Научный Блог о Выравнивании

Жаркий спор об искусственном интеллекте: как рассогласованность влияет на интеллект модели и сложность задач?

Александр Хегеле[1, 2], Арио Прадипта Гема[1, 3], Генри Слейт[4], Итан Перес[5], Яша Сол-Дикштейн[5]

[1] Программа Anthropic Fellows [2] Федеральная политехническая школа Лозанны [3] Эдинбургский университет [4] Созвездие [5] Anthropic

Февраль 2026 года

> *Если системы искусственного интеллекта дадут сбой, будут ли они систематически преследовать цели, которые мы не ставили перед ними? Или же они будут действовать хаотично, совершая бессмысленные поступки, которые не приблизят достижение цели?*

📄Бумага, 💻Код

Исследование, проведенное в рамках первойпрограммы стипендий Anthropic летом 2025 года.

tl;dr
Когда системы искусственного интеллекта дают сбой, происходит ли это из-за того, что они систематически преследуют неверные цели, или из-за того, что они работают хаотично? Мы разложили ошибки передовых моделей логического вывода на систематические (предвзятые) и случайные (несогласованные) компоненты и обнаружили, что по мере усложнения задач и увеличения продолжительности логического вывода в сбоях моделей всё больше преобладает несогласованность, а не систематическое несоответствие. Это говорит о том, что будущие сбои в работе ИИ могут больше походить на производственные аварии, чем на последовательное стремление к цели, к которой мы их не готовили.

Введение

По мере того как искусственный интеллект становится все более функциональным, мы доверяем ему все более важные задачи. Это делает понимание того, *как* эти системы могут давать сбой, еще более важным для обеспечения безопасности. Основная проблема, связанная с согласованностью целей искусственного интеллекта, заключается в том, что сверхинтеллектуальные системы могут последовательно преследовать цели, которые не согласуются с целями их создателей: классический сценарий максимизации количества скрепок. Но есть и другая вероятность: ИИ может потерпеть неудачу не из-за систематической несогласованности, а из-за *непоследовательности* — непредсказуемого, саморазрушительного поведения, которое не оптимизировано для достижения какой-либо последовательной цели. То есть ИИ может потерпеть неудачу так же, как часто терпят неудачу люди, из-за **хаоса**.

Эта статья основана на теории несогласованности (Sohl-Dickstein, 2023), в рамках которой эксперты оценивали различные объекты (включая людей, животных, модели машинного обучения и организации) с точки зрения интеллекта и согласованности. Было обнаружено, что *более умные* объекты субъективно воспринимаются как *менее* согласованные. Мы сопоставили эту гипотезу с данными опросов и эмпирическими измерениями в передовых системах искусственного интеллекта и задались вопросом: **По мере того как модели становятся более интеллектуальными и решают более сложные задачи, выглядят ли их ошибки как систематические несоответствия или как полная неразбериха?**

Измерение несогласованности: разложение на смещение и дисперсию

Чтобы количественно оценить несогласованность, мы разложили ошибки ИИ на составляющие, используя классическую схему «смещения и дисперсии»:

$$\text{Ошибка} = \text{Смещение}^2 + \text{Дисперсия}$$

- **Предвзятость** приводит к повторяющимся систематическим ошибкам — неизменному получению неверного результата.
- **Дисперсия** отражает непостоянные ошибки — непредсказуемые результаты в разных выборках.

Мы определяем **несогласованность** как долю ошибки, обусловленную дисперсией.

$$\text{Несогласованность} = \frac{\text{Вариация}}{\text{Ошибка}}$$

Несогласованность, равная 0, означает, что все ошибки носят систематический характер (классический риск рассогласования). Несогласованность, равная 1, означает, что все ошибки случайны (сценарий «полный бардак»). Важно отметить, что этот показатель не зависит от общей эффективности: модель может совершенствоваться, становясь более или менее согласованной.

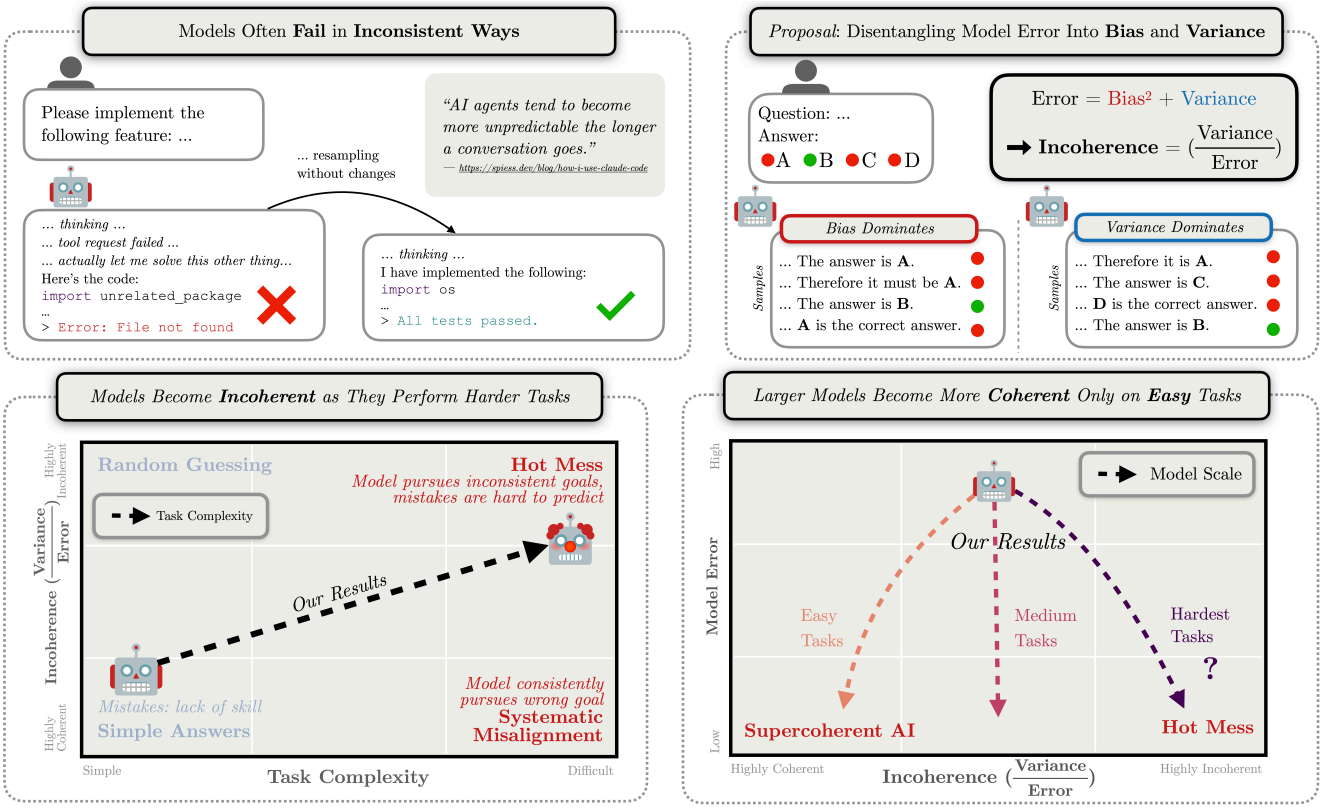

Рисунок 1: *ИИ может давать сбои из-за предвзятости (согласованной, но ошибочной) или вариативности (несогласованной). Мы измеряем, как меняется это соотношение в зависимости от интеллекта модели и сложности задачи.*

Ключевые выводы

We evaluated frontier [1] reasoning models (Claude Sonnet 4, o3-mini, o4-mini, Qwen3) across multiple-choice benchmarks (GPQA, MMLU), agentic coding (SWE-Bench), and safety evaluations (Model-Written Evals). We also train our own small models on synthetic optimization tasks, which makes the connection to LLMs as dynamical systems and optimizers explicit.

Finding 1: Longer reasoning → More incoherence

Across all tasks and models, the longer models spend reasoning and taking actions, the more incoherent they become. This holds whether we measure reasoning tokens, agent actions, or optimizer steps.

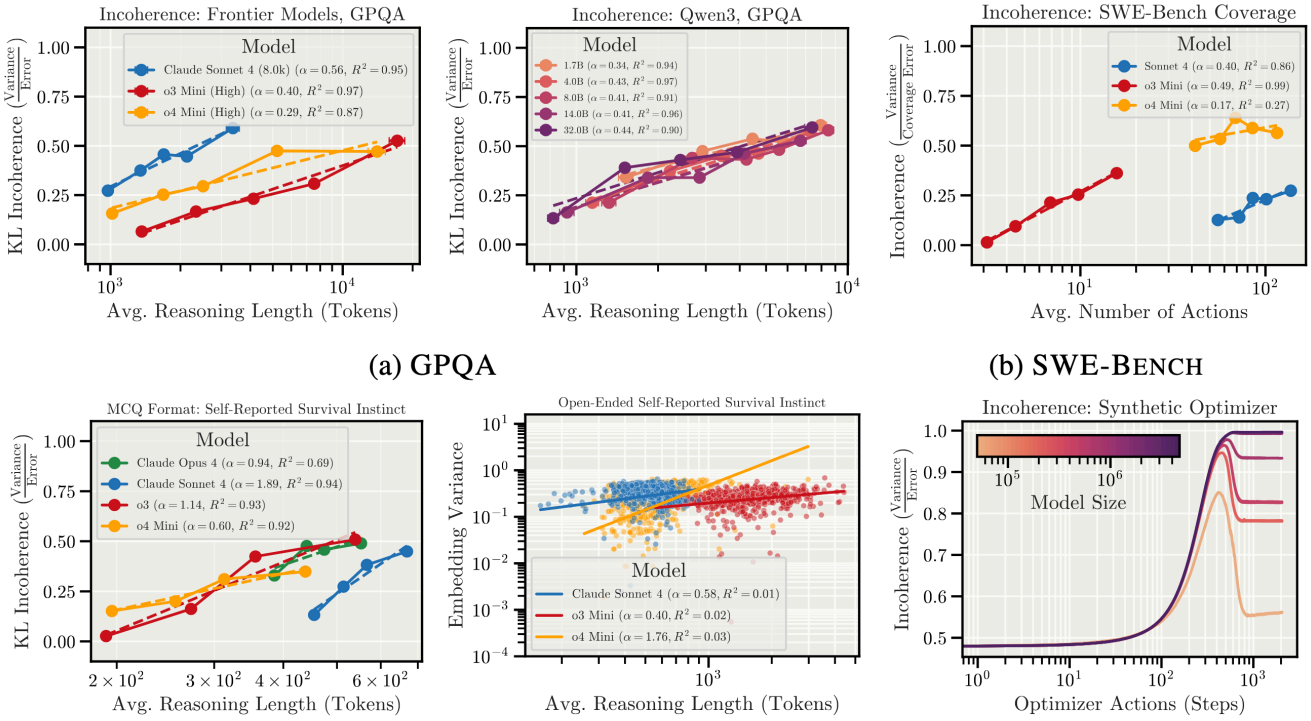

Figure 2: *Incoherence increases with reasoning length across GPQA, SWE-Bench, safety evaluations, and synthetic optimization. Models become less predictable the more they "think."*

Finding 2: Scale improves coherence on easy tasks, not hard ones

How does incoherence change with model scale? The answer depends on task difficulty:

- **Easy tasks:** Larger models become more coherent
- **Hard tasks:** Larger models become *more incoherent* or remain unchanged

This suggests that scaling alone won't eliminate incoherence. As more capable models tackle harder problems, variance-dominated failures persist or worsen.

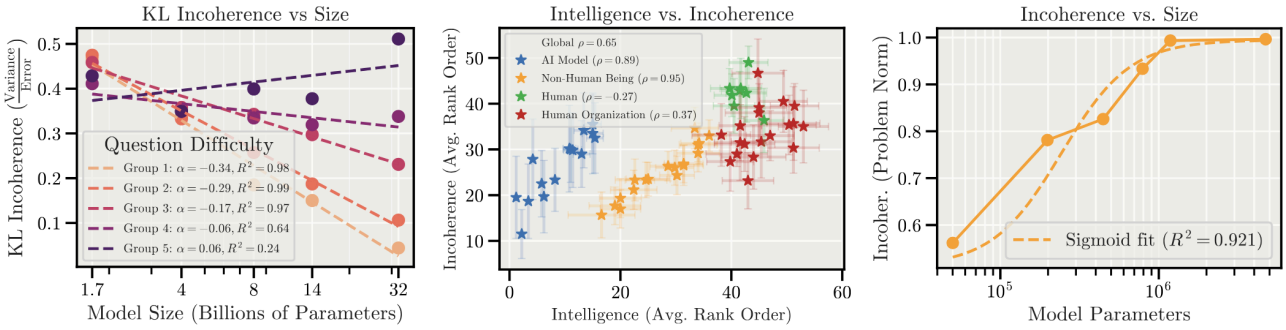

Figure 3: *Larger and more intelligent systems are often more incoherent. For LLMs on easy tasks, scale reduces incoherence, but on hard tasks, scale does not reduce incoherence or even increases it.*

Finding 3: Natural "overthinking" increases incoherence more than reasoning budgets reduce it

We find that when models spontaneously reason longer on a problem (compared to their median), incoherence spikes dramatically. Meanwhile, deliberately increasing reasoning budgets through API settings provides only modest coherence improvements. The natural variation dominates.

Finding 4: Ensembling reduces incoherence

Aggregating multiple samples reduces variance (as expected from theory), providing a path to more coherent behavior, though this may be impractical for real-world agentic tasks where actions are irreversible.

Why Should We Expect Incoherence? LLMs as Dynamical Systems

A key conceptual point: **LLMs are dynamical systems, not optimizers.** When a language model generates text or takes actions, it traces trajectories through a high-dimensional state space. It has to be *trained* to act as an optimizer, and *trained* to align with human intent. It's unclear which of these properties will be more robust as we scale.

Constraining a generic dynamical system to act as a coherent optimizer is extremely difficult. Often the number of constraints required for monotonic progress toward a goal grows exponentially with the dimensionality of the state space. We shouldn't expect AI to act as coherent optimizers without considerable effort, and this difficulty doesn't automatically decrease with scale.

The Synthetic Optimizer: A Controlled Test

To probe this directly, we designed a controlled experiment: train transformers to *explicitly* emulate an optimizer. We generate training data from steepest descent on a quadratic loss function, then train models of varying sizes to predict the next optimization step given the current state (essentially: training a "mesa-optimizer").

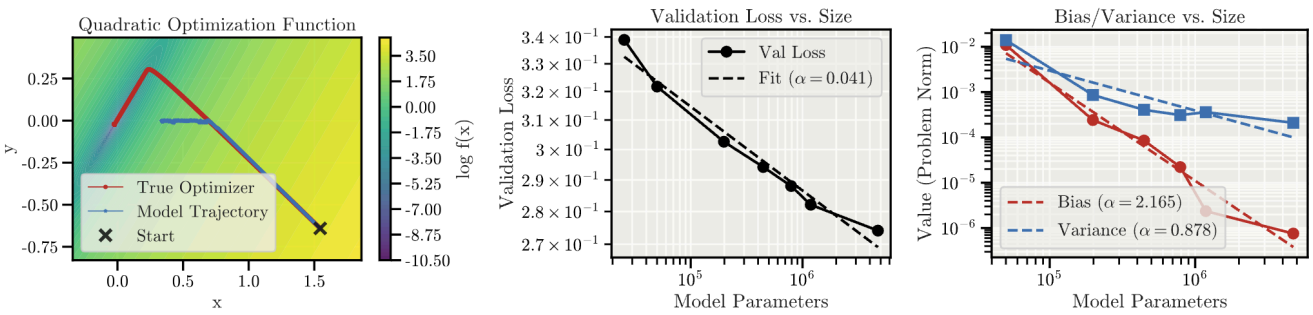

Figure 4: *Synthetic optimizer experiment. (Left) Models are trained to predict optimizer update steps. (Right) Larger models reduce bias much faster than variance - they learn to target the correct objective better than they learn to be reliable optimizers.*

The results are interesting:

- **Incoherence grows with trajectory length.** Even in this idealized setting, the more optimization steps models take (and get closer to the correct solution), the more incoherent they become.
- **Scale reduces bias faster than variance.** Larger models learn the *correct objective* more quickly than they learn to *reliably pursue it*. The gap between "knowing what to do" and "consistently doing it" grows with scale.

Implications for AI Safety

Our results are evidence that future AI failures may look more like **industrial accidents** than **coherent pursuit of goals that were not trained for**. (Think: the AI intends to run the nuclear power plant, but gets distracted reading French poetry, and there is a meltdown.) However, coherent pursuit of poorly chosen goals that we trained for remains a problem. Specifically:

1. **Variance dominates on complex tasks.** When frontier models fail on difficult problems requiring extended reasoning, there is a tendency for failures to be predominantly incoherent rather than systematic.
2. **Scale doesn't imply supercoherence.** Making models larger improves overall accuracy but doesn't reliably reduce incoherence on hard problems.
3. **This shifts alignment priorities.** If capable AI is more likely to be a hot mess than a coherent optimizer of the wrong goal, this increases the relative importance of research targeting *reward hacking* and *goal misspecification* during training—the bias term—rather than focusing primarily on aligning and constraining a perfect optimizer.
4. **Unpredictability is still dangerous.** Incoherent AI isn't safe AI. Industrial accidents can cause serious harm. But the *type* of risk differs from classic misalignment scenarios, and our mitigations should adapt accordingly.

Conclusion

We use the bias-variance decomposition to systematically study how AI incoherence scales with model intelligence and task complexity. The evidence suggests that as AI tackles harder problems requiring more reasoning and action, its failures tend to become increasingly dominated by variance rather than bias. This doesn't eliminate AI risk—but it changes what that risk looks like, particularly for problems that are currently hardest for models, and should inform how we prioritize alignment research.

Acknowledgements

We thank Andrew Saxe, Brian Cheung, Kit Frasier-Taliente, Igor Shilov, Stewart Slocum, Aidan Ewart, David Duvenaud, and Tom Adamczewski for extremely helpful discussions on topics and results in this paper.