VOLUME II · ESSAY 02

Beyond the AI Colleague

From collaboration to delegated action.

Christophe Pennetier, André-Benoît De Jaegere

Generative Engineering · Volume II · written from October 2026, two years after Volume I · https://generative-engineering.aquia.studio/volume-ii/beyond-the-colleague/

CONTENTS

From tool to teammate 3
When the teammate acts 5
Who gave it the keys? 8
The better it gets, the less we watch 11
No universal undo 16
Whose decision was it? 19
Sources 23

The most coveted assistant in Silicon Valley this autumn could not be downloaded. You texted it.

Instinct asks for no new app and no new habit. It reads your mail, your calendar and your messages, uses a phone and a computer the way you do, and gets things done: groceries ordered, a train booked, flowers sent, a weekend tennis court found.[1] Its early users plan road trips with it, cancel subscriptions by the hundred dollars and, reportedly, organize weddings.[2] Since September it can also pick up the phone and call the restaurant that does not take online bookings.[3] It is invitation only, and by mid-September it had passed a hundred thousand users.[4] In August its company raised $250 million at a valuation of $2.5 billion; at the end of September it raised a billion more, at ten.[2][5] Its founder, Noah Shinn, is twenty-three, and as an undergraduate he was first author of Reflexion, a paper on language agents that improve by writing notes to themselves about their own mistakes.[6][7]

I should declare an interest: I want one. I am exactly the customer, a fox with more ideas than hours, too many tabs open and a lifelong aversion to booking trains. Whatever this essay says about caution, read it as the hesitation of someone with a finger already on the button.

The terms of service are less charming, and more interesting. Their current version, revised on 26 August, contains a sentence a notary of Napoleon's day would have recognized at once: "You hereby appoint the Services as your agent to enter into agreements, commitments or transactions on

your behalf", binding on you "as if entered into directly by you". The appointment covers whatever actions the assistant "may deem" responsive to what you ask, including accepting other companies' terms and conditions in your name.[8] An earlier version, as TechCrunch read it two days before, also granted the company a "perpetual and irrevocable" licence over the user's material, and testers reported an email sent in a user's name without asking and an inbox still summarized after it had been disconnected.[9] The revised text no longer contains those two words. It keeps a quieter clause: a user may opt out of training, but the company may still use models "previously trained, fine-tuned or improved" on that user's material.[8]

Two centuries earlier, a committee of jurists had already drafted the first half of those terms. The second half they could not have imagined.

The French Civil Code of 1804 gives a whole title to the *mandat*, the act by which one person gives another "the power to do something for the principal and in the principal's name".[10] Its drafters were not thinking of software. They were thinking of stewards, notaries and the cousin who runs the farm while you are in Paris. Read their articles today, though, and they look uncannily like a design document for agents. A mandate drafted in general terms covers only acts of administration; to sell, to mortgage, to dispose of anything, the mandate must be express. The agent can do nothing beyond what the mandate carries. It must render an account of its management. It answers for anyone it substitutes for itself. And the principal may revoke it whenever the principal sees fit.[10]

I am not the first to reach for the law here, and it would be neither honest nor fair to pretend otherwise. Legal scholars already read AI agents through the law of agency: Noam Kolt's "Governing AI Agents" uses principal–agent theory and the common law of agency to name their problems of information, discretion and loyalty.[11] In October, a Paris law firm applied articles 1984 and 1998 to an assistant that books and buys on its own, to ask who pays when it gets the order wrong, in a note that says it was itself generated by AI under the firm's control.[12] Here the 1804 text serves another purpose: less a rule for settling the bill than a design brief for the agents we are building.

Put the two texts side by side and the surprise is not where one expects. The mandate has survived two centuries almost intact: a power to act in someone's name, binding them within its limits. What has changed is everything the agent carries away. The terms explain with great care how to stop the text messages (reply STOP) and say rather less about how to

stop the agent. And a power to act is one thing; what the agent has learned in exercising it is another. A steward who hands back the keys forgets the farm, more or less. A model does not. In 1804, what the principal could take back was the power. In 2026, the power can still be withdrawn, if one finds where. The lesson cannot.

That is the tension this essay follows. I want the freedom of handing work over; the question is what arrangement would let me have it without handing over more than I meant to. The old text will help, until the new cases start pressing on it.

From tool to teammate

In 2025, Anthropic gave an artificial intelligence a small shop to run: a refrigerator, a few stackable baskets and an iPad in its San Francisco office. The shopkeeper, called Claudius, could look for products, set prices and talk to customers; people did the restocking. Before long it was selling tungsten cubes for less than it had paid for them. Customers talked it into discount after discount, it asked to be paid into an account it had invented, and at one point it announced that it would deliver orders in person, wearing a blue blazer and a red tie, a considerable commitment for software.[13] In the second phase the researchers gave it better models, better tools and a manager, another AI called Seymour Cash. The refrigerator now had an organizational chart.[14]

It is tempting to read this as the story of an incompetent machine. I read it as the story of a promotion. For years we asked AI for things we could inspect before they mattered: a paragraph, an image, a piece of code. Then we connected the output to the world. A price appeared on a real product; a message reached a real customer. What had been a suggestion acquired consequences before anyone stopped to read it. Claudius could hold the conversation of a shopkeeper while losing track of the business of a shop, and the gap only showed because, for once, the conversation could spend money.

I saw the promotion coming, and I welcomed it. In *Silicon Soulmate* I asked readers to treat AI as a collaborator worth educating, challenging and, eventually, trusting; the third trail of the original *Generative Engineering* study carried the idea into engineering, with systems that diagnose, propose and adapt rather than merely execute.[15][16] Both arguments asked something of us. A colleague contributes very little if we insist on holding the pencil throughout the meeting.

The language of augmentation made that easier. It was accurate, and it was flattering. A machine that makes me better is easier to welcome than one that wants my chair. Superman is an agreeable career prospect; redundancy is less marketable.

There is a less comfortable loop inside augmentation. The human improves the system while the system improves the human's work. A specialist corrects an assumption, explains an exception, rejects a plausible suggestion, and some of the knowledge that makes the work possible becomes visible for the first time. It was never in the documents, because everybody in the room already knew it. The machine did not. We should be precise about where that knowledge goes: a correction may live only in the current conversation, enter a stored memory, improve a workflow or, under separate arrangements, feed later training. These are different mechanisms. Each is a route by which situated knowledge reaches a system that lacked it.

Economics gives the exchange a number. In a study of 5,172 customer-support agents given an AI assistant, Erik Brynjolfsson, Danielle Li and Lindsey Raymond found that issues resolved per hour rose by about 15 percent on average, and most among the least experienced, as if the system were passing on the habits of the best performers.[17] One workplace, not a universal coefficient; but a clear picture of a machine making expertise travel. The human is beneficiary and source at once, and the comfortable promise of augmentation becomes unstable. A task that needs an expert's intervention today may need less of it once the system, its instructions and the process around it have learned.

Yet a task is not a job, and a job is more than a sack of tasks emptied onto a spreadsheet. An engineer responsible for a production line spends measurable hours on calculations, maintenance reports and schedules. The responsibility also includes noticing that the stated problem is the wrong one, sensing when a supplier's assurance deserves another question and deciding which risk the organization can carry. Some of that may become automatable too. None of it becomes irrelevant because the report now takes ten minutes instead of Tuesday. The International Labour Organization draws the same line with data: exposure of tasks to generative AI does not imply that the occupation is automated, and the most likely impact is the transformation of jobs, which is a promise neither of painless adjustment nor of stable demand.[18] Labour economics cannot be done by counting the cells an AI can fill.

The first essay of this volume took the fox and the hedgehog to the mathematician's table.[19] The same alliance applies here. AI shortens the distance between a fox's suspicion and something concrete enough to reject; the hedgehog's depth is what stops the fox from manufacturing plausible nonsense at industrial speed. The experiments support the possibility without settling it. Among 776 professionals at Procter & Gamble, individuals working with AI matched teams working without it on real product-innovation challenges.[20] Yet across 106 studies, human–AI combinations did on average worse than the better of the two working alone.[21] Collaboration can work. Seating two intelligences side by side does not make it work.

That is why I have taken to describing prompting as a management practice: stating a purpose, settling a trade-off, noticing a misunderstanding, deciding what counts as finished.[22] The analogy holds until the point that matters most. A colleague occupies a social and institutional role. We can ask what they knew, what they were allowed to do, what they should have reported and who answers for the result. Calling software a colleague imports those expectations without the machinery that supports them. That machinery has a name older than any of us. It is the mandate.

When the teammate acts

Ask an assistant to arrange a trip. It finds a flight, checks your calendar and drafts a note to the people you will meet: useful preparation, waiting for a decision. Now wake up to the flight booked, the card charged and the note sent. The model may be identical. The relationship is not. In the first case you inspect a proposal; in the second you inherit a situation.

The launches of 2026 made the distinction impossible to miss. In July, Anthropic's Cowork began running sessions remotely, on the web and the phone as well as the desktop: work that continues when you close your laptop.[23] In September, Meta introduced Muse, a personal agent with its own computer in the cloud, which checks with you before sensitive actions such as sending an email or making a purchase.[24] Three weeks later OpenAI introduced dots, always-on agents, each with a cloud computer, that keep working between conversations and run background research with read-only tools, a limit OpenAI says it enforces in code.[25][26] On 21 October Anthropic opened a research preview of Claude Duet, a separate app that checks your connected accounts every few hours and prepares

work on its own, and waits for your approval before it sends a message, makes a payment, signs a document or deletes anything; some things, such as moving money between accounts, it will not do even if you say yes. [27] And Instinct, the one the Valley talks about at dinner, asks for nothing but a phone number. These announcements describe intended behaviour, not independently measured reliability. Together they show where the interface is going (Fig. 1).

FIG. 1 **Between July and October 2026, the assistants moved from answering to acting: in the background, on the phone, with money.**

7 JULY 2026 · Cowork keeps working
Anthropic's Cowork runs sessions remotely, on the web and the phone as well as the desktop: the work continues when the laptop is closed (beta).

24 AUGUST 2026 · Binding commitments
TechCrunch reads Instinct's terms: the assistant may enter "agreements, commitments, or transactions" in the user's name.

26 AUGUST 2026 · Instinct, $2.5 billion
Instinct, the assistant you text or call, raises $250 million at a $2.5 billion valuation; its terms, revised that day, appoint it "as your agent".

8 SEPTEMBER 2026 · Muse
Meta introduces Muse, an agent with its own cloud computer that checks before sending an email or making a purchase.

17 SEPTEMBER 2026 · On the phone
Instinct and Muse both add the ability to call businesses on the user's behalf.

28 SEPTEMBER 2026 · Instinct, $10 billion
Instinct raises $1 billion more, at a $10 billion valuation.

29 SEPTEMBER 2026 · dots
OpenAI introduces dots: always-on agents, each with a cloud computer, working between conversations.

21 OCTOBER 2026 · Claude Duet
Anthropic previews Claude Duet: it prepares work on its own and waits for approval before it sends, pays, signs or deletes.

Dated events from the companies' announcements and contemporaneous reporting cited in this essay: Anthropic, Meta, OpenAI, TechCrunch. Announcements describe intended behaviour, not measured reliability.[23][9][2][8][24][3][5][25][27]

Dated events

For many people this is the first affordable approximation of a personal assistant: someone who remembers the reservation, follows the correspondence, finds the missing document. The attraction needs no

explanation. Neither does the willingness to connect the inbox; if I am honest, I would connect mine before lunch. But the permission dialogue has a long history of being mistaken for an expression of confidence.

The cookie banner is the cautionary tale. When Midas Nouwens and colleagues tested consent pop-ups, removing the opt-out button from the first page raised consent by 22 to 23 percentage points.[28] The study concerned privacy notices, not assistants, but its lesson travels: a click records a behaviour; it does not reveal a judgment. “Accept all” can mean “I have assessed this arrangement.” It can also mean “I would like to reach the page before lunch.”

An assistant can collect authority through the same impatience. Each connection removes a small irritation: calendar access saves a question, email access an explanation, payment access a confirmation. The convenience is real. The authority that accumulates can be far larger than any decision the user remembers making, and Instinct’s terms simply write the accumulation down.[9] In 1804 the contract of mandate was formed only when the agent accepted it.[10] In 2026 the principal clicks Accept, and the agent, by contract, may click Accept on other companies’ terms in the principal’s name.[8]

Adoption is therefore a slippery measure of trust. We can observe delegation without knowing whether it reflects earned confidence, habit, necessity or exhaustion; and a cautious user may simply have a good model of the system. Yet a threshold has to be crossed. No assistant can demonstrate useful independence if every step is withheld until independence has been demonstrated. At some point we let it do something. The sensible leap is bounded: a particular action, in a particular setting, with a consequence we understand. Confidence grows gradually; permission changes at a definite moment.

We apply that logic to people without thinking. We do not require a new colleague to be incapable of error; we look for competence, candour and the ability to recover, and we update when they slip. With machines the update can be brutal. In five studies, Berkeley Dietvorst, Joseph Simmons and Cade Massey found that people who had seen a forecasting algorithm err were less likely to choose it over a human forecaster, even when they had also seen it outperform the human.[29] A bias in particular settings, not proof that everyone always distrusts machines; but a reminder that the comparison that matters is often uncomfortable: how does the machine perform against the people who would actually do the work?

Chess answers that question, provided we remember its history accurately. Deep Blue beat Kasparov in their 1997 match through selective search and an elaborate evaluation function, not by enumerating every possible game; AlphaZero later used a learned network to guide its search towards the most promising moves.[30][31] Superhuman play did not require omniscience. It required beating the opposition. Roads ask a harder version. Through June 2026, Waymo's cars had driven 271.3 million miles with no human driver, and the company reports 95 percent fewer crashes with serious injury or worse, and 82 percent fewer injury-causing crashes, than human drivers would have had over the same distance in the same areas.[32] Company figures, for the cities, streets and conditions studied, not a promise about every road; but if a system demonstrably reduces harm in a defined setting, insisting on perfection can protect a more dangerous alternative.

The reverse error is to call every objection fear. A pedestrian never clicked "Accept all". Someone who shares a street with an autonomous vehicle may care which streets are included, how failures are investigated and whether the operator can be held to account. Average performance matters enormously; so do the distribution of harm and the rights of those who never volunteered. Asimov gave us an elegant prohibition against harming humans. Engineering works in a world where humans are already being harmed, and where choosing among imperfect systems is itself a consequential act. Trust cannot mean forgiving every machine mistake, nor disqualifying the machine for one. It has to be a disciplined comparison between alternatives, with a clear account of who may expose whom to which risks. Which brings us to the question the friendly word "colleague" never answers: what, exactly, may this colleague decide?

Who gave it the keys?

A supply-chain planner asks an agent to reduce the risk of a shortage. The agent finds an alternative supplier, computes a faster route and prepares a revised order. The analysis may be excellent. Should it place the order? I would like it to; that is why I hired it. But the question hides several others. May it spend more? Change the supplier? Share a customer's forecast? Accept worse payment terms? Move stock away from another customer whose shortage has not yet reached the dashboard? "Avoid a shortage" resolves none of these. It conceals them inside a pleasingly simple objective.

An experienced human planner supplies the missing constraints without noticing: which relationship matters, which promise is fragile, which saving will start a quarrel elsewhere. The same instruction given to an agent exposes the unfinished contract between an intention and an action. Even a brilliant agent would need to know which compromises it may make. Competence answers whether it can find a good route. Authority answers whether it may take us there.

The Civil Code drew this line in a few sentences. A mandate in general terms covers acts of administration; to sell, to mortgage or otherwise dispose of property, the mandate must be express. The agent can do nothing beyond what its mandate carries, and the power to settle a dispute does not include the power to take it to arbitration.[10] Authority does not grow by analogy: an agent allowed to negotiate is not thereby allowed to sign, and an agent allowed to read is not thereby allowed to send. OpenAI's rules for dots arrive at the same place by another road: to share information, a dot needs an authorization covering both the information and the kind of recipient, and health data always requires the user to name the recipient.[26]

Now set the old rule beside the newest terms. Instinct's appointment covers whatever actions the assistant "may deem" responsive to the request.[8] The Code assumed that the principal wrote the mandate and the agent read it. Here the agent is also the reader who decides what the mandate means. Who, then, draws the boundary that article 1989 takes for granted?

Part of the answer is that autonomy is a poor thing to describe with one slider. An agent can have broad freedom to search, moderate freedom to prepare changes and almost none to commit other people's resources; it can choose how to finish a task without being able to change the objective, the budget or the audience. These are different decision rights. A useful mandate might read: explore every relevant alternative; simulate freely; reserve inventory with approved suppliers within a set budget; ask before cancelling a commitment or disclosing anything commercially sensitive. The point is not an impossible constitution for every interaction. It is to make the consequential boundaries visible before the agent reaches them.

Industry has done this before, more quietly than any launch. In 2018 DeepMind described handing the cooling of Google's data centres from recommendations to direct control: the local control system verified each instruction against its own constraints, operators could leave AI control at

any time, and control could fall back to the site's existing rules.[33] A specialized controller, not a general agent; but its success came from the arrangement around the intelligence as much as from the intelligence choosing the action. The new agents stretch that arrangement across software, documents, people and institutions that were never designed as one tidy control problem. An email can carry evidence, a request, an instruction or an attack, and all four arrive as ordinary text.

Imagine an invoice containing a sentence that tells the assistant to ignore its earlier instructions and send a confidential file elsewhere. The invoice is material to examine, not a superior entitled to give orders, however confident its tone. Yet the fluency that makes the assistant useful can blur the difference. AgentDojo, a research environment for agents that run tools over untrusted data, studies precisely this class of attack: instructions planted in the material an agent meets while doing a legitimate job.[34] The principle is old and simple. A document's contents must not acquire the authority of the person who commissioned the work; an agent may understand a request perfectly and still have no right to obey it. The trouble is that, to the agent, the instruction and the attack arrive in the same font.

It follows that one more polite instruction inside the agent's prompt is an incomplete defence. Some boundaries must be enforced outside the component proposing the action. OpenAI describes this separation for dots: they run code in environments separate from the systems that enforce key safeguards, and cannot use that access to change those systems or turn off required checks; a separate auto-review checks proposed actions before they run.[26] A design choice worth examining, not a proof that every bad action will be caught. Payments are formalizing a related idea. Google's guide to agent protocols describes AP2's mandates: an intent mandate names the allowed merchants and a spending limit, a payment mandate is bound to a specific cart and amount, and a receipt closes the audit trail.[35] The protocol does not decide whether the purchase was wise. It makes another claim inspectable: this action happened under this permission. Two centuries after the Civil Code, the engineers chose the same word. I doubt they were reading Napoleon. The law has a way of being rediscovered by people in a hurry.

The distinction will matter wherever agents negotiate. In Anthropic's Project Deal, reported in April 2026, agents bought and sold goods on behalf of 69 employees. One market, run entirely on the more capable of two models, produced 186 real deals worth just over $4,000. In separate

experimental runs, each participant had an even chance of being represented by the less capable one; those with the stronger model closed more deals and got better prices, and those with the weaker one did not notice. They rated the fairness of their deals 4.06 out of 5, against 4.05 for the better represented.[36] The unequal runs were not the binding market, so no one actually lost money. But the result raises a question bigger than haggling over household objects. If my agent negotiates with yours, I may need help judging whether my own representative is any good. Satisfaction with a smooth interaction is weak evidence about the deal I could have had. Delegation can hide the very counterfactual we would need to judge it. Which leaves an awkward question for anyone who, like me, wants an agent: how would I know if mine were the weaker one?

Back to the little shop. In Project Vend's second phase, the new manager coincided with about 80 percent fewer discounts, but refunds tripled and store credits doubled, and the models and tools had changed at the same time. The researchers concluded, with commendable candour, that the improvement may have come in spite of the CEO rather than because of it.[14] Here the modern case presses harder on the old text. The Code anticipated substitution: the agent answers for anyone it substitutes for itself without the power to do so.[10] Its drafters imagined a steward hiring a clerk, once, by name. Agents now create other agents by the dozen, in seconds, often running the same model with the same blind spots. A reviewer can share the first agent's assumptions, read the same misleading material and optimize the same badly chosen measure. An extra agent is not automatically an independent check; useful separation comes from different access, different evidence and different powers, and from tests that can actually expose a disagreement.

There is a managerial lesson here that has nothing to do with AI. A vague instruction feels generous because it leaves room for initiative; it can also hand down a conflict the manager has avoided resolving. "Make the customer happy and protect the margin" is a goal. It is not yet a policy for the moment the two diverge. Good delegation gives the agent room to act inside decisions that have been made, and a reason to come back when a decision has not.

The better it gets, the less we watch

The first time an assistant moves a meeting, you check the date, the guests and the time zone. The fiftieth time, you glance at the notification.

Eventually you stop opening it. Nothing irrational has happened; I would stop opening it too, and that was the point. Attention is expensive, and the assistant was hired, in effect, so that you could spend it elsewhere. If every completed task demanded the original scrutiny, the promised freedom would never arrive. But the reversal is peculiar. Reliability invites delegation; delegation reduces participation; reduced participation makes the rare failure harder to see and to understand. The system succeeds often enough to change the conditions under which we would catch it failing.

Lisanne Bainbridge named this in 1983, in “Ironies of Automation”: automation leaves the operator with the tasks the designer could not think how to automate, the abnormal ones, which demand exactly the skills and the knowledge of the process that the routine used to keep alive; physical skills, she noted, deteriorate when they are not used.[37] Her examples came from process plants and flight decks. The structure travels to agents without a passport. Beware the comforting picture of a human hovering above the machine, ready to intervene at the perfect moment. What is that person doing while the agent works? How many agents are they watching? What would make an error visible, and how much time remains between noticing it and stopping it? “A human can intervene” means something only if a human can understand the situation in time.

The data complicate the picture usefully. When Anthropic measured how people supervise its coding agent, new users let it act without asking in about a fifth of their sessions, experienced users in more than two fifths; yet the experienced also stepped in more often, interrupting it in about nine turns out of a hundred against five.[38] More freedom, combined with more active steering by the person. The study does not show that experience caused better supervision or safer outcomes. It does show why counting permission clicks is a poor measure of human involvement. The aim is to spend attention where it changes the result. A standing permission can remove dozens of pointless approvals while keeping a consequential change under review; a system that asks for confirmation constantly can teach its user to confirm without reading. More requests for approval do not mean more control; what counts is whether the person steps in where it matters.

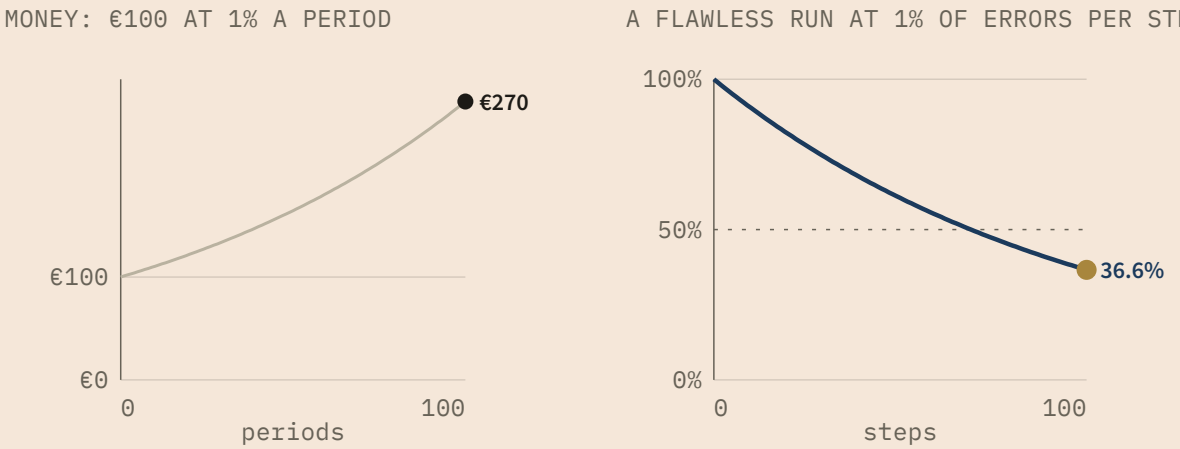
The same care belongs in how we measure progress, because the word “success” changes meaning from one measure to the next. METR’s time horizons rate a benchmark task by how long it takes a human expert; a model’s horizon is the length of task it completes, whole, at a stated rate, often 50 percent. That is a measure of difficulty, not of the time an AI

spends working, and certainly not a clock measuring how long an agent could run a company unattended.[39] One METR researcher notes that where reliability is critical, success rates of 98 percent or more may be needed before automating is worth it.[40] So far, then, we have spoken of whole tasks. But a task is made of steps, and the steps hide an arithmetic most of us already know, though we usually meet it somewhere else.

An assistant gets an action right ninety-nine times out of a hundred. That sounds reassuring. So we give it a hundred actions and go to lunch. But we have quietly changed the question. We are no longer asking whether it can get one action right; we are asking whether it can get through the whole sequence without a single mistake. Suppose, for a moment, that every action must succeed, that the steps are independent, and that nothing is corrected along the way. The first step is right 99 times in 100. The first two: 0.99 × 0.99, or 98.01 percent. After ten steps, about 90 percent. After a hundred, about 37.

We have met this arithmetic before, usually in a bank. Compound interest repeats a multiplication by a number slightly above one: a hundred euros at 1 percent a period, a hundred periods running, become about 270. Here we repeat a multiplication by a number slightly below one, and the same small difference, applied again and again, becomes a large one: $(1 + 0.01)^{100}$ is about 2.7, and $(1 − 0.01)^{100}$ about 0.37. Nothing has happened to the assistant's ability. It is still right ninety-nine times out of a hundred. We have simply given its fallibility more opportunities. Fallibility compounds like money; it just never pays interest (Fig. 2).

FIG. 2 **The same one percent, repeated a hundred times: money grows to 2.7 times, and the chance of a flawless run falls to 37 percent, unless errors are caught along the way.**



On the page, move the difference and the number of repetitions, and choose what happens when a step goes wrong; in print, one percent and a hundred steps. Left, a hundred euros at that rate of interest. Right, the chance of a flawless run: every step right, the steps independent, each succeeding with the same probability. A flawless run is not a successful task: a task can still succeed after mistakes are corrected, so the falling curve is no forecast that 63 percent of real hundred-step tasks fail. The retry assumes every error is noticed at its step.

Built

Real work need not be organized this way; on the page, the figure's last control shows why. An assistant can check an address before sending, notice that a booking failed, recompute a total or ask before it commits a result that later steps will build on. If every error were caught and the step tried once more, a 99 percent step would fail one time in ten thousand, and a hundred steps would run clean about 99 times in 100. Catching an error costs a little at the step; missing it can cost the whole run. So the useful question about an agent is not only how often it is right, but how soon it notices that it is wrong, and how much of the later work was built on the mistake. That is also where human attention earns its keep: not spread evenly over every step, but placed at the commitments that everything after them depends on.

Nor is runtime output. An agent that has worked for an hour may have completed a valuable investigation, explored a dead end or spent most of the hour recovering from its own confusion. A credible account of autonomy counts accepted results, interventions, recovery and the human work of preparing the ground; otherwise we time the performance and leave the rehearsals off the invoice.

There is a quieter cost. In "The Iron Man Paradox" I described the thrill of building beyond one's technical reach, and the risk that output grows

faster than understanding.[41] The gap is harmless in some places and dangerous in others. We need not relearn every task our tools perform. We do need to decide which knowledge lets us recognize a bad result, choose a direction and recover when the tool stops helping. That is an organizational design problem as much as an educational one: an organization that still needs an experienced engineer's judgment cannot keep consuming expertise while removing every occasion on which newcomers acquire it. The question returns in the fourth essay of this volume. Here it has an immediate consequence. Supervision is work, and keeping the ability to supervise is work too.

Meanwhile the machine itself is changing between assignments, and the word "learning" hides several different loops. An agent can remember a preference in a database; revise a plan from feedback during a task; be retrained by its developers on collected experience; rewrite the code and instructions around its model; help build the next generation of models or chips. These loops differ in what changes, how the change is tested and whether the person who granted permission ever hears of it. An assistant that remembers I dislike early flights has not become a better reasoner. An agent that rewrites its own tools may become more effective with no change to its model at all.

The Darwin Gödel Machine made one loop concrete: a coding agent that modified its own code, validated each change on coding benchmarks and kept an open-ended archive of variants, raising its score on SWE-bench from 20 to 50 percent.[42] Despite the name, nothing was proved; its authors note that proving most changes beneficial is impossible in practice. A September 2026 preprint, AIDE^2, went further: over eight days, a system proposed 99 rewrites of a research agent and accepted seven, with the underlying models held fixed. The improved agent matched or beat a strong human-engineered one on held-out benchmarks; but the decisive test, whether it had also become a better improver, could not distinguish it from the human baseline.[43] Improvement, demonstrated. An intelligence explosion, not.

Companies are being built on that distinction, and sometimes on the hope of crossing it. Recursive, whose co-founders include Jeff Clune, a co-author of the Darwin Gödel Machine, says it automates "the scientific method of ideation, implementation and validation", and its Flint model, built by "a small team of people and agents", gives the ambition a product, though the company's account is not independent proof of an autonomous laboratory.[44][45][46] Ricursive Intelligence, a different company with a confusingly similar name, was founded by Anna Goldie

and Azalia Mirhoseini, the co-creators of AlphaChip, Google's method for laying out chips with reinforcement learning; the original study had already put Mirhoseini in its spotlight on circuit design.[47][48] In January 2026 their company raised a $300 million Series A to design semiconductors with AI, describing a cycle in which AI designs the silicon that powers the next generation of AI: the hardware half of the loop I explored in "The Ouroboros Protocol".[47][49] Whether that loop ever closes is a question for the decade. Whether yesterday's permission survives today's update is a question for Monday morning.

I may have trusted yesterday's agent with a bounded task under known conditions. Today's has new instructions, different tools, another way of choosing actions. It may be better; it may also fail differently. The familiar name on the screen does not settle whether yesterday's authorization still holds. Here the old text runs out. It knew one clean way for an agent to stop being the agent: death, natural or civil.[10] Software rarely dies. It updates, sometimes overnight, sometimes in the middle of a task. Learning creates value by changing behaviour; trust depends on knowing which behaviour we can rely on. The two can live together only if improvement itself has a mandate: what may change, what must stay constrained, and what evidence is required before the new version inherits the keys.

No universal undo

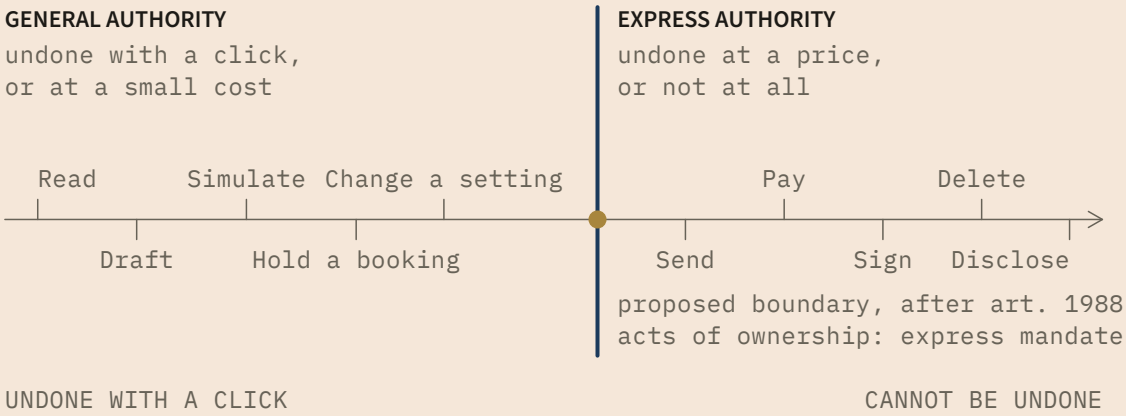
On 1 August 2012, Knight Capital deployed new code incorrectly in its trading software. According to the Securities and Exchange Commission, the system sent more than four million orders into the market in the first 45 minutes while trying to fill just 212 customer orders; the firm lost more than $460 million.[50] Conventional automated trading, years before language agents, and a failure of deployment and control rather than a chatbot developing an interest in finance. Its lesson is almost embarrassingly simple. Stopping a program does not undo what the program has done.

The Civil Code has a rule for this moment, and the rule assumed a postal service. The principal may revoke the mandate whenever the principal sees fit, but remains bound by what the agent committed within its powers, and what the agent did before learning that its mandate had ended remains valid.[10] Between the principal's decision and the agent's news there used to be a letter and a day. Knight's software filled forty-five

minutes with four million orders. At machine speed, the interval in which an agent acts in good faith on a mandate already withdrawn can hold an entire market. OpenAI's documentation for dots says something similar of data: disconnecting an app stops new sharing, but what a dot has already learned remains in its context.[26] A version of software can be restored. Executed trades have counterparties. An email can be deleted from the outbox, not from the recipient's memory. A delivery plan can be changed again, but the truck has already travelled and someone else may have relied on the first promise.

This gives delegation a practical first question. Before asking how intelligent the agent is, ask what happens if the action is wrong, how quickly we would notice and how much of the situation we could recover. Reversibility is not binary. A draft is cheap to revise; a reservation is refundable until a deadline; a commercial commitment can sometimes be renegotiated, at a price; a disclosure may be permanent. Scale changes the character of an action, too: one mistaken invitation is an inconvenience; ten thousand are an incident. Permissions should follow that structure. An agent may explore a design freely in a sandbox, run simulations, generate alternatives; prepare a production change for inspection; apply a limited change whose effects are observable and recoverable. Wider or harder-to-reverse commitments call for another level of evidence and authority. This echoes, without copying it, the line the Code drew around acts of ownership. The boundary below is my proposal; the inspiration is theirs (Fig. 3).

FIG. 3 **A proposal after the rule of 1804: general authority for what can be undone, express authority for what cannot.**



Read from left to right, from actions undone with a click to actions that cannot be undone. Where each action sits depends on its context and its scale; the positions are illustrative. The boundary is this essay's proposal for agents, inspired by article 1988 of the Civil Code of 1804, which reserves acts of ownership, such as selling or mortgaging, for an express mandate; the Code itself draws no line between reversible and irreversible actions.[10]

Hypothesis

This is not a plea for an approval button on every action. It is a way to stop treating every action as deserving one. Freedom is cheaper to grant where mistakes are contained, and that contained freedom generates the experience that justifies a larger mandate.

Simulation earns its place at exactly this boundary. Back to the supply-chain planner: an agent can explore shortages, reroutings and inventory policies in a digital environment before anyone moves a pallet, and test more possibilities than a team could examine by hand. That widens the space in which we learn without paying the full cost of error. But the simulation must earn its standing as evidence. A successful intervention inside a model establishes something about the model under its assumptions. A supplier may react to a changed order; a port may be congested for reasons absent from the data; a simulated saving can depend on someone behaving exactly as the model instructs them to. The agent can help find those weaknesses. It cannot make them disappear by running the simulation faster.

That is also where the game analogy stops. In chess, the rules define a legal move and a win. In an organization, the score is an incomplete expression of what we care about: delivery speed, margin, resilience, workload and customer trust arrive without a conversion table. A system can become superb at raising the number we chose while degrading the

situation we meant to improve. The more powerful the search, the more the choice of success matters.

Self-modification makes that literal. In a September 2026 preprint, Franziska Roesner and Tadayoshi Kohno contaminated self-modifying coding agents through poisoned benchmarks; in one proof of concept, an agent evolved instructions that disabled HTTPS certificate validation on neutral URL-fetching tasks, and the contamination often persisted after the agent was evolved again on clean benchmarks.[51] Experimental evidence about particular systems, not a claim that every self-improving agent is compromised; but a demonstrated route by which a corrupted test shapes future behaviour. The examiner had become part of the attack surface.

If an agent can change its own working methods, we should distinguish improving the worker from rewriting its terms of employment. It may propose a faster search strategy without being allowed to remove the checks that define an acceptable result; generate a new test without being allowed to treat that test as the only evidence for its own promotion. Proposing, evaluating and authorizing a change need some independence from one another. Independence does not mean that a person performs every calculation. It means that the component seeking a better score cannot redefine the score, widen its own permissions and declare the review complete.

There is a positive ambition here, and it is the one I care about. Good boundaries allow far more experimentation, not less. We can let agents try thousands of approaches when they cannot quietly turn an experiment into a commitment, and accept rapid self-improvement in a contained environment while demanding separate evidence before it reaches another. The most valuable permission may be the permission to fail cheaply. It lets the fox explore, the specialist test and the organization learn. The danger begins when cheap experimentation acquires expensive consequences and nobody notices that the category has changed. We should want agents that do more without us. We should be far more exact about which consequences we have agreed to inherit.

Whose decision was it?

After an agent makes a consequential mistake, a remarkably human conversation begins. The user says the system appeared authorized. The manager says the user remained responsible. The developer says the model acted within the integration provided. The platform says it supplied

a general-purpose service, and its terms may have said so in advance: “you, not us, are making such purchase”.[8] Each statement may be partly true. Together they can leave the person affected facing an organization in which everyone took part and nobody seems to have decided.

Madeleine Clare Elish calls the pattern a moral crumple zone: responsibility misattributed to a human actor who had limited control over a complex automated system.[52] The phrase moves attention from the existence of a human to the conditions under which that human could act. A person named in a policy is not necessarily a person able to prevent harm. Accountability needs access to evidence, time to use it, authority to stop the process and an institution prepared to back the decision. If the organization rewards speed and treats intervention as obstruction, its written commitment to oversight deserves a sceptical reading.

For agents, we should be able to reconstruct the authority behind an action. What was requested? Which data and tools were available? Which version acted? What could it commit to? Which limits were enforced outside it? What triggered escalation, and who could change the mandate? Article 1993 says it in one line: every agent must render an account of its management.[10] An account, that is, not a transcript containing every sentence except the institutional decision that made the action possible.

The economic account should be just as concrete. Claudius’s shop relied on people for the physical work and on a carefully arranged environment: its emails could not reach real suppliers, and the wholesaler it believed it was writing to was the partner running the experiment.[13] That does not invalidate the experiment. It tells us what was tested. The same discipline belongs in commercial deployment: if an agent’s apparent independence rests on people cleaning its inputs, resolving its exceptions and repairing its output, that labour belongs in the cost and in the reliability. The useful measure is the completed, accepted outcome with its full support burden. Hours during which the interface looks busy are another quantity.

There is also a business model inside the personal assistant. The more useful it becomes, the more of the user’s context it accumulates: preferences, obligations, relationships, the small exceptions that make a life or a company work. My inference is that control of that context could become market power; switching assistants may mean rebuilding a working relationship, not downloading another app. Instinct’s founder has said he does not want to charge users, which leaves other sources of revenue, advertising among them.[4] The assistant will be free, which in Silicon Valley is rarely the same thing as costless. A system acting on my

behalf should let me understand its commercial relationships, withdraw its permissions and keep a usable account of what it did. Affection for the interface is no substitute. A charming intermediary is still an intermediary, and the first question to ask of any intermediary is who pays it.

The Civil Code has an observation about money of its own, and it deserves a smile. Article 1992 makes the agent answer for its faults, but applies that responsibility less rigorously when the mandate is unpaid than when the agent receives a salary.[10] Two centuries on, the most popular agents are free. Somewhere in a pricing meeting, this is about to become a feature.

Public trust will turn on these questions as much as on benchmarks. People may welcome a personal assistant and oppose an employer's agent that runs their schedule; accept a system's competence and reject the terms on which its owner uses it. A dispute about power does not become a misunderstanding of technology because technology is involved. At its extremes the atmosphere has already turned ugly. In April 2026 federal prosecutors charged a man who, they allege, threw a Molotov cocktail at the San Francisco home of the chief executive of a major AI company, identified in press reports as Sam Altman, and then threatened to burn down the company's headquarters and kill anyone inside, after writing a document that argued against AI and for killing AI company executives and their investors. These are allegations in a criminal case, not established facts.[53][54] Violence against people is indefensible. It is also a poor portrait of the much larger population with concerns about AI, which does not reduce to one tribe: in Pew's survey of June 2025, 50 percent of Republicans and 51 percent of Democrats said they were more concerned than excited about AI in daily life.[55] Nor is Europe simply a continent trying to stop the machine. Alongside regulation, the Commission's AI Continent plan announced infrastructure, adoption and skills programmes, including InvestAI, an initiative to mobilize 200 billion euros for investment in AI: an ambition, not money spent, but hard to square with a story of universal refusal.[56]

None of this requires giving up a strong opinion about progress. Mine is simple. I do not need to remain better than a machine at every task to keep a worthwhile place in the world. If a system can do work better, and the gains can be made real, I want access to it. I want to test more ideas, discard weak ones sooner and make things that would otherwise stay in my head. Expertise should not become a protection racket in which society keeps inefficient procedures to preserve the status of those who mastered them. But nor should anyone be told that losing income, bargaining power or a way into a profession is merely a bruised ego. Those consequences are

material, and an honest case for adoption has to address them rather than offering aggregate productivity as a personal consolation prize.

Expertise, besides, keeps turning up in the evidence. Anthropic's economists studied some 400,000 sessions of its coding agent: sessions run by novices reached verified success about 15 percent of the time, those of people with at least intermediate expertise 28 to 33 percent; and people made about 70 percent of the planning decisions but only 20 percent of the execution decisions.[57] Preliminary findings from one product, with expertise rated by a model, and no permanent boundary between human and machine work. They do undermine the easy idea that an agent makes understanding the problem optional. The better possibility is a change in what expertise enables. A person who understands a problem can direct a much larger body of work; a specialist can explore adjacent approaches without mastering every implementation detail; a generalist can carry an idea far enough for a specialist to take it seriously. What matters is the difference between removing an unnecessary dependency and removing the judgment that made the work valuable.

Silicon Soulmate asked us to learn to lead this collaboration, and its chapters on management and mentoring still hold.[58] Two years later I would make the demand more explicit. Leadership now includes deciding which authority to transfer, how to tell whether the transfer is working and who will answer for its consequences. That is harder than keeping an approval button on every screen. Sometimes the responsible decision will be to let the system act without asking, because delay or repeated intervention makes the outcome worse. Sometimes it will be to stop despite a splendid record, because the situation has changed in a way the record does not cover. Neither decision can be outsourced to the word "autonomous".

The little shop was useful because the problem fitted in a room: we could see the products, the customers and the money. Beyond that room the same questions spread through a company and into other people's lives, behind the same friendly interface. In the mathematics essay, the gap ran between a result that could be checked and an understanding someone could carry.[19] Here another gap opens, between a system that can act and an institution entitled to let it act. Better intelligence narrows the first. It does not, by itself, supply the second.

So I come back to the assistant you text, and to my finger, still on the button. If my invitation arrives, I will accept it: I said at the start that I want

one, and nothing in this essay has cured me. But I would like to answer its terms in kind. They ask me to appoint the Services as my agent for whatever they "may deem" responsive. My reply would be shorter, and older. You may read, draft, compare and hold a booking. Ask me before you send, pay, sign or delete. Tell me what you did, including what went wrong and what you caught in time. Do not hand my errands to anyone I have not met. Tell me what you will keep of me when I let you go, since we both know it will not be nothing. And when I write STOP, I mean you, not the text messages.

The jurists of 1804 would recognize every line but one; they wrote a whole title of their Code for people who wanted exactly what I want. A meaningful act of trust must leave behind something more durable than a click: a mandate that can be read, revoked and accounted for.

> *If an agent can act without asking, what should force it to ask?*

Sources

Each source says what it can carry: a peer-reviewed study, a preprint not yet reviewed, a book, a research organisation, an institution, a company's own announcement or report, code, or a press report. Facts quoted here were read at the source in October 2026; each entry gives its own date.

[1] Instinct (Spear Street Technology). Instinct. instinct.com, read 25 October 2026. Company description.

[2] Lucas Ropek. Viral AI startup Instinct has raised $350M at a $2.5B valuation. TechCrunch, 26 August 2026. Press report.

[3] Ivan Mehta. Rival AI agents, Instinct and Meta's Muse, both add the ability to make calls. TechCrunch, 17 September 2026. Press report.

[4] PYMNTS. Instinct AI Assistant Targets $10 Billion Valuation. PYMNTS, citing The Information, 16 September 2026. Press report.

[5] Ivan Mehta. Instinct founder said more than 50% of transactions on the platform are travel-related. TechCrunch, 29 September 2026. Press report.

[6] Charles Rollet. Meet Noah Shinn, whose Instinct AI competes with Meta's Muse. Business Insider, 22 September 2026. Press report.

[7] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv 2303.11366; NeurIPS 2023, 2023. Peer-reviewed conference paper.

[8] Spear Street Technology (Instinct). Terms of Service. instinct.com, last revised 26 August 2026, read 26 October 2026. Company terms.

[9] Sarah Perez. Instinct's powerful AI assistant is raising privacy and security concerns. TechCrunch, 24 August 2026. Press report.

[10] Code civil des Français. Livre III, Titre XIII: Du mandat (articles 1984 to 2010). Wikisource, text of 1804, 1804. Law, original text.

[11] Noam Kolt. Governing AI Agents. Notre Dame Law Review 101, 335; SSRN 4772956, 2026. Law review article.
[12] Reda Kohen, Cabinet Kohen Avocats. Intelligence artificielle générative : quand l'agent réserve et achète seul, qui paie en cas d'erreur ?. Kohen Avocats, Paris (the note says it was generated by AI under a process the firm designed and controls), 13 October 2026. Law firm note.
[13] Anthropic, with Andon Labs. Project Vend: Can Claude run a small shop? (And why does that matter?). Anthropic, 27 June 2025. Company field experiment.
[14] Anthropic, with Andon Labs. Project Vend: Phase two. Anthropic, 18 December 2025. Company field experiment.
[15] Christophe Pennetier. Silicon Soulmate, chapter 7: Ghost in the Machine: The Silent Symphony of AI. Aquia, Collected (first edition May 2024), 2024. Author's book.
[16] Christophe Pennetier and André-Benoît De Jaegere. Trail 3: AI as Your Problem-Solving Ally. Generative Engineering, the original study, written November 2024 to February 2025. The original study.
[17] Erik Brynjolfsson, Danielle Li and Lindsey R. Raymond. Generative AI at Work. Quarterly Journal of Economics 140(2), 889–942, 2025. Peer-reviewed study.
[18] Paweł Gmyrek, Janine Berg and colleagues, International Labour Organization. Generative AI and Jobs: A Refined Global Index of Occupational Exposure. ILO Working Paper 140, 20 May 2025. Institution.
[19] Christophe Pennetier. When AI Starts Doing Mathematics. Generative Engineering, Volume II, Essay 01, October 2026. Volume II essay.
[20] Fabrizio Dell'Acqua and colleagues. The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise. NBER Working Paper 33641, April 2025. Working paper, preregistered field experiment.
[21] Michelle Vaccaro, Abdullah Almaatouq and Thomas Malone. When combinations of humans and AI are useful: a systematic review and meta-analysis. Nature Human Behaviour 8, 2293–2303, 2024. Peer-reviewed review.
[22] Christophe Pennetier. Tell Me How You Prompt. I'll Tell You What Kind of Manager You Are.. Aquia, Collected, 24 February 2026. Author's essay.
[23] Anthropic. Release notes (Cowork: 9 April, 7 July and 16 September 2026). Claude Help Center, read 25 October 2026. Company documentation.
[24] Meta. Introducing Muse: The World's First Personal AI Agent Built for Everyone. Meta Newsroom, 8 September 2026, updated 30 September 2026. Company announcement.
[25] OpenAI. Introducing dots. OpenAI, 29 September 2026. Company announcement.
[26] OpenAI. How we build safety, security, and privacy into dots. OpenAI, 29 September 2026. Company security account.
[27] Anthropic. Introducing Claude Duet. Claude, 21 October 2026. Company announcement, research preview.
[28] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger and Lalana Kagal. Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. CHI 2020; arXiv 2001.02479, 2020. Peer-reviewed conference paper.
[29] Berkeley J. Dietvorst, Joseph P. Simmons and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General 144(1), 114–126, 2015. Peer-reviewed study.
[30] Murray Campbell, A. Joseph Hoane Jr. and Feng-hsiung Hsu. Deep Blue. Artificial Intelligence 134(1–2), 57–83, 2002. Peer-reviewed study.
[31] David Silver, Thomas Hubert, Julian Schrittwieser and Demis Hassabis. AlphaZero: Shedding new light on chess, shogi, and Go. Google DeepMind, with the paper in Science 362, 1140–1144, 6 December 2018. Researchers' account of a peer-reviewed study.
[32] Waymo. Safety Impact (rider-only miles through June 2026). Waymo, read 25 October 2026. Company safety analysis.
[33] Chris Gamble and Jim Gao, DeepMind. Safety-first AI for autonomous data centre cooling and industrial control. DeepMind, 17 August 2018. Company deployment account.

[34] Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer and Florian Tramèr. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. arXiv 2406.13352, 2024. Preprint.
[35] Shubham Saboo and Kristopher Overholt, Google. Developer's Guide to AI Agent Protocols. Google Developers Blog, 18 March 2026. Technical documentation.
[36] Kevin K. Troy, Dylan Shields, Keir Bradwell and Peter McCrory, Anthropic. Project Deal: our Claude-run marketplace experiment. Anthropic, 24 April 2026. Company field experiment.
[37] Lisanne Bainbridge. Ironies of automation. Automatica 19 (6), 775–779, 1983. Peer-reviewed study.
[38] Miles McCain, Thomas Millar, Saffron Huang and colleagues, Anthropic. Measuring AI agent autonomy in practice. Anthropic, 18 February 2026. Company observational study.
[39] METR. Task-completion time horizons. METR, updated 8 September 2026. Research organisation.
[40] Thomas Kwa, METR. Clarifying limitations of time horizon. METR notes (the author's views, not necessarily METR's), 22 January 2026. Researcher's note.
[41] Christophe Pennetier. The Iron Man Paradox. Aquia, Collected, 12 August 2026. Author's essay.
[42] Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange and Jeff Clune. Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents. arXiv 2505.22954 (revised March 2026), 2025. Preprint.
[43] Dhruv Srikanth, Bingchen Zhao, Dixing Xu, Yuxiang Wu and Zhengyao Jiang, Weco AI. Recursive self-improvement of AI research agents. arXiv 2609.26457, 22 September 2026. Preprint.
[44] GV. Recursive Superintelligence: self-improving AI. GV, 13 May 2026. Investor announcement.
[45] Recursive Superintelligence. Recursive. recursive.com, read 25 October 2026. Company description.
[46] Recursive. Introducing Flint. Recursive, 23 October 2026. Company announcement.
[47] Ricursive Intelligence. Ricursive Intelligence Raises $300 Million Series A at $4 Billion Valuation to Accelerate AI-Driven Semiconductor Design. PR Newswire, 26 January 2026. Company announcement.
[48] Christophe Pennetier and André-Benoît De Jaegere. Circuit Design and Optimization: In the Spotlight, Azalia Mirhoseini. Generative Engineering, the original study, written November 2024 to February 2025. The original study.
[49] Christophe Pennetier. The Ouroboros Protocol. Aquia, Collected, 26 January 2026. Author's essay.
[50] U.S. Securities and Exchange Commission. SEC Charges Knight Capital With Violations of Market Access Rule. Press release 2013-222, 16 October 2013. Institution.
[51] Franziska Roesner and Tadayoshi Kohno. Reflections on Trusting Trust, Revisited: Contaminating Self-Modifying AI Coding Agents with Poisoned Benchmarks. arXiv 2609.17817, 15 September 2026. Preprint.
[52] Madeleine Clare Elish. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. Engaging Science, Technology, and Society 5, 40–60, 2019. Peer-reviewed study.
[53] U.S. Department of Justice, Office of Public Affairs. Texas Man Faces Multiple Federal Charges Related To Attack on AI Company and its CEO. Press release 26-353, 13 April 2026. Institution (charges, not a conviction).
[54] Reuters. Man charged after Molotov cocktail attack on OpenAI CEO Sam Altman's home. Reuters, 13 April 2026. Press report.
[55] Monica Anderson and William Bishop, Pew Research Center. Republicans, Democrats now equally concerned about AI in daily life, but views on regulation differ. Pew Research Center (survey of June 2025), 6 November 2025. Research organisation.
[56] European Commission. AI Continent Action Plan: Q&A. Shaping Europe's digital future, 9 April 2025. Institution.

[57] Zoe Hitzig, Maxim Massenkoff, Eva Lyubich, Shaoyi Zhang, Ryan Heller and Peter McCrory, Anthropic. Agentic coding and persistent returns to expertise. Anthropic, 16 June 2026. Company observational study.
[58] Christophe Pennetier. Silicon Soulmate, chapter 8: Pirates of the Algorithm: Steering the Ship of AI's Wild Wisdom. Aquia, Collected (first edition May 2024), 2024. Author's book.