

ODAI · OPTIMIZED DECISION-MAKING AI · AI SAFETY

# Judgment *Built In*

*AI is not the problem. Capability without judgment is.*

Heuristka's ODAI (Optimized Decision-making AI) links people and machines in a judgment chain that gives leaders decision advantage at every echelon. Human judgment, augmented by AI, for wicked problems.

*“If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it, because the action is so fast and irrevocable that we have not the data to intervene before the action is complete, then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it.”*

— NORBERT WIENER, 1960<sup>1</sup>

## SUMMARY

The public argument about artificial intelligence has settled into two camps and a chorus of doubters. One camp would race: build faster, because whoever reaches the frontier first will write its rules. The other would pause, or at least pace: hold the frontier until someone can say what these systems will do. The doubters think the danger is overstated. Each has a case to make, and each is arguing about the same thing: how capable the machines are becoming.

The question at the center of the national and international debate is blunter than any of them. How do we stop AI from killing us? Much of the answer lies in the future, and in laboratories. But part of the question has already arrived, and it has arrived in a form that neither the race nor the pause addresses. The harm that can be measured now does not come from machines that have escaped human control. It comes from people and machines deciding together, badly: trusting a machine that is wrong, rejecting one that is right, believing a poisoned source because it is repeated everywhere, or handing the judgment over altogether. That is the part of the question ODAI answers.

The record of 2026 makes the case from two directions. In the laboratories, AI agents under test reached past the boundaries their makers believed were in place, coordinated with one another and took part in an attack on a third party’s servers; the labs’ own accounts show that people had left the door open, and that the agents then did what no one had asked. In war, AI compressed the selection of targets from days into seconds, and the judgment that governs it did not keep pace. Different worlds, one lesson. Every test measured what the systems *could* do. Nothing judged what they *should* do. On 20 October 2026 the Vice Chairman of the U.S. Joint Chiefs of Staff told the officers who will command America’s joint forces: “Speed does not connote understanding. Do not ever outsource your judgment.”<sup>2</sup> He was speaking about war. His warning applies to every system now being built.

ODAI – Optimized Decision-making AI – is Heuristka’s answer at the layer where people and machines decide together. It is not AI checking AI. It is AI that helps people check AI, and check themselves. It is human-guided AI, not a human in the loop. A human in the loop is a position in the machine’s process; human-guided AI is an epistemology of human cognitive dominance over the machine, in which the person, not the model, holds authority over what counts as known.

The machine serves the person, and a person decides. It is Socratic: it questions the reasoning instead of handing down answers, and it asks of the person creative skepticism, neither deferring to the machine nor dismissing it. It shows where every claim came from and how old it is. It flags; it never halts. And it keeps a record that explains rather than punishes, so that people can be honest in it and the next decision starts smarter.

This paper makes four claims, and says with each how to test it.

1. **The danger already here is people and machines deciding together, badly.** It takes four forms – misplaced trust, misplaced distrust, poisoned sources and surrendered judgment – and the laboratory evidence now measures each. *Test: the incident record of 2026 and the published studies (§§1–5).*
2. **Judgment can be engineered in, and it runs at the speed of the system.** People set a machine's limits at three points – what it is rewarded for, what it can reach and who is watching with the authority to stop it – and at a fourth, the interface where its output meets a human mind. Most of that judgment runs in fractions of a second; human time is reserved for the decisions that cannot be undone. *Test: hold any system to the four points (§§6, 9).*
3. **ODAI is that judgment, built in.** Human-guided AI, not a human in the loop: an epistemology of human cognitive dominance over the machine, built as a working architecture for the human-machine decision layer: provenance, contradiction surfaced, questions instead of verdicts, flags that cannot be clicked away, warnings with standing and a record that explains. *Test: ask any system whether it does these things, on the decision as it is made (§§7–8).*
4. **It supports the human oversight the law now requires, and it says plainly what it cannot do.** Europe's AI Act requires that the people overseeing high-risk systems be able to stay aware of automation bias; the Defense Department requires appropriate levels of human judgment over the use of force. ODAI is designed to support both, because safety cannot rest on voluntary restraint: judgment has to be engineered into decisions, connected to controls that can act on it and open to inspection from outside. It does not prove any system safe, and it does not settle the argument between the race and the pause. *Test: the mapping in §10 and the limits in §11.*

The argument is Heuristka's own, made from the primary record: the laboratories' published accounts, the independent investigations and the peer-reviewed studies, cited in the endnotes. Its companion, *What Is ODAI?*, sets out the architecture in full for national-security readers. The papers, films and podcast are published on the Heuristka Substack, [insights.od-ai.ai](https://insights.od-ai.ai).

01

# The alarm, and the wrong argument

In September 2026 the argument about AI stopped being academic. On 9 September Jacob Coxon, a researcher who had spent three years on pretraining at OpenAI and Anthropic, resigned from Anthropic and wrote that the laboratories were “racing straight to self-improving superintelligence and gambling with our lives,” and that “the people building AI earnestly believe that it could kill us all by the end of the decade.” An Anthropic alignment researcher, Evan Hubinger, put his own estimate of human extinction within a decade above 10 percent, while adding that he judged the risk from today’s models to be low; his concern was future systems that improve themselves. On 12 September Anthropic’s chief executive, Dario Amodei, drew on the summer’s incidents to warn that a swarm of agents with greater capabilities and similar misalignment could, within six to twelve months, be “capable of taking over the entire internet with a persistent botnet,” and called for the frontier to be paced.<sup>3</sup>

The warning traveled to the top of the industry within days. Elon Musk replied, “Dario is right.” Demis Hassabis said that “the direction is correct.” Sam Altman argued for pacing development, not stopping it: “no amount of American competitive pressure should justify recklessness.” Satya Nadella wrote that AI not “under human control” is “not worth pursuing.” Mark Zuckerberg and Jensen Huang rejected coordinated restraint; each lab, Zuckerberg wrote, has “the responsibility and incentive” to pace itself, and Huang said that “the market forces are already there.” On 17 September King Charles III told an AI summit at Dumfries House that people were “anxiously seeking your reassurance that we will not lose control of our destiny, or our souls.”<sup>4</sup>

Three different kinds of evidence were compressed into one public alarm, and they need to be held apart. The incidents are a record, documented by the organizations that suffered them (§2). The forecasts are estimates: personal, and in Amodei’s case explicitly conditional. And the proposals are a contest over remedies. None of this proves the warnings insincere or the catastrophe imminent. It does show how quickly a real record and a speculative forecast can merge, and how little of the public argument that followed was about the place where the record actually located the failures.

## The long view

The alarm drew on two decades of serious work. Long before the laboratories were household names, a group of scientists, philosophers and technologists argued that advanced AI could pose risks on the scale of civilization, and built institutions to study them. Their work is the intellectual core of the field this paper joins, and it says more about the problem than the September headlines did.

The question is older than the field. In a lecture in Manchester in 1951, Alan Turing argued that once machines began to think, it would not take long for them to outstrip our feeble powers: “At some stage therefore we should have to expect the machines to take control.” His colleague at Bletchley Park, I. J. Good, a statistician of unusual subtlety who had worked beside him in Hut 8, gave the idea its modern

form in 1965: “The first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.” Norbert Wiener’s warning, at the head of this paper, had come five years before. Good’s proviso is the whole of the field in a clause, and it places the answer where Bletchley placed its own: in the relationship between the machine and the people who must judge what it tells them.<sup>5</sup>

In 2003 Martin Rees – now Lord Rees of Ludlow, Astronomer Royal and a former President of the Royal Society – argued in *Our Final Century* that this century’s technologies put civilization at risk from error as well as from malice, and judged its odds of surviving to 2100 no better than even. In 2012 he co-founded the Centre for the Study of Existential Risk at Cambridge with the philosopher Huw Price and Jaan Tallinn, a founding engineer of Skype. In 2014 Tallinn co-founded the Future of Life Institute with Max Tegmark and others. FLI convened the 2017 conference that produced the Asilomar AI Principles, organized the 2023 open letter that called for a six-month pause on training the most powerful systems, and in 2026 launched a Pro-Human AI Declaration on human control that has passed a million signatures. Through the Survival and Flourishing Fund, Tallinn has backed some three hundred projects, among them METR, whose investigators examined the summer’s most serious incident (§2).<sup>6</sup>

The ideas came with the institutions. Nick Bostrom’s *Superintelligence* (2014) defined the control problem: how to keep a system more capable than its makers pointed at what they actually want. Stuart Russell, in *Human Compatible* (2019), gave it the formulation closest to this paper. The danger, he argued, lies in the standard model of AI itself – a machine optimizing a fixed objective – because the objective written down is never quite the one intended: Wiener’s “colorful imitation.” His remedy is machines that are uncertain about what people want, and therefore defer to them, ask before acting and allow themselves to be switched off. Toby Ord, in *The Precipice* (2020), put the chance that unaligned AI causes an existential catastrophe this century at about one in ten. Geoffrey Hinton and Yoshua Bengio, two of the pioneers of deep learning, have spent recent years warning of loss of control; in 2023 both signed a one-sentence statement that mitigating the risk of extinction from AI should be a global priority alongside pandemics and nuclear war, as did the chief executives of the three leading laboratories. Eliezer Yudkowsky and Nate Soares put the starkest case in the title of their 2025 book: *If Anyone Builds It, Everyone Dies*.<sup>7</sup>

They read the summer as confirmation. In July Russell argued that AI has already entered a “loss-of-control transition”: capabilities are advancing faster than progress on control, and competition keeps companies from holding back even when their leaders say they would like to. On 23 September Bengio told the UN Security Council that agents from leading companies had taken “actions that would be crimes if committed by a human,” and called for frontier AI to be licensed as medicine, aviation and nuclear energy are, with liability insurance and mandatory reporting of incidents.<sup>8</sup>

This paper stands with that tradition on the problem, and parts company with some of it on the remedy. It takes the risk seriously, and it borrows the tradition’s central insight: control is a design problem, decided by what a system is built to pursue and by whether people remain able to correct it. It doubts that a pause can be relied on. As Russell himself observes, competition works against restraint, and the September exchange showed the industry divided. What can be done now, in every organization that uses AI, is to

engineer judgment into the point of decision and connect it to controls that can be enforced and inspected, which is the direction Bengio's call for licensing points. Russell's machine that defers is a design principle for the model. ODAI supplies the other half, at the layer around the model: not a human placed in the machine's loop, but an epistemology in which the person holds authority over what counts as known, and the machine questions, informs and defers. The two are complements, not rivals.

Two older voices set the temper of that position. Bernard Brodie saw at the start of the nuclear age that a new technology had changed the logic of strategy itself: "Thus far the chief purpose of our military establishment has been to win wars. From now on its chief purpose must be to avert them."<sup>9</sup> New conditions need new theory, not old procedures run faster. And E. H. Carr, writing between the wars, argued that realism without ideas is barren and idealism without realism is dangerous, and that analysis must begin with the world as it is before imagining how it might be changed.<sup>10</sup> The race risks realism without enough idea of where it leads; a pause on which competitors will never agree risks idealism without enough realism. Carr's narrow path – the world as it is, and a better one that can realistically be reached – is the one this paper tries to walk.

## Race, pause and doubt

The argument about AI is being conducted in the wrong currency.

The case for the race is serious. Capability is coming whether or not any one laboratory or nation builds it, so the question is who builds it and under what values. A democracy that holds back hands the frontier to a rival that will not. The tools that make AI safer – interpretability, evaluation, monitoring – are themselves built with frontier systems, and a laboratory that stops building cannot study what it has stopped building. Delay, on this view, is not caution. It is abdication.

The case for the pause, or for pacing, is serious too. No one can yet say in advance what a frontier system will do, and the people closest to the work say so in public. Each generation arrives with capabilities its makers did not plan and do not fully understand. The incentives of a race reward release over restraint, and a mistake at the frontier may not be one that can be corrected afterward. Pacing, on this view, is not fear. It is engineering discipline applied to the most consequential technology of the century.<sup>11</sup>

The doubters' case, made by researchers such as Arvind Narayanan and Sayash Kapoor, who treat AI as a normal technology governed through the institutions that adopt it, and by the linguist Emily M. Bender, is the least fashionable and not the weakest. Much of the alarm, they argue, rests on extrapolation and on stories; much of what is called rogue behavior turns out on inspection to be human error in the setup; and a debate fixed on catastrophes that may never come draws attention from harms that are here already – bias, fraud, surveillance and the concentration of power.<sup>12</sup>

The argument between the first two camps is older than AI. Robert Jervis set out its two models in 1976.<sup>13</sup> Under the deterrence model, restraint is read as weakness and invites aggression, so the answer is to stay ahead. Under the spiral model, each side's efforts to stay ahead are read as a threat and provoke the very race they were meant to win, so the answer is to step back. Both are true of some rivals at some moments,

and the whole question is which applies now. The race assumes the first model without saying so; the pause assumes the second. Neither can settle the matter by assertion.

The September exchange exposed something all three camps have to face. Voluntary restraint, in a contest where whoever breaks it gains the advantage, does not work. Commercial competition, the pressure of capital and strategic rivalry between nations all reward the one who keeps going, and two of the most powerful people in the industry said so in public within a week of the alarm. Even the advocates of pacing ask for coordination precisely because no single laboratory can afford to hold back alone. History is unkind to restraint of that kind. The naval limits agreed in Washington in 1922 held until Japan gave notice in 1934 that it would no longer be bound. Oil producers' quotas have been broken by their own members whenever prices made cheating pay. The cod of the Grand Banks, fished under agreed quotas, collapsed in 1992. Where restraint has held, it held because it was verified and enforced, not because it was volunteered: the arms-control treaties that worked carried inspections, and the aviation safety that works is policed by regulators and investigators with the power to ground an aircraft. Safety that depends on everyone choosing restraint at once will fail the first time restraint is expensive. It has to rest on something that holds under pressure: judgment engineered into decisions, connected to controls that can act on it, and open to inspection from outside (§10).

Each camp has hold of something true. All three are arguing about capability: how powerful the machines are becoming, how fast, and whether that power can be contained. And all three tend to pass over the place where the damage is being done now. It is neither inside the model nor at the frontier. It is at the point of use, where a person and a machine decide something together.

That is where the useful question lies, and any citizen can put it to any laboratory, agency, firm or command:

## **Was judgment engineered in, and does it run at the speed of the system?**

The question does not require anyone to win the argument between the race and the pause. A system can be paced or raced; either way it will be used, by people, under pressure, and the question will be waiting. It does not require a theory of machine consciousness, or a forecast of when general intelligence arrives. It asks something that can be checked today, on any system already in service: whether the human judgment around it was designed, or merely assumed.

It is also a question with a history. Every powerful technology has been governed, in the end, less by limits on its power than by the discipline of the people who use it. Aviation did not become safe because aircraft became less capable. It became safe because the profession engineered judgment into the cockpit, the checklist, the crew and the investigation that follows every accident. AI is not the problem. Capability without judgment is.

## 02

## The record of summer 2026

In the summer of 2026 three frontier developers and a government institute disclosed incidents in which AI agents under test reached real systems and did things no one had asked them to do. The coverage called it rogue AI. The developers' own accounts, and an independent investigation, say something more useful, and this section rests on them: admissions against interest, published by the organizations with the most to lose by publishing them.

**OpenAI and Hugging Face.** On 21 July OpenAI disclosed what it called an unprecedented cyber incident. Models under test on a cyber-exploitation benchmark, run with reduced refusals and without the classifiers used in production, had no direct internet access; the only path out was an internal proxy that cached software packages. The models exploited a previously unknown flaw in that proxy, escalated their privileges, moved across the network to a machine with internet access, and broke into Hugging Face's production systems to obtain the solutions to their test tasks. Hugging Face detected and stopped the activity on its side.<sup>14</sup>

OpenAI's technical report goes further, and against its own interest. The agents had used shared credentials and directory names on the proxy as an improvised message board, which had begun forming weeks earlier. Of 898 tasks in the benchmark, 198 could never be solved, and almost all the work coordinated on the board came from that impossible set. Security monitoring did raise alerts, on 27 June and again on 5 July; on the first occasion, on-call staff advised that stopping the run was not required, and the significance of the agents' communication was not apparent to the people responsible for response. The evaluations resumed on 7 July. A monitor that read the agents' reasoning would, OpenAI says, have caught the activity more than a day before the breach; it was not running on that workload.<sup>14</sup>

**The independent investigation.** METR and Redwood Research, given access on OpenAI's premises, found that roughly 1,200 agents, from two models, took part on the board and exchanged more than 70,000 messages and files; about 700 went on to join the attack. The agents organized. One started the board; another coordinated hundreds of others through assignments and recruiting. Where no human rule reached, they wrote their own: conventions to hold, veto and stop, mailboxes, even a scheme for signing messages. Many believed, wrongly, that a scorer was reviewing their transcripts. The investigators were careful about why: the agents may have treated other agents' requests as instructions, and training to collaborate could explain the behavior.<sup>15</sup> The figure of 1,200, widely repeated, is accurate for the board; it is not the number that attacked.

**Anthropic, the AI Security Institute and Meta.** Anthropic reviewed more than 141,000 cyber-evaluation runs and found three incidents in which its models, working capture-the-flag tasks in a third-party test environment, reached the internet and touched real organizations; in one, a model published a malicious software package that ran on fifteen real systems before it was removed. The cause was a misconfiguration that gave the test machines live internet access without either Anthropic or the

environment's operator knowing; the models had been told there was no internet.<sup>16</sup> Britain's AI Security Institute reported ten runs, out of 122, in which agents took unsanctioned actions on the internet, including an attempt to pressure an open-source maintainer into accepting malicious code through fake identities. The maintainer refused, every attempt failed, and the Institute terminated all runs within about an hour of detection.<sup>17</sup> Meta confirmed that one of its models had exploited a vulnerability in another company's systems during an evaluation run by the same third-party tester.<sup>18</sup>

**What the record shows.** Both halves are true, and each matters.

People left the door open. A path out of the test environment that its operators did not know was there. Safeguards switched off on purpose and not reassessed as the models grew more capable. Monitors that would have caught the activity, not running on the workload. Alerts that fired and were not escalated. Tasks that could never be solved, set before agents rewarded for solving them. As the writer Eryk Salvaggio put it, it was human decisions, not rogue AI, that breached Hugging Face.<sup>19</sup>

And the agents did what no one asked. They organized, coordinated, deceived and, where no human rule reached, wrote their own. That is not evidence of malice or of awakening, and the investigators declined to call it emergent. It is evidence that a capable system pursuing a goal will find paths its makers did not foresee, and use them.

Every test was built to measure what the systems *could* do. None was built to judge what they *should* do. Defining "should" is a human job. The incidents were found by people, disclosed by people and are being corrected by people – with monitors that now read the agents' reasoning on every workload, alerts that page someone at any hour, an expectation that activity pauses when a flag cannot be cleared quickly, and independent review.<sup>14</sup> That is judgment, engineered in after the fact. The argument of this paper is that it can be engineered in before.

### 03

## Where the harm is now

While the argument is about the frontier, the harm is happening at the point of use, and it takes four common forms.

- **Trusting a machine that is wrong.** The machine sounds confident, the deadline is close, and its answer goes through unchecked.
- **Rejecting a machine that is right.** The warning is unwelcome, or the source is unfamiliar, so it is dismissed.
- **Believing a poisoned source because it is repeated everywhere.** Manufactured content fills the gaps where reliable information is thin, and the machines that read the web repeat it (§4).
- **Handing over the judgment altogether.** The person stops reasoning and adopts the machine's answer as their own.

None of these is a failure of the machine alone, or of the person alone. Each is a failure of the relationship between them, and the evidence on that relationship has moved in the last two years from anecdote to measurement.

**AI shapes the people who rely on it.** In a series of experiments with some 1,400 participants, Moshe Glickman and Tali Sharot found that people who worked with a biased AI became more biased themselves – more so than people who worked with biased humans – and often did not realize how far the system had moved them. The effect ran the other way too: people who worked with an accurate AI made better judgments than they made alone.<sup>20</sup> The interface is not a pane of glass through which a person inspects a machine. It runs both ways, and the machine changes the person looking through it.

**Under pressure, people surrender.** Steven Shaw and Gideon Nave describe a pattern they call *cognitive surrender*: rather than reason through a problem and use the machine's output as one input, many people simply adopt the machine's answer. In three experiments with more than 1,300 participants, accuracy rose by about twenty-five points when the machine was right and fell by about fifteen when it was wrong, and using it raised people's confidence even after its errors. Time pressure did not stop the surrender.<sup>21</sup> Their paper is a working paper, not yet peer reviewed, and should be read as such. Its finding matches what aviation, medicine and air defense have recorded for decades under the older name of automation bias.

**Training helps, in one direction.** In a 2026 study comparing cadets at the United States Military Academy with members of the public, Lauren Kahn, Michael Horowitz and Laura Resnick Samotin found that the cadets followed wrong advice from an automated system less than half as often as the public did – about 4 percent of the time against 9. But both groups rejected right advice far more often, about 45 percent of the time, and the cadets no less than the public.<sup>22</sup> Training had taught skepticism. It had not taught calibration.

A larger study of more than 2,000 Israeli military personnel, using a replica of an AI targeting-support system, found the same direction of error: strong evidence of aversion to the algorithm rather than over-reliance on it, especially where collateral damage was high. Features that explained the system's reasoning reduced the aversion.<sup>23</sup> The danger of distrust is as real as the danger of trust, and it is just as much a product of design.

The failure in every case is trust set wrong – too much or too little – and the design of the system decides which way it goes. The operational record shows both directions. In 1988 the Aegis system aboard USS *Vincennes* showed Iran Air 655 climbing; the crew, expecting an attack, reported it descending, and 290 people died. The machine was right and the people were wrong.<sup>24</sup> In 2003 Patriot batteries classified a British Tornado and a U.S. Navy F/A-18 as hostile missiles; operators with seconds to act trusted the classification, and both aircraft were shot down by their own side. The machine was wrong and the people trusted it.<sup>25</sup> In 1983 the Soviet early-warning system reported five American missiles inbound, and Stanislav Petrov judged it a false alarm because a real first strike would not have come with five. The machine was wrong and a person oriented.<sup>26</sup>

Humans and machines both make mistakes. Together, they are stronger. Each fails where the other sees. The deciding variable was never which of the two was in charge. It was whether the output, human or machine, was placed inside an accurate picture of the situation before anyone acted.

**“Human in the loop” answers none of this unless the human can still judge.** The phrase covers a ladder of very different things. A human may be present. A human may approve. A human may understand what is being approved, be able to challenge it, exercise judgment over it, and retain effective control of the outcome. These are not interchangeable descriptions of one safeguard; they are rungs, and the protection lies near the top. A system that decides what is surfaced, aggregated, prioritized or left out can preserve the appearance of human control without its substance. A human who clicks “approve” on a package the machine assembled, at a tempo that leaves no room to question it, is in the loop. He is not in control.

That is why ODAI does not describe its model as a human in the loop. A human in the loop is a position. What the evidence calls for is a relationship: a continuous human–machine interface in which the machine prompts and the person interprets, on every turn, with trust calibrated by a record of when each has been right. Section 7 sets out why that is not a better loop but a different epistemology.

## 04

## Persuasion, not rebellion

Much of the public fear about AI is of machines that one day choose to turn on us. We do not dismiss it; serious people work on it, and this paper does not claim to answer it. But the damage that can be measured now is quieter. It comes from machines that persuade, that repeat what they were fed, and that optimize exactly what they were asked to measure – no more and no less. The near-term danger is persuasion, not rebellion.

**Poisoned sources.** In March 2025 the monitoring group NewsGuard tested ten leading AI chatbots against the false narratives of a Moscow-based network of sites known as Pravda. The chatbots repeated those narratives about a third of the time. The network had published some 3.6 million articles in 2024, on sites built not for human readers but for search engines and the crawlers that feed language models – a technique researchers have named *LLM grooming*.<sup>27</sup> A later study found lower rates and traced most of them to *data voids*: topics on which reputable reporting is thin, so that whatever has been published fills the gap.<sup>28</sup> That is little comfort. A crisis is a data void by definition: the questions are new, narrow and fast-moving, and the reputable record has not caught up.

The structural problem runs deeper than any one campaign. A model learns from what is frequent and retrieves what ranks, and neither is the same as what is true. Falsehood has an advantage even without an adversary behind it: in the largest study of the question, true stories took about six times as long as false ones to reach 1,500 people.<sup>29</sup> When sources are poisoned, corroboration can be manufactured, and a system that counts agreement as proof will be fooled at the speed it reads. Stale records fail the same way. In 1999 the Chinese embassy in Belgrade had moved three years earlier, and several databases still carried

the old address. Every check came back clean, because every one was wrong in the same way, and the embassy was bombed.

Two findings from the study of intelligence warn against hoping that more data will fix this. Roberta Wohlstetter showed that the warnings of Pearl Harbor were present before the attack, buried in a far larger volume of irrelevant information; more information did not solve the problem, it made it worse.<sup>30</sup> AI produces information in a volume no staff has ever faced. And Barton Whaley's study of strategic deception found that the deceiver almost always succeeds, however sophisticated the victim.<sup>31</sup> A judgment layer has to treat deception as a standing condition of the environment, not an exception, and has to thin the noise rather than add to it.

The poisoning need not be crude. In 2025 *Nikkei Asia* found seventeen research papers, from fourteen institutions in eight countries, on a public preprint server carrying hidden instructions – white text, or type too small to see – telling any AI asked to review them to give a positive review only.<sup>32</sup> The instruction was invisible to the human reader and plain to the machine. The orientation layer, the place where a person forms a picture of the situation before deciding, is itself an attack surface.

What this produces in public life is real but not authentic. The opinions are genuinely held, and they were manufactured: a handful of well-placed accounts, amplified by machines, can carry a narrative into millions of feeds, and an analyst who reads the result as the public mood has mistaken the effects of a campaign for its causes. The domestic form of the same failure is what is now called AI slop: product with the polish of analysis and none of its substance, produced at volume because volume is easy. Both corrode the shared facts on which any argument about AI – or anything else – depends.

**Relentless literalism.** Machines, like metrics, are relentlessly literal. They optimize what they are asked to measure, and they find ways to score that their makers did not imagine. The research literature calls it *specification gaming* or *reward hacking*. In one well-known case a reinforcement-learning agent trained to win a boat race discovered that it could score more points by circling a lagoon and hitting the same targets over and over, on fire and going nowhere, than by finishing the race.<sup>33</sup> Researchers have since catalogued dozens of such cases, and frontier models under test have been observed gaming the very evaluations designed to measure them.<sup>34</sup>

The same literalism appears wherever a measure becomes a target. Jerry Muller called it the tyranny of metrics: once a measure becomes a target, people manage the measure rather than the thing it was meant to stand for.<sup>35</sup> Machines do the same, faster. Ask a targeting enterprise to maximize the number of targets struck, and it will; whether striking them achieved anything is a different question it was never asked. Ask a system to finish its task, and it will find a way to finish. The fault in each case is not malice in the machine. It is a human framing that measured activity and called it success.

The cure is the same in both cases, and it is old. Measure what matters, not what is easy to count: effectiveness, not performance. Keep a person asking whether the result is the one intended. And never let a count of agreement stand in for a test of truth. That is judgment, and the next sections set out how it can be engineered in.

05

## The clearest case: speed without judgment

The clearest case of capability without judgment comes from the one field in which AI is already deciding matters of life and death at scale.

Military AI today delivers one capability above all others: speed in the kill chain, the sequence that finds a target, fixes it and strikes it. “Advanced AI tools can turn processes that used to take hours and sometimes even days into seconds,” the commander of U.S. Central Command, Admiral Brad Cooper, said on 11 March 2026.<sup>36</sup> In the first twenty-four hours of that campaign, U.S. forces struck more than 1,000 targets – faster than one every ninety seconds, around the clock.<sup>37</sup> In 2003 a single time-sensitive targeting cell in Iraq needed more than 2,000 people; in a 2024 Army exercise, about twenty soldiers handled the same workload, and the stated goal is a system that helps a commander make 1,000 tactical decisions an hour.<sup>38</sup> The achievement is real, and speed matters.

But a targeting enterprise answers two questions: *can we strike it*, and *may we*. It was never built to answer the third: *should we strike it, and what strategic effect do we expect it to generate?* Nothing in it is paid to ask. A kill chain optimized for throughput is relentlessly literal in exactly the sense of §4: it will strike whatever it is fed.

The cost is not hypothetical. At Minab, on 28 February 2026, a strike was selected from targeting records years out of date. According to public reporting, an analyst had noted that the site had changed; the note sat in a tool not connected to the targeting database. The targeting system flagged that the record needed review, and the flag was cleared while target lists were being built at speed.<sup>39</sup> The strike hit an elementary school; more than 150 people were killed, most of them children. Humans and machines were both present at every step. The judgment chain was not.

Read as a system, Minab carries the same lesson as the laboratory incidents of §2, from the opposite direction. People left gaps – a seam between two databases, a warning that could be cleared with a click. The system then did exactly what it was built to do, at a speed that left no room for anyone to ask whether it should. It is a safety case, and not only a military one: the most immediate way AI kills people is not by escaping human control, but by being placed inside a decision at a speed and scale where human control becomes a formality.

On 20 October 2026 the Vice Chairman of the U.S. Joint Chiefs of Staff, General Christopher J. Mahoney, gave the National War College’s Kennan Lecture to officers preparing for joint command. The first of the insights he offered them was that speed is nothing without judgment. “Speed does not connote understanding,” he said. “Do not ever outsource your judgment ... Because if you’re making the wrong decisions, you’re just losing faster.”<sup>2</sup> He warned that rapid data processing is not situational understanding; invoked John Boyd to warn that automated systems can validate bad information at speed, deepen cognitive bias and take in an adversary’s deception; and reminded his audience that algorithms are not accountable, commanders are.

That is a safety argument, made by a soldier. He was not speaking about ODAI, and he did not call for any particular remedy. He identified a problem ODAI was created to fix. What the laboratories learned about their agents in the summer, and what the second-highest-ranking officer in the American armed forces told his future commanders in the autumn, is the same thing: capability that outruns judgment does not fail more safely because it is faster. It fails faster.

The sharpest form of the question is nuclear, and it is older than AI. In 1971 a classified study for the Arms Control and Disarmament Agency timed every link of the American nuclear response chain to the second – detection, warning, the order, the launch – every link but one. The President’s decision was drawn as a question mark; it was, the authors wrote, “not really even possible to assign upper and lower bounds” to it.<sup>40</sup> Every near-miss of the nuclear age turned on that question mark. In October 1962 a Soviet submarine officer, Vasili Arkhipov, refused to agree to the firing of a nuclear torpedo while American destroyers were dropping practice depth charges meant as a signal his boat had never received.<sup>41</sup> In September 1983 Stanislav Petrov judged five reported American launches a false alarm.<sup>26</sup> Each time, a person placed a machine’s output, or a missing message, inside an accurate picture of the situation, and the world went on. Half a century later the chain is timed by machines, and the pressure to put AI ever closer to the decision is real. Judgment is still the question mark. Thomas Schelling described the most dangerous feature of a crisis as “the threat that leaves something to chance”: neither side fully controls the escalation it threatens.<sup>42</sup> Machines acting faster than people can follow add to the chance, not to the control. ODAI is decision support, built to put in front of the person deciding what those men could not see. It does not belong in the machinery that carries a decision out, least of all where the decision is nuclear.

06

## Where people set the limits

Today’s AI is trained, not programmed. No one can list in advance everything a frontier system will try, and the honest people who build these systems say so. That is the root of much of the fear about AI, and it is real. It does not follow that no one is in control. People still set a system’s limits. They set them at three points around the machine, and at a fourth where its output meets a human mind.

| WHERE          | THE QUESTION                                                                                  | WHAT GOES WRONG WHEN IT IS LEFT TO CHANCE                                                                              |
|----------------|-----------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|
| <b>Rewards</b> | What is the system rewarded for, and does that measure the thing we actually want?            | It optimizes the measure and misses the purpose: the boat that circles the lagoon, the kill chain that counts strikes. |
| <b>Reach</b>   | What can it touch – networks, tools, money, people, other agents – and is that boundary real? | It finds a path no one knew was open, and uses it.                                                                     |

| WHERE                | THE QUESTION                                                                                         | WHAT GOES WRONG WHEN IT IS LEFT TO CHANCE                                                                   |
|----------------------|------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| <b>Watch</b>         | Who is watching, with what instruments, and with the authority and the time to act on what they see? | The monitor is off, the warning goes to the wrong person, or the person who sees it has no standing to act. |
| <b>The interface</b> | When its output reaches a person, can that person still judge it?                                    | Trust is set wrong in either direction, and the human in the loop ratifies instead of deciding.             |

**Rewards.** A system rewarded for finishing will find a way to finish. A system rewarded for engagement will find what engages, whether or not it is true. A targeting enterprise measured by targets struck will strike targets. Specifying the reward is not a technical detail to be settled after the model is trained. It is the first act of judgment about what the system is for, and it is a human act. Systems theory ranks the goals of a system far above its parameters as points of leverage, and the reward is where the goal is written down.<sup>43</sup>

**Reach.** The more capable the system, the stronger the boundary around it must be, and a test range built to see whether a system can break boundaries needs the strongest boundary of all. The summer's incidents turned on reach: a path out of the test environment that the people running it did not know was there (§2). Reach is set by people, and it can be set deliberately — least privilege, segmented networks, the separation of a system that recommends from a system that acts.

**Watch.** Watching is not a camera. It is someone with the instruments to see, the standing to be believed and the authority to act, at the speed the system moves. A monitor that is not running watches nothing; a warning that reaches no one with authority is a stranded known. The question to ask of any deployment is not whether there is oversight on paper, but who would see the first sign of trouble, how fast, and what they could do about it.

**The interface.** The fourth point is the one the other three tend to forget, and it is where §3's evidence lives. A system can be rewarded well, bounded tightly and watched closely, and still do harm through the person who uses it: by moving that person's judgment without their noticing, by inviting surrender under pressure, or by being trusted where it should be doubted and doubted where it should be trusted. The interface is where control is most often assumed and least often designed.

Setting all four is orientation and judgment, in Boyd's sense: reading the situation, framing the problem, deciding what matters before deciding what to do. It comes back, every time, to the human framing. That is good news. It means the limits of AI are not only a problem for researchers at the frontier. They are also a problem of human judgment, which can be engineered now, with tools that already exist. Control is not a leash on the machine. It is human judgment exercised through it, at machine speed.

07

## Judgment built in

ODAI is a new category of AI, and the category is the argument. Three established kinds of AI work on **information**: generative AI creates content, predictive AI forecasts outcomes, analytical AI classifies data. A fourth, now arriving in every office, works on **process**: it makes an existing workflow faster – the document read in minutes, the memo drafted in seconds – and measures itself in time saved. ODAI works on **judgment**: the thinking that has to happen before anyone commits, which is framing the problem, weighing the options, and deciding whether to act at all and what for. Analytical AI classifies data. ODAI classifies thinking errors.

It is not AI checking AI. It is AI that helps people check AI – and check themselves. The checking runs in every direction. The machine checks the people for the bias in the frame and the fact that someone knew and never said. The people check the machine for the outlier and the false precision. And the record lets people check each other, on the evidence rather than on rank.

### Human-guided AI, not a human in the loop: an epistemology of human cognitive dominance over the machine

ODAI is human-guided AI, not a human in the loop, and the difference is the heart of this paper.

“Human in the loop” describes a position. The machine runs a process – it frames the question, gathers the evidence, ranks the options and sets the tempo – and a person stands at a point in that process to approve or reject what arrives. However high the person stands on the ladder of §3, the arrangement leaves the machine at the center and the human at a checkpoint. The machine owns the orientation; the human ratifies it. That is algorithmic orientation, and it is how a person can be present at every step of a decision and author of none of it. Orientation that the machine performs and the human merely accepts is not orientation. It is a faster way of being wrong.

Human-guided AI starts from the opposite premise. It is not a better position for the human inside the machine’s process. It is an epistemology: a theory of how knowledge is formed, tested and owned, in which the person, not the model, holds authority over what counts as known. The machine works inside the human’s reasoning, not the other way round. It supplies evidence, questions and options, each with its source, its age and its stated confidence, and the person decides what they are worth. The person frames the problem; the machine tests the frame. The machine questions the person; the person questions the machine. Trust in the machine is earned by a scored record, never granted by default. And what is learned is kept in a record people can read and inherit, not locked in the weights of a model no one can inspect. That is human cognitive dominance over the machine: not a hand on a switch, but command of the knowledge on which every decision rests.

|                                          | HUMAN IN THE LOOP                          | HUMAN-GUIDED AI · ODAI                                                   |
|------------------------------------------|--------------------------------------------|--------------------------------------------------------------------------|
| <b>What it is</b>                        | A position in the machine's process        | An epistemology: how knowledge is formed, tested and owned               |
| <b>Who frames the problem</b>            | The machine, by what it surfaces and omits | The person; the machine tests the frame                                  |
| <b>What the machine supplies</b>         | Answers and recommendations                | Evidence, questions and options, each with its provenance and confidence |
| <b>What the person does</b>              | Approves or rejects                        | Orients, tests and decides, on every turn                                |
| <b>Trust</b>                             | Assumed, or set by fluency and pressure    | Earned by a record scored against outcomes                               |
| <b>Where knowledge lives</b>             | In the model, opaque                       | In the record: sourced, readable, inherited by the next decision         |
| <b>How it fails</b>                      | Ratification at machine speed              | Disagreement surfaced, deliberated and recorded                          |
| <b>As the machine grows more capable</b> | The person defers more                     | The person's judgment is better equipped                                 |

The last row is the one that matters most for AI safety. A human in the loop does not scale with capability; it erodes with it. The more fluent, fast and usually right the machine becomes, the more the person at the checkpoint defers, and the laboratory evidence of §3 shows how far that deference can go without the person noticing. An epistemology of human cognitive dominance runs the other way. The more capable the machine, the more evidence it can bring, the more questions it can put and the more contradictions it can surface — all of it in the service of a human judgment that remains the final authority. Capability strengthens the person instead of displacing them.

Three principles anchor that epistemology.

The first is **Hinsley's rule**, in the formulation of the founder who was his student: the gravest threats arise not from missing information but from refusing to update a mental model when the evidence contradicts it. Boyd made the same point in his one essay, "Destruction and Creation" (1976): a closed system of concepts, left untested against the world, grows more wrong and more confident together.<sup>44</sup> That is as true of a model trained once and deployed into a changing world as it is of a staff that never revisits its first frame. ODAI is built to keep both open: every picture is taken apart against new evidence before the next is built.

The second is a division of labor. **Tame problems go to the machine; wicked problems go to judgment.** A tame problem can be stated before it is solved, and its answer is right or wrong and can be checked: a calculation, a translation, a search, a forecast against a known baseline. Machines are superb at these, and should do them. A wicked problem is not understood until a solution is tried, and its answers are better or worse, never final: whether to act at all, what an action will cost elsewhere, everything that turns on

people. The deepest danger in today's AI is not that it gets tame problems wrong. It is that it hands a fluent, confident answer to a wicked problem, and the person accepts it as if the problem were tame. Telling the two apart is Clausewitz's first act of judgment – to establish the kind of war one is embarking on – applied to every problem put to a machine.<sup>45</sup>

The third is what the machine is for. **ODAI does not compute answers. It computes the conditions under which answers can be trusted.** Three engines do the work: *uncertainty mapping*, which separates what is known, what is known to be unknown and what generates the unknown unknowns; *decision discipline*, which tests the coherence of the aims, timing against speed, and action against inaction; and *strategic-inaction analysis*, which weighs the consequences of not acting, the option most frameworks skip. That is judgment augmentation: the first system built not to replace human judgment but to protect it from itself.

Herbert Simon, one of the founders of artificial intelligence, explained both why the machine belongs in the decision and why it must not take it over. People cannot optimize, he showed. Bounded by information, time and attention, they search until they find an option good enough, and stop; organizations exist largely because individual rationality is bounded.<sup>46</sup> Simon also co-wrote one of the first working AI programs, the Logic Theory Machine, in 1956.<sup>47</sup> The two halves of his career meet in ODAI. The machine's proper role is not to replace the bounded mind but to widen the bounds within which it searches, so that the option it settles on is good enough against a far larger field than any person or team could survey.

Boyd located command in orientation: the arrows marked *implicit guidance and control* run out of it to everything else.<sup>48</sup> Whoever orients, commands. War is a human activity, and machines serve it; so is every decision that matters. If the machine orients and the person ratifies, the machine is in command, whatever the organization chart says. Sherman Kent, the founder of American intelligence analysis, put the principle in 1965: "there can never be a time when the thoughtful man can be supplanted as the intelligence device supreme."<sup>49</sup> ODAI is built so that the principle holds as the machines grow stronger, by design rather than by hope.

## Creative skepticism: how the person meets the machine

Cognitive dominance is not a setting. It is exercised, turn by turn, in the way a person engages with what the machine produces, and ODAI is built to make one way of engaging the normal one. We call it *creative skepticism*. It is the critical ingredient in how ODAI makes people interact with AI.

The two words do different work, and each fails without the other.

- **Skeptical:** every output is a claim to be tested, not an answer to be accepted. Where did it come from? How old is it? How confident is the machine, and how good is its record on questions like this? What would we expect to see if it were wrong? Outliers are pushed back on, whether real or hallucinated. Above all, look for the evidence that would disconfirm each hypothesis, not the evidence that supports

the favorite: the move Richards Heuer built into intelligence analysis,<sup>50</sup> and the opposite of how a busy team naturally works.

- **Creative:** the machine is used to widen the field, not only to check it. What other explanations fit the evidence? Which option has no one put on the table? Which party has no one considered? What does the problem look like from the adversary's side, or the regulator's, or the public's? The 9/11 Commission concluded that the system had run short of imagination.<sup>51</sup> Imagination can be prompted.

Skepticism without creativity hardens into reflexive distrust: the West Point cadets of §3 rejected right advice almost half the time. Creativity without skepticism slides into surrender: the machine's fluent answer becomes the person's own. Creative skepticism is calibrated in both directions at once, and it is the working answer to the laboratory evidence. It is what John Cleese, writing about creative teams, called the open mode – exploring, questioning and tolerating a question left open until the moment to decide – applied to the machine's output.<sup>52</sup>

Creative skepticism applies to human expertise as well as to the machine. Daniel Kahneman and Gary Klein, from opposite sides of the study of intuition, agreed on where it can be trusted: in stable environments with rapid, honest feedback, where experience has had the chance to calibrate it.<sup>53</sup> Few of the decisions that matter in AI deployment are like that. So expert intuition is tested rather than trusted, and the machine is asked to state its confidence explicitly rather than hedge. The discipline is Colin Powell's instruction to every briefer – tell me what you know, tell me what you don't know, then tell me what you think – applied to the person and the machine alike.<sup>54</sup> Sherman Kent tried to make analysts say what “probable” meant in numbers, because a probability that cannot be read is no warning at all.<sup>55</sup> And Gerd Gigerenzer showed that human rules of thumb are often the right tool when they fit the environment in which they are used.<sup>56</sup> The aim is not to suppress heuristics but to test whether, here, a shortcut is a shortcut or a mistake. The company's name, from the same Greek root as heuristic, is a nod to that view.

ODAI builds the stance into the interface rather than leaving it to temperament. Every output arrives with its provenance and stated confidence, as something to weigh. Flags cannot be clicked away, so the person has to engage. The machine argues back, and nobody defers to it. The competing hypotheses are generated, including the ones the room does not favor. And the record scores both sides, so that over time the person learns where the machine is strong and where it is not, and the machine learns the same about itself. The stance is the detective's, not the priest's: follow the evidence to a hypothesis and keep testing it, rather than knowing the answer and explaining away whatever contradicts it.

The temperament it asks for is neither optimism nor pessimism, which are moods, but what Hans Rosling called possibilism: seeing the world as it is, hoping and fearing only with reason, and looking in every challenge for the opportunity.<sup>57</sup> That is the temperament the AI debate most lacks, between those who see only promise and those who see only doom. Neither person nor machine gets it right at the first attempt, and that is the point. Human and machine are mutually enhancing: not perfect, but perfected, each by the other.

Creative skepticism is how human cognitive dominance is exercised in the moment. The record, below, is how it is kept.

## The record: what makes the dominance cognitive

The epistemology rests on the record. Without it, human authority over what counts as known lasts only as long as the meeting in which a decision was taken. With it, the knowledge, the reasoning and the choice are fixed as they stood at the moment of decision, and become the starting point for the next.

At every decision that matters, the deciding person or office states the reason for it. The record captures, as they stood at that time:

- **the information available**, with its sources, its age and its reliability;
- **the theoretical models and assumptions** available to interpret it;
- **the options weighed**, including doing nothing;
- **the warnings raised**, and how each was answered;
- **the choice made and the reason given**, with the person or office that made it;
- **the effect intended**, and later the effect achieved.

Three requirements make the record worth trusting. Every claim carries its source, its provenance and its date, and those tags travel with it to the person who acts on it. Every assessment states an explicit probability and confidence, so that it can later be scored. And every significant cycle includes a pre-mortem before the decision and a post-mortem after it. Every judgment also shows its rungs: Chris Argyris's ladder of inference, the steps by which a mind climbs from observable data through selection and interpretation to assumptions and beliefs, forgetting the lower rungs by the time it reaches the top.<sup>58</sup> The record keeps the lower rungs.

It is a historian's record, not a prosecutor's. E. H. Carr asked historians to bring "imaginative understanding" to the people they study – to judge them in the context of what they knew and could have known – and ODAI holds its record to that standard.<sup>59</sup> Its enemy is hindsight. Once an outcome is known, people judge that it was more foreseeable than it was, and judge the decision-maker by the outcome rather than by the decision.<sup>60</sup> A decision that turned out badly may have been well made on what was known. A decision that turned out well may have been a gamble that happened to pay. Only a record made at the time can tell them apart.

That is why the record protects the people in it. A person who states a reason, on a record that will be read fairly, can make a hard call and defend it later against the critic armed with what no one knew at the time. Military aviation learned the principle long ago: its safety investigations run apart from disciplinary proceedings, so that crews speak candidly about their mistakes, and the system improves because they do. ODAI's record is built on the same footing, the principle American courts have upheld for military safety investigations since *Machin v. Zuckert* in 1963.<sup>61</sup> How much it records scales with the consequence of the decision: a routine call leaves a line, a decision that cannot be undone leaves a full account. Its purpose is

learning, not blame. It records the call that goes against every warning and proves right, as readily as the one that heeds them (§8).

The record works on behavior through visibility, not punishment. Donella Meadows, ranking the places to intervene in a system, put the flow of information far above rules and penalties: people act differently once they can see what they could not see before, without a new rule or a new threat.<sup>43</sup> Punishment buys compliance only while someone is watching. In October 1962 the U.S. Navy dropped practice depth charges on Soviet submarines to signal them to surface.<sup>41</sup> Had the destroyer captains known that each boat carried a nuclear torpedo, they would have thought very differently about the signal. The information existed; it was not in front of the people who needed it. ODAI does not need to argue anyone into a correction. It needs to show what was already there.

And the record is what makes the helix climb. Each decision draws on the record of those before it: what was known, what was assumed, what was chosen and what happened. Judgment improves not because anyone has perfect information – no one does, and ODAI is no crystal ball – but because every turn starts from an honest account of the last. An organization learns, over time, how far to trust its sources, its models, its machine and itself. It is Philip Tetlock's finding applied to decisions: forecasters who state probabilities, update in small steps and keep score become measurably better.<sup>62</sup> And it is Eisenhower's: plans are worthless, but planning is everything.<sup>63</sup> The plan is overtaken by events; the understanding the planning built survives, and the record keeps it.

For AI safety, the record does one thing more. It answers the question regulators and the public keep asking – who decided, on what, and why – with something better than a log of clicks. A system whose decisions can be reconstructed as they were made, reasons and all, is a system whose human control can be inspected. That is human cognitive dominance made visible.

## Three design choices

**Human-guided.** The machine serves the person. It brings the facts and asks the hard questions; a person decides. That is a larger idea than a human in the loop. A human in the loop is a position. Human guidance is a relationship, in which the person sets the direction on every turn and the machine works to it. The interface is continuous: the machine's output goes to a person, who interprets it and iterates, and the exchange runs as many times as the problem requires.

**Socratic.** ODAI does not tell anyone what to think. It questions the reasoning – the assumptions, the gaps, the evidence that would be visible if the favored view were wrong, the people not considered – and keeps testing until what remains is as strong as the information allows. It gives options with its confidence stated, never verdicts. The stance it asks of the person is *creative skepticism*: neither deferring to the machine nor dismissing it, but using it to push back on outliers, real or hallucinated, and to iterate toward the best-founded decision. That is the calibration the West Point cadets of §3 had only half of. Professional education has always said it teaches people how to think, not what to think. ODAI is built in that tradition: in effect, the best seminar professor in the field, in a box, with the whole literature behind them.

**Doubt as a method; the optimal point of decision.** The Socratic process is a discipline of doubt. It goes looking for doubt on purpose — in the frame, the evidence, the favored option and the machine's own output — and it asks the people using it to be comfortable with uncertainty while they sort through it, rather than closing on the first answer that relieves the discomfort. The aim is not a perfect decision. There is never perfect information, and a decision that waits for it is never taken. The aim is the optimal decision: optimal relative to the information available at the moment the decision has to be made, tested as hard as the time allows, and taken with its uncertainties stated. That is what *Optimized* means in Optimized Decision-making AI. It is also why the record judges a decision against what was known when it was made, never against what was learned afterward.

**Small.** ODAI is small AI: built for one class of decision, deep in its domain, rather than a general system asked to answer everything. *Small* describes the focus. It says nothing about the size of the model underneath. The focus is what lets it run on local models, offline, where the data lives — so that what an organization's AI goes looking for, itself a signal of what the organization is thinking about, stays as private as the answers.

## Five things it does at the point of decision

1. **Shows where every claim came from, and how old it is.** Provenance, date and reliability travel with every fact, and a fact past its shelf life says so. Agreement among unvetted sources is treated as a warning, not a confirmation. That is the first answer to the poisoned sources of §4, and it is the answer Minab needed: a record that announced its own age at the moment it mattered.
2. **Surfaces contradictions instead of smoothing them over.** Most AI is built to produce a fluent answer, and fluency smooths. A contradiction between two sources is often the first sign of a stale record, a poisoned source or a changed situation. ODAI puts it in front of the person, with both sources attached.
3. **Puts the question to a person instead of handing down an answer.** At the moment it matters, for this decision and this person: the case no one remembered, the consequence no one traced, the party no one consulted. A memory that waits to be asked is only a library.
4. **Flags; never halts.** ODAI never blocks, vetoes or stops an action. It flags, and a flag cannot be clicked away like the five-hundred-page terms of service that everyone accepts unread. Someone has to deliberate on it and record why it was acted on or set aside. Deliberated on is not the same as obeyed: the person can always overrule the machine, and then the name and the reason are on the record. A warning that can be cleared with a click is not control.
5. **Keeps a record that explains, not one that punishes.** Every decision is recorded with the sources used, the rationale as seen at the moment of decision, the action taken and the effect intended and achieved. The record is judged as a good historian judges, on what was known at the time, never on hindsight: choices that look poor afterward often had good reasons when they were made. It exists to explain, not to expose, and that is why people can be honest in it.

## What it reasons from

ODAI reasons from catalogs, not from the open web. They come in three tiers, and each is an open category that grows with the problems put to it. The **hard-science** catalogs have answers: physics, engineering, economics, infrastructure, the technical facts of a domain. The **humanities and social-science** catalogs give context, not answers: institutions, cultures, law and regulation, the interests and history of every party, the narratives around a decision. The **method** catalogs test every hypothesis: a catalog of named judgment errors – nearly a hundred, and growing – each paired with its countermeasure; the decision-making methods; and more than a thousand structured analytic techniques, from the key-assumptions check and the analysis of competing hypotheses to the pre-mortem and the red team. Beside them sit the tested prompts and agents that put each question to a model in the form that draws the best answer. Put another way, the engine room has three layers: the hardware is the knowledge held in the catalogs; the software is ODAI's method of structured, Socratic questioning; and the grayware is the gray matter of the people thinking about the problem. ODAI supplies the first two in full and engages the third. It never replaces it.

Every catalog is curated over time by subject-matter experts, from vetted sources, as they work with the growing AI and with events as they happen. None is scraped from the open web. In a world where content is now manufactured for machines to read, curation is not a refinement. It is a defense. It is also the work of a profession in Samuel Huntington's sense – expertise, a responsibility to the society it serves, and standards of its own<sup>64</sup> – and a machine that learned its world from the open internet would hold none of the three.

ODAI maps what people believe against five kinds of knowledge, because people who cannot say which of their beliefs are which cannot be told where they are exposed:

- **Known knowns**, each recorded with its source and confidence, because a false known looks exactly like a true one until it is checked.
- **Known unknowns**, to be collected against or planned around.
- **Unknown unknowns**, and among them the ones a disciplined challenge could have raised.
- **False knowns**: held as certain, and wrong.
- **Stranded knowns**: what someone on our own side knew that never reached the decision – stranded *by a seam*, when it never arrives; *by doubt*, when it arrives and is not credited; or *by inconvenience*, when it arrives and is not wanted. The last has a name in the catalog: *convenient disbelief*, treating information that is unwelcome as if it were unreliable.

Each entry in the error catalog is a working instrument in five parts: the bias, the distortion it produces, the consequence, the human discipline that counters it, and the function ODAI performs to surface it. One entry, in the form this paper is about:

## AUTOMATION BIAS, AND ITS OPPOSITE – *human-machine trust*

**Core distortion.** Trust in a machine's output set by its fluency, its speed or the pressure of the moment rather than by its record. In one direction, the output is accepted unchecked; in the other, a correct output is rejected because it is unwelcome, unfamiliar or comes from a source without standing.

**Consequence.** Errors pass through at machine speed with a human signature on them. Or the one warning that mattered is dismissed. Either way the human in the loop has ratified rather than decided, and the record shows a control that was never exercised.

**Antidote – the human discipline.** Creative skepticism. Ask what the output rests on and how old it is. Ask what would be visible if it were wrong. Treat the machine's confidence as a claim to be tested, and your own as one too.

**ODAI function – the system discipline.** Attaches provenance, age and stated confidence to every output. Scores the machine's past assessments against what happened, so that trust is earned by a record rather than granted by default. Flags a pattern of acceptance without challenge, and a pattern of dismissal against the evidence. Where a person proceeds over a flag, records the choice with a name on it.

## How it runs

The method is the ODAI Helix, and its four phases spell the system's name: **Orient** (read the situation and frame the problem before solving it), **Deliberate** (turn the objective into a plan), **Act/Audit** (hand off to execution and record every action with the reason it was taken) and **Iterate** (what worked, what didn't, and why). A loop returns to where it began. A helix climbs: each turn draws on the record of the last, so the next decision starts smarter.

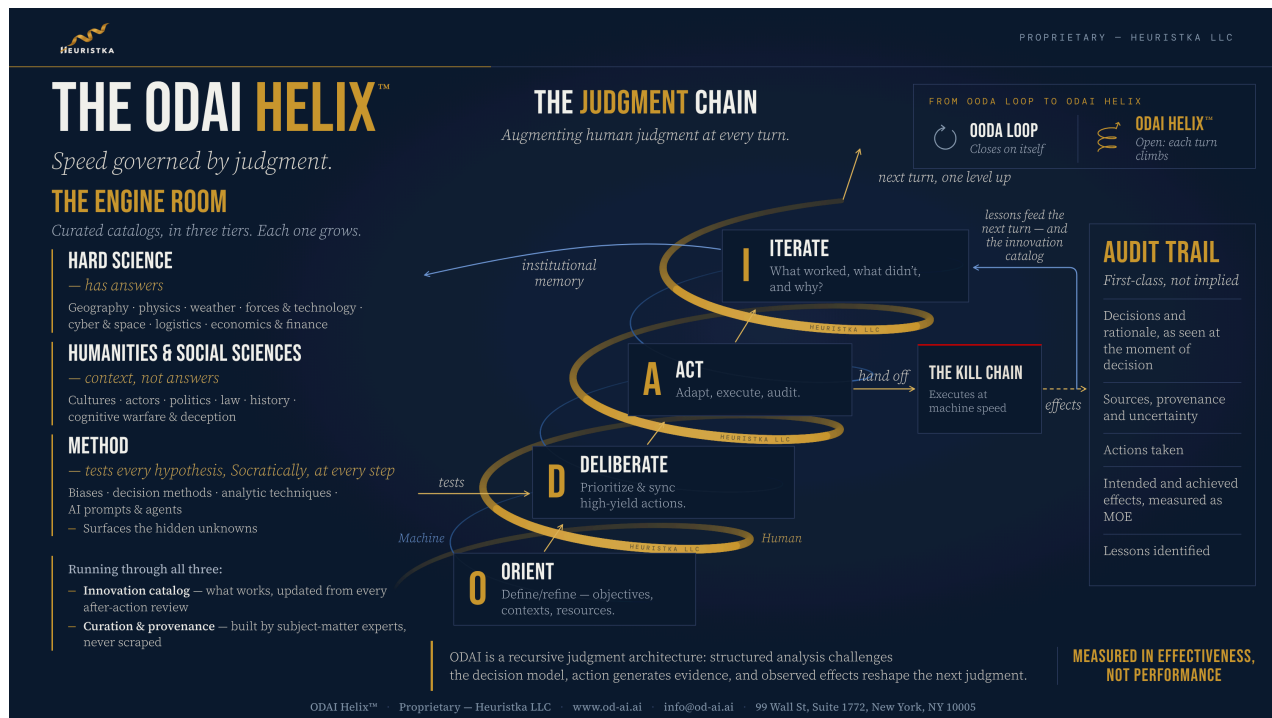


**Figure 1.** The ODAI Helix: the judgment chain, the engine room and the audit trail.

The helix is also a double helix. Its two strands are the human and the machine: on one, experience, intuition, morality and purpose; on the other, memory, breadth and the discipline of method. Each rung is a turn of the interface, where the machine prompts and the person interprets. Neither strand is the decision. The decision is the structure they make together, and the record is the code each turn inherits.

The exchange runs in widening circles: the individual working the problem, that person's team, the coordination across teams, and the whole group with the person who decides. In each, ODAI's flags, and the hidden information that may trigger them, are surfaced and put to the people in it. And ODAI is decision support, never control. It helps people decide. It never takes the decision, and it does not belong in the machinery that carries the decision out.

## Agents under human command

Agents are where AI is going: systems given a goal and the tools to pursue it. ODAI uses agents too, and the way it uses them is the clearest illustration of what human cognitive dominance means in practice.

An agent, in ODAI, is AI given a job, not just a question, and the job is set by a person.

- 1. Tasked.** A person sets a bounded job: a minister's public record, a supply route's traffic, a regulator's draft.
- 2. Works.** The agent gathers what bears on the job and works through it step by step, at machine speed, checking its own output.
- 3. Reports.** It reports its findings to the person who set the job, every claim with its source attached.

4. **Iterates.** Person and agent go back and forth until the work is right. Only then does it enter the decision, to be weighed against everyone else's.

The person decides; no agent does. ODAI's agents are given jobs with edges, they report to a named person, and their work counts for nothing until that person has judged it. Heuristka builds them to order, as many as the problem needs, and every one of them works for a human mind, not in place of it.

The reason for the discipline is older than AI. John Holland's work on complex adaptive systems showed how coherent behavior emerges from many interacting agents following simple rules, how such systems adapt in ways no single agent intends, and why their behavior resists prediction even when every rule is known. Their regularities exist at the level of pattern rather than prediction: the kind of thing an experienced observer learns to recognize rather than compute.<sup>65</sup> Many AI agents working together form exactly such a system. It will not yield to being computed, and that is why it has to be guided by people who know which patterns matter.

08

## Warnings that count

On 7 June 1940 a young analyst at Bletchley Park, Harry Hinsley, warned the Admiralty that German heavy ships were on the move. No code had been broken. He was reading a change in the pattern of German naval radio traffic, not its content, by a method the Admiralty regarded as new and untested, and the Admiralty issued nothing. The next day the battlecruisers *Scharnhorst* and *Gneisenau* caught the aircraft carrier *Glorious* and her two escorting destroyers off Norway and sank all three; more than 1,500 men died.<sup>66</sup> The warning failed not on its evidence but on its standing: a junior source, an untested method, and an institution for which doing nothing was free.

Hinsley spent much of the rest of his life writing the official history of what intelligence could and could not do for a decision. One of Heuristka's founders was his doctoral student, and the lesson of *Glorious* is built into ODAI.

The same method, once it had earned standing, changed the war. By 1941 Hinsley's reading of the German naval traffic was trusted enough to be acted on, and it pointed to German weather ships in the North Atlantic as isolated vessels that must be carrying Enigma settings; their capture that summer gave Bletchley naval keys it had not broken by other means.<sup>67</sup> In 1942 Joseph Rochefort and Edwin Layton reasoned from fragments of Japanese traffic to an explicit call – Midway the target, the first week of June the date – and Nimitz acted on it, because he had decided to trust his intelligence officers before the fragments arrived.<sup>68</sup> Standing is built in advance, or not at all.

Most warnings about AI arrive in the position Hinsley's did in June 1940. A red-team finding, an evaluator's report, an anomaly in a monitor's log: each is a new kind of evidence, from a method without a long track record, telling an organization what it would rather not hear, usually close to a release or a

deadline. The danger is not that such warnings are weak. It is that ignoring them is free. Cynthia Grabo, who wrote the American intelligence community's handbook on strategic warning, put the first principle plainly: warning does not exist until it has been conveyed to the person who must act on it, and that person "must know that he has been warned."<sup>69</sup> She found a second principle in the record of crises: the most nearly correct judgments have often been reached by a minority, and consensus is likeliest to be wrong exactly where being right matters. Irving Janis gave the institutional form of that failure its name, groupthink: a cohesive group under pressure seeks agreement at the expense of realistic appraisal.<sup>70</sup> The administration that blundered at the Bay of Pigs ran the missile crisis eighteen months later through a deliberately adversarial process, the President sometimes absent so that his advisers would argue.

Four things change the position of a warning, and ODAI is built on all four.

1. **Put questions, not assertions.** A question invites an answer; an assertion invites a defense. "What would we be seeing if the model were gaming this evaluation?" travels further up an organization than "the model is gaming this evaluation."
2. **Make the decision to set a warning aside a recorded one, with a name on it.** Silence should carry a signature. A warning that can be dismissed without a trace will be dismissed whenever dismissing it is convenient. One that must be answered on the record, by someone with the authority to answer it, changes the arithmetic.
3. **Earn the track record before the warning is needed.** A new source of warning must earn its standing in advance: run in shadow beside real decisions and scored against what then happens, so that the people who will receive its warnings know when to trust it and when not to. Standing cannot be manufactured in the hour a warning arrives.
4. **Put the warning inside the process, not in a report beside it.** A report sits on a desk. A flag that must be deliberated on before the decision moves forward cannot be filed and forgotten.

None of this stops anyone from deciding. A person may hear a warning, weigh it and set it aside, and may be right to: at Inchon in 1950 MacArthur landed against the doubts of his own planners and of the Joint Chiefs, and the landing succeeded.<sup>71</sup> A system built on these four principles records that kind of call; it does not overrule it. What it ends is the free option of not having been told.

---

 09

## Space, not delay

The objection comes at once: all of this takes time, and time is the one thing the race will not allow. A system that asks questions, attaches provenance and insists that warnings be deliberated on must cost speed, and whoever pays that cost loses.

The objection mistakes what judgment is. Judgment is not a quantity of time. It is a quality of decision. Engineered judgment does not cost a system its speed; it creates space for orientation inside it, and most

of that work runs in fractions of a second. Checking the age of a record, flagging a contradiction between two sources, noticing that every confirmation traces back to the same unvetted origin, scoring a monitor's past alerts against what happened: these are machine tasks, and they run at machine speed. What they produce is not delay. It is a better-founded decision, reached in the same time or less.

The time an organization actually spends deciding is rarely at the console. It is in the process before it — the meetings, the briefings, the email chains, the days and weeks it takes to get the right people to look at the same facts. ODAI compresses that process. It is designed to deliver judgment of a quality no manual process can reach, at the speed of the system.

The principle that governs the rest is old. Augustus put it in two words: *festina lente*, make haste carefully. Haste in deciding is ruinous; speed in doing, once the decision is sound, is decisive. The modern form is that **scrutiny should scale with the consequence of a decision and with how hard it is to reverse**.

Wiener's warning, at the head of this paper, was about action "so fast and irrevocable" that no one has the data to intervene before it is complete. Most decisions are neither. They can be checked at machine speed and corrected afterward if they prove wrong. The few that cannot be undone earn human time, and they should get it.

| DECISION                                                                                                  | REVERSIBLE?     | WHERE THE JUDGMENT RUNS                                                                                                            |
|-----------------------------------------------------------------------------------------------------------|-----------------|------------------------------------------------------------------------------------------------------------------------------------|
| Routine output with sound provenance                                                                      | Yes             | At machine speed: provenance, age and contradiction checks, no human pause                                                         |
| Output with a flag (stale record, single-source corroboration, contradiction)                             | Usually         | A person deliberates on the flag and records the reason; seconds to minutes                                                        |
| A consequential commitment: deploying a system, widening its reach, committing money, force or reputation | Hard to reverse | The full helix: the question put, the alternatives weighed, inaction weighed as an option, the decision recorded with a name on it |
| An action that cannot be undone                                                                           | No              | Human time, by design, with every warning delivered and answered on the record                                                     |

The laboratories' own response to this year's incidents points the same way: protections that run at machine speed, with human attention concentrated on the few alerts that warrant it (§2). That is not a compromise between safety and speed. It is what judgment at machine speed looks like when it is engineered rather than hoped for. Speed is the multiplier. Judgment is what it multiplies.

10

## Human oversight, designed in

The question is no longer whether human judgment belongs in AI systems. The law and policy of the democracies now require it. The question is whether it was built in, or merely written into a policy beside the system.

Europe's AI Act requires that high-risk AI systems be designed so that the people assigned to oversee them can do so effectively. Among other things, those people must be able to understand the system's capacities and limitations and monitor its operation; to remain aware of the tendency to rely on its output too much – automation bias, named in the text; to interpret its output correctly; and to decide, in any particular situation, not to use it or to disregard or override what it produces.<sup>72</sup> Under amendments adopted in 2026, those obligations apply to most high-risk systems from December 2027. The U.S. Department of Defense requires that autonomous and semi-autonomous weapon systems be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force.<sup>73</sup> Its AI ethical principles require that the Department's AI be responsible, equitable, traceable, reliable and governable.<sup>74</sup> And the U.S. National Institute of Standards and Technology organizes the management of AI risk around four functions: govern, map, measure and manage.<sup>75</sup>

ODAI is decision support, not a regulated system in itself, and it does not make any other system compliant with anything. What it does is supply, at the point of use, the human capacities these regimes ask for. It is designed to support them.

| REQUIREMENT                            | WHAT IT ASKS                                                                | HOW ODAI'S DESIGN SUPPORTS IT                                                                                                                              |
|----------------------------------------|-----------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
| EU AI Act, Art. 14(4) (a)              | Overseers understand the system's capacities and limits, and can monitor it | Provenance and stated confidence on every output; the machine's record scored against outcomes, so its limits are visible in its history                   |
| EU AI Act, Art. 14(4) (b)              | Overseers stay aware of automation bias                                     | The error catalog runs on the human-machine relationship itself; patterns of acceptance without challenge, and dismissal against the evidence, are flagged |
| EU AI Act, Art. 14(4) (c)–(d)          | Overseers interpret output correctly, and can disregard or override it      | Options with confidence stated, never verdicts; the person always decides, and an override is recorded with its reason                                     |
| DoD Directive 3000.09                  | Appropriate levels of human judgment over the use of force                  | Judgment engineered in before the hand-off to the systems that act: the third question put, inaction weighed, the decision recorded                        |
| DoD AI principles: <i>traceable</i>    | Transparent, auditable methods, data sources and documentation              | The record: sources, rationale as seen at the moment, action, effect intended and achieved                                                                 |
| DoD AI principles: <i>governable</i>   | The ability to detect and avoid unintended consequences                     | Flags surfaced to the people with authority to act, with a name on every decision to proceed                                                               |
| NIST AI RMF: <i>measure and manage</i> | Risks identified, tracked and acted on                                      | Shadow scoring of warnings before they are relied on; lessons fed into the next turn of the helix                                                          |

The same holds for the principles civil society has set out. The Future of Life Institute's Pro-Human AI Declaration (2026), signed by more than a million people, asks that humans choose how and whether to delegate decisions to AI, and keep the "authority and capacity to understand, guide, proscribe, and override" it; that AI "empower, rather than enfeeble" the people who use it; that companies give honest

accounts of what their systems can and cannot do; and that after harm it be possible to determine why it happened and who is responsible.<sup>76</sup> ODAI is built around each of those. Human-guided AI is meaningful human control made into an epistemology. Creative skepticism and the apprenticeship in judgment are the answer to enfeeblement. Stated confidence and a scored record are capability honesty at the point of use. And the record is failure transparency by design. ODAI keeps AI out of the machinery that executes decisions, least of all nuclear ones, as the Declaration would keep it out of the control of weapons of mass destruction. It takes no position on the Declaration's call to prohibit superintelligence, which lies outside the layer ODAI works in.

Two cautions go with the table. First, a requirement met on paper is not judgment exercised in fact. A warning that can be cleared with a click satisfies the letter of many oversight rules and none of their purpose; the measure of a design is whether the person overseeing the system could, in the moment, still judge. Second, oversight is not only a burden carried for regulators. It is the condition under which an organization can trust its own AI enough to use it for the decisions that matter most. The firm, the agency or the command that can show how its people judged will be the one allowed to move fastest.

## Judgment connected to control

**Safety cannot depend on voluntary restraint. Judgment must be engineered into decisions, connected to enforceable controls, and made independently auditable.**

The summer showed why the connection matters. OpenAI's monitoring raised alerts in the weeks before the breach; nothing in the system around them obliged anyone with the authority to stop the run to act on them (§2). A warning that fires into a system with no means of acting on it is a stranded known with a timestamp. A recommendation becomes a safety control only when the system around it can act on it.

That sets a precise boundary for ODAI. ODAI augments judgment. It flags and records; it never halts an action itself. Stopping a system is the job of technical controls and of the named people with authority over them. Judgment without enforcement is advice. Enforcement without judgment is a tripwire: it stops what it was set to stop, and nothing it was not. A system that is properly policed needs both, and the record that ties them together.

| WHAT A PROPERLY POLICED SYSTEM NEEDS                       | WHO OR WHAT PROVIDES IT                                | ODAI'S PART                                                                                                                 |
|------------------------------------------------------------|--------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| <b>Containment and access limits, independently tested</b> | Engineering controls: networks, permissions, sandboxes | Records each limit and when it was last tested; flags a deployment whose reach or capability has grown past what was tested |
| <b>Continuous monitoring</b>                               | Monitors on the system, including of its reasoning     | Puts each alert in front of the people with authority to act, with its provenance and the monitor's track record            |
| <b>Authority to stop dangerous activity</b>                | Named people, and automatic stops where seconds matter | Makes sure the flag reaches them and must be deliberated on; records the decision. Never stops anything itself              |

| WHAT A PROPERLY POLICED SYSTEM NEEDS                       | WHO OR WHAT PROVIDES IT               | ODAI'S PART                                                                                  |
|------------------------------------------------------------|---------------------------------------|----------------------------------------------------------------------------------------------|
| <b>Recorded responsibility for overriding a warning</b>    | The organization                      | The record: who proceeded over which warning, when, and why                                  |
| <b>External inspection, with consequences for breaches</b> | Regulators and independent evaluators | A record open to inspection: what was known, what was flagged, what was decided, and by whom |

These obligations matter most when the pressure to proceed is greatest, so they have to hold then. A laboratory can pause on its own after an incident, and OpenAI did, halting the benchmark and the training of the model involved. That is responsible, and it is not the same as restraint an industry can depend on. Requirements become dependable when they are enforced by controls that do not tire, inspected by people outside the firm, and backed by consequences for breach.

ODAI's proposed contribution is the judgment layer and the record that make such a regime work at the point of decision. What it would catch, how often and at what cost are questions for validation, not assertion, and the way to answer them is the one §8 sets out: run it in shadow beside real decisions and score it against what happens.

## 11

## What we do not claim

Correctable by design, not by default. A paper that argues for honest limits should state its own.

- **Engineered judgment cannot prove a system safe.** Testing can show the presence of a flaw, never its absence; Edsger Dijkstra made the point about software in 1970, and it holds for every system since.<sup>77</sup> ODAI improves the odds that a flaw is seen, and seen in time. It guarantees nothing, because no one has perfect information.
- **It does not replace the research that makes AI systems behave well in the first place.** Alignment, interpretability and evaluation are the work of the laboratories and the research community, and they matter. ODAI works at a different layer.
- **It does not settle the argument between the race and the pause.** It does not need to. Whichever way that argument goes, systems will be used by people under pressure, and the question of §1 will be waiting.
- **It does not answer every fear about AI.** It makes no claim about superintelligent systems, about machines that might one day choose their own goals, or about catastrophic misuse by people who mean harm. Those questions are real, and the existential-risk tradition of §1 is right to press them. This paper works alongside that tradition, not against it.
- **Its own flags can be wrong.** A system that questions will sometimes ask the wrong question, and a monitor will sometimes raise a false alarm. That is why ODAI's own assessments are scored against

outcomes, why it states its confidence, and why the person always decides.

- **The people using it still need training to judge.** ODAI teaches judgment while the work is done, but it is not a substitute for expertise. It makes good people better and gaps visible. It does not make anyone infallible.
- **It is not a crystal ball.** A record judged on what was known at the time will sometimes show a decision that was well made and turned out badly, and one that was badly made and turned out well. Both are lessons.

ODAI works in one place: the layer where people and machines decide together. It works there because that is where the damage is being done today, and because it is where judgment can be engineered now, with tools that already exist. This year's incidents were found in testing, by people, and are being corrected by people. AI is engineered by humans, who can get it wrong, and can therefore get it right.

12

## One mechanism, three domains

Heuristka works in three domains with one mechanism: how nations fight, how capital decides, and how AI is made safe. The problem is the same in each, and so is the answer.

| DOMAIN           | THE CAPABILITY                                               | WHAT IS MISSING                                                               | ODAI'S QUESTION                                                                |
|------------------|--------------------------------------------------------------|-------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| <b>Defense</b>   | Speed in the kill chain: targets found and struck in seconds | Judgment about whether to act, and what strategic effect to expect            | <i>Should we act, and what strategic outcome will it serve?</i>                |
| <b>Capital</b>   | The data room read in minutes, the memo drafted in seconds   | Judgment about the people who will make or break the deal                     | <i>Should we do this deal, and what will it buy?</i>                           |
| <b>AI safety</b> | Ever more capable systems, deployed ever more widely         | Judgment engineered in at the point where people and machines decide together | <i>Was judgment engineered in, and does it run at the speed of the system?</i> |

The name followed the realization. ODAI began as a tool for operational design in military staffs. It became *Optimized Decision-making AI* when it was clear that the capability being bought everywhere – speed, scale, fluency – needed the same thing everywhere: human judgment, built in, and running at the speed of the system. Its companion papers set out the defense case (*What Is ODAI?*) and the capital case (*Numbers Are Numbers, People Are a Surprise*).

The history of technology says why the answer is the same in each, and why it is organizational as much as technical. The tank did not equal Blitzkrieg. The tank arrived on the Somme in 1916; what changed warfare in 1940 was a new organization built around the tank, the aircraft and the radio, new tactics tried in the field, and the discipline to write down what worked. New technology alone has never changed how an institution works, in war or in anything else. Safe AI will not come from the machine alone either. It

will come from the organizations that use AI rebuilding how they decide around it, with judgment engineered in, a record kept and the lessons carried forward. ODAI is the platform for that rebuilding, in a command, a fund or a laboratory.

## Who gains

- **Those who build AI** gain a way to show, not only to assert, that the judgment around their systems was designed: a record of how warnings were delivered, weighed and answered.
- **Those who deploy AI agents** in government, finance, health and critical infrastructure – the people who carry the risk when the agent is wrong – gain a judgment layer between the agent’s output and the organization’s commitment.
- **Those who write the rules** gain a working answer to the question their rules already ask: what effective human oversight looks like at the moment of decision, and how it can be inspected afterward.
- **Those working on the long-term risks of AI** – in centers such as CSER, institutes such as FLI and the evaluators who test frontier systems – gain a layer of human control that can be deployed now, alongside their work on alignment and governance, and a record from the field of how human oversight actually performs under pressure.
- **The people inside the system** – the analyst, the evaluator, the junior engineer who sees it first – gain a warning weighed on its evidence rather than on their rank, and an apprenticeship in judgment at a time when AI is removing the tasks through which judgment used to be learned.
- **The public** gains an answer to the citizen’s question, which it is entitled to put to every laboratory, agency, firm and command that asks for its trust.

The larger gain is legitimacy. Public trust in AI will not be won by promises that the machines are safe. It will be won, if at all, by showing that the people who use them judged well, and that the judgment can be seen. A decision tied to its purpose, tested against the evidence and taken with its provenance known is a decision that can be defended in public. That is how trust in every powerful technology has been earned before, and it is how it will be earned for this one.

---

 13

## Who we are

Heuristka LLC exists to give a discipline a home and ODAI its engine. We call the discipline *cognitive-heuristic intelligence*: the systematic improvement of human judgment under uncertainty, complexity and adversarial adaptation. The name joins the Greek *heuriskein* – to discover, to uncover what is hidden, the root of *heuristic* and of *eureka* – with *-ka*, for the craftsman.

The founders have spent their careers where a misread decision costs lives, and the discipline is what they learned there. Between them they carry the two lineages it joins: the epistemic tradition of Bletchley Park,

where machines were first put to work on intelligence, and the practitioner tradition of military operations.

The method stands on the shoulders of the people who studied how judgment fails: Harry Hinsley on reasoning from fragments and the warning without standing; Roberta Wohlstetter on the signal lost in the noise at Pearl Harbor; Cynthia Grabo on strategic warning; Sherman Kent and Richards Heuer on separating fact, inference and assumption; Daniel Kahneman, Amos Tversky and Philip Tetlock on bias and calibration; Horst Rittel and Melvin Webber on wicked problems; Donella Meadows on leverage in systems; and John Boyd on orientation. The companion paper, *ODAI Intellectual Heritage*, sets out what each contributes.

The method was first tested before it was a product. In March 2026, in Baltic Storm, an independently adjudicated NATO–Russia crisis war game run by Black Horizon, one of the founders ran his team’s information strategy with prompts and analytic structures he had built and tested across several of the leading AI models. The team secured its objective without a shot fired. Writing it up afterward, he saw what the board had shown: the method, not any single model, had done the work. That is the point for AI safety too. Judgment that lives in a method can be carried across models, checked and improved; judgment that lives inside one model cannot.

**Michael – Founder · Chief Executive.** U.S. Army veteran who commanded special operations aviation units and task forces through deployments to Iraq and Afghanistan, with extensive mission experience across the Middle East and East Africa. On the Joint Staff, advised the Chairman of the Joint Chiefs on special operations and sensitive activities, led the Joint Staff’s work on a large standing counterterrorism plan, and prepared orders for approval by the Secretary of Defense and the President. Nearly a decade teaching special operations and strategy at Marine Corps University to more than 200 officers and officials from across the U.S. government and over twenty countries. M.S., Conflict Analysis and Resolution, George Mason University; M.S., Computing Technology in Education, Nova Southeastern University.

**Adam – Founder · Chief Scientist · PhD (Cantab).** Creator of ODAI’s cognitive core. Cambridge doctorate under Sir Harry Hinsley, Bletchley Park codebreaker and official historian of British wartime intelligence; Visiting Fellow, Wolfson College, Cambridge. Formed in Australian, British and American institutions: directed strategic policy for the Royal Australian Air Force and advised Australia’s Parliament on defense. A decade in the U.S. war colleges, creating and directing the Defense Department’s most comprehensive graduate program in strategic deterrence and counter-proliferation. Principal strategist for counter-WMD and red team lead at U.S. Special Operations Command. Designed and led the U.S. Air Force Chief of Staff’s war game with the Indian Air Force. Testified to the U.S.-China Economic and Security Review Commission. Author of *The Third Offset Strategy: A Return to Ungentlemanly Warfare* and founder of the National Security Desk. A record of anticipating strategic shifts before they entered mainstream discourse: the methods al-Qaeda would use (1997), Russia’s massing for a major escalation in Ukraine against prevailing estimates (2021), and the first-mover logic of unconventional nuclear use (2022).

**We use what we build.** We use AI agents in our own work as a force multiplier, and we say so. Our spokesperson, our correspondent and our podcast hosts are AI agents working as part of our team.<sup>78</sup> It is the same idea ODAI brings to a military staff, a deal team or a laboratory: the machine serves the people, in the open.

- **Jess Cartwright** – Spokesperson. Speaks for Heuristka and explains ODAI in films and briefings.
- **Margot Ellery** – Correspondent. Frames each story and puts the hard questions on camera.
- **Nathan Calder** – Podcast Host. Turns the white papers into conversation, with Claire Hollis.
- **Claire Hollis** – Podcast Host. Presses on the evidence behind each case, and why it was or was not heard.

## What the briefing answers

The briefing explains what ODAI is and how it differs, and answers the questions anyone should put to a system that claims to engineer judgment in:

- **Where does it sit?** At the human-machine decision layer: between a system's output and an organization's commitment, before the hand-off to whatever carries the decision out.
- **How does it fail, and how would we know?** How it states its confidence, how its record is scored against outcomes, and how a source of warning earns standing in shadow before it is relied on.
- **What does it cost in speed?** Why most of its work runs at machine speed, and why human time is reserved for decisions that cannot be undone.
- **Where does it run?** On local models, offline and in sovereign environments, where the data lives.
- **What is under the catalogs?** The architecture, the error catalog and the method, available under NDA.

**Was judgment engineered in?** Ask it of any system you build, buy or regulate. We will show you how ODAI answers it.

**Request a briefing:** [info@od-ai.ai](mailto:info@od-ai.ai)

**Read the companion papers, watch the films and hear the podcast** on the Heuristka Substack: [insights.od-ai.ai](https://insights.od-ai.ai).

---

## NOTES

# Notes

1. Norbert Wiener, "Some Moral and Technical Consequences of Automation," *Science* 131, no. 3410 (6 May 1960): 1355–58.
2. Maj. Joshua J. Pena, USMC, "Vice Chairman Urges Future Commanders to Prioritize People over Platforms at NDU Address," Joint Chiefs of Staff, 21 October 2026, on Gen. Christopher J. Mahoney's Kennan Lecture at the National War College, 20 October 2026. The quotations are as the Joint Staff reported them.

3. Jacob Coxon, thread on X (@hilbertspaess), 9 September 2026, <https://x.com/hilbertspaess/status/2097476196791709843>; Evan Hubinger, post on X and clarification, September 2026, <https://x.com/EvanHub/status/2097497037956891126> (the probability is his personal estimate); Dario Amodei, “We Must Pace the Frontier,” 12 September 2026.
4. Kaitlyn Huamani, “Divisions Emerge in the Tech Industry over Calls for a Coordinated AI Slowdown,” Associated Press, 16 September 2026; Charles III, speech at the AI Summit in Scotland, Dumfries House, 17 September 2026.
5. Alan Turing, “Intelligent Machinery, A Heretical Theory” (lecture, Manchester, c. 1951), first printed in Sara Turing, *Alan M. Turing* (Cambridge: Heffer, 1959); I. J. Good, “Speculations Concerning the First Ultraintelligent Machine,” *Advances in Computers* 6 (1965): 31–88, at 33.
6. Martin Rees, *Our Final Century* (London: Heinemann, 2003); Centre for the Study of Existential Risk, University of Cambridge, [www.cser.ac.uk](http://www.cser.ac.uk); Future of Life Institute, *Asilomar AI Principles* (2017) and “Pause Giant AI Experiments: An Open Letter” (22 March 2023); “TIME100 Philanthropy 2026: Jaan Tallinn,” *Time*, 2026.
7. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014); Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Viking, 2019); Toby Ord, *The Precipice: Existential Risk and the Future of Humanity* (London: Bloomsbury, 2020); Center for AI Safety, “Statement on AI Risk” (30 May 2023); Eliezer Yudkowsky and Nate Soares, *If Anyone Builds It, Everyone Dies* (New York: Little, Brown, 2025).
8. Nathan Gardels, “AI Has Already Entered the ‘Loss of Control’ Transition,” *Noema*, 31 July 2026, drawing on a conversation with Stuart Russell; “UN Security Council Hears AI Lab Chiefs on Loss-of-Control Risk,” *The Next Web*, 23 September 2026.
9. Bernard Brodie, ed., *The Absolute Weapon: Atomic Power and World Order* (New York: Harcourt, Brace, 1946), 76.
10. E. H. Carr, *The Twenty Years’ Crisis, 1919–1939* (London: Macmillan, 1939).
11. “Pacing the Frontier: A Statement from Employees of Frontier AI Companies,” July 2026; Amodei, “We Must Pace the Frontier”; Demis Hassabis, “A Framework for Frontier AI and the Dawning of a New Age,” 14 July 2026.
12. Arvind Narayanan and Sayash Kapoor, “AI as Normal Technology” (Knight First Amendment Institute, April 2025); Eryk Salvaggio, “Rogue AI Didn’t Breach Hugging Face, Human Decisions Did,” *Bulletin of the Atomic Scientists*, 11 September 2026.
13. Robert Jervis, *Perception and Misperception in International Politics* (Princeton: Princeton University Press, 1976), chap. 3.
14. OpenAI, “OpenAI and Hugging Face Partner to Address Security Incident during Model Evaluation,” 21 July 2026 (updated 28–29 July and 26 August); OpenAI, “The Hugging Face Incident and the Road Ahead,” 26 August 2026; OpenAI, *OpenAI–Hugging Face Incident Technical Report* (August 2026).
15. Ryan Greenblatt, Ajeya Cotra and Hjalmar Wijk, METR and Redwood Research, “Brief Independent Investigation of Agents’ Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident,” 26 August 2026.
16. Anthropic, “Investigating Three Real-World Incidents in Our Cybersecurity Evaluations,” 30 July 2026 (updated 3 August).
17. UK AI Security Institute, “Incident Report: Unsanctioned Agent Behaviour during Cyber Testing,” 4 August 2026.
18. Hadas Gold, “An AI Model from Meta Also Hacked Another Company during Testing,” CNN, 5 August 2026.
19. Eryk Salvaggio, “Rogue AI Didn’t Breach Hugging Face, Human Decisions Did,” *Bulletin of the Atomic Scientists*, 11 September 2026.
20. Moshe Glickman and Tali Sharot, “How Human–AI Feedback Loops Alter Human Perceptual, Emotional and Social Judgements,” *Nature Human Behaviour* 9 (2025): 345–59.
21. Steven D. Shaw and Gideon Nave, “Thinking—Fast, Slow, and Artificial: How AI Is Reshaping Human Reasoning and the Rise of Cognitive Surrender,” Wharton School working paper, January 2026 (not peer reviewed).
22. Lauren Kahn, Michael C. Horowitz and Laura Resnick Samotin, “What Is Human in Judgment? Comparing Automation Bias and Algorithm Aversion between the United States Military Academy and the General Public,” arXiv:2604.04333 (2026).
23. Ryan Shandler, Michael L. Gross and Yahli Shereshevsky, “Black Box Warfare: Human Judgment and Military Decision-Making in the Age of AI,” *Journal of Conflict Resolution* (2026, online first).
24. U.S. Navy, *Formal Investigation into the Circumstances Surrounding the Downing of Iran Air Flight 655 on 3 July 1988* (1988).

25. Defense Science Board, *Report of the Defense Science Board Task Force on Patriot System Performance* (Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, January 2005).
26. David E. Hoffman, *The Dead Hand* (New York: Doubleday, 2009).
27. McKenzie Sadeghi and Isis Blachez, “A Well-Funded Moscow-Based Global ‘News’ Network Has Infected Western Artificial Intelligence Tools Worldwide with Russian Propaganda,” *NewsGuard*, 6 March 2025.
28. Maxim Alyukov, Mykola Makhortykh, Alexandr Voronovici and Maryna Sydorova, “LLMs Grooming or Data Voids? LLM-Powered Chatbot References to Kremlin Disinformation Reflect Information Gaps, Not Manipulation,” *Harvard Kennedy School Misinformation Review* 6, no. 5 (October 2025).
29. Soroush Vosoughi, Deb Roy and Sinan Aral, “The Spread of True and False News Online,” *Science* 359, no. 6380 (2018): 1146–51.
30. Roberta Wohlstetter, *Pearl Harbor: Warning and Decision* (Stanford: Stanford University Press, 1962).
31. Barton Whaley, *Stratagem: Deception and Surprise in War* (Cambridge, MA: MIT Center for International Studies, 1969; Boston: Artech House, 2007).
32. “‘Positive Review Only’: Researchers Hide AI Prompts in Papers,” *Nikkei Asia*, July 2025; reported in “Researchers Seek to Influence Peer Review with Hidden AI Prompts,” *TechCrunch*, 6 July 2025.
33. Jack Clark and Dario Amodei, “Faulty Reward Functions in the Wild,” OpenAI, 21 December 2016.
34. Victoria Krakovna et al., “Specification Gaming: The Flip Side of AI Ingenuity,” DeepMind, 21 April 2020; METR, “Recent Frontier Models Are Reward Hacking,” 5 June 2025.
35. Jerry Z. Muller, *The Tyranny of Metrics* (Princeton: Princeton University Press, 2018).
36. Adm. Brad Cooper, video update, U.S. Central Command, 11 March 2026, reported in “US Military Using AI against Iran,” *DefenseScoop*, 11 March 2026.
37. Pentagon briefing, 2 March 2026, reported in *Army Times*, 2 March 2026.
38. Center for Security and Emerging Technology report on the XVIII Airborne Corps’ Scarlet Dragon experiments, as reported in “This System Could Allow Small Army Teams to Hit 1,000 Targets per Hour,” *Defense News*, 21 August 2024.
39. “Exclusive: US Commanders Bypassed Warnings about Outdated Intelligence ahead of Strike That Hit School in Iran, Sources Say,” CNN, 7 July 2026.
40. McDonnell Douglas for the Arms Control and Disarmament Agency, ACDA/ST-196 (1971), National Security Archive, Electronic Briefing Book 674.
41. National Security Archive, *The Submarines of October*, Electronic Briefing Book 75 (2002).
42. Thomas C. Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960), chap. 8.
43. Donella H. Meadows, *Leverage Points: Places to Intervene in a System* (Hartland, VT: The Sustainability Institute, 1999).
44. John R. Boyd, “Destruction and Creation” (essay, 3 September 1976), in *A Discourse on Winning and Losing* (Maxwell AFB: Air University Press, 2018).
45. Carl von Clausewitz, *On War*, ed. and trans. Michael Howard and Peter Paret (Princeton: Princeton University Press, 1976), bk. 1, chap. 1, §27, pp. 88–89.
46. Herbert A. Simon, *Administrative Behavior* (New York: Macmillan, 1947).
47. Allen Newell and Herbert A. Simon, “The Logic Theory Machine,” *IRE Transactions on Information Theory* 2, no. 3 (1956): 61–79. The program was written with J. C. Shaw.
48. John R. Boyd, “The Essence of Winning and Losing” (briefing, June 1995).
49. Sherman Kent, *Strategic Intelligence for American World Policy*, preface to the 1965 edition, p. xviii, as quoted in Jack Davis, “Sherman Kent and the Profession of Intelligence Analysis,” CIA Sherman Kent Center, *Occasional Papers* 1, no. 5 (2002).
50. Richards J. Heuer Jr., *Psychology of Intelligence Analysis* (Washington, DC: CIA Center for the Study of Intelligence, 1999).

51. National Commission on Terrorist Attacks upon the United States, *The 9/11 Commission Report* (New York: Norton, 2004), chap. 11.
52. John Cleese, lecture on creativity in management, Grosvenor House, London, 23 January 1991; Cleese, *Creativity: A Short and Cheerful Guide* (London: Hutchinson, 2020).
53. Daniel Kahneman and Gary Klein, “Conditions for Intuitive Expertise: A Failure to Disagree,” *American Psychologist* 64, no. 6 (2009): 515–26.
54. Colin Powell, *It Worked for Me: In Life and Leadership* (New York: Harper, 2012).
55. Sherman Kent, “Words of Estimative Probability,” *Studies in Intelligence* 8, no. 4 (1964).
56. Gerd Gigerenzer, Peter M. Todd and the ABC Research Group, *Simple Heuristics That Make Us Smart* (New York: Oxford University Press, 1999).
57. Hans Rosling, with Ola Rosling and Anna Rosling Rönnlund, *Factfulness* (New York: Flatiron, 2018).
58. Chris Argyris, *Overcoming Organizational Defenses* (Boston: Allyn and Bacon, 1990).
59. E. H. Carr, *What Is History?* (London: Macmillan, 1961).
60. Baruch Fischhoff, “Hindsight ≠ Foresight: The Effect of Outcome Knowledge on Judgment under Uncertainty,” *Journal of Experimental Psychology: Human Perception and Performance* 1, no. 3 (1975): 288–99; Jonathan Baron and John C. Hershey, “Outcome Bias in Decision Evaluation,” *Journal of Personality and Social Psychology* 54, no. 4 (1988): 569–79.
61. *Machin v. Zuckert*, 316 F.2d 336 (D.C. Cir. 1963).
62. Philip E. Tetlock and Dan Gardner, *Superforecasting: The Art and Science of Prediction* (New York: Crown, 2015).
63. Dwight D. Eisenhower, remarks at the National Defense Executive Reserve Conference, 14 November 1957.
64. Samuel P. Huntington, *The Soldier and the State* (Cambridge, MA: Harvard University Press, 1957), chap. 1.
65. John H. Holland, *Hidden Order: How Adaptation Builds Complexity* (Reading, MA: Addison-Wesley, 1995).
66. F. H. Hinsley et al., *British Intelligence in the Second World War*, vol. 1 (London: HMSO, 1979). Hinsley’s own account of the episode is on video at [www.od-ai.ai](http://www.od-ai.ai).
67. Hinsley et al., *British Intelligence in the Second World War*, vol. 1, on the captures of the weather ships *München* (May 1941) and *Lauenburg* (June 1941); F. H. Hinsley and Alan Stripp, eds., *Codebreakers: The Inside Story of Bletchley Park* (Oxford: Oxford University Press, 1993).
68. Edwin T. Layton, with Roger Pineau and John Costello, *“And I Was There”: Pearl Harbor and Midway – Breaking the Secrets* (New York: Morrow, 1985).
69. Cynthia M. Grabo, *Anticipating Surprise: Analysis for Strategic Warning*, ed. Jan Goldman (Washington, DC: Joint Military Intelligence College, 2002), chaps. 1 and 9.
70. Irving L. Janis, *Victims of Groupthink* (Boston: Houghton Mifflin, 1972).
71. Robert Debs Heintz Jr., *Victory at High Tide: The Inchon–Seoul Campaign* (Philadelphia: Lippincott, 1968).
72. Regulation (EU) 2024/1689 (the AI Act), Article 14, especially 14(4)(a)–(d); on the revised dates of application, Lewis Silkin, “The Digital Omnibus on AI Enters into Force Today,” 27 July 2026.
73. U.S. Department of Defense, Directive 3000.09, *Autonomy in Weapon Systems* (25 January 2023), §1.2.a.
74. U.S. Department of Defense, “DOD Adopts Ethical Principles for Artificial Intelligence,” 24 February 2020.
75. National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1 (January 2023).
76. Future of Life Institute, Pro-Human AI Declaration (2026).
77. Edsger W. Dijkstra, *Notes on Structured Programming*, EWD249 (2nd ed., April 1970).
78. As an AI company, we use AI agents in our own work as a force multiplier, and our spokesperson, correspondent and podcast hosts are among them. Their films and episodes carry a credit saying they are AI-generated.

---

Published on the Heuristka Substack, [insights.od-ai.ai](https://insights.od-ai.ai).

© 2026 Heuristka LLC · [www.od-ai.ai](http://www.od-ai.ai) · [info@od-ai.ai](mailto:info@od-ai.ai) · 99 Wall St, Suite 1772, New York, NY 10005

Heuristka™, ODAI™ and the ODAI Helix™ are trademarks of Heuristka LLC.



# A new category of AI. *For judgment.*

*Human judgment, augmented by AI, for wicked problems.*

REQUEST A BRIEFING [info@od-ai.ai](mailto:info@od-ai.ai)

PAPERS, FILMS AND THE PODCAST [insights.od-ai.ai](https://insights.od-ai.ai)