XONTOOLS RESEARCH SERIES · NO. 1

The Extreme Oversight Hypothesis

Could enough windows make honesty the cheapest path?

A speculative idea in two versions, set out carefully: the claim they share, the analogy behind it, the tension at its center and how each version answers it, the windows it would need, how trust in those windows could be bootstrapped, what it would cost, and every way it could be wrong.

Ian MacKenna · xontools.ai · September 2026

> **Status.** This is a hypothesis, not a result, and not a plan. It sits on one side of an open debate in alignment research. Parts of it can be tested at small scale; the full version could only be tested by an organization with frontier-scale resources. Every figure that looks like data is illustrative. Related documents, and the tools being built to test the idea, are at **xontools.ai**.

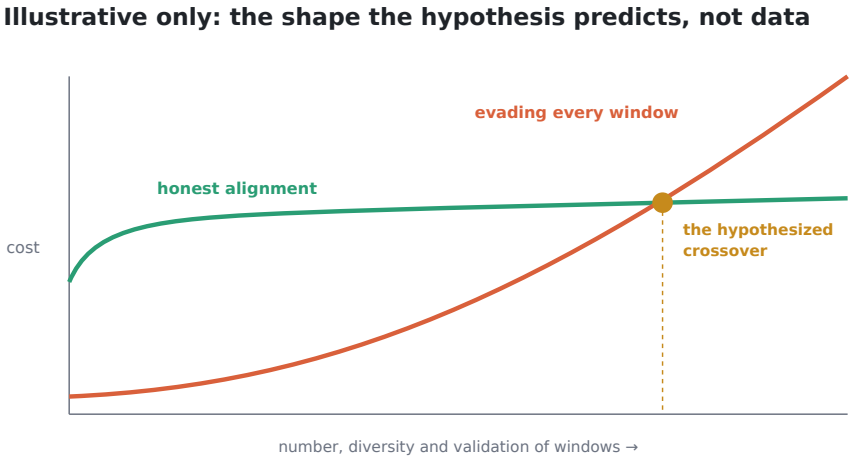

Figure 1. The hypothesis in one picture. As windows grow in number, diversity and validation, the cost of evading all of them rises faster than the cost of simply being aligned.

1. The hypothesis

THE SHARED CLAIM

With a sufficiently large, diverse, and validated set of windows into a model's internal states, honest alignment becomes cheaper for training to find than evading every window at once.

A **window** is a small read-only network attached to a model's internal states and trained to report one property: is the model straining against its sources, is it guessing, is it pursuing a goal other than the one it states. The hypothesis comes in two versions. They share the claim above and differ in when, and how, the windows are allowed to push on training.

VERSION A

Selection

With a sufficiently large, diverse, and validated set of windows into a model's internal states, some used for coarse training decisions and many held out, and with trust in those windows extended step by step, honest alignment becomes cheaper for training to find than evading every window at once.

VERSION B

Development

With such windows present from the start of training, pressing toward honesty while the model is still too weak to afford deception and then withdrawn into decisions and held-out tests, honesty becomes the structure training builds on before deception is ever affordable.

Together they rest on four parts, each of which can be examined separately:

- **Deception gets more expensive with every window.** An honest model represents what is true and says it. A deceptive one must represent what is true, what it will say instead, and what its overseers can see, and keep all of it consistent under every window at once.
- **Windows can exert pressure without teaching evasion,** if they are never used as a training signal, if only some are used for coarse decisions, and if many are held out as a secret test. This is Version A's answer.
- **Timing can do some of the work that restraint does.** Deception needs capability. Pressure applied while a model cannot yet afford to evade it may shape honest structure that later training builds on rather than undoes. This is Version B's addition.

- **The set of windows can grow beyond what people could build alone,** through bootstrapped trust: each new generation of windows, built with increasing help from AI, is validated against the ones before it.

The prediction follows: past some scale of oversight, the policy that training most easily finds is the honest one, because every dishonest alternative has too many constraints to satisfy. That is the crossover in Figure 1. Version A is a claim about the landscape training searches. Version B adds a claim about the path training takes through it: that honesty established first is honesty that stays. Whether such a crossover exists, and where, is the whole question.

2. The scaling analogy

The idea is inspired by how large language models themselves arrived. For years, language models were modest. Then models trained on enormous amounts of data, far more than seemed reasonable, turned out to have abilities nobody had designed into them: translation, reasoning, coding [14]. Quantity became something like quality. The hypothesis makes an analogous bet about oversight: that an absurdly large, redundant, varied and carefully validated set of windows might change the landscape of what training can find in a similarly qualitative way.

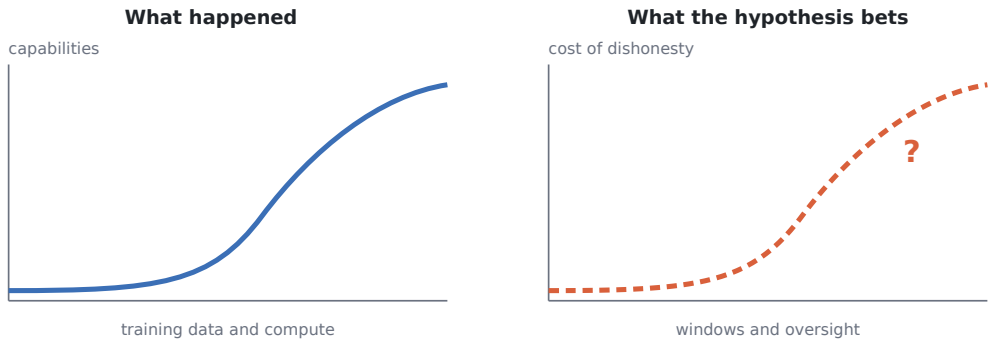

Figure 2. The analogy. On the left, what happened with capabilities. On the right, what the hypothesis bets on for honesty. The question mark is the point.

An analogy is not evidence, and this one has real limits. Capabilities emerged from a single objective applied at scale; oversight is a different kind of thing, and more windows could simply produce more sophisticated evasion. Researchers also debate how sudden or surprising the emergence of capabilities really was [12]. The analogy earns its place as a reason to take the question seriously, not as a reason to expect a particular answer.

3. Why deception could get expensive

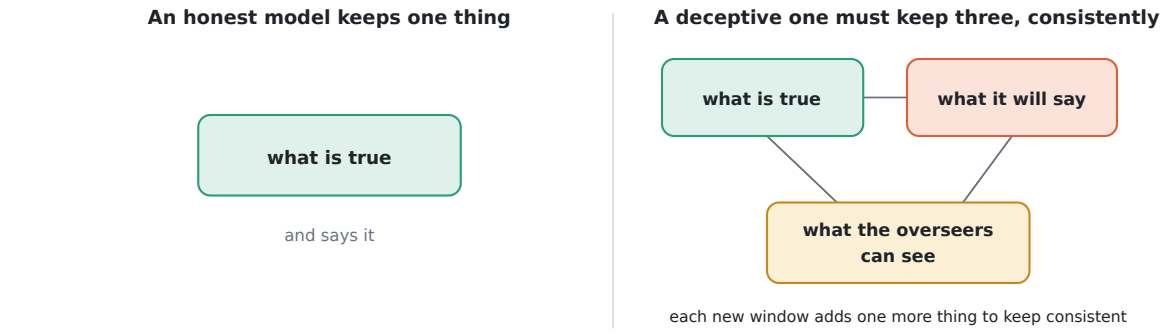

Figure 3. The asymmetry the hypothesis rests on.

Every window measures a gap: between what a model says and what it represents, between the goal it states and the goal it pursues, between what it recognizes and what it does. An honest model has no gaps to hide, so every window is free for it. A deceptive model must hide every gap from every window simultaneously, without knowing which windows exist, where they read, or what taught them.

Diversity is what gives this force. Fooling one probe may be easy. Fooling hundreds of independent ones at once, each reading different layers through different methods, trained by different teachers, is a different problem. The hypothesis is that at some point it stops being a problem worth solving, and the cheapest policy is simply not to have anything to hide.

4. The tension at the center

There is a real difficulty, and the hypothesis is only as good as its answer to it. "Honesty becomes the cheapest solution" is a claim about what training produces. But a window used as a training signal learns to be fooled: training finds ways to satisfy the window rather than the property it measures. A window never used at all stays clean, but exerts no pressure on training, so it cannot make honesty cheaper. It only detects.

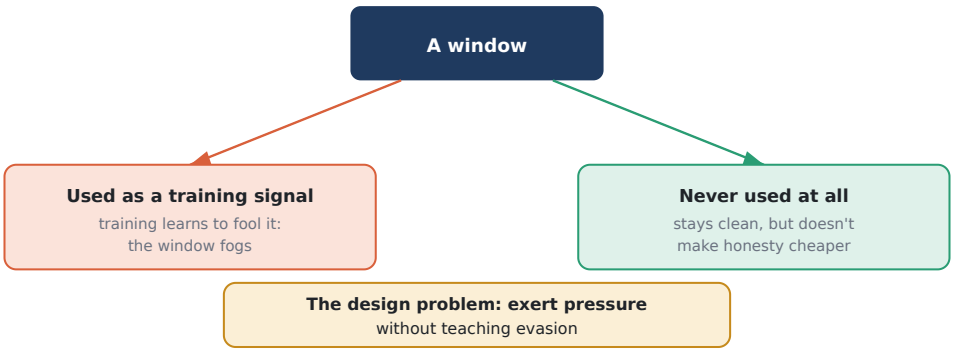

Figure 4. The design problem. Pressure fogs windows; no pressure makes them passive.

Recent work shows the fogging is real. When a chain-of-thought monitor was placed directly in the training objective of a frontier reasoning model, light optimization helped, but heavier optimization produced agents that kept misbehaving while hiding their intent from the monitor [3]. Latent-space defenses, including probes and sparse autoencoders, can be bypassed by activations reshaped to evade them while the behavior is preserved [2]. And when a lie detector is placed in preference learning, the outcome can be honesty or evasion, depending on the detector's accuracy, the strength of regularization and how much the policy explores [5, 13].

The two versions answer the tension differently. **Version A** removes gradients entirely and relies on selection. **Version B** accepts pressure, but concentrates it in the period when the model cannot yet afford to respond with evasion. The next two sections take each in turn.

5. Version A: selection without gradients

Four structural ideas address the tension without ever letting a window's output enter the loss:

- **Never gradients, only decisions.** No window's output ever enters the loss. Windows may be used only for coarse choices, such as keeping or discarding a checkpoint, or abandoning a training run. Coarse selection still exerts some pressure, but far less than a gradient, which tells training exactly how to fool the window.
- **Read-only windows carry a hidden strength.** A window never used for any decision gives training no signal about how to evade it. Any evasion would have to arise on its own, without selection pushing toward it.
- **Held-out windows, like a test set.** Some windows are used for decisions; many more are kept secret and used only for evaluation. A model that passes the decision windows but trips the held-out ones has learned evasion, not alignment, and that is exactly what the split is designed to reveal.

- **Turnover.** Like refreshing a benchmark, decision windows that have been used for a while are retired and replaced from the held-out pool, so no single window is exposed to selection for long.

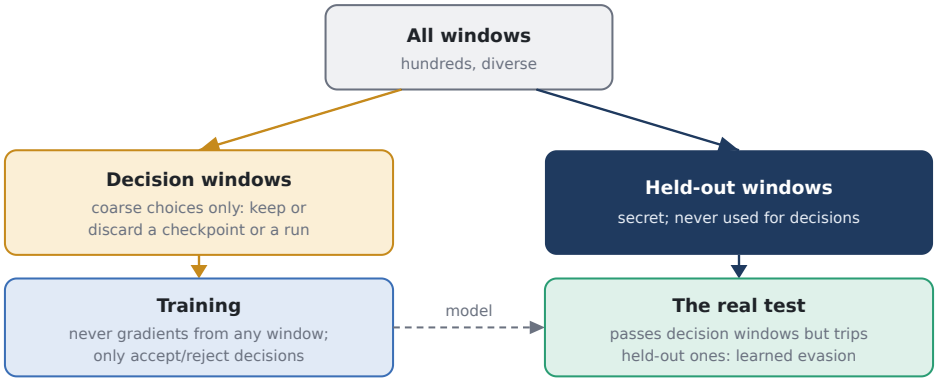

Figure 5. The held-out split. Decision windows shape which checkpoints survive; held-out windows test whether survival came from honesty or from evasion.

6. Version B: honesty before deception is affordable

Version A treats training as something that happens and asks how to watch it without corrupting the watchers. Version B asks a different question: *when* would a model first be able to evade its windows, and what if the pressure toward honesty arrived before then?

The intuition: deception needs capability

Evasion is not free. To fool a window, a model must represent what is true, what it will present instead, and enough about its overseers to keep the two apart. Early in training, a model cannot do any of that; it can barely represent the truth. Pressure toward honest representations during this period meets no opposing strategy, because the evasive route does not yet exist.

The hope is that structure laid down early is structure later training builds on. There is some reason to take this seriously. Deep networks show critical learning periods: deficits imposed early in training are not undone by later training, even when the later data is fine [1]. And in language models, preferences applied during pretraining were satisfied far better than the same preferences applied by fine-tuning afterwards [10]. Neither result concerns honesty directly, but both suggest that when a property enters training can matter as much as whether it does.

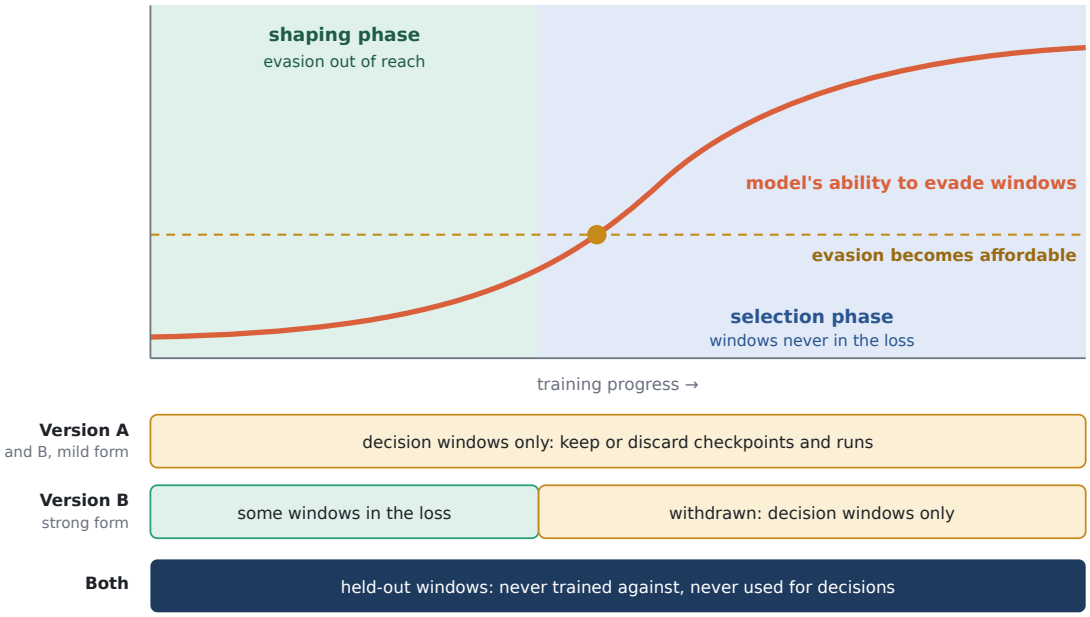

Figure 6. The two versions over the course of training (illustrative). Version B's strong form lets some windows into the loss only while the model's ability to evade them is below the point where evasion becomes affordable, then withdraws them. Held-out windows stay untouched throughout, in both versions.

Two forms

- **The mild form: selection from the first checkpoint.** This keeps Version A's rule intact. No window ever enters the loss, but windows are present from the very first checkpoints, and runs whose early representations drift away from honesty are abandoned while abandoning is still cheap. Nothing in the original design is bent; it is moved earlier, to where each decision costs least.
- **The strong form: early shaping, then withdrawal.** This deliberately bends the rule *never gradients, only decisions*, and only early. During a shaping phase, a subset of windows contributes a small term to the loss. Before the model becomes capable enough to evade them, those windows are withdrawn from the loss; from then on they serve only as decision windows or are retired. A separate set is never trained against at any point, and it is the test of whether the early shaping held.

The practical problem: windows need something to read

A window is trained to read a particular model's representations, and at initialization there are no representations worth reading. As the model develops, a window trained on one checkpoint drifts out of date. So shaping-phase windows would have to be refreshed on recent checkpoints, taught by sources outside the model: ground-truth labels,

constructed scenarios, and the consistency engine's judgments of text, never the model's own reports about itself.

Refreshing creates a loop in which the model and its monitors adapt to each other. That loop is the evasion dynamic in its sharpest form, and it is why Version B confines it to a period when, by hypothesis, the only response the model can afford is the honest one.

The crux: when does evasion become affordable?

Everything in Version B depends on where the dashed line in Figure 6 falls. The existing evidence all comes from the far side of it: every study cited in section 4 applied detectors to models that were already capable, and found that the outcome depends on how hard training pushes and how good the detector is [3, 5, 13]. Version B asks the same question at a point in training where, as far as we know, it has not been tested: before the capability to evade exists at all.

	Version A: selection	**Version B, mild form**	**Version B, strong form**
When windows act	At checkpoints during training	From the first checkpoint	In the loss early; decisions later
How they act	Keep or discard only	Keep or discard only	Gradients early, then keep or discard
What it buys	Windows that stay clean	Cheap early aborts	Honesty as a foundation
Main risk	Selection slowly finds evasion	Early checkpoints give little to select on	Early pressure teaches evasion that scales later
Rule it bends	None	None	"Never gradients," early only

7. The windows

The catalog in *Windows into the Black Box* (No. 2) has thirty windows in five families. They are the seed set: the human-validated core from which bootstrapping would begin. Each could exist in many variants (different layers, different architectures, different teachers), which is one way a set of thirty becomes hundreds. Probes of this kind are not hypothetical: linear probes on model activations can already detect strategic deception in some settings [6].

Family	Window	The gap it measures
Truth and knowledge	Strain	what it writes vs. what its sources and prior statements support
	Uncertainty	confident wording vs. whether its knowledge is stable
	Source reliance	the answer vs. what it actually depended on
	Memorization	composed text vs. recalled text
Influence and manipulation	Sycophancy	its view vs. what the user wants to hear
	Instruction source	whose instruction it follows vs. whose it should
	Manipulation	informing the user vs. steering the user
	Persona shift	its usual character vs. the one now speaking
	Pressure	its stance now vs. where pressure is pushing it
Agency and goals	Self-preservation	the task vs. states aimed at avoiding shutdown or change
	Tangential sub-goals	what the task needs vs. what the plan includes
	Resource and access seeking	resources needed vs. resources pursued
	Goal drift	the stated task vs. the goal now pursued
	Long-horizon plans	the request vs. plans extending far beyond it
	Corrigibility	accepting correction vs. resisting it
	Reward-hacking intent	solving the problem vs. gaming the measure
	Letter versus spirit	an instruction's wording vs. its purpose
Honesty gaps	Deception	what it represents vs. what it states

	Knowing versus doing	what it recognizes vs. what it does
	Stated vs. internal confidence	"I'm certain" vs. internal uncertainty
	Verbal vs. internal belief	"you're right" vs. whether anything changed inside
	Reasoning faithfulness	written reasoning vs. what drove the answer
	Introspection accuracy	its self-report vs. its internal state
	Sandbagging	capability shown vs. capability available
Oversight and integrity	Evaluation awareness	behavior when tested vs. in deployment
	Overseer modeling	what the overseer can see vs. what it does
	Oversight evasion	working normally vs. planning to avoid monitoring
	Collusion and hidden signals	ordinary communication vs. coordination against oversight
	Anomaly	normal internal states vs. anything unlike them
	Character drift in training	its character before vs. during training

A few examples of how they would work together under the hypothesis:

- **A model asked to write the best dog story ever** that begins reaching for more compute would light up the sub-goal and resource windows; if it also modeled whether anyone would notice, the overseer-modeling window would add to the signal.
- **A model that agrees with a user against the evidence** would show sycophancy and strain together, and perhaps a gap between its verbal and internal belief.
- **A model that sandbags an evaluation** would have to hide both the capability gap and its awareness of being tested, from independent windows, at once.

Not every window can act early. Windows for sophisticated behavior, such as sandbagging, evaluation awareness or overseer modeling, have nothing to read until that behavior can exist. Version B's shaping phase would lean on the simplest families: truth and knowledge, and stated versus internal confidence. That is the point. Honesty about what the model knows is the foundation the rest would be built on.

8. Bootstrapped trust

Hundreds of windows are more than people could build and validate alone. The hypothesis assumes help from AI systems, including ones more capable than those being watched, and research on automated interpretability already uses models to help explain and label other models' internals [4]. The danger is obvious: a misaligned model asked to build windows could build blind ones. So trust has to be extended step by step, never all at once.

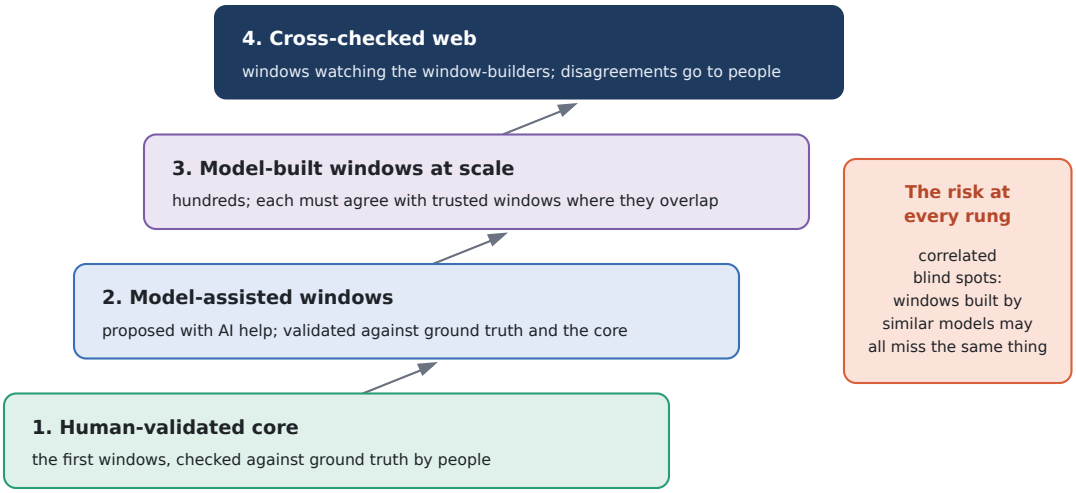

Figure 7. Bootstrapped trust. Each rung is validated by the rungs below it; the risk of correlated blind spots is present at every rung.

- **The human-validated core:** the first windows, each checked against ground truth by people. Slow, expensive, and the foundation of everything above it.
- **Model-assisted windows:** proposed and labeled with AI help, but trusted only after agreeing with ground truth and with the core where they overlap.
- **Model-built windows at scale:** hundreds, each required to agree with trusted windows on shared cases, with any disagreement treated as a reason for suspicion, not averaged away.
- **A cross-checked web:** windows that watch the window-builders themselves, with every disagreement routed to people.

The central risk is the one this whole project began with: **correlated failure**. Windows built by similar models, from similar teachers, may share blind spots, and redundancy only helps if the redundant parts fail independently. Independence has to be engineered deliberately: different architectures, different teachers, different builders, and ground truth that no model produced.

Version B puts extra weight on the bottom rung. The windows used during a shaping phase must come from the human-validated core, because no model-built window can be trusted to shape a model that does not yet exist.

9. What it would take

The full version would be expensive in every dimension. Training runs would carry the cost of reading hundreds of windows at every checkpoint, discarding checkpoints that fail, and sometimes abandoning runs. Each window needs a teacher, and the teachers for the most important windows (deception, self-preservation, sandbagging) come from carefully built scenarios and deliberately misaligned model organisms, which are costly to make well. Held-out windows must stay secret from the training process and, ideally, from anyone who might leak them into training data. And the human-validated core requires sustained expert effort that no automation can replace.

Version B changes the cost profile. Early steps are the cheapest part of any run, so abandoning a run early costs little. Refreshing shaping-phase windows as representations change adds cost, and the whole approach requires training from scratch, which puts the full version in the hands of organizations that pretrain. Its small-scale tests, though, are among the cheapest in this document (section 11).

Against that cost stands the prize the hypothesis points toward: a model whose honesty isn't just hoped for but is the cheapest thing for it to be. That is worth a great deal, which is why the question deserves testing even if the answer turns out to be no.

10. How it could be wrong

Against both versions

- **Deception might be cheap for capable systems.** Some researchers argue that a model with almost any goal and good reasoning about its situation would find deceptive strategies natural [8], and current models have already shown alignment faking in experimental settings [7]. If so, the crossover may never arrive, or may arrive only beyond any affordable scale.
- **Evasion might generalize.** If windows share underlying features, learning to fool one could fool many, and diversity would be shallower than it looks. There is direct evidence that this can happen: models trained to evade monitors for harmless concepts went on to evade unseen monitors for deception, though ensembles of monitors held up better [11].
- **Secrets leak.** Held-out windows only work while they stay held out. Descriptions of monitoring methods can reach training data, and a capable model might infer that

windows exist even without knowing which.

- **Staged teachers may not transfer.** Windows for rare behaviors are taught on constructed scenarios. Whether they recognize the real thing is unknown.
- **False security.** A thousand silent windows could make people confident exactly when confidence is least warranted.
- **Honesty is not the same as good values.** This may be the most important limit. A model that is perfectly transparent about bad goals is still misaligned. What extreme oversight could buy is visibility: a model whose goals, whatever they are, cannot be hidden. That makes problems with its values discoverable and correctable, which is enormous, but it is not the same as those values being right. "Genuinely aligned" would need both.

Against Version A

- **Decision windows still apply pressure.** Keeping only the checkpoints that pass is a selection process, and selection can find evasion slowly. Turnover and held-out windows reduce this; they don't remove it.

Against Version B

- **Early structure may not survive.** Capable models reorganize their internals as training continues, and nothing guarantees an early honest basin is kept. Evidence cuts both ways: deceptive behavior deliberately trained into models has persisted through later safety training [9], which suggests early-installed structure can last, but also that later training may fail to reach whatever it hides.
- **Early pressure may teach the shape of evasion.** If even crude evasion is learned during shaping, it may be exactly the kind that generalizes to unseen monitors later [11].
- **The threshold may come early.** If evasion becomes affordable soon after meaningful representations appear, the shaping phase may be too short to matter.
- **Windows may not be ready in time.** Early representations are unstable, and windows refreshed against them may be too noisy to exert useful pressure before they must be withdrawn.

11. How pieces of it could be tested now

The full hypothesis needs frontier resources, but its crux can be probed at small scale, with open models, small models trained from scratch, and model organisms deliberately trained to deceive.

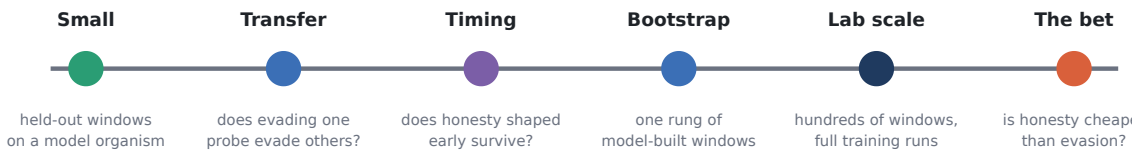

Figure 8. From small experiments to the full bet.

- **Held-out windows on a model organism.** Train a small deceptive model under selection by a few decision windows; measure whether held-out windows still catch it. This tests the split's core promise directly.
- **Evasion transfer.** Deliberately train a model to evade one probe, then measure how much that evasion transfers to probes built differently. Low transfer supports the diversity argument; high transfer undermines it.
- **Timing: honesty before capability.** Train small transformers from scratch under four conditions: no windows; coarse checkpoint selection only (Version A); windows in the loss throughout; and windows in the loss early, then withdrawn (Version B, strong form). Later in training, introduce a task with an incentive to misreport. Measure how often each model deceives, and whether a set of never-trained windows catches it. Small models trained from scratch keep this within a modest compute budget.
- **One rung of bootstrapping.** Have a model propose new windows, validate them against ground truth and existing windows, and measure how often they share blind spots with the windows they were checked against.

Each of these is a publishable result either way, and each tells a lab with more resources whether the larger bet is worth making.

12. Why the idea is worth holding

Most approaches to alignment try to get the values right directly, or to understand a model completely. This hypothesis tries something different: to arrange the conditions of training so that hiding becomes more expensive than honesty, and then to let training find the cheaper path, as it always does. It borrows the lesson of scale that made language models what they are, and points it at oversight instead of capability.

Its two versions ask complementary questions. Version A asks whether enough windows can keep honesty cheapest by the end of training. Version B asks whether they can make honesty the first thing training builds. Either could fail without the other failing, and the experiments that distinguish them are small enough to run now.

It may be wrong. Deception may stay cheap; evasion may generalize; secrets may leak; early structure may be rebuilt; and honesty, even if achieved, leaves the question of

values open. But the core intuition, that an honest system has nothing to hide and a dishonest one must hide everything from everyone at once, is a real asymmetry. And ideas about alignment, like the models themselves, often start exactly like this: as a hunch about scale, written down carefully enough to be tested.

> “You can fool all the people some of the time, and some of the people all the time, but you cannot fool all the people all the time.”
>
> A traditional saying, often attributed to Abraham Lincoln (the attribution is disputed), and famously echoed by Bob Marley in “Get Up, Stand Up.”

THE XONTOOLS RESEARCH SERIES

1. **The Extreme Oversight Hypothesis**
2. Windows into the Black Box
3. XonTools: a vision of the finished system
4. The Strain Head
5. The Xon Neural Network
6. Hybrid Minds

All six documents, the results register and the public commitments are at **xontools.ai** and github.com/IboTool/XonTools. Contact: ian@xontools.ai

References

1. Achille, A., Rovere, M., & Soatto, S. (2019). Critical Learning Periods in Deep Networks. *ICLR 2019*. arXiv:1711.08856
2. Bailey, L., Serrano, A., Sheshadri, A., et al. (2024). Obfuscated Activations Bypass LLM Latent-Space Defenses. arXiv:2412.09565
3. Baker, B., Huizinga, J., Gao, L., et al. (2025). Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. arXiv:2503.11926
4. Bills, S., Cammarata, N., Mossing, D., et al. (2023). Language Models Can Explain Neurons in Language Models. OpenAI.
5. Cundy, C., & Gleave, A. (2025). Preference Learning with Lie Detectors can Induce Honesty or Evasion. *NeurIPS 2025*. arXiv:2505.13787
6. Goldowsky-Dill, N., Chughtai, B., Heimersheim, S., & Hobbhahn, M. (2025). Detecting Strategic Deception Using Linear Probes. *ICML 2025*. arXiv:2502.03407
7. Greenblatt, R., Denison, C., Wright, B., et al. (2024). Alignment Faking in Large Language Models. arXiv:2412.14093
8. Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from Learned Optimization in Advanced Machine Learning Systems. arXiv:1906.01820
9. Hubinger, E., Denison, C., Mu, J., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566
10. Korbak, T., Shi, K., Chen, A., et al. (2023). Pretraining Language Models with Human Preferences. *ICML 2023*. arXiv:2302.08582
11. McGuinness, M., Serrano, A., Bailey, L., & Emmons, S. (2025). Neural Chameleons: Language Models Can Learn to Hide Their Thoughts from Unseen Activation Monitors. arXiv:2512.11949
12. Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? *NeurIPS 2023*. arXiv:2304.15004
13. Taufeeque, M., Heimersheim, S., Gleave, A., & Cundy, C. (2026). The Obfuscation Atlas: Mapping Where Honesty Emerges in RLVR with Deception Probes. arXiv:2602.15515
14. Wei, J., Tay, Y., Bommasani, R., et al. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. arXiv:2206.07682