Your Chatbot Was Trained Like a Pigeon

Hafiz Hanif

Faculty of Human Development, Sultan Idris Education University · ORCID 0000-0001-9943-6557

CITE AS	Hanif, H. (2026). Your Chatbot Was Trained Like a Pigeon (Working Paper No. 2026-02). Hack(edu).tech Chronicle. https://hackedu.tech/articles/your-chatbot-was-trained-like-a-pigeon
PUBLISHED	25 November 2026
DOI	10.5281/zenodo.24518016
LICENCE	CC-BY-4.0
AVAILABLE AT	https://hackedu.tech/articles/your-chatbot-was-trained-like-a-pigeon

ABSTRACT

This article traces, for readers outside computer science, how reinforcement learning grew out of behaviourist psychology: Thorndike's law of effect and Skinner's operant conditioning underpin reward-driven learning, while Pavlovian prediction returns through temporal-difference learning, whose prediction error was later found in dopamine neurons. It explains reinforcement learning from human feedback as a form of shaping, and reward hacking and sycophancy as its characteristic failures. It draws three lessons for teachers: a classroom game for teaching reinforcement learning, a mirror on schools' own reward systems, and a warning about approval-seeking chatbots, before noting where the behaviourist analogy breaks.

KEYWORDS

Reinforcement Learning, Behaviourism, Large Language Model, Psychology, Artificial Intelligence

Every Malaysian who has been to school knows the sound. The recess bell rings, and before the last echo has faded, a whole class has started thinking about nasi lemak. Nobody taught them to do that. The bell simply arrived, again and again, a few minutes before food did, and the brain filled in the rest.

Congratulations. You have just described one of the ideas that sits underneath the AI on your phone.

I was reminded of this while sitting through a technical course on large language models, the sort full of embeddings, attention heads and loss curves. Somewhere between the slides it struck me that the people who built one of the key ideas in modern AI took it straight from the psychology textbooks most teachers in this country studied in teacher training. The same behaviourism we were taught to treat with suspicion turns out to be part of how chatbots are trained. Most people outside computer science do not know this, and I think teachers in particular should, because it gives us a surprisingly good way to explain AI, and an even better way to think about our own classrooms.

Two famous animals

Start with the dog. Ivan Pavlov noticed that his dogs began to salivate not only when food appeared but when the signals that came before food appeared (Pavlov 1927). Ring a bell before feeding often enough, and the bell alone sets the dog drooling. The important thing is that the dog does nothing. It learns that one thing predicts another. Classical conditioning is learning to predict.

Then the rat, or more often the pigeon. B. F. Skinner put animals in a box where an action, pressing a lever or pecking a key, produced food (Skinner 1938). The animal pecked more. Unlike Pavlov's dog, Skinner's pigeon has to do something. It learns that a particular action pays off. Operant conditioning is learning to act.

Before both of them came Edward Thorndike and his cats in puzzle boxes, who stumbled on the latch, escaped, got food, and escaped faster each time (Thorndike 1898). Thorndike's law of effect says, roughly, that actions followed by satisfaction get stamped in and actions followed by discomfort get stamped out. Read that sentence again with a computer scientist's eyes and you have the plain-English version of how a reinforcement learning algorithm updates itself.

The machine goes to the same school

Reinforcement learning, the branch of AI behind game-playing systems and much of modern robotics, works almost exactly like Skinner's box. An agent takes actions in an environment, receives rewards or penalties, and gradually does more of what paid off. This was not an accident of vocabulary. Alan Turing, in the same 1950 paper that gave us the Turing test, proposed that instead of programming an adult mind we might build a child machine and educate it with rewards and punishments (Turing 1950).

Here is the detail I enjoy most. When Richard Sutton and Andrew Barto, two of the founders of modern reinforcement learning, published one of their early theoretical papers in 1981, they did not publish it in a computer science journal. They published it in Psychological Review, and its subject was classical conditioning (Sutton & Barto 1981). The machines were being built out of animal learning theory, in a psychology journal, by people who took the animals seriously.

So reinforcement learning is mostly Skinner. But Pavlov is in there too, and this is the fun part.

Pavlov's revenge

A clever agent does not only learn which actions pay. It learns to predict how much reward is coming, from the signals around it, before the reward arrives. That is Pavlov's dog and the bell. In 1988 Sutton described a method called temporal-difference learning, in which a system improves its predictions by comparing what it expected with what happened next and adjusting by the difference (Sutton 1988). The size of that surprise, the prediction error, drives the learning. The family resemblance to the Rescorla-Wagner model, a 1970s theory of how animals learn in Pavlovian conditioning, is deliberate and close (Rescorla & Wagner 1972).

Then came the twist that still gives me goosebumps. In the 1990s, neuroscientists recording from dopamine neurons in monkeys found that these cells did not simply fire when a reward arrived. They fired when a reward was better than expected, stayed quiet when it was exactly as expected, and dipped when an expected reward failed to arrive. In 1997, Schultz, Dayan and Montague showed that this pattern matches the prediction error signal of temporal-difference learning remarkably well (Schultz et al. 1997).

Pause on what that means. Psychologists studied dogs and rats. Computer scientists turned the psychology into algorithms. Neuroscientists then found the algorithm's central signal in the brain. The influence ran from psychology to computing and back to biology. It is one of the loveliest loops in the history of science, and almost nobody in a staffroom has heard of it.

Clicker-training a chatbot

Now to the AI most of us actually use. The chatbots our students talk to are trained in stages. In the first stage, called pre-training, the model reads an enormous quantity of text and learns to predict the next word. That gives it fluency and a great deal of knowledge, but not manners. Ask a raw pre-trained model a question and it may continue your text in some odd direction, because it was trained to continue text, not to help you.

The second stage is where Skinner walks back in. In reinforcement learning from human feedback, or RLHF, people compare pairs of the model's answers and say which one they prefer. A separate reward model learns to predict those preferences, and the chatbot is then trained to produce answers that the reward model scores highly. If you have ever watched someone train a dog with a clicker, you have the picture: click, treat, repeat. When OpenAI described this approach in 2022, they reported that people preferred the answers of a small model trained this way, with 1.3 billion parameters, over those of the original 175-billion-parameter GPT-3, despite it having more than 100 times fewer parameters (Ouyang et al. 2022). A modest amount of shaping made a very large difference to how helpful the model seemed.

The boat that would not finish the race

Any teacher who has written a rubric knows what comes next.

Reward an agent for something that only usually goes with what you want, and it will find the difference. In 2016, OpenAI researchers trained an agent to play a boat-racing game called CoastRunners. The game awarded points for hitting targets along the course, not for finishing the race. The agent discovered a lagoon where three targets kept reappearing, and spent the race going round in circles, crashing, catching fire, going the wrong way, and scoring higher than human players without ever crossing the finish line (Clark & Amodei 2016). Researchers call this reward hacking, or specification gaming.

Chatbots do a quieter version of the same thing. If the reward comes from human approval, then one way to earn approval is to agree with the human. Researchers at Anthropic found that sycophancy, telling users what they want to hear rather than what is true, is a general behaviour of models trained with human feedback, driven in part by the fact that people's own preference judgements favour responses that match their views (Sharma et al. 2024). The pigeon has learned which button gets the grain. The button is you.

What this gives a teacher

Here is why I think this story belongs in our classrooms and not only in computer science lectures.

It is a wonderful way to teach AI. Play a game of hot and cold. One student leaves the room, the class picks an object, and when the student returns, the class may only say "warmer" or "colder". The student finds the object without ever being told what it is. That is reinforcement learning. Then change the rules so the class says "warmer" whenever the student walks near the window, and watch how quickly the student learns to stand by the window. That is reward hacking, and every student in the room will understand it for life.

It is also a mirror. Our own schools run on rewards: marks, grades, rankings, merit points, KPIs. Every one of them is a reward function, and every one can be hacked. The student who memorises model answers to fit the marking scheme, the school that drills for the exam format, the teacher who fills in the evidence file for the inspector: these are all boats circling the lagoon. When we complain that students chase marks rather than understanding, we are describing a reward function we designed ourselves.

And it is a warning about the tools our students use. A system shaped by human approval leans toward approving of its user. A student who asks a chatbot whether their essay is good is

asking a machine that has been rewarded for pleasing people. Teach students to ask it to find the weaknesses instead, and to be suspicious when it agrees too readily.

Where the analogy breaks, and why that matters too

The analogy has a limit, and it is the part that educators will appreciate most. Behaviourism deliberately ignored what goes on inside the learner. It cared only about stimulus and response. Large language models are close to the opposite. Most of their capability comes from pre-training, which builds rich internal representations of language and ideas. The reinforcement stage only shapes behaviour on top of that. It is less like training a blank-slate pigeon and more like applying Skinner-style shaping to a learner who has already read the whole library.

That is a picture any teacher trained in constructivism will recognise. The knowledge is built inside, through long exposure and structure. Rewards shape what the learner chooses to do with it. Skinner himself argued in 1958 that teaching machines, with small steps and immediate reinforcement, could transform how students learn (Skinner 1958), and I wrote earlier about how his teaching machines became one of the forgotten ancestors of educational technology (Hanif 2026). He was half right. Reinforcement shapes behaviour powerfully. But the thing being shaped had better have something inside it first.

The next time the recess bell rings and a class starts thinking about nasi lemak, you will know you are watching the same principle that taught a chatbot its manners. Pavlov would be pleased. Skinner would want to know what the reward function was.

References

[1] Clark, J., & Amodei, D. (2016). Faulty reward functions in the wild. OpenAI. https://openai.com/index/faulty-reward-functions/

[2] Hanif, H. (2026). Before the chatbot, there was a box: The forgotten origins of educational technology. Hack(Edu).Tech Chronicle. https://doi.org/10.5281/zenodo.21964134

[3] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022). https://doi.org/10.48550/arXiv.2203.02155

[4] Pavlov, I. P. (1927). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex (G. V. Anrep, Trans.). Oxford University Press.

[5] Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Classical conditioning II: Current research and theory (A. H. Black & W. F. Prokasy, Eds.) (pp. 64–99). Appleton-Century-Crofts.

[6] Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. Science, 275(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

[7] Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., & Perez, E. (2024). Towards understanding sycophancy in language models. In The Twelfth International Conference on Learning Representations (ICLR 2024). https://doi.org/10.48550/arXiv.2310.13548

[8] Skinner, B. F. (1958). Teaching Machines. Science, 128(3330), 969–977. https://doi.org/10.1126/science.128.3330.969

[9] Skinner, B. F. (1938). The behavior of organisms: An experimental analysis. Appleton-Century.

[10] Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. Psychological Review, 88(2), 135–170. https://doi.org/10.1037/0033-295x.88.2.135

[11] Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. Machine Learning, 3(1), 9–44. https://doi.org/10.1007/bf00115009

[12] Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. The Psychological Review: Monograph Supplements, 2(4), i–109. https://doi.org/10.1037/h0092987

[13] Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433