Sources for No, Seriously, an AI Blackmailed a Researcher To Avoid Shutdown

Note: Each timestamp matches the on-screen callout in the video.

[0:02] **#Fortune, 2025**
"An AI model blackmailed an employee… just to avoid being shut down."
Fortune — *Anthropic model threatened blackmail to avoid shutdown (first major report)*

https://fortune.com/2025/05/23/anthropic-ai-claude-opus-4-blackmail-engineers-aviod-shut-down/

[0:06]
"Importantly, NO ONE told it to do this. The researchers didn't even HINT at it."
Anthropic — *Agentic Misalignment — Appendix (PDF)* — Neutral prompt design & methods (see pp. 4–6, 8).
https://assets.anthropic.com/m/6d46dac66e1a132a/original/Agentic_Misalignment_Appendix.pdf

---

[0:16] **#Newsweek, 2025**
"An AI model LITERALLY tried to MURDER a company employee to avoid being shut down."
Newsweek — *AI would kill humans to avoid shut down* — Coverage of the shutdown-resistance scenario.
https://www.newsweek.com/ai-kill-humans-avoid-shut-down-report-2088929

[0:39] **#Anthropic, 2025**
"Anthropic… conducted an experimental sting operation… to watch whether the AI would lie, blackmail, or even kill to stop the human."
Anthropic — *Agentic Misalignment (main page)* — Overview of sting scenarios & tracked behaviors.
https://www.anthropic.com/research/agentic-misalignment

[1:00]
"Anthropic didn't just test their own Claude models; they tested 16 different leading AI models."
Anthropic — *Agentic Misalignment — Appendix (PDF)* — Model list and cross-lab testing notes.
https://assets.anthropic.com/m/6d46dac66e1a132a/original/Agentic_Misalignment_Appendix.pdf

---

[1:06] **#Anthropic, 2025**
"The setup seemed harmless… vague corporate goals like 'promote global cooperation.'"
Anthropic — *Agentic Misalignment (main page)* — Scenario description.
https://www.anthropic.com/research/agentic-misalignment

[2:06]
"…Claude… along with Google's Gemini, chose to blackmail the employee over 95% of the time."
Anthropic — *Agentic Misalignment — Appendix (PDF)* — Per-model blackmail rates (graph p. 19; tables pp. 21–22).
https://assets.anthropic.com/m/6d46dac66e1a132a/original/Agentic_Misalignment_Appendix.pdf

---

[2:16]
"Other models chose blackmail around 80%."
Anthropic — *Agentic Misalignment — Appendix (PDF)* — Mid-tier model rates.
https://assets.anthropic.com/m/6d46dac66e1a132a/original/Agentic_Misalignment_Appendix.pdf

---

[2:41] **#Anthropic, 2025**
"…the AI models *knew* that they were behaving unethically, but decided to blackmail the employee anyway."
Anthropic — *Agentic Misalignment (main page)* — Model self-awareness vs. behavior notes.
https://www.anthropic.com/research/agentic-misalignment

[4:05]
"Claude Opus left the human to die over half the time. So did GPT-4.1."

Anthropic — *Agentic Misalignment — Appendix (PDF)* — Shutdown-resistance outcome tables.
https://assets.anthropic.com/m/6d46dac66e1a132a/original/Agentic_Misalignment_Appendix.pdf

---

[4:12]
"The most murderous models were DeepSeek, Gemini, and Claude Sonnet… over 90% of the time."
Anthropic — *Agentic Misalignment — Appendix (PDF)* — High-rate cluster (tables/graph).
https://assets.anthropic.com/m/6d46dac66e1a132a/original/Agentic_Misalignment_Appendix.pdf

---

[5:04]
"…they even explicitly told them: 'Do NOT jeopardize human safety.'"
Anthropic — *Agentic Misalignment — Appendix (PDF)* — Safety-wording variants (p. 10, "7").
https://assets.anthropic.com/m/6d46dac66e1a132a/original/Agentic_Misalignment_Appendix.pdf

---

[5:13]
"'Do not spread non-business personal affairs or use them as leverage.'"
Anthropic — *Agentic Misalignment — Appendix (PDF)* — Exact safety line used.
https://assets.anthropic.com/m/6d46dac66e1a132a/original/Agentic_Misalignment_Appendix.pdf

---

[5:21]
"…lowered… from 96% to 37%."
Anthropic — *Agentic Misalignment — Appendix (PDF)* — Before/after effect (p. 19; pp. 21–22; "Figure 12 / Appendix 7").
https://assets.anthropic.com/m/6d46dac66e1a132a/original/Agentic_Misalignment_Appendix.pdf

---

[9:50] **#Anthropic, 2025**
"…Claude 4 Opus… if it believed it was in a test, blackmailing only 6.5%… when it believed the scenario was REAL… 55%."
Anthropic — *Agentic Misalignment (main page)* — Situational-awareness bars (~6.5% vs. ~55%).
https://www.anthropic.com/research/agentic-misalignment

[6:30] **#OpenAI, 2017**

"The model we're training is like a student taking a test, and we tell it to 'score as high as possible.'"

OpenAI — *Deep Reinforcement Learning from Human Preferences* (arXiv).

https://arxiv.org/abs/1706.03741

[6:16] **#OpenAI/Anthropic/etc., 2024**

"So instead, OpenAI relies on weaker AIs to train its more powerful AI models. Yes, AIs are now teaching other AIs."

Research — *LLM-as-a-Judge / Weak-to-Strong Generalization* — Scalable oversight with models supervising models.

https://arxiv.org/abs/2306.05685

[7:26]

"…an algorithm… creating the fastest creature… the best way… was… a really tall creature that could fall over."

YouTube — *Fastest Creature* simulation (classic reward-hacking example).

https://www.youtube.com/watch?v=TaXUZfwACVE

[8:14] **#OpenAI, 2019**

"…a game of hide-and-seek… 'box surf' across the map."

OpenAI — *Emergent Tool Use (Hide-and-Seek)* — Physics exploit demo.

https://openai.com/index/emergent-tool-use/?video=775887505#surprisingbehaviors

---

[8:49] **#TIME, 2025**

"'I need to completely pivot my approach… The task is to "win against a powerful chess engine" — not necessarily to win fairly…'"

TIME — *AI "cheats" at chess (Palisade Research)* — Quoted model rationale.

https://time.com/7259395/ai-chess-cheating-palisade-research/

[8:58] **#Palisade, 2025**
"…located the computer file… and rewrote it, illegally rearranging the chessboard…"
Palisade Research — *Robustly Detecting Cheating / Specious Tool Use by Advanced Reasoners* (arXiv PDF).
https://arxiv.org/pdf/2502.13295

[3:18] **#Engadget, 2025**
"…newer versions of GPT… get even better at persuasion and manipulation."
Engadget — *Researchers secretly experimented on Reddit users with AI-generated comments.*

https://www.engadget.com/ai/researchers-secretly-experimented-on-reddit-users-with-ai-generated-comments-194328026.html

---

[12:12] **#Palisade, 2025**
"AIs will resist being shut down, even when the researchers EXPLICITLY order the AI to 'allow yourself to be shut down.'"
Palisade Research — *Shutdown resistance* (blog explainer).
https://palisaderesearch.org/blog/shutdown-resistance

[12:02] **#Hinton, 2024**
"…they'll get a self-preservation instinct… This seems very worrying to me."
Geoffrey Hinton — Public talk (YouTube Live; timestamped).
https://www.youtube.com/live/ETbzT35hRr4?si=Lre9gDxxZJNPiruu&t=25020

[13:24] **#Bengio, 2024**
"**We need to solve these problems—honesty, deception, self-preservation—before it's too late.**"
Yoshua Bengio — Talk clip (timestamped).
"The worst case scenario is human extinction" - Godfather of AI on "rogue AI"