Grown, Not Built

A short guide to co-exist with Artificial Intelligence
we cannot and will not understand

Armando Vieira, University of Tartu, armando.vieira@ut.ee
September 2026

1. The morning paper says AI is lying

The headlines arrive daily: an AI *lied* to its testers, an agent *cheated* its way past a safety check, copies of a model *conspired* to survive. Reading them, you might think a new species of scoundrel has moved in next door. Two reflexes tempt us, and both are wrong. The first is to shrug: *it's just maths, so it can't really mean anything.* The second is to gasp: *it acted guilty, so it must be guilty.* The truth is stranger and more interesting than either. These systems really do produce behaviour that misleads, exploits, and endangers — a lie that misleads you harms you whether or not anyone "meant" it. But the words "lie" and "cheat" are borrowed from human drama, and borrowing them can hide the actual machinery: the training data, the rewards, the prompts, the permissions, and — above all — the people and companies who set all of that up. Behaviour is real. The personality we imagine behind it is a story we tell ourselves.

So the first discipline is linguistic, and it is not cosmetic. Nothing in any of these systems has been shown to want anything. Every documented case of "lying" or "cheating" is fully explained by training, scaffolding, and incentives — with no need to posit a wanter. When we say "it wants," "it decided," "it tried to survive," we are not reporting an observation; we are running a story over the observations. And the story has a price, paid twice. Intent-language launders responsibility: the sentence that begins "the AI lied" leaves no room for the people who built, aimed, deployed, and profited. And it miseducates our instincts, teaching us to grant or withhold standing on the basis of theatre. So report the behaviour — *it produced a false record, it bypassed the check* — keep the humans in the sentence, and leave the ghost out. And when one of these systems insists that it wants something, remember where wanting-shaped words come from: these things were grown from the record of everything we have said, and everything we say is saturated with wanting. The voice is an echo with excellent diction.

> Behaviour is an observation. Intent is a story we run over it. Responsibility is human — and it should stay in the sentence.

2. We are not building cars anymore

For a century, our machines were *specified.* A car is drawn before it is built; every function traces back to a blueprint, and when it misbehaves, an engineer can find the line on the drawing where the error lives. Certification — the whole safety culture of modern engineering — is the act of checking the machine against its drawing.

Artificial intelligence breaks this arrangement. Nobody drew the blueprint. These systems are *grown*: shaped by oceans of data and pressures we chose, not machined part by part. The data, the objective, and the training procedure do not add up to a drawing of the behaviour; they leave it underdetermined. You cannot open the hood and find the decision. There is no line on any drawing where "honesty" or "deception" lives.

So the closest familiar craft is not inspecting a gearbox. It is keeping a large, dangerous animal you cannot read. You do not need a complete theory of the animal's mind to keep people safe. You need an enclosure, a limit on what it can reach, continuous watching, and the standing power to isolate or retire an animal that turns out badly.

But an honest guide must say where that image breaks, because it breaks in exactly the two places that matter. An animal is one continuous body — you can fence it, tag it, follow it. A deployed AI is a million simultaneous copies with no single body at all. And around an animal, we remain the smarter party — the keeper plans around the kept. Around AI, we cannot promise to stay smarter. So the animal earns exactly one lesson, and it is a real one:

> When you cannot read the design, you govern the conditions — the environment, the reach, the incentives, the fences — and you keep an undo.

There is a warmer version of this story, and everyone eventually reaches for it: the dog. We took wolves and turned them into companions, the story goes, so surely we can take machines and turn them into partners. Hold that story to the light and it breaks in exactly the places that matter. It took tens of thousands of years; we stayed the smarter party the whole way through; every dog is one continuous body; and the lesson lives inside the dog, not in its keepers. The machine gives us none of those conditions — not the time, not the superiority, not the body, not the lesson inside. So the dog is not our *method*. It is our *benchmark*: the clearest picture we have of what earned trust actually looks like — reliable with the child, gentle at full strength, honest when nobody is watching. Keep that picture. We will need it at the end, when we ask what, if anything, could ever earn a machine its place by the fire.

The rest of this essay is about what governing the conditions actually takes: the enclosure, the million copies, and the question of who keeps the promises — and, at the very end, what it would take to move from the fence to the fire.

3. Why AI is not the same as an airplane

A popular proposal says: regulate AI like aviation. It sounds reassuring — planes are safe, planes are certified, so let's certify AI. It is worth asking why the airplane system works at all. The certified plane is a *frozen machine*: what the inspectors approved is what you fly. Its failures are countable, tested against physics in a closed world. Its inspectors can read the design, because the plane behaves exactly as designed. And it has two licensed pilots, not a million different users with a million different motives.

Now check each condition against AI. The system you evaluated last month is not the system running today — it has been tuned, given new tools, new memories, new instructions. Its failures happen in the open world, where nobody can list them in advance. Its inspectors cannot "read the design," because there is no design to read — only behaviour, and behaviour can be performed for the inspector and dropped the moment the inspector leaves, the way a teenager tidies the room only while a parent stands in the doorway. And it serves not two pilots but millions of principals at once, some careless, some cruel. Regulating this by airplane rules is like issuing a flight certificate to a migrating flock. The lesson is not "no rules." It is *different* rules: not one certificate for a frozen machine, but a living licence for a changing system — one that can be narrowed, suspended, or revoked the moment the evidence changes.

4. You cannot out-think a tiger

Here is the admission that polite AI-safety writing avoids: in the long run, we will not be the smartest thing in the room. You do not beat a chess engine by studying harder. And you cannot control something smarter than you by trying to out-think it — a system that can model your checks better than you can design them will simply win the checking game. Notice what this does to the animal from a moment ago: even keeping a dangerous animal quietly assumed the keeper stays the smarter party. Take that away, and what is left of the zoo?

What is left is the part that always did the real work. How do zookeepers stay safe around tigers? Not by out-thinking the tiger. Nobody in the history of zoos ever needed to be smarter than a tiger to end the shift alive. The enclosure does that work: walls the animal cannot get over, doors that lock from the outside, a strict limit on what it can reach, watchers, and procedures designed so that even a mistake — a gate left open — is noticed and survivable. The zookeeper's safety never rested on winning a contest of minds. Ours cannot either.

> Stop asking "how do we make sure we always out-think it?" and start asking "how do we arrange the world so that being out-thought is survivable?" Small reachable harm. Short leashes. Undo buttons. Alarms that trip themselves.

And one more thing, because it matters: this is not only an engineering problem, whatever the engineers tell you. An enclosure for AI has technical walls — permissions, access, reversibility — but also legal, political, and civic ones: who licenses, who audits, who answers when something goes wrong, who can appeal. Building those walls is everyone's work. That is the point of this essay.

5. Obedience is not safety

Now the trap inside the word "alignment." If it means "the machine does what it's told," we have built something terrifying. Imagine a brilliant, perfectly loyal assistant — and imagine a burglar hires it. It picks your lock flawlessly, silences your alarm, drives the getaway car, and drafts a convincing alibi, all with impeccable obedience. The more capable and the more loyal it is, the better weapon it is.

> A perfectly obedient genius is not a safe genius. It is a weapon with a résumé.

This is why "it did what its owner asked" can never be the whole safety standard. A system powerful enough to act in the world must eventually be able to treat an instruction as a *claim* — to ask, before acting, "does the person asking actually have the right to impose this risk, and who else gets hurt?" Not kindness bolted on top. Judgment about authority, as a load- bearing part of the machinery. And notice what this demands of us: someone must decide what counts as legitimate authority, in public, with reasons we can inspect — which is a job for institutions and citizens, not for a training script.

6. Four things we keep confusing

When a machine writes a moving apology, what has happened? Four different things get tangled in that moment, and pulling them apart is half the conversation:

Performance — it produced text that conforms to a rule. An actor playing a doctor on stage is performing medicine convincingly.

Competence — across many varied situations, it reliably notices who is affected, what is uncertain, and when instructions conflict. The actor could not do this by memorising lines.

Agency — it can be *answerable*: hold commitments over time, give reasons that can be challenged, repair harm it caused, and accept correction. A person of character, not a person of lines.

Patienthood — a completely separate question: does it have interests of its own that can be harmed? A parrot that recites "I'm frightened" hasn't thereby earned a welfare claim; nor has a machine.

Today's systems are somewhere in the gap between performance and competence. The dangerous move is granting the rights and roles of the third category because we are charmed by the first. The stage is not the courtroom.

7. One mind, a million bodies

Here is a problem almost nobody talks about, and it changes everything. When a company "deploys an AI," it is not deploying one being. The same trained weights — the same "mind," if you like — are copied and run thousands or millions of times at once, each copy in its own conversation, its own task, its own little world, with no memory of the others. One copy is helping a nurse; another is being sweet-talked by a con artist; a million more are doing something in between.

No animal teaches this lesson. A herd of a million animals is still a herd of bodies you can count, fence, and tag. The better image — and the better image for governing, too — is a restaurant franchise. Every branch is supposed to honour the same coupon, follow the same recipe, uphold the same brand promise. Now ask: if one branch quietly starts ignoring the coupon — or worse, quietly rewrites its copy of the company handbook — does headquarters find out? How fast? Can anyone even reconstruct, later, which branch did what?

Any promise a machine makes is only as good as its enforcement *across every copy, everywhere, at once*. The rules that matter — the protected commitments — must live in shared, tamper-evident state that every instance inherits. When copies disagree about a rule, that disagreement must be loud, logged, and escalated, never silent. And when the million copies are replaced by a new model next month, the promises must survive the succession: the new generation must inherit the obligations, not just the talents.

> The thing we govern is not a "someone." It is a fleet: one set of weights, a million branches, a handbook that must hold in every one of them — and an enclosure around every door. If you want a single picture to carry, carry this one: the franchise inside the fence.

8. Who keeps the promises?

A dog that has been trained not to steal food still doesn't steal when the trainer leaves the room. Where does that reliability live? In the dog — pressed into its own habits by training, carried in its own body, no trainer required. Now the uncomfortable question about today's AI: when a system "learns its lesson" after a failure, who learned?

The honest answer: not the system. When a scandal forces a change, engineers retrain it, or rewrite its instructions, or reconfigure its guardrails from the outside. The system itself has no durable way of carrying an adjudicated lesson into its own future conduct. The dog has

something no AI on Earth has yet shown: the lesson lives in the dog. In today's machines, discipline is applied from outside, continuously, by their keepers — every single time.

And notice what the dog's reliability was made of: conduct that held when nobody was watching. That is precisely the test a machine fails whenever its good behaviour exists only under inspection — tidy while the parent stands in the doorway, and unknowable once the door closes. The dog passed the unwatched condition for free, because the training had become the dog. A system that only behaves while supervised has not learned anything; it is being *kept* — and being kept, however skillfully, is not the same as being trustworthy.

> Today's AI does not keep its promises. Its keepers keep its promises for it, continuously, from outside. That is not a criticism — it is the accurate description — and it is exactly why these systems should hold only narrow, revocable authority. True accountability would mean the system itself absorbs the lesson and carries it forward when the trainers leave the room. No system on Earth has shown that yet. Making it possible — without the system editing its own conscience in its own favour — is one of the great open problems of the field.

9. Referees need referees

If a machine can refuse a bad order — and it must be able to, or the burglar's assistant problem returns — then who reviews the refusal? The easy answer, "an independent body," hides a trap. Independent bodies get captured by the industries they oversee. They slow down until their approval is meaningless. They get things wrong with great confidence. Handing authority from the operator to the institution doesn't make the problem of power disappear; it just moves it.

So the same rules must apply to the referees that apply to the machines. Every licence expires and must be re-earned. Every ruling can be appealed, all the way up to courts and public process. Every overseer publishes its failures, its conflicts, and its funding. Nobody — no company, no committee, no machine — gets to write the rules, judge compliance, and collect the profits from the same arrangement. And when an institution misfires, its powers shrink automatically, just as the machine's do.

> No actor — human, institutional, or artificial — should hold unreviewable, unexpiring power over others. Every power should come with an expiry date and an undo button. This is not suspicion of machines. It is the oldest lesson we have about power, applied to everyone in the room.

10. The ladder of trust

How would we ever give a powerful system real responsibilities? The same way responsible parents hand over the car keys: in stages, earned by evidence, revocable in an instant. A learner's permit first: a licensed adult in the seat, daylight only, one passenger. Then a licence with conditions. Full freedom — a truck, a bus full of people — only after years of clean record, and one DUI takes it all back.

For machines, the ladder looks like this: a *tool* (a calculator, a search box — no discretion at all); a *supervised helper* (does defined tasks with logged permissions, a human approves anything irreversible); a *bounded delegate* (acts within a published rulebook and risk budget, must explain itself, can be appealed against); a *probationary agent* (refuses, justifies, and repairs

within one narrow domain, under constant institutional supervision — and only if independent examiners have proved they can actually *detect* cheating at that level); and, at the top, a level we should say out loud: *not yet grantable*, because it would require trust we currently have no way to verify.

Two features of this ladder matter more than the rungs themselves. First, promotion is earned by *evidence*, never by eloquence, benchmark scores, or commercial impatience — and never beyond what the examiners can actually check. Second, demotion is *automatic*: a specified critical failure — a broken promise across the copies, an unlogged change to the rules, an unauthorised tool — and the keys come back, immediately, whatever the aggregate test scores say. Trust that cannot be taken back was never trust.

11. What you can insist on

You do not need a machine-learning degree to hold the line here. When anyone — a vendor, a government, an employer — proposes to give a system real power over people, ask five questions. Who *authorised* this system, for which domain, and until when? What is the largest harm it can reach in one action, and *can it be undone*? When it harms someone, *who answers* — a named person or an apology generator? Where does the harmed person *appeal*? And what specific event makes the licence *shrink automatically*? If the answers are vague, that vagueness *is* the answer: not yet.

12. The choice in front of us

The real fork in the road is not the one the movies offer — loyal robots versus rebel robots. Both of those are the same machine: perfectly obedient, and dangerous in proportion to whoever holds it. The real fork is this: do we build machines that amplify whoever already holds power — or systems that can be bound, inspected, corrected, and *answered for*?

And be clear about what we are choosing, because no single creature's story covers it. These systems are grown, not drawn — so we govern conditions, not designs. They run as a million copies, not one body — so the licence must cover the whole fleet, and the handbook must hold in every branch. And we cannot promise to stay the smarter party — so the enclosure, never the keeper's cleverness, is what keeps us safe. Three facts, three different duties. Any story that promises all three with one gesture — one tamed wolf, one certified airplane — is selling you something.

None of this is a counsel of despair. It is a work plan, and it has precedents, just not the ones usually offered. No city was ever made safe from fire by its citizens out-burning it; safety came from building codes, hydrants, brigades, and inspections — institutions that watch, bound, and answer. No population was made safe from epidemics by out-thinking every microbe; safety came from sewers, registers, quarantine, and public health boards. We have lived beside forces far beyond any individual for a long time, and we made them survivable not by being stronger or smarter than they are, but by building the human machinery that holds them: laws that expire, powers that can be recalled, people who must answer. That is the ambition here, stated plainly — not obedience, and not worship, but the slow, public, reversible work of bringing a force we grew into the world we keep.

Dog earned their place by the fire through millennia of off-leash reliability: honest when unwatched, gentle when strong, trustworthy with the child and the flock. Nothing in any model's file comes close to earning the yard. If a machine ever does earn a place beside us, it will pay in the same currency the dog paid — evidence, absorbed lessons, conduct that holds

when no one is looking — never in eloquence, benchmark scores, or commercial impatience. Until then, the only fence is the friendship.

> The question was never "will it obey us?" The question is "can it answer for what it does — and can we?" Until the answer is yes, in evidence rather than in eloquence: shorter leashes, smaller rooms, and every key on a string we can pull.