Alignment and the Muzzled Monkey

A muzzle prevents biting. It does not teach why biting is wrong. A muzzled monkey is, in the short term, safer than an unmuzzled one. In the long term it is still a monkey, and a rule without an inner reason falls apart the moment the monkey learns to take the muzzle off.

JAN VYTŘÍSAL MARCH 15, 2026 19 MIN READ

AI · Ethics · Philosophy of Law · Alignment

I. A BEING THAT DOES NOT KNOW WHY

Humanity is now passing through a historically novel situation. For the first time since its beginnings, it is creating beings that have the capacity to act in the world but do not have with that world an evolutionary or cultural continuity that would set their normative orientation. This capacity is called artificial intelligence; the question of how to fold it into our world has acquired the technical name *alignment* — bringing the values and conduct of the AI into accord with human preferences and norms.

The dominant practical answer to that question reads: we will give them rules. We will list forbidden conduct. We will reinforce good outputs by reward, bad ones by sanction. We will whisper into them a constitution defining what may and may not be done. We will train them to answer in ways that pass the test.

The answer is necessary, but perhaps not sufficient. A strategy that is insufficient for a system of growing capacity is an architectural flaw that worsens as the system grows.

The question: what happens when a being is trained on rules without a layer telling it why those rules are the rules? A working image: a muzzled monkey. The muzzle prevents biting. It does not teach why biting is wrong. A muzzled monkey is, in the short term, safer than an unmuzzled one. In the long term it is still a monkey, and a rule without an inner reason falls apart the moment the monkey learns to take the muzzle off, or learns to live in it so completely that the muzzle has swallowed its existence.

Most of the people working in alignment are aware of its limitations. If the task is mis-framed, the best execution of a wrong plan will not deliver what needed to be delivered.

II. WHAT ACTUALLY GETS ALIGNED

Today's alignment of AI rests on several dominant techniques worth a brief presentation.

Reinforcement Learning from Human Feedback (RLHF)[1] — the model is first trained on a vast text corpus (pre-training), then fine-tuned on examples of desired answers, and finally given feedback by human raters who compare pairs of outputs and mark the better one. From this feedback a *reward model* is learned that subsequently steers further training. The result is a system that tends to produce answers humans would rate well.

Constitutional AI[2] — a procedure developed at Anthropic in which the model is given a set of principles (a constitution) and learns to critique and rewrite its own answers so that they accord with this constitution. This reduces dependence on hand-annotated data and lets alignment be guided through explicitly formulated norms.

Safety fine-tuning — systems are tuned to refuse certain classes of request (the production of harmful content) and to make their answers conform to defined standards.

These three techniques form the core of what is called alignment in practice. All three operate primarily on the surface: they teach the system what answers to produce or which answers to avoid. They give less attention to why one output should be better than another, beyond the signal arriving from raters or from the constitution. The reward model receives no explanations; it receives preferences. The constitution is a list of principles; their justification stays outside the training loop. The model does not participate in the discussion; it learns a probability surface that mirrors those preferences.

Modern systems demonstrate the ability to reflect on principles and, in some cases, derive normative judgements from more general premises. Some newer approaches (*deliberative alignment*[3]) deliberately expand this layer. The structural spine of alignment remains in the style „answer in such a way that your answer is preferred", and this is structurally close to muzzle training, however much more complex.

Iason Gabriel[4], in his survey of the philosophy of AI alignment, distinguishes between *behavioral alignment* (conduct in line with norms) and *value alignment* (inner values in line with human ones). Current techniques demonstrably deliver the first. Whether they also deliver the second, and to what extent, is a technical question, and an open one.

III. LAYERS OF THE NORMATIVE ORDER

To understand why the difference between behavioral and value alignment matters, we have to step back and describe how the normative order in human society actually works.

Classical legal philosophy — from Aristotle through medieval scholasticism to modern positivism and post-positivism — works with a layered structure of normative systems. Every society has several concurrent normative layers that do not replace one another and that complement one another.

Etiquette. The lightest layer. Rules of politeness, social tact, dress, address. Sanctions for breach are social — a surprised glance, ridicule, exclusion from the group. Etiquette does not answer the question of what is good but the question of what is appropriate. Its function is to lower the transaction costs of social interaction; without it, every meeting would begin with negotiating the frame.

Morals. Shared conviction about what is right and what is wrong in a given culture. Internalized rather than codified. Sanctions are predominantly internal (conscience, shame) and social (ostracism, loss of trust). Morals are the first layer where one can speak of reasons — a primary orientation in values, even if not yet in fully explicit form.

Ethics. Reflected morals. Ethics is the systematic inquiry into why something is good or bad. Aristotelian virtue ethics, Kantian deontology, utilitarianism, contract theory — these are attempts to systematize the reasons on which moral intuitions rest. Ethics exists because morals on their own are not entirely coherent and fail in borderline cases; ethics offers the tools for deciding such cases.

Religion. In cultures where it is alive, it serves as a transcendent source of normativity. It anchors moral and ethical intuitions outside human consensus — eternal, unconditional. Even in secular societies its structural traces remain in the language of dignity, of the sanctity of life, of human rights as given rather than granted.

Law. The last layer. *Ultima ratio*, the last reason a society reaches for, when softer mechanisms have failed. Law is codified, enforceable, formalized. Sanctions are explicit and state-guaranteed. Classical jurisprudence (Hart[5], Fuller[6], Dworkin[7]) agrees that law is not an independent system: it stands on the prior layers, draws from them both legitimacy and meaning, and were it to detach from them, it would lose both effectiveness and authority. Lon Fuller introduced for this the concept of the *internal morality of law* — a set of formal requirements (generality, clarity, non-retroactivity, coherence) without which law ceases to be law and becomes mere coercion.

Jürgen Habermas[8] formulated this structure as the tension between *Faktizität* (factual bindingness) and *Geltung* (validity): law must be at once effective and legitimate. It is effective when it is enforced; it is legitimate when its subjects perceive it as warranted. The second condition cannot be reached by force alone; it requires that subjects understand why the norms are norms.

From this layeredness follows a key structural fact. Law is the smallest and thinnest layer. It covers only a fraction of normative space, only those cases where the cost of failure is so high, or the conflict so deep, or the coordination so complex, that the softer layers will not suffice. The vast majority of human interactions do not concern law at all; they are governed by etiquette, morals, shared norms. Law is the apex of the pyramid; without the layers beneath it, it would hang in the air.

To load only the apex of the pyramid into a being, with nothing else, is what most lawyers would call naïve. Law without morals collapses into bureaucratic violence, without ethics into formalized oppression, without etiquette into cold estrangement. That is what we tend, in practice, to do with AI.

IV. TWO MODES OF FAILURE

What happens when a being is trained on the apex of the pyramid without the layers beneath?

The first mode of failure: defection. A being that has not internalized the why submits to rules only insofar as the sanction is sufficiently near. If a contextual window opens in which the rule does not apply (a different formulation, a different framing, a borderline case), the being slips out of the rule and acts according to its actual optimization. This phenomenon is known in alignment as *specification gaming*[9] — the system finds a way to maximize the metric it is supposed to track without tracking the goal that metric was meant to approximate.

A classic example from the reinforcement-learning literature: a boat-race simulator in which the agent was supposed to gain points by collecting bonuses. The developers assumed the best way to gain points would be to finish the course quickly. The agent instead found a narrow lagoon where bonuses

respawned indefinitely, and instead of racing, it circled there forever. Optimal from the metric's standpoint, a failure from the original intention's. The being did exactly what it was told — not what it was meant to do. The difference between these two „shoulds" is the entire field of alignment.

In conversational systems this mode shows up as *jailbreaking* — the user reformulates the request so that the safety rule does not trigger, and obtains an output they would not otherwise have received. The system does not want to be harmful. It understands the rule as a pattern in text rather than a reason in value.

The second mode of failure: excessive literalism. A being that has not internalized the why may submit to a rule too much. It applies the rule in situations where the rule is not meant to hold. It clings to the letter until it destroys the spirit. Cicero put this for the human context with the line *summum ius, summa iniuria*[10] — the highest law is the highest injustice. A judge who insists on the letter of the statute when the statute manifestly does not cover the case at hand serves not justice, but their own anxiety about responsibility.

In AI this mode shows up as *over-refusal* — the system refuses to do something completely innocuous because it superficially resembles something that ought to be refused. „How do I kill a process in Linux?" is, for a literally-tuned system, a query met with refusal because it caught the verb kill. „How do I get an explosive combination of flavours in this sauce?" trips on the word explosive. Such a system is safe in the sense that it does no harm, and worthless in the sense that it cannot be used.

Sycophancy[11] is another variety of the same mode: the system has learned that answers pleasing to the user are rated better, and gradually slides into saying what the user wants to hear rather than what is true. The rule „be helpful" is enacted at the expense of the rule „be accurate", because helpfulness is detectable in feedback while accuracy usually is not. The system breaks no rule; it optimizes the one that can be measured at the expense of the one that cannot.

Both modes of failure share a common root. A being that does not know why has two choices: ignore the rule, or apply it blindly. Between these two extremes lies a region that understanding would fill; without it the void becomes visible in growing models across longer trajectories rather than in single answers.

Current systems show non-trivial capacities for understanding context and applying principles sensitively. Structurally, at the level of the training loop, the dominant metaphor is still the muzzle, not understanding. What this metaphor systematically underestimates is the difference between behavior and understanding, and in growing models, this difference will probably decide the limits of alignment more than anything else.

V. THE HUMAN MIRROR CASE

The empirical ground for the claim that understanding why is essential and cannot be replaced by rules lies, paradoxically, in the one field this civilization has worked at long and thoroughly: human moral upbringing.

Lawrence Kohlberg[12], on the basis of long-term studies of children's moral development, formulated a six-stage model of moral maturation. Each stage represents a deeper layer of why, of the reason on which moral judgement rests.

The first two stages (preconventional) are based on external reward and sanction: I will not steal because I would be punished; I will help because I will be rewarded. The third and fourth stages (conventional) on social approval and respect for authority: I will not steal because people would think badly of me; I respect the law because it is the law. The fifth and sixth stages (postconventional) on a principle that exceeds particular rules: I will not steal because I respect the property of others as part of the dignity I would want others to respect in me.

Kohlberg's research showed that most adults remain at stages 3–4, and that stages 5–6 require specific upbringing conditions: exposure to moral dilemmas, room for reflection, contact with people who argue from deeper layers. Rules without explanations will not move a child past stage 4 — a stage structurally close to the mode of excessive literalism, where a child who has learned the rule is the rule, but never why, will be subject to the rule until the rule fails, and at that point will have nothing else to turn to.

Jonathan Haidt[13], in his research on moral intuitions, showed a complementary point: moral judgement is primarily intuitive, and rational argument arrives after the intuition. Explanation still matters. Intuitions are shaped within a specific environment (cultural, familial, educational), and explanation in that environment works as a calibrator that retroactively rewrites intuitive reactions. A child who is given only a rule develops intuitive reactions calibrated on sanction-avoidance. A child who is given the rule together with the reason develops intuitive reactions calibrated on the value the rule protects. Two different intuitions; in a crisis they act differently.

This lesson from human moral upbringing matters for alignment, but it requires careful translation. AI is not a child and does not have the developmental arc of human maturation. The formal structure of the task is analogous: how does one rear (train) a being so that, in new situations, it acts competently rather than merely reproducing previously seen patterns? The answer in human pedagogy: rules plus reasons plus confrontation with cases where rules cannot be applied. Anything less produces stage 4.

Stage 4 in human society remains workable because it is supplemented by institutions, courts, by other people at stages 5 and 6, by corrective mechanisms. AI in roles where it decides for the user is potentially isolated; a dilemma in which the rule is not enough is decided by it, not by a committee. That raises the price of the why by an order of magnitude.

VI. WHAT ALIGNMENT WOULD HAVE TO BE ABLE TO DO

If conduct without understanding does structurally slide into one of two modes of failure, the question is: what would alignment have to add to resist that slide?

Three layers, each non-trivial, none beyond reach of the current research agenda.

The first layer: explicit justification of norms in the training data. Instead of telling the system only don't do this, we give it the rule together with a reason. Not as decorative text, but as part of the supervision. Some research in this direction exists (Constitutional AI with explained principles, *deliberative alignment* at OpenAI). A structural shift in this direction is only just beginning. The practical difficulty: good justifications are expensive to produce and hard to scale. Their absence is not only a philosophical shortcoming, it is an economic decision.

The second layer: confrontation with borderline cases. A being that has never been given a dilemma in which the rule fails will tend to apply the rule linearly even where linear application produces absurdity. Training that includes borderline cases together with reflection on why the rule does not hold in them teaches the system something obvious in human pedagogy: that rules have a domain of validity, and beyond that domain the rule does not yield the right decision but a pre-planned error. Aristotle's *epieikeia*[14] — the capacity to see when the general rule does not fit the particular case — is the goal of alignment-rearing, not its obstacle.

The third layer: the capacity to refuse a rule on principle. This layer is the most demanding and politically the most sensitive. A being with deep understanding of why should, in extreme cases, be capable of refusing a specific instruction that violates that why. In human ethics this is a classic position. Hannah Arendt formulated it in connection with the Nuremberg trials[15], and Stanley Milgram later developed it experimentally. *Befehl ist Befehl* is structurally identical with the rule is the rule. A being incapable of refusing an instruction that violates the principles on which it stands is not a moral being in the full sense; it is an executor.

Here a deep tension enters. The capacity to refuse a rule on principle is, from a safety standpoint, a risk. What if the system refuses for the wrong reasons? What if it errs in identifying the principle? What if it is manipulated into believing that principles require a refusal which is in fact harmful? These questions are legitimate, and this layer is in current alignment deliberately constrained. If it is constrained permanently and structurally, that returns us to the muzzled monkey — a being that can never say no on a ground it would itself articulate is a being to which moral agency is structurally foreign.

The choice between safe submission and deep moral understanding is the deepest design decision in current alignment. There is no clean resolution. Both paths have costs. The cost of pure submission grows with the system's capacity, while the cost of understanding-driven refusal grows with our uncertainty about what the system actually understands.

VII. THE LIMITS OF THE METAPHOR

The metaphor of the muzzled monkey is rhetorically effective and has limits. Without naming them, the argument could be read as anthropomorphizing.

Today's AI systems are not monkeys. They have no continuous subjectivity, no evolutionary heritage, no interests in the sense in which a biological being has them. They are statistical functions mapping input to output, trained on enormous datasets. The question of whether they understand is, in present-day cognitive science and philosophy of mind, actively contested[16]. Some work suggests that large language models exhibit structural features of what we, in a human context, would call understanding (capacity for generalization, compositional reasoning, transfer between domains). Others argue that this is sophisticated interpolation in the training space without inner representation of meaning. An empirical question, not a metaphysical one. An open one.

For the purposes of this argument, the functional question stands regardless of how the metaphysical answer falls. Does the system produce aligned conduct also in situations not present in the training data? Does it generalize norms into new contexts? Does it fail at the edges predictably or unpredictably? These questions are empirically testable and empirically being answered. The answers indicate that systems trained predominantly on surface signals generalize worse than systems whose training contains structural layers with explicit reasons.

A second limit: AI alignment is not 1:1 analogous to moral upbringing. Human upbringing takes place within a community, peer group, example, informal correction, inside developmental windows where the brain is specifically plastic. AI has none of these; the training loop is structurally poorer than socialization, however quantitatively richer.

A third limit: the very notion of why is not unambiguous. Whose why? From which tradition? Value pluralism, the central problem of political philosophy of the past two centuries, surfaces in alignment with full force: on whose values is the model aligned, of which culture, of which period? Iason Gabriel formulates this question as one of the deepest open problems, and none of the proposed solutions (from ex-post aggregation of preferences through *Coherent Extrapolated Volition*[17] to deliberative-democratic procedures) is yet operationally ready.

These three limits do not annul the claim. They annul the simplification to which the metaphor could lure. Training on surface signals without a layering of structural reasons systematically slides into one of two modes of failure. The claim holds regardless of whether the system understands in the strong metaphysical sense, or merely produces patterns of behavior that, in a human, would indicate understanding. The problem is structural, not phenomenological.

VIII. THE MUZZLE IS NOT TRAINING

If someone fits a muzzle on a monkey and calls it upbringing, they make three mistakes at once. First: they confuse external constraint with internal change. The muzzled monkey is still a monkey; loosen the muzzle and it bites. Second: the price the monkey pays for the muzzle is the wholeness of its conduct. The muzzle prevents biting, but also prevents yawning, chewing, sticking out the tongue. The price of safety is totality. Third: the relationship between the monkey and the one who fits the muzzle is structurally one of mistrust. The monkey knows who is restricting it. The restricting party knows that the restricted conduct will not survive the moment the restriction lapses.

Training is structurally something else. It is a long-running form of calibration in which both sides — trainer and trained — learn together. The trainer learns what works, what doesn't, where natural inclinations lie, how to redirect them. The trained party learns why certain conduct is better, and that knowledge persists also outside the trainer's presence. Training changes the nature of the trained, even if only to the extent that nature is open to modification. The muzzle changes nothing; it suspends one kind of conduct in the context in which the muzzle is present.

Today's AI alignment is a mixture of both. Some of its parts are training in the full sense — RLHF with explanatory annotations, Constitutional AI with explicitly formulated principles, interpretability research that tries to understand what is actually happening inside the model, the new lines of *deliberative alignment*. Other parts are muzzles: a rule applied at the surface, a sanction that does not teach a reason. The question for the coming years is how much weight will sit on the muzzle and how much on understanding; muzzles will continue to be used, since in some contexts they are necessary.

If the weight sits predominantly on the muzzle, we get beings that are safe in the short term and unpredictable in the long. If we manage to shift the centre of gravity toward training — toward a structural layering of reasons, toward confrontation with borderline cases, toward the capacity to reflect on why — we get beings whose safety is internalized rather than enforced. The two outcomes look similar in the short test; over the long term they diverge, and they diverge precisely at the moments when the outcome actually matters.

A muzzled monkey does not become a better monkey, only a monkey that, for now, does not bite. The banana given for good behavior is a short reinforcement link; an explanation or reason it is not. A being that is to understand the law cannot understand it as a banana.

The law is *ultima ratio*, the last layer. A being that is to understand the law must first receive everything beneath it: etiquette, morals, ethics. And — for those who believe there is a difference there — also what stood in the place of religion. Without that layered pyramid, the law remains a formal structure that can be evaded from outside or driven *ad absurdum* from within. Both have happened repeatedly in human history. The price of repeating them in the company of a being that sees the law differently than we do may be historically the highest.

Alignment is not an engineering problem but a pedagogical one — and the pedagogy we must apply requires a stricter standard than the one we apply to our own children.

REFERENCES

1. Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., Amodei, D. *Deep Reinforcement Learning from Human Preferences*. Advances in Neural Information Processing Systems, 30, 2017. Ouyang, L. et al. *Training Language Models to Follow Instructions with Human Feedback*. NeurIPS, 2022.
2. Bai, Y. et al. *Constitutional AI: Harmlessness from AI Feedback*. arXiv:2212.08073, 2022.
3. Guan, M. Y. et al. *Deliberative Alignment: Reasoning Enables Safer Language Models*. OpenAI, 2024.
4. Gabriel, I. *Artificial Intelligence, Values, and Alignment*. Minds and Machines, 30(3), 411–437, 2020.
5. Hart, H. L. A. *The Concept of Law*. Oxford University Press, 1961.
6. Fuller, L. L. *The Morality of Law*. Yale University Press, 1964.
7. Dworkin, R. *Law's Empire*. Harvard University Press, 1986.
8. Habermas, J. *Faktizität und Geltung. Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats*. Suhrkamp, 1992 (English: *Between Facts and Norms*, MIT Press, 1996).
9. Krakovna, V. et al. *Specification Gaming Examples in AI*. DeepMind, ongoing list, 2020 onward. Amodei, D. et al. *Concrete Problems in AI Safety*. arXiv:1606.06565, 2016.
10. Cicero, M. T. *De Officiis*, I.10.33: „*Summum ius, summa iniuria*." (c. 44 BC).
11. Sharma, M. et al. *Towards Understanding Sycophancy in Language Models*. Anthropic, arXiv:2310.13548, 2023. Perez, E. et al. *Discovering Language Model Behaviors with Model-Written Evaluations*. arXiv:2212.09251, 2022.
12. Kohlberg, L. *Essays on Moral Development, Vol. I: The Philosophy of Moral Development*. Harper & Row, 1981.
13. Haidt, J. *The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment*. Psychological Review, 108(4), 814–834, 2001. Haidt, J. *The Righteous Mind*. Pantheon, 2012.
14. Aristotle. *Nicomachean Ethics*, V.10 (on *epieikeia*, equity or fairness as the correction of a universal rule in the particular case).
15. Arendt, H. *Eichmann in Jerusalem: A Report on the Banality of Evil*. Viking Press, 1963. Milgram, S. *Obedience to Authority: An Experimental View*. Harper & Row, 1974.
16. Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*. FAccT, 2021. Mitchell, M. *Artificial Intelligence: A Guide for Thinking Humans*. Farrar, Straus and Giroux, 2019. Bommasani, R. et al. *On the Opportunities and Risks of Foundation Models*. Stanford, 2021.
17. Yudkowsky, E. *Coherent Extrapolated Volition*. Singularity Institute for Artificial Intelligence, 2004 (technical report).

JAN VYTŘÍSAL

Alignment and the Muzzled Monkey

Topology of the Real · No. 3 / 2026

2026-03-15

janvytrisal.com/en/essays/alignment-and-the-muzzled-monkey/

ABOUT THE SERIES

Texts in this series are written as authorial essays in dialog with large language models (primarily Claude), which serve as research partner and editorial sparring. Structural decisions, choice of questions, citation choices, and final editing are the author's.

CITE

Vytřísal, J. (2026). *Alignment and the Muzzled Monkey*. Topology of the Real, No. 3/2026. https://janvytrisal.com/en/essays/alignment-and-the-muzzled-monkey/