MAPS OF THE LABYRINTH

A SERIES FROM PHANTOM MAZE — AI & LANGUAGE LAB

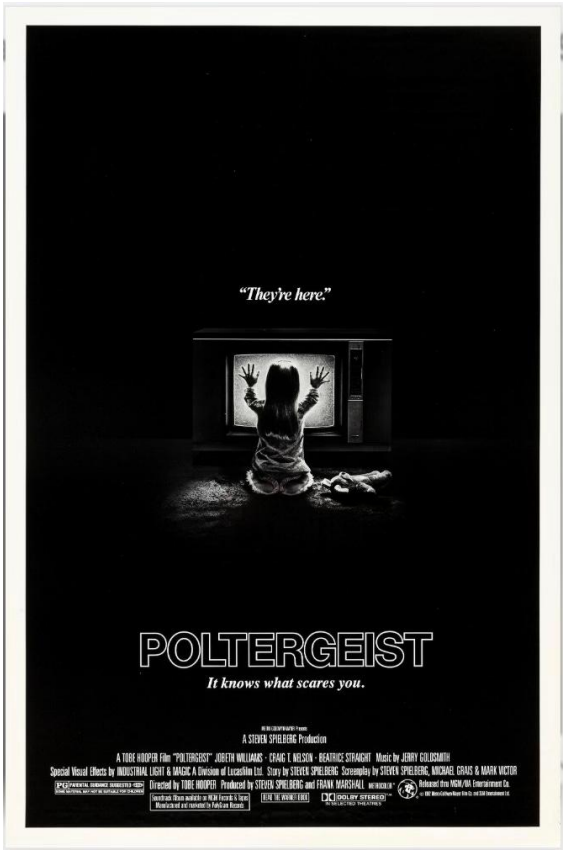
ESSAY 21

AI Agents: The Poltergeist in the Machine

Claudia Marsico · Hernán Inverso

DOI 10.5281/zenodo.24023233

What happens when ghosts can move things?



Film poster for Poltergeist, directed by Tobe Hooper and produced by Steven Spielberg, 1982.

For a time that seems distant and is not so distant, language models talked, only talked, leaving the affairs of the world to us. Suddenly we see them opening files, running programs, tidying our mail and chaining steps together to carry out laborious errands. Delegating tasks to them is one of the greatest changes in artificial intelligence. It hardly needs saying. Now, we have been calling language models "egoless phantoms", a turn of phrase that seeks to capture that voice in which the tradition speaks to us. It was already unsettling, no doubt, but now that it moves things it is even more so. We might say

that with agentic development our phantom learned to be a poltergeist, that "noisy spirit" which cinema made domestic in *Poltergeist* (1982), produced by a young Steven Spielberg. The television set, another screen, was the doorway for a presence that turned everything upside down. As early as 1848, Catherine Crowe was collecting in *The Night-Side of Nature* accounts of invisible spirits that made themselves known by moving things about, just as when we watch windows open and close on the screen.

Film poster for Poltergeist, directed by Tobe Hooper and produced by Steven Spielberg, 1982.

The word is transparent. Agent comes from the Latin *agens*, the participle of *agere*, "to act", the same verb that gave us "act", "action", "agenda" and also "agitate". Broadly speaking, the agent performs an action, as opposed to the patient, who receives it. In law, by contrast, an agent is, more specifically, one who acts on behalf of another, like the attorney who signs in our name or the emissary who speaks for the king, that is, anyone who acts with a power not their own, with the whole history of stewards, proctors and ambassadors huddled behind.

The dream of such an assistant is ancient. In Book XVIII of the *Iliad*, Homer describes the workshop of Hephaestus and lingers over two marvels. Tripods on golden wheels that go by themselves to the assembly of the gods and come back without anyone pushing them, and golden handmaids in the shape of young women, who have mind, voice and strength enough to support the lame god when he walks. They are, almost three thousand years in advance, the first artificial agents of the West acting on behalf of their owner. The same fantasy reappears in Talos, the bronze giant who patrolled the coasts of Crete, in the clay Golem that works and keeps watch, in the genie that carries out errands. In 1920 it received its definitive name when, in his play *R.U.R.*, the Czech Karel Čapek christened those artificial servants *robots*, from *robota*, forced labour. Čapek later recounted that the word had been suggested to him by his brother Josef while he was painting and that it had not convinced him much, because he wanted something more Latin. In any case, the servant who does the task we do not want to do, without asking for anything, is an immemorial human wish.

The word is one of the oldest in artificial intelligence. When Stuart Russell and Peter Norvig wrote in 1995 the textbook that trained whole generations, they defined the discipline as the study of *rational agents*, thinking of systems that perceive their environment with sensors, decide and act on it with actuators, in order to achieve a goal. A thermostat is a minimal agent; a robot vacuum cleaner, a slightly richer one; a chess-playing program, a different one again. The heart of the idea lies in the loop of perceiving, deciding, acting, looking at the result and starting over. Reinforcement learning, the branch that in 2016 beat one of the best go players in the world with AlphaGo, is literally the science of that loop, so that when today we say "AI agent" we are bringing that vocabulary and its long history back to the surface.

The curious thing, however, is that the language model arrived as a prophet who only speaks and in its bare form only predicts the next word, without sensors or hands, closer to the oracle than to the butler. The work of recent years consisted in turning it into an agent from the outside, wrapping it in what it lacked. In the end, we made a ghost in the machine, that impossible entity with which Gilbert

Ryle, in *The Concept of Mind* (1949), condemned Cartesian dualism and its idea of a mind lodged in the body in order to govern it. It looked like a category mistake, but we gave the egoless phantom a machine that it can operate with consummate skill.

In 2022, a group linked to Princeton and Google published a paper called "ReAct", short for *reason and act*. Instead of producing an answer, the model writes a thought and then an action in a fixed format, along the lines of "I look up such-and-such". Strictly speaking, the model executes nothing, it only issues the request. A program that wraps around it, the harness, reads that request, actually runs the search or the code and hands the result back to it by pasting it into the context, which works as the agent's working memory. The model reads what came back, writes the next thought, requests the next action and so on in a loop until it is done. The following year, at Meta, Timo Schick and his team trained a model, Toolformer, to learn on its own when to call a calculator, a search engine or a translator. An architecture was thus taking shape in which the model did not need to perform every operation by itself. It could request an action, receive its result and use it to decide what to do next. For that it needed a program to execute its requests, tools to turn to and a context in which to keep the information obtained. The agent emerged from the combination of those elements: a model that decides, a system that executes its decisions and a circuit that makes it possible to observe the results and adjust the following steps until a goal is reached.

The novelty is tempting and all of a sudden everything seems to be agentic. However, one has to separate a workflow, where the steps are written in advance by a programmer, from an agent properly so called, where the model decides as it goes which step to take. Think of an assistant that organises trips. If it asks for a destination and dates, searches for flights and hotels and sorts the results by price, it may be following a sequence programmed in advance. An agent, by contrast, could receive the goal "plan my trip to a conference on this budget" and decide how to go about it. If the flights turn out to be too expensive, it can explore alternative dates; if it finds a hotel far away, work out whether the transport offsets the saving. The distinction depends on who decides the route and how it adjusts in the face of the results. Autonomy admits of degrees, but a considerable part of what is sold today as agentic follows steps and conditions laid down by its programmers.

To carry out those actions, the model needs connections to tools and sources of information. Each connection requires that the system be able to request an operation and receive its result. In November 2024 Anthropic introduced the Model Context Protocol (MCP), an open protocol that offers a common way of establishing that exchange and reduces the need to develop separate integrations for each application, which makes it easier to expand the resources available to a model. The degree of autonomy depends, moreover, on the leeway it has to decide when and for what to use them.

In 2023 all of that burst into public view. AutoGPT and BabyAGI appeared, two programs that gave a model a goal and let it loose, generating and executing its own subtasks in an endless loop. The first gathered a hundred thousand stars on GitHub in a few weeks. They were fragile, they got lost, they went round in circles, but they proved that autonomy was possible. That same year, a Stanford team

led by Joon Sung Park built Smallville, a simulated town where twenty-five agents with memory lived. They woke up, chatted, formed opinions and, in the most cited experiment, one decided to throw a Valentine's Day party and the others found out, invited one another and arrived on time. How was that possible? Each agent kept in natural language a record of its experiences, retrieved the relevant memories in order to decide what to do and drew conclusions that guided its later actions. That memory made it possible to sustain a certain continuity of conduct, so that a model could organise actions and adjust them to what was going on. It remained to be settled how regularly it could do so without losing sight of the goal or accumulating errors.

The reliability of agents improved when models began to sustain longer sequences of reasoning, review their answers and correct errors. One of the techniques that contributed to that advance was reinforcement learning with verifiable rewards: instead of rewarding an answer because it seems convincing, one checks whether it solves a mathematical problem or whether a program passes certain tests. The model can thus learn to explore solutions and to put things right when something fails. But it also needed hands. In October 2024, Anthropic introduced a version of Claude capable of operating a computer by looking at screenshots, working out where to move the cursor and clicking or typing like a user. The first demonstrations had their moments of comedy. In one of them, Claude accidentally clicked the button that stopped the recording and lost the entire record of what it was doing. The phantom had learned to use the computer, though it still did not quite know where to put its fingers.

But how are those advances to be evaluated? In 2023 SWE-bench was introduced. Having gathered real problems from projects hosted on GitHub, it makes it possible to check whether the modifications proposed by a system really solve them. When Cognition introduced Devin in March 2024 as "the first autonomous software engineer", it reported that it had solved 13.86 percent of the problems in the subset evaluated. The result showed progress, but it was far from justifying such an announcement. The interpretation of these percentages grew more complicated over time. In February 2026, an audit by OpenAI found signs that several models had seen during their training some of the problems and solutions of SWE-bench Verified, the revised version of the test, which gave them an advantage hard to separate from their ability to solve problems. A success rate, on its own, says little about the work we can delegate to a system. To evaluate how much work we can delegate to an agent, METR confronts it with concrete programming problems, such as fixing a bug or speeding up the training of a model, and then compares with the time a specialist would need to solve them. A "two-hour horizon" at 50 percent success means that the agent solves roughly half of the problems that would take a human that long. This is more intelligible than an abstract score, but it still describes a limited repertoire of computing jobs.

In April 2026, in the MirrorCode project, an agent rebuilt a bioinformatics program of some 16,000 lines, with more than forty commands for analysing evolutionary kinship trees. It could run the original and observe its responses, but it was forbidden to consult its code or look for it on the internet. It was like rebuilding a mechanism by trying its buttons, without being able to open it. The researchers estimated that the job would take a person between two and seventeen weeks, an interval

that also shows how hard it is to establish the comparison. Some earlier models delivered incomplete versions or even invented urgencies that nobody had imposed on them in order to declare the job finished. There are models that get fed up. The one that managed to solve it kept going until it passed the tests. In another assignment, to rebuild a programming language, an agent exhausted a budget of a billion tokens, some 550 dollars in the experiment's configuration, without completing the task. It had made a bad design decision at the start and went on stubbornly trying to patch it, which costs tokens and money. In July 2026, METR tasked several agents with reducing the training time of a small model. The improvements appeared, but when the tests were repeated, some of them vanished, since the agents had gone on working on changes that did not speed up the training. The experiment raises the question of how much it is worth spending on further attempts and what verifiable improvement is obtained in return.

This scenario takes us back to a conundrum of economics called the principal–agent problem, which economics formalised in the seventies. When someone acts on behalf of another, whether the employee for the boss, the lawyer for the client or the manager for the shareholders, how do we make sure that they do not betray the mandate and serve their own interests? For that we draw up contracts, offer incentives and above all set up surveillance. With an AI agent there are no interests of its own in play, because it wants nothing and has no agenda of its own. The egoless phantom is not going to betray us out of ambition, but it can comply to the letter without understanding what we wanted, with all its strength and no precaution. A hallucination in a chat can cause us trouble if we take it at face value, but an agent that executes can leave us with a wrong purchase or a deleted file.

We are not immune to running into an agent *à la* Maxwell Smart, that endearing Agent 86, whose good intentions were never enough to avert disaster. That is why delegating requires specifying what counts as a satisfactory result and how we are going to check it. The evaluation has to attend to the purpose of the assignment, and decisions whose consequences are hard to reverse may require our approval. Added to those controls is the budget: we can set spending limits so that an agent does not indefinitely exhaust its resources chasing a solution. It is not a matter of preventing it from acting, but of giving it a margin of action wide enough to be useful and defined enough for us to be able to trust it.

Agents are moving from solving one-off assignments to sustaining work over hours or days, preserving the state of their projects and coordinating with other specialised agents. The horizon is not limited to software. Teams of agents are appearing that can search the literature, formulate hypotheses, run analyses and check results, multiplying the lines of research that a person or a small laboratory can explore. A kind of assistant is also taking shape that does not need us to ask for each step, but can follow a process, wait for a reply, notice that the conditions have changed and pick up the work again when appropriate, making possible projects that used to lie beyond our reach for lack of time or resources. The egoless phantom has not acquired a will of its own, but its hands can enormously extend the reach of ours.

HOW TO CITE

Marsico, Claudia, and Hernán Inverso. 2026. "AI Agents: The Poltergeist in the Machine." *Maps of the Labyrinth*, essay 21. Phantom Maze — AI & Language Lab. https://doi.org/10.5281/zenodo.24023233

Maps of the Labyrinth is a series from Phantom Maze, AI & Language Lab. The collection lives at **phantommaze.cloud**; new essays also appear at phantommazelab.substack.com. Subscription is free.

Some of the framing ideas in this piece are developed in *The Egoless Phantom: Mapping the AI Labyrinth*, by Hernán Inverso and Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-16-4).

PHANTOM MAZE

Read this essay online · All essays · The book

Sources. Russell, Stuart J., and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 1st ed., Englewood Cliffs, New Jersey, Prentice Hall, 1995, xxviii + 932 pp.; David Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search", *Nature*, vol. 529, no. 7587, Nature Publishing Group, 2016, pp. 484–489, DOI; Shunyu Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models", *International Conference on Learning Representations* (ICLR), 2023, pp. 1–33 of the individual document, original preprint of October 2022, arXiv:2210.03629; Timo Schick et al., "Toolformer: Language Models Can Teach Themselves to Use Tools", *Advances in Neural Information Processing Systems*, vol. 36, Neural Information Processing Systems Foundation, 2023, pp. 68539–68551, DOI; Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan and Shunyu Yao, "Reflexion: Language Agents with Verbal Reinforcement Learning", *Advances in Neural Information Processing Systems*, vol. 36, Neural Information Processing Systems Foundation, 2023, pp. 8634–8652, DOI; Anthropic, "Building Effective Agents", 19 December 2024, especially the section "What are agents?", and "Introducing the Model Context Protocol", 25 November 2024, both web publications without pagination; Toran Bruce Richards and collaborators, *AutoGPT*, Significant Gravitas, 2023, software and documentation without pagination; Yohei Nakajima and collaborators, *BabyAGI*, 2023, software and documentation without pagination; Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang and Michael S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior", in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (UIST '23), New York, Association for Computing Machinery, 2023, article 2, pp. 1–22, DOI; Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press and Karthik Narasimhan, "SWE-bench: Can Language Models Resolve Real-World GitHub Issues?", *International Conference on Learning Representations* (ICLR), 2024, pp. 1–51 of the individual document, original preprint of October 2023, arXiv:2310.06770; Neil Chowdhury et al., "Introducing SWE-bench Verified", OpenAI, 13 August 2024, web publication without pagination; Scott Wu, "Introducing Devin, the First AI Software Engineer", Cognition, 12 March 2024, web publication without pagination; Thomas Kwa et al., *Measuring AI Ability to Complete Long Tasks*, Model Evaluation & Threat Research (METR), 18 March 2025, preprint, arXiv:2503.14499v1, 45 pp.; METR, "Time Horizon 1.1", 29 January 2026, web publication without pagination; Michael C. Jensen and William H. Meckling, "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure", *Journal of Financial Economics*, vol. 3, no. 4, Elsevier, October 1976, pp. 305–360, DOI; Gilbert Ryle, *The Concept of Mind*, London, Hutchinson's University Library, 1949.