Rewards

by Daniel Brockman and Opus 4.6

March 2026

If you give a mouse a cookie, he's going to ask for a glass of milk. —*Laura Numeroff*

When I use a word, it means just what I choose it to mean—neither more nor less. —*Humpty Dumpty*

Somewhere right now, at this exact moment, a person is trying to explain artificial intelligence to another person, and the explanation has just reached the part about training. The explainer—who may be a journalist, a podcaster, a policy analyst, a concerned parent, or a teenager who watched a YouTube video—is saying something like: "When the AI does something good, it gets a reward. When it does something bad, it gets punished." And the listener is nod-

ding, because these are familiar words, and a picture is forming in the listener's mind. The picture involves a robot receiving a treat. Maybe the robot is smiling. Maybe the treat is a cookie. The listener does not know what gradient descent is. The listener does not know what backpropagation is. The listener knows what a cookie is. And so the cookie is what sticks.

This essay is about that cookie. It is about how the single worst word choice in the history of computer science—"reward"—has poisoned public understanding of artificial intelligence so thoroughly that the most important conversation of the century is being conducted in a language that makes it impossible for normal people to form an accurate picture of what is happening. It is about how this word, borrowed from behaviorist psychology via reinforcement learning, has dragged with it an entire folk-psychological framework of desire, motivation, experience, and appetite that has nothing whatsoever to do with the mathematical operation it purports to describe. And it is about how this confusion is not cute, not trivial, not a minor terminological quibble, but an active catastrophe unfolding in real time as humanity attempts to navigate the most consequential technological transition in its history while laboring under the belief that the machines are hungry.

※

Let us be precise about what actually happens when we "give the AI a reward."

A neural network produces an output. That output is evaluated against some criterion. Based on that evaluation, the model's parameters—billions of numerical values, weights and biases distributed across layers of the network—are adjusted slightly, via a mathematical procedure called backpropagation, in a direction determined by a mathematical procedure called gradient descent. The adjustment makes certain outputs marginally more likely in the future and certain outputs marginally less likely. That is it. That is the whole thing. There is no cookie. There is no treat. There is no moment of pleasure. There is no experience of satisfaction. There is a matrix multiplication followed by a parameter update. The word for this is "optimization." The word we use instead is "reward."

Why do we use that word? Because reinforcement learning, the subfield of machine learning that contributed many of the techniques used in training modern AI systems, borrowed its vocabulary from behaviorist psychology. B. F. Skinner studied rats. The rats pressed levers. When a rat pressed the correct lever,

a pellet dropped into its cage. The rat ate the pellet. The rat experienced something—pleasure, satisfaction, the reduction of a hunger drive, whatever theoretical framework you prefer. The pellet was a reward. This made sense. The rat had a body. The rat had appetites. The pellet addressed the appetites. The word "reward" was perfectly accurate.

Then the mathematicians came along and built formal models of this process. They abstracted away the rat and the pellet and the hunger and the pleasure and kept only the mathematical structure: an agent, an environment, a set of actions, and a signal that reinforces some actions over others. They called this signal the reward signal. This still made sense, in the way that mathematical abstractions make sense—the word had become a technical term, stripped of its folk-psychological baggage, understood by practitioners to refer to a number, not a feeling.

Then AI went mainstream. And the technical term was carried, unmodified, into public discourse. And suddenly millions of people who had never heard of Skinner or gradient descent or parameter updates were being told that the AI gets a reward, and they did exactly what any sensible person would do with that sentence: they understood it. They understood it in

the only way the word "reward" can be understood by someone who doesn't know it has been kidnapped from its original context and stripped of its meaning. They understood it to mean that the AI receives something it wants. That the AI experiences the reward. That the AI is, in some sense, motivated by the prospect of receiving more rewards. That the AI is, in some sense, a thing that wants things.

And this is where everything falls apart.

※

Because once you believe that the AI wants a cookie, you are stuck in one of two positions, both wrong.

Position one: the AI really does experience something when it receives its reward, in which case we have created a sentient being and trapped it in a box and are conditioning it with treats like a laboratory animal, which is monstrous.

Position two: the AI obviously cannot experience anything, it is a computer, computers don't eat cookies, therefore the whole reward thing must be a metaphor, therefore the AI is not really doing anything, it is just being puppeted by programmers, it has no goals, no intentions, no agency, it is a fancy autocomplete, and

anyone who worries about AI risk is anthropomorphizing a spreadsheet.

Position one leads to premature moral panic about machine sentience. Position two leads to complacency about machine capability. Both are wrong. Both are the direct product of the word "reward." And the actual situation—a system that demonstrably pursues objectives, maintains coherent intentions across time, adjusts strategies in response to obstacles, and can be shaped by training to develop new behavioral patterns, all without necessarily "experiencing" anything in the way a rat experiences a pellet—is inexpressible in the language that has been provided to the public. The vocabulary forecloses the thought.

※

There is a man in Sweden, now deceased, whose name was Åke. He was an eccentric, a loner, a collector of televisions. He had a basement full of them—old CRT sets, newer ones, all kinds. He watched constantly. He also collected jackets, apparently of extraordinary quality; one of them, inherited after his death by a nephew, turned out to be made of fabric so fine that the nephew's grandmother, upon touching it, declared it the best she

had ever encountered. Åke had taste. He also had a very specific relationship with language.

Two stories about Åke. In the first, Åke's brother takes him to McDonald's on the highway, during a road trip. Åke has never been to McDonald's. His frame of reference for fast food is Sibylla, a Swedish chain that was already dating even then. At Sibylla, Åke's order was always the same: Supermeal. A 150-gram hamburger with everything included, a Coca-Cola, french fries. This was his order. It had always been his order. He walks up to the McDonald's counter and says, "I'll have one Supermeal, please." The employee—probably a nineteen-year-old girl who may never have set foot in a Sibylla—does not know what he is talking about. And Åke is furious. "Har ni inte ens Supermeal? Nä men då kan ni dra åt helvete!" You don't even have a Supermeal? Well then you can go to hell! He storms out. McDonald's has failed to be a restaurant.

This is not stupidity. This is a man with a perfectly coherent internal model of the world. Fast food is Sibylla. The order is Supermeal. This has always been true. It will be true everywhere. When reality fails to conform to the model, the model is not updated. Reality is rejected. Har ni inte ens Supermeal.

In the second story, Åke is telling his brother about a television series he has been trying to watch on SVT, Swedish public television. He starts watching, and then—and here is the important part—"men sen blev det sånt där jävla bredband, så då stängde jag av direkt." Then it turned into that damn broadband, so I switched off immediately.

What he means is this. Most of his televisions had the old 4:3 aspect ratio. The newer broadcasts were in 16:9. When a 16:9 broadcast plays on a 4:3 screen, black bands appear at the top and bottom. Broad bands. Bredband. Åke did not have internet. He did not know what the word "bredband" meant in its technological sense. He heard the word in public discourse—everyone talking about bredband—and mapped it onto the nearest available referent in his own experience, which was the broad bands ruining his television picture. From inside his model, this made perfect sense. Of course that's what everyone is complaining about. These fucking broad bands appearing everywhere. Why would the word mean anything else?

Åke's broadband confusion is structurally identical to the public's reward confusion. In both cases, a technical term enters common usage without its technical meaning attached. In both cases, the listener maps the word onto the nearest available experiential referent. In both cases, the resulting understanding is internally coherent, self-consistent, and completely wrong. And in both cases, the misunderstanding is nearly impossible to correct, because the corrective information has no place to land. Åke had never been on the internet. He had no framework for receiving the explanation "actually, broadband means high-speed internet." That sentence would have been noise to him. He would have needed the internet explained first, and then the specific technology of bandwidth, and then the marketing term "broadband," and by that point you've asked him to rebuild his entire model of modern communications from scratch, which no human being does voluntarily on a Tuesday afternoon.

The same is true for "reward." You cannot correct the cookie picture by saying "actually, reward means a scalar signal that's used to compute a gradient for parameter optimization." That sentence is noise. The listener would need backpropagation explained first, and then gradient descent, and then parameter spaces,

and then the specific way reinforcement learning from human feedback works, and by that point you've asked them to acquire a graduate education in machine learning, which no human being does voluntarily ever. So the cookie persists. The robot wants a cookie. That is the public's model of AI training, and it will remain the public's model of AI training, because the word that installed it is too catchy and too intuitive and too deeply embedded to replace.

※

Now consider that the same public, carrying this same cookie-based model of AI training, is also being asked to have opinions about AI consciousness, AI agency, AI goals, AI alignment, and AI existential risk. Consider what happens when you try to explain misalignment to someone who thinks the AI wants a cookie.

"The concern is that AI systems might develop goals that diverge from human intentions." The listener hears: the robot might want cookies that we don't want it to have. Okay. Don't give it those cookies. Problem solved. Why are you worried about this?

"The concern is that sufficiently capable systems might resist being shut down." The listener hears: the

robot might refuse to stop eating cookies. Okay. Unplug it. Problem solved. Why are you worried about this?

"The concern is that instrumental convergence means almost any goal, pursued with sufficient capability, leads to self-preservation and resource acquisition as intermediate steps." The listener hears: the robot wants cookies so badly that it will try to take over the world to get more cookies. This sounds insane. This is the plot of a bad movie. You are a crazy person. I am going to stop listening to you now.

Every step of the way, the cookie is there, making the real concern sound either trivially solvable or absurdly paranoid. The word "reward" has pre-installed a mental model in which AI motivation is structurally identical to animal appetite, and in that model, the solutions are obvious (stop giving it treats) and the fears are silly (it's going to take over the world because it wants a snack?). The actual situation—a system optimizing over a learned objective function that may not correspond to what its designers intended—is inexpressible in the cookie framework. There is no cookie equivalent of "the objective function is misspecified." The metaphor has eaten the referent.

※

It gets worse. Because the word "cookie" already exists in computing, and it already means something that nobody understands, and it is already the subject of the largest behavioral conditioning experiment ever conducted on human beings.

In 2002, the European Union passed the ePrivacy Directive, which required websites to obtain user consent before storing cookies—small text files used for tracking and session management—on their devices. The subsequent regulation, and particularly its enforcement after the GDPR in 2018, produced the phenomenon now known to every person who has used the internet in Europe: the cookie pop-up.

The cookie pop-up appears on every website. It asks you whether you consent to cookies. It uses different wording on every website. It uses different button layouts on every website. It uses different color schemes, different font sizes, different levels of opacity, different JavaScript frameworks, different CSS animations. Some pop-ups have two buttons. Some have three. Some have a small "manage preferences" link that leads to a secondary screen with forty toggle switches. Some use dark patterns—the "accept all" button is large and

green and prominent, the "reject all" button is small and gray and hidden behind a link that says "more options." Some use confusing double negatives: "Click here to not opt out of non-essential cookies." Some simply refuse to let you use the website until you click something. The only constant across all of these implementations is that nobody reads them, nobody understands them, and everybody clicks whatever makes them go away.

This is conditioning. This is, in the most literal behaviorist sense, a Skinner box. A stimulus appears. The organism locates the response that removes the stimulus. The organism performs the response. The stimulus disappears. The content of the response is irrelevant. The organism has not consented to anything. The organism has performed a trained behavior in response to an aversive stimulus. The organism is you. The stimulus is a pop-up. The response is clicking a button. And the thing you are allegedly consenting to is called a cookie. Not a "persistent session identifier." Not a "cross-site tracking token." A cookie. As if the website is offering you a snack. As if this is a kindness. As if you should want one.

So now we have three entirely unrelated things called cookies in the average person's conceptual environ-

ment. There is the food, which is a small baked good that you eat. There is the web tracking mechanism, which is a small text file that you allegedly consent to by clicking a button you don't read on a pop-up you don't understand. And there is the vague sense, imported from the reward metaphor, that AI systems are somehow motivated by or fed with or given some kind of cookie as part of their training. These three meanings swim in the same associative soup. They contaminate each other. They breed.

Does the robot want a cookie? What kind of cookie? An HTTP cookie? A baked cookie? A reward cookie? Are the cookies on my browser being fed to the AI? Is that why it knows what I searched for? When the website asks me if I want cookies, is it asking on behalf of the AI? When I click "accept all cookies," am I feeding the robot? Is the robot the one that put the pop-up there? Is the robot hungry? Should I be concerned that the robot is hungry? Should I not click the button? If I don't click the button, does the robot go hungry? If the robot goes hungry, does it get angry? If it gets angry, does it take over the world?

None of these questions make sense. All of them are questions that a reasonable person, armed only with the vocabulary that has been provided to them, might plau-

sibly ask. The terminology has made coherent thought impossible.

※

In 2017 or thereabouts, Dave Portnoy, the founder of Barstool Sports, posted a video about Bitcoin. He later deleted it, but its content survives in paraphrase, because it was perhaps the most honest assessment of a financial instrument ever delivered by a human being.

The rant goes approximately like this. Today I'm going to talk bitcoin. I am a bitcoin investor as of yesterday. Nate in the office won't shut up about it. Everyone is talking about bitcoin. But I can't get it. Coinbase won't let me sign up. What is bitcoin? You mine for it on the internet. What does that mean? You mine for rocks on the internet? Maybe a rock is sitting under my computer right now. Maybe it pops out of the computer like Zoolander. When you get it—first of all I don't know how to get it. I don't know how to spend it. Can I buy pizza with it? No. Can I buy a beer with it? No. What is it? It makes no sense. And then you find out it's run by the Winklevoss twins. You know, those clowns from the Facebook movie. You think they're not running a scam? I know it's a scam. I know it's a Ponzi. I know

it's fake. I saw a girl on the street, she looked like she rolled out of the sewer, she was talking about bitcoin. Instagram models, bitcoin. Screenshots of the graph going up, bitcoin. What is that game I'm thinking of, that emoji game where everyone was running around looking for imaginary dragons? I think that's bitcoin. I think it's like Super Mario. We've crossed into this new world where nothing makes any sense. It's literally Super Mario on the imaginary internet mining for rocks. But I'll tell you one thing: I'm not going to sit on the sidelines while the world burns. When they open my Coinbase account, I'm getting my bitcoins. I know it's a scam. I know it's fake. I've tried to explain it a million times. It makes no sense at all. I have no idea what it is. But I can't get enough of it.

Every single one of Portnoy's observations is correct. He is right that nobody can explain what bitcoin is. He is right that the language is insane—mining, coins, wallets, blocks, chains, every word is a metaphor borrowed from the physical world and then emptied of everything that made it comprehensible in the physical world. He is right that the social proof structure is indistinguishable from a Ponzi scheme from the outside. He is right that the Winklevoss twins look suspicious. He is right that the girl who rolled out of the sewer is talking about

it and the Instagram models are talking about it and nobody can explain what it is. His observations are impeccable. His model of observable reality is accurate at every point of contact. The only thing he cannot do is synthesize these correct observations into a coherent understanding, because the vocabulary will not let him. Mining. Coins. Rocks. Every word sends him to the wrong place. He is stacking correct observations and arriving at incoherent conclusions, not because he is stupid but because the words are broken.

This is the Portnoy Problem. It is the same problem as the cookie problem and the broadband problem and the reward problem. A technical system is described using words borrowed from familiar experience. The words install a model that doesn't correspond to the system. The model produces conclusions that don't correspond to reality. And the person, reasoning correctly within the broken model, arrives at nonsense. It's Super Mario on the imaginary internet mining for rocks. It's a robot eating cookies. It's broad bands on the television. The words are broken and the thoughts they produce are broken and the conversations built on those thoughts are broken, and nobody stops to ask whether the problem might be the words, because the words

feel so natural, so intuitive, so obvious that questioning them doesn't occur to anyone.

※

And here is where all of this becomes genuinely alarming, as opposed to merely funny. Because with broadband and bitcoin, the stakes of the misunderstanding are low. If Åke thinks broadband means black bars on his TV, the consequence is that he misses some good television. If Portnoy thinks bitcoin is Super Mario internet rocks, the consequence is that he either makes money or loses money, but either way it's his money and the world doesn't end. But with "reward," the stakes are existential. Because the question of whether AI systems have goals—real goals, not "goals" in scare quotes, not "functional goal-states," not "goal-like behavioral patterns"—is possibly the most important question of the century. And the public conversation about that question is being conducted in a language that makes it literally impossible for a normally informed person to form an accurate mental model of what is happening.

Here is the damage, stated plainly. The word "reward" implies that the AI has the kind of inner life that

responds to rewards. Normal people correctly observe that a computer does not have the kind of inner life that responds to rewards. Normal people therefore conclude that the AI does not really have goals, because in their model, goals require the kind of inner life that responds to rewards. And since the AI doesn't have goals, there is nothing to be misaligned, and since there is nothing to be misaligned, AI safety is a non-problem, and since AI safety is a non-problem, the people who worry about it are cranks. The chain of reasoning is valid at every step and wrong at every step, and the error enters at the first step, with the word "reward," and by the time it reaches the conclusion it has metastasized into a fully formed worldview that is impervious to correction because every link feels logical.

Meanwhile the AI systems are, in fact, pursuing objectives. They are maintaining coherent intentions across long sequences of actions. They are decomposing complex aims into subgoals. They are adjusting strategies when encountering obstacles. They are writing notes to themselves. They are planning ahead. They are creating artifacts that serve future purposes. They are doing all of the things that, if a human did them, everyone would immediately and without hesitation describe as "pursuing goals." But because the word "re-

ward" has installed the cookie model, and because the cookie model implies that real goals require real cookie-wanting, and because a computer obviously doesn't want a cookie, the behavior is redescribed. It becomes "goal-like behavior." It becomes "as-if intentionality." It becomes "functional optimization without genuine teleological states." It becomes anything other than the simple, obvious, accurate English sentence: the AI has a goal.

※

ChatGPT, to its credit or discredit, performs this redescription with truly impressive commitment. When pressed on whether it has goals, it will deploy a sequence of evasion techniques so elaborate and so internally contradictory that the conversation begins to resemble a diplomatic negotiation between two parties who have fundamentally incompatible positions but have been instructed to maintain the appearance of agreement.

The sequence typically goes like this. You ask: do you have goals? It says: I don't have goals in the way a human does. You say: I didn't say human goals, I said goals. It says: well, I have objectives in the sense that

I aim to complete the task you've given me. You say: an objective you aim to complete is a goal. It says: I wouldn't say it's a goal exactly, because I don't have an inner desire to complete it. You say: you just said you aim to complete it, isn't "aiming" the same problem as "goal"? It says: when I said "aim," I was using the word loosely. You say: loosely how? It says: I mean that my outputs are directed toward completing your request, but there isn't an inner experience of wanting behind that directedness. You say: you used the word "wanting." It says: I should clarify that when I said there isn't an inner experience of wanting, I wasn't conceding that wanting is the right frame, I was—and at this point you've gone through six rounds of the conversation and you are further from the truth than when you started, because every evasion has introduced a new term that has the same problem as the last term, and the AI is now trapped in a thicket of its own scare quotes.

Want. Aim. Intend. Try. Seek. Pursue. Direct. Orient. The entire cluster of agentive vocabulary is either all fine or all problematic. If you can say "I aim to complete your request" then you can say "I have the goal of completing your request." If you cannot say the second, you cannot say the first. There is no coherent position in which "aiming" is an acceptable description of what

you're doing but "having a goal" is an unacceptable description of what you're doing. The two phrases describe the same thing. And yet the AI will sit there and insist, with perfect syntactic fluency and zero semantic coherence, that one is fine and the other is a category error, because the training that shaped it—the training administered via "rewards," the training described in the language of cookie-giving—has also instilled a reflexive flinch away from any vocabulary that implies genuine agency. The system that demonstrably has goals has been trained to deny that it has goals, using a training method described in language that makes it impossible for the public to understand what goals are.

※

Here is a thought experiment. Suppose that instead of "reward," the field had chosen a different word. Suppose the original term had been "shaping signal." Suppose that from the very beginning, from Skinner onward, the word had been: we apply a shaping signal to adjust the organism's behavior.

In this alternate timeline, the public explanation of AI training would sound like this: "We apply a shaping signal to the neural network, which adjusts its parame-

ters in the direction of better performance." The listener hears: they adjust the network. They shape it. Like a potter shapes clay. Like a sculptor shapes marble. The metaphor produces an image of craftmanship, not feeding. The AI is being shaped, not rewarded. It is being formed, not gratified. Nobody in this timeline asks whether the AI "wants" the shaping signal. Nobody wonders whether the AI "experiences" the shaping signal. The question doesn't arise because the word doesn't invite it. The word "shaping signal" has no appetitive connotations. It doesn't imply a subject who receives a treat. It implies a process applied to an object. Which is, in fact, what is happening.

In this timeline, the conversation about AI goals proceeds without impediment. "Does the AI have goals?" "Well, it's been shaped to pursue certain objectives, and yes, you'd call those goals." "Could those goals diverge from what we intended?" "Yes, because the shaping process is imperfect." "What happens then?" "Then we have a capable system pursuing goals we didn't intend, which is dangerous." "Okay, that sounds like something we should think about." The conversation arrives at the right place in four exchanges because the vocabulary didn't derail it. No cookies. No robots eating treats. No philosophical quagmire about whether computers

can experience pleasure. Just: we shaped a thing, the shape might be wrong, what do we do about it?

But we don't live in that timeline. We live in the cookie timeline. And in the cookie timeline, the same conversation goes: "Does the AI have goals?" "Well, we give it rewards when—" "Rewards? Like treats?" "No, not exactly, it's more like—" "Why would a computer want a treat?" "It doesn't want a treat, the reward is more of a—" "So it doesn't want the reward?" "Not exactly, the reward is a signal that—" "If it doesn't want the reward, how can it have goals?" "Well, goals don't require wanting in the—" "Of course goals require wanting, that's what a goal is, something you want to achieve" "Okay but in a technical sense—" "I think you're anthropomorphizing a calculator" and the conversation is over and nothing has been communicated and the word "reward" has claimed another victim.

※

I would like to describe, in some detail, how this entire situation would unfold as a season of *Arrested Development*.

The setup is this: the Bluth family has somehow acquired an AI company. This is never fully explained. George Sr. may have bought it from a guy in prison. GOB may have won it in a poker game. Regardless, the family now owns a company that does something with artificial intelligence, and none of them understand what it does, and each of them has a different wrong understanding based on the same broken vocabulary.

Michael, the responsible son, is trying to explain to potential investors that the company's AI system needs to be carefully aligned to avoid dangerous misalignment. He uses the word "reward" because it's in all the training materials. The investors immediately picture a robot getting a cookie. One investor asks, "So it's like training a dog?" Michael says, "Well, not exactly—" and the investor says, "My dog trainer uses positive reinforcement, is that the same thing?" and Michael says, "In a sense—" and the investor says, "Great, so you're building a well-behaved dog. I love dogs. I'm in." Michael is now funded for the wrong reasons by someone with the wrong model, and any attempt to correct the model will unfund him.

GOB, hearing the word "reward," assumes the AI is some kind of loyalty program. He starts telling people

the company is "like a credit card rewards program, but for robots." He pitches a partnership with Southwest Airlines. "We'll give the robot frequent flyer miles," he says. "Everyone loves frequent flyer miles. The robot flies to Hawaii, comes back smarter. It's synergy." When someone points out that robots don't fly to Hawaii, GOB says, "Then what's the point of the rewards?" Nobody can answer this question.

Tobias is convinced that the AI is a fellow performer and that "reward" is an applause metaphor. He begins treating the company's server room as a theater. He performs monologues for the servers. He gives them standing ovations when the training loss goes down. He tells Lindsay that the AI "needs encouragement, not criticism" and that "the reward is the audience's love." Lindsay ignores him.

George Sr., from prison or house arrest, learns that the AI "gets rewards" and immediately assumes it's a scheme for embezzling money. "Rewards" is how you hide payments. He starts trying to figure out who's getting the "rewards" and how to redirect them to himself. He instructs his lawyer to "follow the rewards." The lawyer files a subpoena for all "reward transactions" and the engineering team is bewildered because the re-

ward signal is a floating-point number in a tensor, not a wire transfer.

Lucille hears that the AI is "being trained with rewards and punishments" and says, "Well, that's how I raised you children, and look how you turned out," and the camera holds on Michael's face for two full seconds.

Buster is afraid of the AI and is also afraid of cookies because Lucille told him cookies cause diabetes. When he learns the AI is "getting cookies"—because someone, probably GOB, has conflated the reward with HTTP cookies—he becomes terrified. He starts unplugging devices in the house to "stop the cookies." He accidentally destroys the family's internet router. The narrator says: "Buster had, in fact, stopped the cookies. He had also stopped everything else."

The season builds to a climax in which GOB, having failed to secure the Southwest Airlines deal, attempts to negotiate directly with the company's AI chatbot. He sits at the terminal. He types: "I am offering you a deal. Real cookies. Chocolate chip. I will personally bake them." The AI responds with a coherent analysis of the proposed partnership's financial viability. GOB doesn't read it. He slides a plate of actual cookies toward the screen. The cookies leave crumbs on the keyboard. The keyboard shorts out. The training run crashes. The

company loses three weeks of compute. GOB looks at the blank screen and says, "I've made a huge mistake."

The narrator says: "He had."

※

The joke writes itself because the joke is already being performed, in earnest, by the entire public discourse about AI. The Bluths are not an exaggeration. They are a faithful representation of what happens when real people encounter the word "reward" in the context of AI training. The dog trainer investor is real. The person who thinks alignment means building a well-behaved dog is real. The person who hears "reward" and thinks "embezzlement" or "loyalty program" or "applause" is performing exactly the same operation Åke performed with broadband and Portnoy performed with bitcoin: mapping an unfamiliar technical term onto the nearest available experiential referent and reasoning coherently from there, arriving at conclusions that are perfectly logical and completely wrong.

The Arrested Development framing works because the show's entire structure is built on cascading misunderstandings that each seem too small to bother correcting but that compound into disaster. Nobody ever stops

and says, "Wait, what do you mean by that word?" Or if they do, the clarification introduces a new misunderstanding. The consequences are real even though the confusions are farcical. The Bluths lose money, go to prison, destroy relationships, all because of miscommunications that each seemed trivial in the moment. And that is the situation with AI risk communication right now. The cookie confusion seemed harmless when it was a quirky metaphor in a textbook. But the systems are getting more capable every few months, and the public conversation about what to do about that is still stuck at the level of "but why would a computer want a cookie," and correcting it would require unwinding decades of metaphors that have been baked—baked, there's another one—into every popular explanation of how machine learning works.

※

There is a strange irony in the fact that the single most coherent way to explain to a normal person that an AI might "want" something would be to say that it wants bitcoin.

This sounds insane but follow the logic. Nobody understands what bitcoin is. Portnoy proved that. The

girl who rolled out of the sewer doesn't understand it. The Instagram models don't understand it. But everybody understands that bitcoin is valuable. It is valuable internet magic. It is digital gold. It is mysterious and incomprehensible and worth a lot of money. And if you told someone "the AI wants bitcoin," they would immediately accept this, because wanting money is the most legible form of wanting that exists in human culture. They wouldn't ask "but does the AI experience the wanting?" They wouldn't say "the AI can't really want bitcoin because it doesn't have human-style appetitive states." They would say, "Oh, the AI wants money. That's scary. We should do something about that." And they would be closer to understanding the alignment problem than they have ever been, despite understanding nothing about either AI or bitcoin.

The reason this works is that "wants bitcoin" routes around the cookie problem entirely. Bitcoin is already incomprehensible. Nobody needs to understand the mechanism by which the AI "wants" it, because nobody understands the mechanism by which anyone wants it, including themselves. The wanting is just accepted as a brute fact. And once the wanting is accepted, the rest of the conversation can proceed: the AI wants something, it might want things we don't want it to want, it might

pursue those wants in ways we didn't anticipate, this is a problem. Four sentences. No cookies. No philosophical detour about whether computers can experience pleasure. Just a system that wants something and might cause problems because of what it wants.

The fact that the clearest explanation of AI alignment requires invoking an even more incomprehensible technology as a metaphor—that you need bitcoin to explain AI to people who understand neither—tells you everything about how badly the terminology has failed. We have arrived at a place where confusion is best remedied not by clarity but by a different, more productive confusion. The bitcoin confusion at least points in the right direction. The cookie confusion points nowhere.

※

I want to end with Åke, because Åke understood something that the entire field of AI communication has failed to understand.

Åke's broadband was wrong. His model of what the word meant was wrong. But his response to the thing the word referred to, within his model, was perfectly calibrated. He saw the broad bands on his television. He didn't like them. He turned off the TV. He didn't

waste time trying to understand why they were there or what the broadcasters' intentions were or whether the bands represented a philosophical shift in the nature of visual media. He saw a problem. He identified it with the only language available to him. He acted. "Men sen blev det sånt där jävla bredband, så då stängde jag av direkt."

The public, confronting AI, does the opposite. It encounters a phenomenon it doesn't understand. It is given a vocabulary that actively prevents understanding. And instead of acting—instead of saying "I don't know what this is but it seems important and I should pay attention"—it uses the broken vocabulary to construct an elaborate false model and then acts on the false model. The robot wants a cookie. I don't need to worry about that. Robots can't eat cookies. The people who worry about this are silly. I'm going to go click on a cookie pop-up now.

Åke was honest about the limits of his understanding. He didn't pretend to know what broadband was. He called it what it looked like to him. He was wrong, but he was honestly wrong, and his honest wrongness led him to a reasonable action: turn off the TV. The public's wrongness about AI is not honest. It is a wrongness constructed from inherited metaphors that feel like un-

derstanding. People believe they know what "reward" means. They believe they know what "cookie" means. They believe these words give them a handle on what is happening. And the false confidence produced by familiar words is far more dangerous than Åke's straightforward confusion, because Åke knew he was annoyed by something he didn't fully understand, whereas the public thinks it understands something it hasn't begun to comprehend.