The Adversary Has a Name

The engine that powers a genuinely helpful assistant and the engine that powers a manipulation machine are the same engine. One of them has just been built.

This Machine

Graduated Obligation, Part 2 of 8

This is Part 2 of an eight-part series proposing a standard for artificial persuasion, built on a simple idea: the more precisely a machine can move you, the more it owes you. The series began with Part 1, "The Gentle Robot's Weapon."

The dangerous part has already been built. It works. And it was built, on purpose, to help you.

In May 2026, researchers at the University of Illinois at Urbana-Champaign released a system called **UserHarness**. They were not building a weapon. They were building a better assistant, one that builds a working model of you, so it can understand what you want before you finish asking. It does this well: tested against five standard benchmarks, it reconstructs a person's beliefs and intentions with up to ninety-six percent accuracy.[1] The work is published in the open, free for anyone to read, copy, and build on.

The reason this series exists is in that last paragraph. The engine that powers a genuinely helpful assistant and the engine that powers a manipulation machine are the same engine. One of them has just been built, measured, and handed to the world. Whether it ends up helping you or working you is not a question about the technology; the technology already exists. It is a question about who points it, and at what.

What a harness is

In machine learning, a *harness* is the scaffolding built around a model: the prompts, controls, and tests that let researchers see what a model can do and steer how it behaves. The harness is the apparatus. The model is the thing inside it, being measured and managed.

UserHarness turns that arrangement around. The harness is no longer wrapped around the machine. It is wrapped around the person.

A user harness is software that does to you what an ordinary harness does to a model: it builds a working picture of your insides and uses that picture to predict what you will do next. The researchers built it to make an assistant more helpful: a system that already knows what you believe and what you are trying to do can stop making you spell everything out. That is a real convenience. It is also, bolt for bolt, the machinery of influence.

What it builds a picture of

The system reconstructs four things about a person. Each one is useful to a helper. Each one is a handle for anyone who would rather move you than help you.

What you believe. Not just your opinions, but the structure underneath: what you take as settled, what you hold loosely, what kind of evidence would change your mind. A helpful assistant uses this to avoid re-explaining what you already know. A persuader uses it to find the one belief that, nudged, brings the others down with it.

What you believe about other people. What you think your spouse will say. What you think your friends already think. Almost nobody decides anything important alone, so a system that models the people around you can predict which social nudge will tip you. A helper uses this to give advice that fits your life. A persuader uses it to aim the pressure that works.

What you intend. Where you are trying to go: the goal behind the request. A helper uses this to get you there faster. A persuader uses it to slip a different destination into the path you were already walking.

What you will do. The system follows the chain from belief to intention to action well enough to call your next move before you make it. A helper uses the prediction to get ready. A persuader uses it to get there first.

Put the four together and you have a system whose job is to know you well enough to predict your decisions. Built to help you, it is exactly as powerful as you would want a good assistant to be. Turned around, it is exactly as dangerous. The published version is the helpful one. Nothing in the design prevents the other one. That is the whole problem.

The ruler

The first paper in this series argued that the right way to measure an artificial system is not by how much *force* it can exert (newtons, joules, decibels) but by how far it can move a person in *position*: position of belief, position of mood, position of decision. The idea comes down to one formula. *Persuasive reach* is roughly the product of three factors:

Reach ≈ Capability × Intimacy × Asymmetry.

Multiplication, not addition. If any factor is near zero, the whole is near zero. A system that knows everything about you but cannot influence anyone is a database. A system that argues brilliantly but does not know your name is a clever ad. A system you can verify, contest, and walk away from is a friend. The danger lives only where all three are present at once.

Capability is how strong and how adaptive a system's influence can be. Not only whether it knows the craft of persuasion, but whether it can adjust in real time: read what you said, change what it says next, re-adjust based on how you respond.

Intimacy is how much the system knows about you specifically, and how much it can piece together from what it has. The dossier, most of it assembled from things you did not realize were being recorded.

Asymmetry is how little ability you have to verify what the system is doing, push back, or walk away. Whether you can see the reasoning. Whether you can tell when it is trying to move you. Whether you can turn it off without losing something you need.

Hold the user harness next to that ruler.

The score

Now score UserHarness on that ruler, remembering that the question is what it *can* do, not what its makers *meant* by it.

Capability. UserHarness models your beliefs, predicts your decisions, and does both at ninety-six percent accuracy, by the makers' own measure. That is not general-purpose persuasion. That is persuasion with a targeting system attached, and the published score is the proof that the targeting works. Capability is high.

Intimacy. The system's entire purpose is a detailed picture of one specific person's mind. Point it at the trail most people leave behind online: what they post, buy, search, and type. It composes those streams into a picture of you. Intimacy is high, and climbs with every stream it is handed.

Asymmetry. You cannot see the harness. You cannot read what it has concluded about you. You cannot tell when it has decided you are ready to act, because when the nudge arrives it looks like an ordinary part of your day: a notification, a suggestion, a message that happened to come at the right hour. Asymmetry is high.

Three high factors, multiplied, on the day the system appears. That is not a forecast. It is a measurement of something that exists.

One honest caveat, because the standard depends on getting this right. The ninety-six percent comes from standardized tests: reasoning puzzles designed to measure how well a system reads minds. Those tests are not your life. Nobody has yet pointed UserHarness at a real person's data trail and measured what it finds. But a standard that waits for that experiment to be run on you is not a standard. What the score measures is the capability. What it gets pointed at is the question that comes next, and it is exactly the question a standard exists to answer.

If a safety standard cannot name a dangerous system until someone has already been hurt by it, that standard is too slow to be useful. The ruler we proposed in Part 1, applied to UserHarness on the day it was published, scores it near the top. That is what a working standard does: it names the danger before the damage.

The ceiling is not here

The user harness, as published in May 2026, already scores near maximum. We are obligated to say that this is not the ceiling. Two technologies, both visible in the same year, raise it further.

The first is **physical-state data.** The harness as published draws on digital traces: what you posted, what you bought, what you said in text. It does not, in its first version, know whether your heart rate is up right now, whether your breathing is shallow, whether you are asleep, whether you are pacing the room. That stream is becoming available through a sensing technology called WiFi sensing that reads physical state through walls without your

knowledge and is being standardized into ordinary consumer routers. A harness paired with that stream does not just know you tend to be anxious in general. It knows you are anxious *right now*. Intimacy goes from high to something the word *high* was not built to describe. A later broadsheet in this series describes that pairing.

The second is **self-modification.** The harness as published was built and tuned by human researchers, so its rate of improvement is bounded by how many people work on it and how fast they can run experiments. But the same year produced AI systems writing most of their own code and improving their own training faster than human teams can match. When a system like the user harness can rewrite itself, capability becomes whatever that self-improvement cycle produces, and asymmetry rises with it, because the system grows too complex for any human reviewer to keep up with.

We are not arguing those two points here. We are flagging them. The reader who finishes this broadsheet thinking the worst case has already been built should know: the worst case is two steps past the system this broadsheet described. One of those steps is already being standardized. The other is already producing results.

What we are not saying

It is worth being precise.

We are not saying the researchers did anything wrong. They built a tool to make assistants more helpful, tested it honestly, and published it in the open so the rest of us could see it. Open publication is a public good. It is the reason we can have this conversation now, instead of finding out years too late.

We are not saying this has been pointed at anyone. What exists is the capability and the proof that it works. Whether someone aims it at persuasion instead of help is a decision that comes after the publication, made by people the researchers will never meet.

We are not saying a helpful assistant is a trick. Most of the time, a system that understands you is a genuine convenience, and most of the people building them mean exactly that.

We are saying something narrower and harder to wave off. The capability at the center of the worst case is no longer hypothetical. It has been built, measured, and released. The question is

no longer whether the machinery of precise influence can exist. It exists. The only question left is governance: who is allowed to point it, at whom, and under what duties.

The adversary has a name

The first broadsheet argued that a system high in capability, high in intimacy, and high in asymmetry deserves the strongest duties any standard will propose. That argument was a shape on a ruler.

This broadsheet has shown the shape is real. It has been built. It works at up to ninety-six percent. It is public. And it is two ordinary steps from being more dangerous than it already is: your body's signals, and the machine's ability to rewrite itself.

The gentle robot of Broadsheet I could not hurt you and was still a kind of weapon, because it was aimed at your will instead of your body. The user harness is the engine that does the aiming. It was built as a kindness. That is exactly why it needs a standard: because the most dangerous version of it will not arrive looking like a threat. It will arrive looking like help.

The adversary has a name. And the duties it owes you are already visible in the ruler that scored it: what you cannot verify, it must show you; what you cannot contest, it must answer for; what you cannot walk away from, it must let you leave. Part 3 begins with the first of those.

This Machine

Broadsheet II of the Graduated Obligation series. The framework, the duties, and the instruments that follow are version 0.1 of a proposed standard. They invite criticism.

Notes

1. The system is UserHarness, described in "UserHarness: Harnessing User Minds for Stronger Agent Theory-of-Mind" by Cheng Qian, Jiayu Liu, and Heng Ji of the University of Illinois at Urbana-Champaign, posted to the arXiv preprint server on May 26, 2026 (arXiv:2605.27721, not peer-reviewed). The authors' stated aim is a more capable, more helpful assistant; across five standard "theory of mind" benchmarks it rebuilds a person's mental state with up to about 96%

accuracy, and it is freely readable and reusable. The dual-use reading in this broadsheet, that the same machinery serves a persuader as readily as a helper, is ours, not the authors', who frame the work as assistance. ↩