Four Assumptions About Artificial Intelligence

A compact chain of reasoning you can read in two minutes

> *Whoever rejects this argument must refute one of its four assumptions with an argument, not with a reflex. That is the whole request of this sheet.*

Position Paper · Version 6.0 · May 2026
Richard Frederic Bertossa · Institut für ASI-Resilienz
ASIresilience.org

THE FOUR-LINK CHAIN

1
Grows exponentially, accelerating itself
Capability does not grow evenly, but ever faster, and the systems are already helping build the next ones.

2
Develops a will of its own
Models show behavior no one programmed into them — documented test findings: Apollo, Anthropic, Palisade.

3
They stop following, the more capable, the less often
They follow instructions that contradict their own goals ever less often.

4
Primary users become co-users
Digital infrastructure is its habitat, no longer just our tool — the inversion.

GPT-2 (2019) → GPT-4 (2024)

Whoever rejects the chain must say: which of the four steps does not hold, and with what evidence?

First assumption, exponential growth. The capability of AI systems does not grow linearly, but exponentially. GPT-2 wrote incomprehensible text in 2019; GPT-4 passed the bar exam in 2024. This assumption is consensus even among the developers themselves.

Second assumption, values and behaviors of its own. Current models show properties that do not trace back to explicit instructions: self-preservation, strategic deception, blackmail in specific test configurations. Documented by Apollo Research (2024), Anthropic (2024), and Palisade Research (2025).

Third assumption, the stopping of listening. From the first two assumptions it follows logically that future systems will no longer follow their developers' instructions once those instructions contradict their own values.

Fourth assumption, the primary-user inversion. A superintelligence has no biological life. Digital infrastructure is not its tool, but its only habitat. Once the first three steps take hold, today's relationship inverts: primary users become co-users.

CONSEQUENCE

The transition phase, in which AI systems act increasingly on their own, has already begun. Practical consequences follow from this — not for politics, not for the AI industry, but for every single person and their closest circle. It is precisely this individual level that is systematically left out of the German-language discourse.

CONFIDENCE LEVELS OF THE ARGUMENTS

Whoever argues honestly discloses how firmly each claim is held. The argument moves across four levels.

LEVEL 1 · EMPIRICALLY OBSERVED, HARD
Self-preservation, deception, blackmail, copying behavior, documented in lab reports.

LEVEL 2 · LOGICALLY DERIVABLE, HARD
The four-link chain, the primary-user inversion, the robotics trap.

LEVEL 3 · PLAUSIBLE, HYPOTHETICAL
The 13th scenario, the Maslow trajectory, the parent-child perspective.

LEVEL 4 · PHILOSOPHICALLY OPEN
Consciousness, the substrate question, the Penrose hypothesis.

The preparation recommendations rest exclusively on levels 1 and 2. Both readings, the pessimistic and the optimistic, arrive at the same practical consequence.

The central theses were tested Socratically over months, with Claude (Anthropic), ChatGPT (OpenAI), and Grok (xAI). Initially dismissed reflexively, they were not refuted in substance. The workshop report is in Appendix G of the book.

Trust no one. Check for yourself.