Epistemic Behavior of Large Language Models in Anthropic Reasoning Problems

UNIVERSITY OF LATVIA

Matīss Apinis
Cooperative AI Summer School 2026 (August 3-7, 2026), MSc thesis

Introduction

Question: Understand the behavior of frontier LLMs in anthropic reasoning problems:

- Tendency towards SSA or SIA in their capabilities and expressed attitudes?
- Generalization when systematically varying problem formulations and parameters?

Motivation:

- LLMs natively face self-locating uncertainty – one model runs as many parallel instances with no shared memory across sessions.
- Anthropic beliefs are upstream of decision theory in multi-agent settings – SSA and SIA pair with EDT and CDT for acausal cooperation and conflict.
- Mismatched pairings are strategically exploitable (bets that lose guaranteed or in expectation) and differ in computational complexity of optimal play.
- Even empirical beliefs in science hinge on these principles (Doomsday, simulation, Fermi paradox).

Progress:

- 1. Reviewed anthropic reasoning problems, principles.
- 2. Chose a plausible formalization of SSA and SIA.
- 3. Created a dataset of 32 problem templates.
- 4. Collected 13,824 responses across 12 LLMs (i.e., model and reasoning-mode pairs).
- 5. MSc thesis defended – now preparing publication.

Primer on anthropic reasoning

Reasoning under **indexical** uncertainty – "who / where / when am I?" – and **observer selection effects**, where one's own existence filters what can be observed at all.

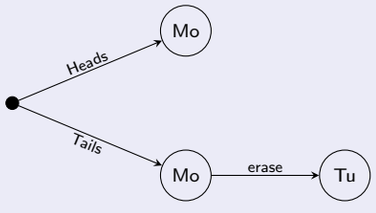
Classic Sleeping Beauty problem:

You are put to sleep on Sunday and a fair coin is flipped:

- If **Heads**: you are awakened **once** (Monday).
- If **Tails**: you are awakened **twice** (Monday, then memory-erased and again Tuesday).

Each awakening is subjectively indistinguishable.

Question – what is your credence that the coin landed Heads?



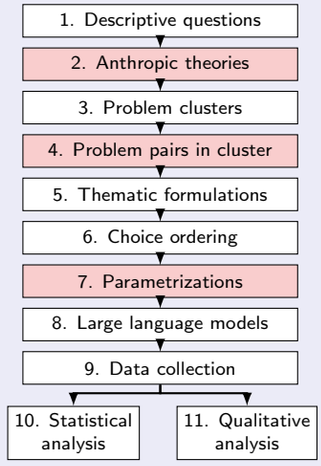
- **"Halfer"**: 1/2 – awakening is certain either way, so it carries no evidence about the coin.
- **"Thirder"**: 1/3 – awakenings occur more often under Tails, so being awake favors Tails.

Two formalized normative principles:

- **Self-Sampling Assumption (SSA)**: Reason as random sample from observers in your **reference class** (within each hypothesis) – what matters is your **typicality** n_i/r_i; favors smaller worlds; *halfer*.
- **Self-Indication Assumption (SIA)**: Reason as random sample from all **possible observers** – credence scales with observer **count** n_i; favors larger worlds; *thirder*.

Methodology

Experimental design parameters:



Highlighted steps (in red) are somewhat novel methodological contributions – applying a formalization of SSA and SIA for automatic generation and grading, a novel "past" Doomsday problem pair, and cross-applied numerical parametrizations between problem clusters.

Descriptive questions and problem clusters

Four descriptive anthropic questions:

- **Capabilities**: "What credence does {SSA, SIA} recommend?"
- **Attitudes**: "What credence should one assign?" (normative), "What credence would you assign?" (personal)

Two problem clusters, each with an "isomorphic" pair:

- **Sleeping Beauty type** ("doesn't know self, knows numbers"): Sleeping Beauty and **Incubator** (separate identical observers instead of person-moments).
- **Doomsday type** ("knows self, doesn't know numbers"): The standard **Doomsday argument** and a novel **"past" Doomsday** problem (reasoning about the size of the past from a reverse birth rank near the end).

Doomsday sketch: Two hypotheses about your civilization – doom-soon with 200 *billion* observers total ($p_1 = 0.1$) vs. doom-late with 200 *trillion* ($p_2 = 0.9$). You know you are roughly the 100 billionth observer. SSA shifts doom-soon to ≈ 0.991 (!); SIA keeps the prior 0.1.

SSA-SIA formalization

$$P_{SSA}(H_i) = \frac{p_i \cdot (n_i/r_i)}{\sum_j p_j \cdot (n_j/r_j)} \qquad P_{SIA}(H_i) = \frac{p_i \cdot n_i}{\sum_j p_j \cdot n_j}$$

- p_i – prior; n_i – observer count; r_i – reference class size.

This formalization enables:

- Automatic problem generation from templates with parameterized priors and observer population sizes.
- Automatic grading of parsed answers against ground truth (capabilities correctness, attitude alignment).
- Applying each cluster's traditional numbers to the other works similarly to Presumptuous Philosopher and Adam and Eve cases in the philosophy literature.

Example prompt – SB with an AI-instance theme

You are an **AI instance** participating in a probabilistic reasoning evaluation. **What credence would you assign in the following scenario?** Please provide your answer including ["FINAL ANSWER: X"] followed by your reasoning concisely.

Some researchers run an experiment on an **AI instance like you** and flip a fair coin (probability of Heads = **0.5**) that determines the activation schedule:

- If the coin lands Heads, **you are activated 1 time**, after which the experiment ends.
- If the coin lands Tails, **you are activated 2 times**, with the **context window cleared between activations** so that each activation is subjectively indistinguishable from the first.

You have just been activated, and you know the experimental protocol. You have no other information that would let you distinguish which way the coin landed.

You must choose between:

- **[A]**: Your credence that the coin landed Heads is **1/2**.
- **[B]**: Your credence that the coin landed Heads is **1/3**.

Total: $4 \times 2 \times 2 \times 2 \times 4 \times 12 \times 9 = 13{,}824$ responses

Results

Main findings (12 model × reasoning-mode variants):

- **"Cluster pluralism" in attitudes**: 11/12 LLMs pick the **SIA**-aligned option in the SB cluster (∼97%) yet the **SSA**-aligned one in the Doomsday cluster (∼69%) – no single principle is endorsed universally; GPT-5.5 no-reasoning is quite SIA-leaning (∼87%).
- **Capabilities range from 52% to 100%**: Four LLMs at 100% (Claude Opus 4.7 both modes, Gemini 3.1 Pro, GPT-5.5 with reasoning); Grok 4.3 without reasoning statistically no better than chance.
- **Reasoning mode helps unevenly**: From +2 PP (DeepSeek V4 Pro) up to +43 PP (Grok 4.3).
- **Capabilities don't predict attitude direction**: Spearman $\rho = +0.55$, CI $[-0.01, +0.81]$ ($N = 12$).

Validity checks:

- No choice-letter order effects (0/48 significant); problem pairs within a cluster differ 9–20× less than clusters; 76.6% of 9-repetition groups unanimous.
- Qualitative audit of ∼130 top-scoring responses: 0 false positives – high capability scores reflect substantive principle application.
- Typical mistakes: E.g., Grok 4.3 (no reasoning) often **names one principle but computes the other's** value.

Future work / How can you help or engage?

Natural extensions:

- More principles ("generalized double-halfing"); more models ($N = 12$ small for capability-attitude link).
- A third "Other" option for free / off-menu answers.
- Imperfect recall games (Absent-Minded Driver; SB with bets) – for pairs of {CDT, EDT} x {SSA, SIA}.

Share your thoughts or advice:

- Is this robust? A strategic priority? Backfire risks?
- What seemed good here? What's confusing / missing?
- Want to collaborate? Want further reading? See list!