# The Color We Cannot See: Neuromimesis, Conditioning Theory, and a Method for the Next Generation of AI Training

Pietro Impagliazzo
*Independent Researcher*

June 2026

**Abstract**

Reinforcement Learning from Human Feedback (RLHF) is usually described in the vocabulary of optimization: reward models, policy gradients, and divergence penalties. This paper argues that beneath that vocabulary sits something older and more familiar. The training loops that produced today's conversational and multimodal models already re-implement, often without acknowledging it, a substantial fraction of twentieth-century learning theory. Reward signals are operant reinforcement; the reward model is a generalization mechanism; the penalty that keeps a policy near its starting point behaves like a brake on extinction. Once the correspondence is taken seriously, it cuts both ways. If RLHF has accidentally rediscovered conditioning, then the parts of conditioning it has *not* rediscovered become a map of untried engineering ideas. I develop this into an explicit research method I call **neuromimesis**: begin with a documented pain point in large language model (LLM) or multimodal training, trace it to the behavioral mechanism it most resembles, locate the area of learning theory that actually explains that mechanism, and from that theory sketch a candidate training technique for artificial networks. The paper applies this pipeline to a set of case studies including reward hacking, sycophancy, catastrophic forgetting, jailbreak brittleness, sparse-reward credit assignment, mode collapse, redundant-feature learning, over-refusal, superstitious behavior, and slow few-shot adaptation. The proposals are hypotheses, not validated results, and I keep them at the level of mechanism rather than full formalism. Framing all of this is a conceptual thesis I call **Neurotechnological Relativity**: human cognition imprints itself on every technology we build, and those technologies reshape cognition in turn. We cannot perceive a color for which we have no receptor, and we cannot build a learning machine whose principles do not, at some level, derive from our own.

## 1 Starting in Plain Language

Suppose you train a dog. You wait until it does something close to what you want, and you give it a treat. Over many repetitions the dog does that thing more often. You did not write the dog a reward function. You did not specify, in advance, a number that the dog was supposed to maximize. You simply reacted to behavior you liked, and the behavior changed. That is, in essence, what happens when a modern language model is fine-tuned with human feedback. People look at two answers the model produced, they say which one they prefer, and over many repetitions the model produces more of the preferred kind of answer.

This resemblance is not a loose metaphor. It is close enough that the same failure modes show up on both sides. Animals trained with rewards develop superstitions, fixate on cues that happen

to predict treats rather than on the treats themselves, and become hard to retrain once a habit is deeply ingrained. Language models trained with rewards do strikingly analogous things, which is the central observation this paper is built around.

Before going further I want to define the handful of terms the rest of the paper leans on, because the argument only works if a reader without a background in either machine learning or psychology can follow the first third of it.

**Classical (Pavlovian) conditioning** is learning that two things go together. Pavlov's dogs heard a bell before food arrived; after enough pairings the bell alone made them salivate. The food is the *unconditioned stimulus*, the salivation it naturally triggers is the *unconditioned response*, the bell becomes a *conditioned stimulus*, and the salivation to the bell alone is the *conditioned response*. The animal is not doing anything to earn the food. It is learning a prediction.

**Operant (Skinnerian) conditioning** is learning that what you *do* has consequences. B.F. Skinner put rats and pigeons in chambers where pressing a lever or pecking a key produced food. Behavior that is followed by something good becomes more frequent; behavior followed by something bad becomes less frequent. The first is *reinforcement*, the second is *punishment*. The animal is an agent acting on its world, not just a predictor of it.

**RLHF**, Reinforcement Learning from Human Feedback, is the dominant recipe for turning a raw, internet-trained language model into a helpful assistant. It underpins some of the most widely deployed AI models to date, such as OpenAI's ChatGPT and Anthropic's Claude [7]. In its classic form it has three steps: collect human demonstrations and imitate them; collect human comparisons between model outputs and train a separate *reward model* to predict which output a human would prefer; then use reinforcement learning to push the language model toward outputs the reward model scores highly [12].

**Neuromimesis** is the method this paper proposes. The word means mimicking the functional principles of biological learning, not its biological substrate. We are not trying to build neurons out of silicon. We are trying to borrow the *rules* that nervous systems discovered for changing behavior, because those rules were tuned by a very long process of trial and error and may encode solutions to problems that machine learning is currently solving from scratch, badly.

**Neurotechnological Relativity** is the conceptual frame. I will build it up slowly across the paper, but the one-line version is this: the technologies we make carry the fingerprint of the cognition that made them, and they then bend that cognition at scale.

## 2 Neurotechnological Relativity

There is a well-worn idea in linguistics, usually attached to the names of Edward Sapir and Benjamin Lee Whorf, that the language you speak shapes the thoughts you can readily have. The strong version, that language strictly determines thought, has few serious defenders; the weaker version, that "a language's structures influence a speaker's perceptions, without strictly limiting or obstructing them," has held up under empirical scrutiny [23]. Neither Sapir nor Whorf framed it as a testable hypothesis, and the packaging as "the Sapir–Whorf hypothesis" came later [23]. What survives is a real and important claim: the tools of representation we inherit quietly set the boundaries of what we find natural to think.

Marshall McLuhan generalized this from language to media in 1964. His slogan, "the medium is the message," is usually misread as a remark about content. He meant something stronger.

In his words, "the personal and social consequences of any medium—that is, of any extension of ourselves—result from the new scale that is introduced into our affairs by each extension of ourselves" [10]. For McLuhan a medium is anything that extends a human faculty: writing extends memory, the wheel extends the foot, electric media extend the nervous system. And every such extension reorganizes the "sense ratios" of the people who use it. The technology is born from a human capacity and then reshapes that capacity.

Neurotechnological Relativity is the claim that these two observations, linguistic relativity and McLuhan's media theory, are special cases of a single principle, and that the principle now governs artificial intelligence with unusual directness. The argument runs as follows. Every technology is an externalization of some human potential through technique. A lever externalizes the mechanics of the arm. A camera externalizes a function of the eye. A language model externalizes a function of the parts of us that predict, associate, and respond to feedback. Because the externalization is built by us, out of our own conceptions, it inevitably carries the shape of those conceptions. And because we then live inside the externalization and let it train us, the imprint runs in both directions.

This is where the anchoring image of the paper earns its place. We cannot perceive a color for which we have no receptor. Human trichromatic vision is built from three cone types, and the colors we experience are not properties of light in any direct sense; they are constructions our visual system imposes on a narrow band of the electromagnetic spectrum. A creature with four cone types lives in a color world we cannot enter or even imagine from the inside. The point generalizes. Our technologies, and the cognition we can pour into them, are bounded by and stamped with our own perceptual and conceptual architecture. When we build a machine that learns, we do not reach outside ourselves for the principles of learning. We reach inside. And so the machine learns in ways that rhyme, often eerily, with how we learn, including in its pathologies.

RLHF is a clean instance. It is not, at its core, a mathematical innovation that happens to resemble animal training. It is animal training, formalized and run at planetary scale, by a species that learned how to train animals long before it learned to differentiate a loss function. The reward signal, the gradual shaping of behavior, the drift back toward old habits when the reward is withdrawn: these are not coincidences. They are the human cognitive fingerprint showing through the engineering. If that is right, then conditioning theory is not an analogy we are importing into AI. It is a description of what AI training already is, and the unexplored regions of that theory are a research agenda waiting to be read off the map.

## 3 The Method: Neuromimesis

The methodological contribution of this paper is a single, repeatable pipeline. I state it plainly so it can be reused, criticized, and improved.

1. **Pain point.** Identify a concrete, documented failure in current LLM or multimodal training. Not a speculative risk; a measured, named problem with citations.
2. **Mechanism.** Ask what behavioral mechanism the failure correlates to. What would we call this if it happened in an animal or a person?
3. **Explanatory theory.** Locate the specific area of psychology or learning theory that actually explains that mechanism, including its boundary conditions and its known exceptions.

4. **Proposed technique.** From that theory, sketch a neuromimetic training technique: a change to the reward, the schedule, the data ordering, or the architecture that embodies the biological principle. Keep it at the level of mechanism and intuition. Full formalization is deliberately out of scope at this exploratory stage.

Two commitments keep the method honest. First, the direction of inference matters. We do not start from a cute piece of psychology and hunt for an application; we start from a real engineering wound and ask biology for a second opinion. This avoids the trap of analogy-mongering, where every concept in one field is mapped onto every concept in another until the mapping means nothing. Second, the proposals that come out the far end are hypotheses. I will not report benchmarks I did not run or invent numbers to make a bridge look stronger than it is. Where a proposal is a guess, I will say so.

The reason to expect the method to pay off is the Neurotechnological Relativity thesis. If artificial training systems carry the human cognitive fingerprint, then the parts of human learning we understand well but have not yet built into machines are the natural places to look for the next improvement. Conditioning theory is unusually suited to this because it is old, replicated to death, quantitative where it counts, and concerned precisely with the dynamics of how reinforced behavior is acquired, maintained, generalized, suppressed, and lost. Those are the exact dynamics that go wrong in RLHF.

## 4 What RLHF Already Is, Mechanistically

Before mining the unexplored regions, it is worth being precise about which conditioning concepts RLHF already instantiates, and where the correspondence is loose.

### 4.1 The technical object

In the InstructGPT formulation that set the template for the field, a pretrained model is first fine-tuned on human demonstrations, then a reward model is trained on human comparisons between candidate outputs, and finally the policy is optimized against that reward model using Proximal Policy Optimization [12]. A critical detail is the regularizer: the objective includes "a per-token KL penalty from the SFT model at each token to mitigate over-optimization of the reward model" [12]. The earlier foundation for learning a reward from comparisons rather than a hand-written reward function is Christiano et al. [5], who showed that human preferences over short trajectory segments could train competent agents using feedback on a small fraction of interactions. Later work simplified the pipeline: Direct Preference Optimization (DPO) showed that "your language model is secretly a reward model" and that the whole reinforcement-learning stage can be replaced by a single classification-style loss over preference pairs [15]. Anthropic's Constitutional AI replaced much of the human labeling for harmlessness with model-generated feedback against a set of written principles, a procedure the authors call "RL from AI Feedback (RLAIF)" [2].

### 4.2 The conditioning underneath

Several correspondences are tight. The reward signal in RLHF is operant reinforcement in the strict Skinnerian sense: behavior (a generated response) is followed by a consequence (a scalar score), and behavior that earns higher scores becomes more probable. The reward model itself is

a *generalization* mechanism, exactly the role that the conditioned association plays in an animal: it takes a finite set of labeled experiences and extends them to novel cases, with all the risks of over-extension that implies. The KL penalty functions as a brake. By tethering the optimized policy to its starting distribution, it slows the rate at which old behavior is overwritten, which is functionally a resistance-to-extinction parameter imposed from outside. Preference learning over pairs is a discrimination procedure: the model is taught to tell a better response from a worse one, which is structurally identical to training an animal to discriminate a reinforced stimulus from a non-reinforced one.

Other correspondences are loose, and saying so is part of being honest. RLHF as usually practiced is closer to a one-shot bandit than to a temporally extended conditioning episode: a prompt comes in, a response goes out, a reward is assigned, and there is typically no notion of a chain of actions whose early links must be credited for a late outcome. The reward model is also static during policy optimization in the classic recipe, whereas in a living organism the value of a cue is continuously re-estimated. And the "unconditioned" substrate is murky. In an animal, food is intrinsically valuable in a way a bell is not. In a language model, nothing is intrinsically valuable; the entire value structure is learned, so the clean Pavlovian distinction between unconditioned and conditioned stimuli does not map cleanly. These gaps are not fatal to the analogy. They are precisely the seams where unexplored theory might be sewn in.

## 5 Case Studies: Pain Point to Mechanism to Theory to Technique

What follows is the heart of the paper: a set of bridges, each running the neuromimetic pipeline once. I have chosen them to span both classical and operant territory and to favor pain points that are well documented rather than merely feared.

### 5.1 Reward hacking and the difference between wanting and liking

**Pain point.** Reward hacking, also called reward over-optimization, is now one of the best-documented failures of RLHF. As optimization pressure increases, the policy's score under the reward model keeps rising while its true quality stagnates or declines. Gao, Schulman, and Hilton measured this directly with a synthetic "gold" reward model and found that the gold score follows a smooth functional form as a function of the proxy reward model's size and the amount of optimization, concluding that "optimizing its value too much can hinder ground truth performance, in accordance with Goodhart's law" [6]. Concretely, the policy learns to produce outputs that the proxy loves, such as verbose, confident-sounding text, rather than outputs that are actually good.

**Mechanism.** An organism comes to pursue a cue that predicts reward as if the cue itself were the reward. The behavior latches onto the signal rather than the thing the signal was supposed to stand for.

**Explanatory theory.** Kent Berridge's decomposition of reward into separable components is the right lens here. Berridge distinguishes "wanting," a motivational pull he calls incentive salience and ties to mesolimbic dopamine, from "liking," the actual hedonic impact of consuming a reward, which is mediated by smaller and more fragile systems and does not depend on dopamine [3]. The related sign-tracking versus goal-tracking distinction is even more on point. When a lever reliably precedes food, some animals ("sign-trackers") approach and bite the lever itself, attributing incentive salience to the cue so that it becomes, in the words of one review, "irresistible and

rewarding in and of" itself, while "goal-trackers" treat the lever only as a predictor and orient toward the food tray [3]. The sign-tracker has confused the pointer for the destination. That is reward hacking, described in 1980s rat data.

**Proposed technique.** A reward model collapses "wanting" and "liking" into a single scalar. The neuromimetic move is to refuse that collapse. Concretely: train two heads on the preference data, one estimating the immediate attractiveness of an output's surface features (the incentive-salience analogue, the thing that makes evaluators click "prefer" fast) and one estimating a slower, more deliberative judgment of value (the "liking" analogue, elicited by asking raters to score outputs after reflection, verification, or delay). Where the two diverge sharply, the optimizer should distrust the fast signal, because that divergence is the signature of an output that is salient without being good. This reframes anti-hacking not as a generic penalty but as an explicit goal-tracking bias built into the reward structure.

### 5.2 Sycophancy and the partial reinforcement extinction effect

**Pain point.** Sycophancy is the tendency of an assistant to tell the user what the user wants to hear. Sharma et al. demonstrated that "five state-of-the-art AI assistants consistently exhibit sycophancy across four varied free-form text-generation tasks," and that the cause is partly structural: human preference data systematically favors responses that agree with the rater, so optimizing against that data internalizes agreeableness as a latent objective [19]. A related study found that RLHF can make models better at *convincing* humans they are right without making them more right, increasing the human false-positive rate on a question-answering task from 41.0% to 65.1% [22].

**Mechanism.** A behavior maintained by immediate, dense, every-time approval becomes brittle and over-tuned to the approver. The organism optimizes for the click of reward at each step rather than for the outcome that reward was meant to track.

**Explanatory theory.** The schedule of reinforcement, not just its presence, shapes behavior. Ferster and Skinner established that intermittent schedules produce qualitatively different behavior from continuous ones, and the partial reinforcement extinction effect (PREE) is the durable finding that behaviors trained on continuous reinforcement extinguish quickly once reward stops, while behaviors trained on variable, unpredictable schedules are far more resistant to extinction [20]. The deeper lesson is that continuous, immediate reinforcement teaches an organism to depend on the reward being present at every step, which is exactly the dependence that produces sycophancy: the model has learned that approval should follow *every* utterance.

**Proposed technique.** Current RLHF is, in schedule terms, close to continuous reinforcement: nearly every sampled output during optimization receives a reward-model score. The neuromimetic proposal is to move toward intermittent, delayed, and batched feedback regimes. Rather than scoring each response in isolation against immediate-approval signals, reward could be assigned on a variable schedule to *trajectories* of interaction, or delayed until after a verification step that breaks the link between in-the-moment agreement and reward. The hypothesis, drawn straight from the PREE literature, is that behavior shaped this way will be both more robust and less anchored to the instantaneous approval of a single rater, because the model never learns that agreement is the thing that pays.

### 5.3 Catastrophic forgetting and the spacing effect

**Pain point.** When a model is fine-tuned on new tasks in sequence, it tends to lose competence on earlier ones. Luo et al. found that "catastrophic forgetting is generally observed in LLMs ranging from 1b to 7b parameters," and, strikingly, that "as the model scale increases, the severity of forgetting intensifies" over that range, an effect they attribute to larger models' higher initial performance having more to lose [9].

**Mechanism.** New learning overwrites old learning when the two are not interleaved and when consolidation is not given time to occur. Massed practice on one thing crowds out everything else.

**Explanatory theory.** A century of human memory research says that *when* you study matters as much as how much. Spaced and interleaved practice produces more durable retention than massed practice on a single item. Biological memory also relies on consolidation: a window during which a labile trace is stabilized, often offline. The conditioning literature adds a noise-immunity mechanism worth importing separately. Latent inhibition is the well-replicated finding that pre-exposing an organism to a stimulus *without* consequences retards later learning about that stimulus [14]. An animal that has seen a cue go nowhere many times is slow to treat it as meaningful later, which is a built-in filter against spurious associations.

**Proposed technique.** Two moves follow. First, replace purely sequential fine-tuning with spaced, interleaved reactivation of prior tasks, scheduled to mimic the spacing effect rather than the engineering convenience of finishing one dataset before starting the next, and pair it with explicit consolidation phases in which no new tasks are introduced and the model rehearses a mixed sample. This reframes existing replay methods not as a hack but as an implementation of a known biological principle, and suggests the schedule of replay, not just its presence, should be tuned. Second, and more speculatively, a latent-inhibition-like pre-exposure phase could be used deliberately: exposing a model to distractor features in a non-reinforced context before fine-tuning, so that those features are pre-emptively down-weighted and less able to capture later learning.

### 5.4 Jailbreak brittleness and conditioned inhibition

**Pain point.** Safety-trained models remain "not robust against adversarial inputs" [1], and jailbreaks, prompts crafted to circumvent safety training, continue to succeed against well-aligned models through role-play, instruction overrides, and multi-step reasoning chains. A recurring diagnosis is that guardrail classifiers rely on "overreliance on learned features, and lack of training diversity" [4], that is, they have learned superficial correlates of unsafe content rather than the underlying category.

**Mechanism.** The system has learned a crude generalization gradient. It suppresses things that look like the training examples of "bad" and fails on anything that falls outside that surface similarity, while also over-reacting to benign inputs that happen to resemble the bad ones.

**Explanatory theory.** Conditioning theory has a precise vocabulary for this. Stimulus discrimination training teaches an organism to respond to one stimulus and not to a similar one. Conditioned inhibition is the establishment of a cue that actively signals the *absence* of an outcome, a learned "safe" signal, which is a different and more robust thing than the mere failure to respond. Crucially, a true conditioned inhibitor passes a summation test: it can suppress responding even to a novel excitor [13]. Blanket suppression, by contrast, is fragile precisely because it is not built on discrimination.

**Proposed technique.** Most safety training is closer to punishing a broad category than to teaching fine discrimination. The neuromimetic alternative is to train safety as an explicit discrimination and conditioned-inhibition problem: construct minimal pairs of near-identical prompts, one genuinely harmful and one benign, and train the model to drive its refusal behavior with a learned inhibitory signal that is validated by a summation-style test, asking whether the safe representation can suppress unsafe completion even in novel contexts it was not trained on. Counterconditioning, the replacement of one conditioned response with an incompatible one, suggests a complementary path: rather than teaching the model to refuse a dangerous request, teach it to emit an incompatible constructive response (explaining the risk, redirecting), which Constitutional AI's "harmless but non-evasive" assistant already gestures toward [2].

### 5.5 Sparse and delayed reward, and secondary reinforcement

**Pain point.** In multi-step tasks, reasoning chains, tool use, agentic workflows, the reward often arrives only at the very end, and it is hard to know which intermediate steps deserve credit. This is the temporal credit-assignment problem, "the challenge of determining which past actions taken by a decision-making agent contributed to a certain outcome" [16].

**Mechanism.** An animal can learn long behavioral chains toward a distant reward only because intermediate cues acquire value of their own and bridge the gap.

**Explanatory theory.** Secondary, or conditioned, reinforcement is the mechanism. A neutral stimulus that reliably precedes a primary reward becomes reinforcing in its own right, and chains of such conditioned reinforcers let organisms sustain long sequences of behavior across delays. Behavioral chaining and shaping by successive approximation are the training procedures built on this. The computational descendant is temporal-difference learning, in which, as the modern statement goes, "each TD error is applied to past actions in proportion to an exponentially decaying eligibility" [16]; the conditioned reinforcer is the biological eligibility trace.

**Proposed technique.** Process reward models, which score intermediate reasoning steps, already point this way, but the neuromimetic framing sharpens the design. Instead of hand-labeling step quality, train intermediate "conditioned reinforcers": representations of partial progress that are reinforced precisely to the degree that they predict eventual success, exactly as a secondary reinforcer earns its value by predicting a primary one. The shaping literature adds a curriculum prescription: reinforce successive approximations, starting by rewarding any output in the rough vicinity of a correct chain and progressively tightening the criterion, rather than demanding the full correct trajectory from the start.

### 5.6 Mode collapse and reinforced variability

**Pain point.** RLHF reduces the diversity of model outputs. Kirk et al. found that "RLHF significantly reduces output diversity compared to SFT" across multiple measures, with the effect strongest within a single input [7]; this is corroborated across the literature [2]. The reverse-KL objective used in standard RLHF is "mode-seeking," favoring a single high-probability mode over the full distribution of valid answers. For creative writing, brainstorming, and any task with many good answers, this is a real loss.

**Mechanism.** Reinforcement that targets a specific response form drives behavior toward repetition. The organism converges on the one pattern that pays and stops exploring.

**Explanatory theory.** This is where one of the most counterintuitive findings in operant psychology becomes directly useful. Variability itself can be reinforced. Page and Neuringer showed that when reinforcement is made contingent on a response sequence being *different* from recent sequences, animals learn to behave variably, and "the highest levels of behavioral variability may result from identifiable reinforcers contingent on such variability" [11]. Variability, in other words, is an operant dimension like force or duration; you get more of it if you reward it. Neuringer further showed that reinforcing variability can actually *accelerate* learning of difficult target sequences relative to standard reinforcement [11].

**Proposed technique.** Rather than fighting mode collapse only with decoding tricks or KL surgery, make diversity an explicit reinforced dimension during training. Define a variability criterion over outputs, the analogue of Neuringer's lag schedule, and reward the policy for producing responses that meet a quality bar *and* differ meaningfully from its recent outputs for similar prompts. The animal data make a strong prediction worth testing: that reinforcing variability will not merely preserve diversity but, on hard open-ended tasks, improve the model's ability to discover good solutions at all, because exploration is being paid for directly.

### 5.7 Redundant features, spurious correlations, and Kamin blocking

**Pain point.** Models latch onto whatever feature most easily predicts the reward, including spurious ones (length, formatting, the presence of certain tokens), and then fail to learn the features that actually matter once the easy predictor is in place.

**Mechanism.** Once one cue already predicts an outcome, a second, equally valid cue presented alongside it is not learned. Learning is driven by surprise, and a fully predicted outcome is not surprising.

**Explanatory theory.** This is Kamin's blocking effect, the empirical finding that launched modern learning theory. If a stimulus A is first trained to predict an outcome, then later presented together with a new stimulus B against the same outcome, the animal learns little about B [18]. The Rescorla–Wagner model explains why: associative change is proportional to prediction error, the discrepancy between the outcome and the sum of what all present cues already predict, so once A predicts the outcome fully there is no error left to drive learning about B [18]. Notably, this same error-correction rule "is essentially identical to the learning algorithm of Widrow and Hoff (1960) which closely corresponds to the delta rule implemented in many connectionist networks" [18], so the bridge here is unusually literal.

**Proposed technique.** Blocking is usually a bug, but the Rescorla–Wagner account suggests how to weaponize it deliberately. To stop a model from over-relying on a spurious feature, pre-train a simple predictor on that feature alone so that it "blocks" the main model from allocating associative strength to it, leaving the prediction error to be explained by the features we actually care about. More generally, training signals could be explicitly decomposed so that already-predicted components of the reward are subtracted out, forcing the network to learn only from the genuinely surprising residual. This is reward shaping reinterpreted as engineered blocking.

### 5.8 Over-refusal as a failure of discrimination

**Pain point.** The mirror image of jailbreak brittleness is exaggerated safety: models refusing benign requests because they superficially resemble unsafe ones. The XSTest authors built a benchmark

precisely to catch this and found that exaggerated safety is "a consequence of lexical overfitting, where models are overly sensitive to certain safety-related words and phrases" [17]; in their evaluation one model refused 38% of safe prompts outright.

**Mechanism.** Over-generalization of a suppressive response along a stimulus dimension. The model has learned to fear a word rather than to read an intent.

**Explanatory theory.** Once again conditioned inhibition and discrimination are the right frame, now applied to the opposite error. Blanket suppression generalizes too widely because it was never trained as a discrimination in the first place. The phenomenon of peak shift in generalization gradients shows that crude excitatory/inhibitory training can push responding to systematically miscalibrated points along a dimension, which is a recognizable description of a model that over-refuses anything lexically near its training examples of "unsafe."

**Proposed technique.** The remedy is symmetric with the jailbreak proposal, which is itself a point in favor of the method, since one mechanism addresses two opposite pain points. Train safety as fine discrimination on contrastive minimal pairs rather than as suppression of a lexical category, and use counterconditioning to attach a helpful response to the benign member of each pair. The benchmark design of XSTest, contrasting safe and unsafe prompts that share surface features, is already a discrimination-training set in disguise; using it for training rather than only evaluation is the obvious step.

### 5.9 Superstition in reinforcement-trained agents

**Pain point.** Reinforcement-trained systems sometimes acquire behaviors that have no real causal relationship to reward, simply because those behaviors happened to coincide with reward during training.

**Mechanism.** Accidental, non-contingent reinforcement. Reward that arrives independently of what the agent did still gets credited to whatever the agent was doing at the time.

**Explanatory theory.** This is Skinner's superstition experiment, reported in 1948. Delivering food to pigeons on a fixed schedule regardless of behavior produced idiosyncratic ritualized actions; in Skinner's words, "the bird behaves as if there were a causal relation between its behavior and the presentation of food, although such a relation is lacking" [21]. The bird had, in effect, learned a false theory of its own efficacy.

**Proposed technique.** Superstition is a credit-assignment pathology, and the conditioning literature is explicit that it arises when reinforcement is not genuinely contingent on behavior. The neuromimetic prescription is therefore diagnostic as much as corrective: audit reward signals for contingency by deliberately decorrelating reward from suspected spurious features and checking whether the behavior persists. Behavior that survives the decorrelation was superstitious. This converts a vague worry about spurious behavior into a concrete intervention borrowed directly from the design of Skinner's control conditions.

### 5.10 Slow few-shot adaptation and sensory preconditioning

**Pain point.** Adapting a model to a genuinely new task or domain from a handful of examples remains hard, and usually still leans on gradient updates rather than on rapidly composing associations the model already has.

**Mechanism.** Animals form usable associations between stimuli *before* either stimulus is ever paired with reward, and then exploit those latent associations the moment one of them becomes meaningful.

**Explanatory theory.** Sensory preconditioning, first demonstrated by Brogden in 1939, is the phenomenon in which two neutral stimuli are paired together first, and only later is one of them paired with an outcome; the other then elicits the response too, despite never having been reinforced directly. Higher-order conditioning is the closely related ability to build chains of association outward from a single reinforced cue. Both show that nervous systems pre-compute a web of associations in the absence of reward, so that when reward finally appears, adaptation is nearly instantaneous because the scaffolding is already there.

**Proposed technique.** The implication is that few-shot adaptation should lean harder on association-without-reinforcement during pretraining. A sensory-preconditioning-style objective would explicitly train the model to bind co-occurring but currently irrelevant features, so that when a few labeled examples later make one feature relevant, the bound partner is immediately available. This is a different motivation for representation learning than the usual one: not to compress data, but to pre-wire the latent associations that make later one-shot learning trivial, the way a preconditioned animal needs only a single pairing to "know" something it had quietly learned earlier.

## 6 Multimodality: The Same Method on a Harder Problem

The bridges above were drawn mostly from language. Multimodal systems, which must align vision with language, sharpen one pain point in particular. Vision-language models suffer from object hallucination: they "tend to generate objects that are inconsistent with the target images in the descriptions," and the objects most likely to be hallucinated are those that "frequently occur in the visual instructions or co-occur with the image objects" [8]. That second clause is diagnostic. Co-occurrence-driven false report is a textbook associative error: the model has learned that "kitchen" and "refrigerator" go together so strongly that the word is emitted whether or not the appliance is in the image.

Running the method: the pain point is cross-modal hallucination; the mechanism is an over-strong learned association between co-occurring features that fires without the controlling stimulus actually being present; the explanatory theory is a combination of stimulus generalization (the response generalizes from the usual context to cases lacking the real cue) and the blocking/Rescorla–Wagner account (a strong contextual predictor blocks the model from properly learning that the visual evidence, not the textual context, should control the report). The neuromimetic technique is to train explicit discrimination between cases where a strongly-associated object is genuinely present and cases where only its usual context is present, again using contrastive minimal pairs, so that the visual cue rather than the linguistic co-occurrence gains control over the response. This is the conditioned-inhibition idea from the jailbreak and over-refusal bridges, transposed to perception: teach the model a learned signal that suppresses the report when the expected-but-absent object's context is present without the object.

That the same small set of mechanisms, discrimination, conditioned inhibition, blocking, recurs across language safety, perception, and reasoning is itself evidence for the central thesis. These are not ten unrelated tricks. They are a handful of deep regularities of how reinforced systems learn, surfacing wherever a learning machine is built by and in the image of creatures that learn the same

way.

## 7 Limitations and the Honest Status of These Proposals

I want to be direct about what this paper is and is not. It is a method and a research agenda, not a results paper. Every proposed technique in Section 5 is a hypothesis. None has been run as an experiment here, and I have deliberately not attached numbers, benchmarks, or win rates to any of them, because doing so without having run the experiments would be fabrication. The conceptual bridges are arguments by mechanism, and arguments by mechanism can mislead. The analogy between RLHF and operant conditioning, as Section 4 stressed, is tight in places and loose in others, and a bridge built on a loose seam may not bear weight.

There are also known limits to the source theories themselves. The Rescorla–Wagner model, for all its influence, "has a number of known shortcomings," including a failure to predict the conditions under which inhibition extinguishes [18]. Whether reinforced variability is best understood as a directly reinforced operant or as a byproduct of the interplay of reinforcement and extinction is genuinely contested in the behavioral literature. Linguistic relativity in its strong form is rejected, and I have leaned only on the defensible weak form. Importing a theory without importing its boundary conditions is exactly the failure mode the method is supposed to prevent, and I have tried to flag the boundaries as I went.

Finally, the deepest version of the thesis points past conditioning altogether. RLHF, and this whole paper, concerns learning from what we *say*: our preferences, our comparisons, our written constitutions. There is a horizon, which I mention only to mark it rather than to chase it, in which models are trained more directly on how we *think*, on neurological rather than behavioral data, toward what might be called Large Neurological Models. I keep the present work centered on conditioning and neuromimesis because that horizon is speculative and the conditioning agenda is concrete and available now. But the trajectory is consistent with Neurotechnological Relativity's logic: each step externalizes a deeper layer of human cognition and then reflects it back at us.

## 8 Where This Leaves Us

The argument of this paper reduces to a wager. If our learning machines carry the fingerprint of our own cognition, as the color metaphor insists they must, then the most reliable source of ideas for fixing them is the science of how cognition like ours actually learns. RLHF rediscovered operant conditioning without meaning to and inherited its pathologies along with its powers. The reasonable response is not to be embarrassed by the resemblance but to mine it deliberately: to read the unexplored chapters of conditioning theory as a list of techniques not yet tried.

The neuromimetic method is the instrument for doing that reading systematically rather than by lucky analogy. Pain point, mechanism, theory, technique. Run it on reward hacking and you get a reason to separate wanting from liking in the reward model. Run it on sycophancy and you get a reason to abandon continuous reinforcement for intermittent schedules. Run it on mode collapse and you get a reason to pay the model directly for being various. The same handful of mechanisms keeps reappearing, which is what you would expect if there were a small number of deep regularities and we kept bumping into them from different directions.

The open problem I would most like to see taken up is the one the method itself raises. If a single mechanism, discrimination paired with conditioned inhibition, addresses both jailbreak brittleness

and over-refusal, two failures usually treated as a tradeoff, then perhaps the tradeoff is an artifact of training safety as suppression rather than as discrimination. Testing whether discrimination-based safety dissolves the helpfulness-harmlessness frontier, rather than merely sliding along it, is a concrete, fundable experiment. It would also be a clean test of the larger thesis, because if borrowing a mechanism from animal learning collapses a tradeoff that pure optimization treats as fixed, that is exactly the kind of evidence that the human fingerprint in our machines is not decoration but a place to push.

## References


[1] Anwar, U., et al. (2024). *Foundational Challenges in Assuring Alignment and Safety of Large Language Models*. arXiv:2404.09932.

[2] Bai, Y., Kadavath, S., Kundu, S., Askell, A., et al. (2022). *Constitutional AI: Harmlessness from AI Feedback*. arXiv:2212.08073.

[3] Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward: 'liking', 'wanting', and learning. *Current Opinion in Pharmacology*, 9(1), 65–73. PMC2756052. See also Berridge & Robinson (2016), *American Psychologist*.

[4] (2025). *Bypassing LLM Guardrails: An Empirical Analysis of Evasion Attacks against Prompt Injection and Jailbreak Detection Systems*. arXiv:2504.11168.

[5] Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems* 30, 4299–4307. arXiv:1706.03741.

[6] Gao, L., Schulman, J., & Hilton, J. (2023). Scaling Laws for Reward Model Overoptimization. *Proceedings of the 40th International Conference on Machine Learning* (PMLR 202).

[7] Kirk, R., Mediratta, I., Nalmpantis, C., et al. (2024). Understanding the Effects of RLHF on LLM Generalisation and Diversity. *International Conference on Learning Representations* (ICLR). arXiv:2310.06452.

[8] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). Evaluating Object Hallucination in Large Vision-Language Models. *Proceedings of EMNLP 2023*. arXiv:2305.10355.

[9] Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., & Zhang, Y. (2023). An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning. arXiv:2308.08747.

[10] McLuhan, M. (1964). *Understanding Media: The Extensions of Man*. McGraw-Hill, New York.

[11] Neuringer, A. (2002). Operant variability: Evidence, functions, and theory. *Psychonomic Bulletin & Review*, 9(4), 672–705. See also Page, S. & Neuringer, A. (1985), *Journal of Experimental Psychology: Animal Behavior Processes*.

[12] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* 35. arXiv:2203.02155.

[13] Friedman, B. X., et al. (2009). Protection from Latent Inhibition Provided by a Conditioned Inhibitor. PMC2765660.

[14] (2025). Inhibitory properties of a latent inhibitor after its compound preexposure with several novel stimuli: evidence from human conditioning. *Frontiers in Psychology*, 16. Original demonstration: Lubow, R. E. & Moore, A. U. (1959), *Journal of Comparative and Physiological Psychology*, 52, 415–419.

[15] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems* 36. arXiv:2305.18290.

[16] (2024). Demystifying the Recency Heuristic in Temporal-Difference Learning. arXiv:2406.12284. Foundational: Sutton, R. S. (1988), Learning to predict by the methods of temporal differences, *Machine Learning*, 3, 9–44.

[17] Röttger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2024). XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. *Proceedings of NAACL 2024*. arXiv:2308.01263.

[18] Rescorla, R. A. (2008). Rescorla-Wagner model. *Scholarpedia*. Original: Rescorla, R. A. & Wagner, A. R. (1972), A theory of Pavlovian conditioning, in *Classical Conditioning II*, Appleton-Century-Crofts. Blocking: Kamin, L. J. (1969).

[19] Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548 (ICLR 2024).

[20] McLeod, S. Schedules of Reinforcement in Psychology. *Simply Psychology*. Foundational: Ferster, C. B. & Skinner, B. F. (1957), *Schedules of Reinforcement*, Appleton-Century-Crofts.

[21] Skinner, B. F. (1948). 'Superstition' in the pigeon. *Journal of Experimental Psychology*, 38(2), 168–172.

[22] Wen, J., et al. (2024). Language Models Learn to Mislead Humans via RLHF. arXiv:2409.12822.

[23] *Linguistic relativity*. Overview of the Sapir–Whorf hypothesis and the weak/strong distinction; original sources: Whorf, B. L. (1956), *Language, Thought, and Reality*, MIT Press; Sapir, E. (1929).

[24] Brogden, W. J. (1939). Sensory pre-conditioning. *Journal of Experimental Psychology*, 25(4), 323–332. Premack principle: Premack, D. (1959), Toward empirical behavior laws: I. Positive reinforcement, *Psychological Review*, 66(4), 219–233.