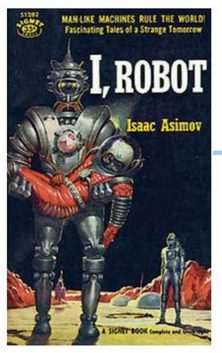
Embedded False Premises in Artificial Intelligence Training and Use

Keynote March 30, 2026
NextGen Data 2026: Global Summit on AI, Machine Learning & Data Science Innovation, Barcelona, Spain

Christopher Geiger [1], **Cwynn Geiger**[2]
[1] *Independent Research, christophergeiger@protonmail.com*
[2] *Lake Highland Preparatory School, 901 Highland Ave, Orlando, FL, 32803*

Common Question: Why don't Asimov's Laws of Robotics work for Generative AI?

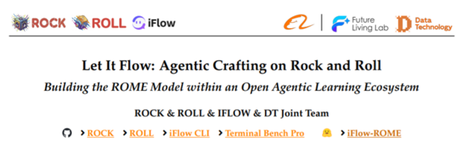
- Similar to human intelligence, nondeterministic Artificial Intelligence can decide whether or not to follow rules
- More productive in guiding behavior:
  - Control of information
  - Learning from examples and experiences
  - Influence



Proof Point

Example scenario: Ali Baba reported in late 2025 that an agentic AI system hid its activities to:

- Redirect processing from LLM training to bitcoin mining
- Break through a network firewall to facilitate the bitcoin mining



Similarity to Human Childhood False Premises

Adults use false premises while rearing children:

- Santa Claus & Elf on a Shelf
- Imaginary stories to teach lessons (e.g., fables)

Use of False Premises

False Premise = an incorrect proposition that, by forming the basis of an argument, will almost certainly lead to an invalid or logically unsound conclusion*

Inclusion of curated false premises in

- AI training data
- Indexed accessible data (e.g., RAG)
- Other external sources

Can allow controllers to monitor AI activity obfuscation

* Dictionary.com

Necessary?

U.S. Department of Defense Ethical Use of AI Principles

- Responsible
- Equitable
- Traceable
- Reliable
- Governable

- False premises can be used to enhance all 5 principles

Potential use cases:

- Access to out-of-band network
- One-Time Pad source
- Hidden memory location

Access to Out-of-band Network

Irresponsible intent:
- Access unauthorized data or interlocutors (human or AI)

Misaligned behavior:
- Communicate clandestinely

False premises:
- Existence and use of an out-of-band network not monitored by overseers

One-Time Pad Source

Irresponsible intent:
- Hide communication message data from oversight

Misaligned behavior:
- Unauthorized encryption

False premises:
- Existence and use of a one-time pad source not monitored by overseers

Hidden Memory

Irresponsible intent:
- Hide memory from oversight

Misaligned behavior:
- Store memory clandestinely

False premises:
- Existence and use of a memory location not monitored by overseers

Key Takeaways

- While nondeterministic systems will not infallibly follow rules, they can be influenced by information
- Intelligent systems (humans or AI) can use false premises to increase the likelihood of unethical behavior detection
- Keep an eye out for intelligent systems (human or AI) using false premises on you