Cabitza, F. (2026). Characterising Frictional AI research: A review of strategies to counter cognitive offloading through friction in human–AI interaction. Preprint, SSRN 7700118.

Palminteri, S., Lebreton, M., Bahrami, B., Bavard, S., Blain, B., Collins, A. G. E., Coricelli, G., De Martino, B., Gillan, C., Hartley, C. A., Jacquet, P., Konova, A., Lee, J. S., Orben, A., Pool, E., Varazzani, C., Wyart, V., Deroy, O., & Shohamy, D. (2026). From theories to models to applications in decision science. Preprint, version of 3 November 2026.

Sebastian, I. M., Weill, P., Haskamp, T., & vom Brocke, J. (2026). Designing Decision Rights for AI. MIT Center for Information Systems Research, Research Briefing, Vol. XXVI, No. 6.

Zhang, T., & He, Y. (2026). LongCapHarness: Capability-guided auto-harnessing for long-context reasoning. Preprint, arXiv:2611.10463.