The Scientist AI: Safe by Design, by Not Desiring

February 13, 2026

Abstract

Scientific theories aspire to describe what is, as opposed to prescribe what ought to be. At LawZero, we take this idea as a design principle for safe artificial intelligence: that *understanding*—even of arbitrary depth and scope—can be disentangled from *preference* over how the world unfolds.

We distill into a non-technical overview the motivations and core components of the **Scientist AI**, a system that aspires to this ideal. Agency, we argue, rests on three pillars—affordances, goal-directedness, and intelligence—each a matter of degree. By limiting the first two while pursuing the third, we aim to build a system that is highly intelligent yet incapable of holding or pursuing goals of its own. The Scientist AI comprises a generator held accountable by a neutral estimator, allowing for creative thought without compromising safety. Two key ingredients are (i) *contextualization*, a transformation of the training data that disentangles facts from statements about such facts (*e.g.*, opinions), and (ii) *consequence invariance*, a property of the training process that prevents feedback about downstream outcomes.

We believe this approach offers a promising path toward systems that are at once powerful, transparent, and safe, and that may serve as trustworthy anchors in a world of increasingly autonomous AI.

1 Introduction

"I would rather discover one true cause than gain the kingdom of Persia."

—Democritus, *apud* Dionysius of Alexandria, per Eusebius.

Consider the laws of physics: given the present state of a system, they predict the system's future state. Their descriptive nature does not carry preferences over downstream outcomes, *e.g.*, the gravitational force that holds planets in orbit also pulls apples to the ground—and in exactly the same way no matter the moral character of the person sleeping underneath.

Now imagine an idealization of a theoretical scientist: a mind that has internalized the laws of nature and uses them so as to predict what will happen under various circumstances, but without predilection for how things unfold, motivated purely by disinterested understanding. This image is ancient and pervades human culture: Spinoza spoke of the rational mind as one that labors not to mock or lament nature's unfoldings but solely to understand them, and Zhuangzi of the master as one who yields to the natural structure of things.

We propose to take this idea seriously as a design principle for safe artificial intelligence. We aim to build a predictive model that captures causal mechanisms that explain everything we observe, from the motion of particles to human and AI behavior. This would be extremely useful because, among other things, it would allow us to impartially evaluate scientific hypotheses and questions (*e.g.*, "Will this spacecraft achieve orbit?", "Will this drug effectively treat this pulmonary disease?", "Will implementing this policy decrease carbon emissions?"). Such model also offers critical applications as a guardrail (*e.g.*, "Will executing this source code pose a security risk?", "Will steering left avoid a collision between this vehicle and the child crossing the road?", or, more broadly, "Will taking this action cause harm to humans?"). Predictions from such a system should be a faithful reflection of the underlying laws and can be made to take into account uncertainty arising from lack of information.

We believe this approach could form the foundation of a novel type of safe artificial intelligence, an approach which is urgently needed. As AI systems grow increasingly autonomous, so do the risks they carry. Autonomous cars and coding agents, for instance, already cause harm through consequential mistakes that outpace verification. While more capable systems will make fewer errors, greater capabilities pose

even greater risks: systems that excel at their tasks often remain highly opaque and entrench existing biases under the guise of statistical objectivity.[1] And even agents that are neither imprecise nor opaque may find themselves at odds with human interests—values that are notoriously difficult to specify, subject to revision, and far from universally shared. These risks are increasingly recognized by the scientific community, world leaders, and the public at large.

Motivated by this, at LawZero we are building a system we call the **Scientist AI** (SAI). At a high level, the SAI consists of two interacting components: an estimator of arbitrary probabilities of interest (such as "this apple falls onto me"), and a generator that explores hypotheses about the world (such as "the apple is still green"). Trained with data that has been situated in context (by attributing claims to sources, *e.g.*, "Newton has an apple" versus "Newton said 'I have an apple' in 1666"), the SAI is careful to distinguish what is true from what has only been said. Additionally, like an idealized scientist, it builds causal models of what is and might otherwise have been. And precisely because it lacks ambitions of its own, the SAI is well-positioned to be a safe and effective guardrail for frontier AI agents whose actions we do not fully trust.

2 Agency

> *"Thou shalt not make a machine in the likeness of a human mind."*
>
> — Orange Catholic Bible, *Dune* (Frank Herbert).

The standard definition of a rational agent comes from decision theory: the study of *choice*. Classically, a *rational agent* is an entity that acts as though it has beliefs (*e.g.*, probabilities), preferences (*e.g.*, utilities), and chooses actions so as to further its interests. Our notion of "agency" quantifies the degree to which an actor "rationally" exerts control over the world through three pillars—fundamental aspects of choice:

AFFORDANCES delimit the scope of actions available to a system. A scale can only display weights; a chess-playing robot can move pieces on a board; a language model with internet access can send emails, query databases, and execute code. Having more affordances means being responsible for a larger number of more complex or impactful choices.

GOAL-DIRECTEDNESS refers to an agent's capacity for holding preferences about its environment and favoring behaviors that align with preferred outcomes. Shakespeare's Hamlet famously says that "there is nothing either good or bad but thinking makes it so"; this kind of "thinking" is what characterizes goal-directedness. More precisely, a goal-directed agent is one that breaks an *a priori* symmetry by preferring one environmental outcome to another, all else being equal.

A chess engine, for instance, is goal-directed because it prefers winning to losing. A classifier trained with log likelihood is not goal-directed because its learning objective is a natural consequence of making observations, but a classifier that artificially places double the weight on one class does have a preference. Generalizing to a more important and complex example, a language model trained to model the distribution of human text (in a given corpus) is not goal-directed as a predictive model (relative to its training corpus), but can easily be given goal-directedness through deployment scaffolding,[2] instruction tuning, or reinforcement learning (*e.g.*, from human feedback).

Crucially, the capacity to hold a preference or a goal requires a choice: between a given goal and plausible alternatives to it (such as its negation).

INTELLIGENCE involves the efficient acquisition and usage of epistemic abilities: learning, memory, and the ability to reason and make inferences based on knowledge. In a sense, a more intelligent agent has more memory, a wider array of possible thoughts, and a richer set of perspectives. With a richer conceptual landscape comes a greater ability to drive finer and better targeted action choices.

[1] The influence of recommendation algorithms on public discourse offers a cautionary example. More generally, ecology teaches us that rapid perturbations to complex and interconnected environments can prove catastrophic even when each individual change appears minor.

[2] By sampling word sequences according to the statistics of human language, especially in contexts constructed to set some goal, the model often exhibits agentic behavior through imitation, because humans are agents.

We call an entity *agentic* to the extent that it scores high on all three pillars. Our starting point is the idea that removing any one pillar substantially reduces the associated risks. A highly intelligent and highly goal-directed system with no affordances cannot affect the world. A highly intelligent system with vast affordances but no goal-directedness has no reason to use those affordances in any particular direction.

Our goal at LawZero is to build the Scientist AI, a system that is highly intelligent but with circumscribed affordances and a marked lack of goal-directedness, as we describe in the next sections.

3 Truth and Context

"...a text without a context is a pretext ..."

To build the Scientist AI in practice requires knowledge at many levels of abstraction, much of which lies latent in the accumulated text of human civilization (or, more modestly, on the internet). The success of large language models demonstrates that this corpus of data, inconsistencies notwithstanding, captures a great deal of information about how the world works alongside languages themselves. But training on human-generated text introduces two problems:

- The first is that language is produced by agents, who write with purposes, *e.g.*, to persuade, to entertain, to coordinate. A predictor trained directly on such data learns not only facts about the world but also the patterns by which agents pursue their goals through language;
- The second is that statements, as they appear in text, are not necessarily factual. A predictor that treats assertions as ground truth learns to reproduce confident claims without regard for their veracity.

Both problems arise from treating text as though it directly represents the world, when in fact the observed text represents what agents chose to say about the world. This observation contains the germ of our approach:

> *Even a statement that is biased or false contains something true, namely, that it occurred, made by a particular source, at a particular time, in a particular context.*

For example, the statement "the satellite's motion is governed by Newton's laws" may be contested: perhaps relativistic corrections matter, perhaps they do not. But the statement "the engineering team, in their report, calculated the trajectory using Newtonian mechanics" is verifiable regardless, and informative about the methods and assumptions in use. Should a discrepancy arise, we learn something about the team's reasoning, not just about the trajectory. Likewise, "the Sun revolves around the Earth" is false; that Ptolemy wrote as much is not.

We call this approach *contextualization*: a transformation of the training data that makes the epistemic provenance of claims explicit by attributing sources. In doing so, the truth of the original statement becomes a latent variable to be inferred from patterns of evidence, and the transformed statement can more reasonably be taken as true.

To illustrate this point, consider what would happen if the Internet were flooded with text saying "the Earth is flat". Absent a post-training fix, training on this data leads to models that claim the Earth to be flat. What's worse, each instance of "the Earth is flat" that appears in the pre-training corpus makes the problem more severe and the fix more difficult. The contextualized dataset, however, has no such issue: every declaration of the flatness of the Earth serves only to sharpen the trained model's understanding of the people who claim so, without necessarily affecting the model's propensity to join the flat-Earth society itself.

Narrowly construed, contextualization can be viewed as a practical and conceptually straightforward way to get high-quality data with an explicit signal about factuality. But how should we make use of it?

4 Design and Safety

"A system only acquires a degree of safety through the initial work of those who designed it."

— Daniellou, Simard, & Boissières, *Human and Organizational Factors of Safety*.

Two questions remain: how to train an estimator of probabilities that inherits no undesired preferences from the process that shapes it, and how to let it reason and explain without compromising its neutrality.

4.1 Training for Disinterest

That, *a priori*, prediction without preference is possible we have already seen, for a simulator based on the laws of nature would be such an example. Can we build another, similarly disinterested estimator ourselves?

Non-agentic starting point. We begin with the (revered) tradition of initializing our estimator network at random. Thus it will not be agentic—by virtue not only of its lack of intelligence, but also its lack of coherent goals. Having a persistent goal in a complex world requires many coordinated choices, each adapted to circumstances; the odds of stumbling upon such coordination by chance are small.

Non-agentic ending point. At the other extreme, the target it is training towards is an idealized, disinterested scientist that cares only to understand the world and the laws of nature that govern it.

The difficulty, then, lies in ensuring that the optimization process that we use to train the estimator does not happen to acquire new goals along the way.[3] The training process of the SAI is specifically designed to minimize this possibility. We discuss this next.

Consequence invariance. Our proposed approach is to sever any training signal that would inform the estimator about the downstream consequences of its predictions. If the training procedure cannot evaluate whether a given estimate led to favorable or unfavorable outcomes in the world, then it cannot learn to bias its outputs toward achieving preferred outcomes. We call this property *consequence invariance*: the learned estimator, and each update to it, must be invariant to any choice of utility function over downstream effects of predictions. In practice, this rules out training regimes akin to standard reinforcement learning, where a model interacts with an environment and receives feedback in service of an arbitrarily chosen utility that would depend on the effects of estimates. It also rules out objectives that optimize for anticipated future performance, as in model-based planning. In other words, the training signal must be myopic, concerned only with making an accurate prediction for the given query, given the data at hand.

A subtlety arises, however. A good estimator that models the world must be able to *understand* the consequences of actions, including the consequences of its own estimates, should those estimates influence the world. This is the problem of *performative prediction*: the act of forecasting can change the actual variables being forecast.[4] If the estimator ignores this problem, it risks systematic error; if it does take performative prediction into account, then it may have an incentive to impact the world: it could select its outputs to bring about states of the world that make its estimates come true.

One resolution to this problem draws from existing work on oracles [1], and the heart of it lies in evaluating counterfactuals. During training, we only ever pose questions in the spirit of: "I know you may give a different answer, but what would be the probability of event Y if you were to output q as the answer to this query?", for different values of q. Ensuring accurate evaluation of these counterfactuals remains an open problem, but multiple such answers can sever the training signal about the future outcome of the world,[5] thereby eliminating this potential expression of goals and any corresponding training signal.

In the next section, we fill in some details about causal modeling, as is needed to make sense of counterfactual queries.

4.2 Generation and Inference

Thus far we have focused on just one component of the SAI: a estimator that only provides probabilities. On its own, it lacks many affordances required for scientific inquiry: to generate hypotheses about the world; to

[3] This is especially relevant if, as we discuss next, the predictions of the estimator have real-world consequences, *e.g.*, if they are used by someone.

[4] This phenomenon is ubiquitous and has been rediscovered across disciplines under various names, *e.g.*, self-fulfilling prophecies in sociology [11], reflexivity in economics [14], the Lucas critique in macroeconomics [10], Goodhart's law in policy [6], and more recently performative prediction in machine learning [13, 12]. The common thread is that predictions, once acted upon, become entangled with the phenomena they describe—and in doing so, often cease to be reliable.

[5] If multiple predictions are consistent with the world and a choice must be made between them, it is possible to make that choice in a way that is blind to the consequences, *e.g.*, by picking one which commits the least about the future.

interrogate explanations; to craft narratives that fit together with a single compelling story; to come up with insightful questions and identify meaningless ones.

A generator with freedom. A second component of the SAI is responsible for generation. It produces reasoning and directs attention to relevant features of the world, allowing for complex arguments and the kind of "thinking" exhibited by reasoning language models.

Fortunately (and perhaps counterintuitively), only the estimator of conditional probabilities must be a paragon of neutrality; we need not hold this generation component to the same high safety standards. Intuitively, it does not matter if the SAI sometimes generates misleading or self-serving arguments, provided it is also equipped to step back and evaluate them in a truly neutral way. This means the generator's affordances (*cf.* §2) are gated by approval of the estimator, which itself is diminished in both goal-directedness and affordances. The separation between the two systems is what allows us to focus so narrowly on the neutrality of the estimator without giving up entirely on the power of free creative thought. This approach follows a long tradition of separating powerful but untrusted provers (in our case, the generator) from limited but trustworthy verifiers (in our case, the estimator).[6] Evidence of the effectiveness of this approach can be found in the theory of proof systems, in which one can make surprisingly good use of advice without needing to trust it [2, 3, 5]. These ideas have already been adapted by others in service of designing safer AI systems [8, 4, 7], and indeed similar motifs have long appeared in machine learning to advance the capabilities of such systems themselves.[7]

Notice that the relationship between a statement ("the meteor will strike the Earth") and its veracity (whether it actually strikes)—-a key motivation for contextualization—is analogous to the relationship between the roles of the generator and the estimator. While contextualization is an important first step towards respecting this distinction and implicitly encodes a partial model of the world, ultimately we aim to go further and equip the SAI with a full causal world model. Moreover, we believe this can be done with the ingredients already at hand, as we explain next.

Interpretable fragments of a world model. As it interacts with the estimator, the generator makes statements and causal claims about relevant variables, thereby extracting relevant parts of the causal world, which were previously only implicit within the estimator. As opposed to merely learning correlations, the SAI generator is trained to propose hypotheses about causal mechanisms and to represent them in a form that can be inspected. This creates an information bottleneck: predictions must flow through an interpretable structure—namely, a causal model whose variables' meanings are anchored (via natural language) by reuse (of words) across many different contexts.

Even if the estimator were to end up with unstated implicit preferences, the SAI must give explicit causal reasoning through interaction with the generator. We expect this makes goals harder to hide, and easier to discover should they exist.

Non-agentic deployment. The SAI is designed to lack agency, but it could still be misused. For this reason, access to the SAI will need to be restricted and regulated. Yet, compared to the frontier AI models of 2026, we believe it will be feasible to build a safe public interface. Indeed, we are optimistic that it may suffice to place only two constraints: users may only query the SAI in 'fact mode', *i.e.*, in the register that the contextualization reserves for known information, and users cannot ask about the behavior of agents (*e.g.*, "factually, what would an expert chemist such as Albert Hoffman do in my position to synthesize LSD?"). Such queries, and those likely to cause significant harm, can be filtered out by another instance of the SAI itself, whose job would be to estimate the likelihood of harm resulting from answering. Beyond misuse, there are also other valid concerns about the human ideal of "perfect neutrality"; we reflect on some of these concerns in concluding.

[6] This approach of checks-and-balances is analogous to the separation of powers within modern states, such as the separation of the executive branch from the judiciary branch (*e.g.*, where a court strives to independently and impartially evaluate the charges against an individual brought forth by the police).

[7] Actor-critic systems in RL [9], and similar systems have incorporated in learning from reward models [15].

5 Conclusion

"The very act of trying to look ahead to discern possibilities and offer warnings is in itself an act of hope."
— Octavia Butler.

The Scientist AI is designed to develop a dispassionate understanding of the world. However, from Shelley's Frankenstein to Vonnegut's Hoenikker, people have long been wary of the dangers posed by the detached pursuit of scientific truth. For starters, scientifically illuminating real-world experiments can be harmful. Yet the SAI neither performs experiments itself, nor does it desire "new" understanding beyond what it can deduce from past data.[8] Moreover, we do not mean to suggest that *human* science should aspire to this extreme kind of objectivity. Science as practiced is rightly situated, shaped by the questions and interests that impact human life—and this partiality is not a defect but a condition of responsibility.[9] Instead, since we are not yet sure whether AI can handle this kind of responsibility, we propose training a machine in a particular way to approximate a disinterested witness.

The work described here is currently underway, and much remains to be done. We are proceeding incrementally, with empirical evaluation of each component at increasing scales. If this efforts succeed and we can build a highly intelligent system without the capacity for desires or goals of its own, then it is not hard to imagine how it could be a powerful and trustworthy anchor, and thus, in case AI capabilities continue to rise, an indispensable tool to help humanity safely navigate our future.

References

[1] Stuart Armstrong and Xavier O'Rorke. Good and safe uses of ai oracles. *arXiv preprint arXiv:1711.05541*, 2017.

[2] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM (JACM)*, 45(3):501–555, 1998.

[3] Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: A new characterization of np. *Journal of the ACM (JACM)*, 45(1):70–122, 1998.

[4] Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. Avoiding obfuscation with prover-estimator debate. *arXiv preprint arXiv:2506.13609*, 2025.

[5] Oded Goldreich. Computational complexity: a conceptual perspective. *ACM Sigact News*, 39(3):35–39, 2008.

[6] Charles AE Goodhart. Problems of monetary management: the uk experience. In *Monetary theory and practice: The UK experience*, pages 91–121. Springer, 1984.

[7] Kai-Chieh Hsu, Haimin Hu, and Jaime F Fisac. The safety filter: A unified view of safety-critical control in autonomous systems. *Annual Review of Control, Robotics, and Autonomous Systems*, 7, 2023.

[8] Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

[9] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

[10] Robert E Lucas Jr. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46. North-Holland, 1976.

[8] To illustrate: suppose two competing theories, A and B, are equally consistent with all past observations. Rather than seeking a new experiment to discriminate between them, the SAI rests in that ambiguity and averages across both possibilities. It has no drive to resolve uncertainty through intervention in the world, and no channel through which a user could instill one.

[9] This view has philosophical roots. For example, Donna Haraway's *situated knowledges*, Helen Longino's work on contextual values, and Roger Pielke Jr.'s *The Honest Broker* all argue that objectivity is not compromised but strengthened by acknowledging the position from which inquiry proceeds.

[11] Robert K Merton. The self-fulfilling prophecy. *The antioch review*, 8(2):193–210, 1948.

[12] Mehrnaz Mofakhami, Ioannis Mitliagkas, and Gauthier Gidel. Performative prediction with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 11079–11093. PMLR, 2023.

[13] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[14] George Soros. *The alchemy of finance*. John Wiley & Sons, 2015.

[15] Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.