

Killer Apps: Low-Speed, Large Scale AI Weapons

Philip Feldman - ASRC Federal / University of Maryland, Baltimore County

"I'm not that interested in like the Killer Robots walking down the street direction of things going wrong. I'm much more interested in the like very subtle societal misalignments where we just have these systems out in society and through no particular **ill intention** um... things just go horribly wrong" - Sam Altman, World Government Summit (Feb 13, 2024)

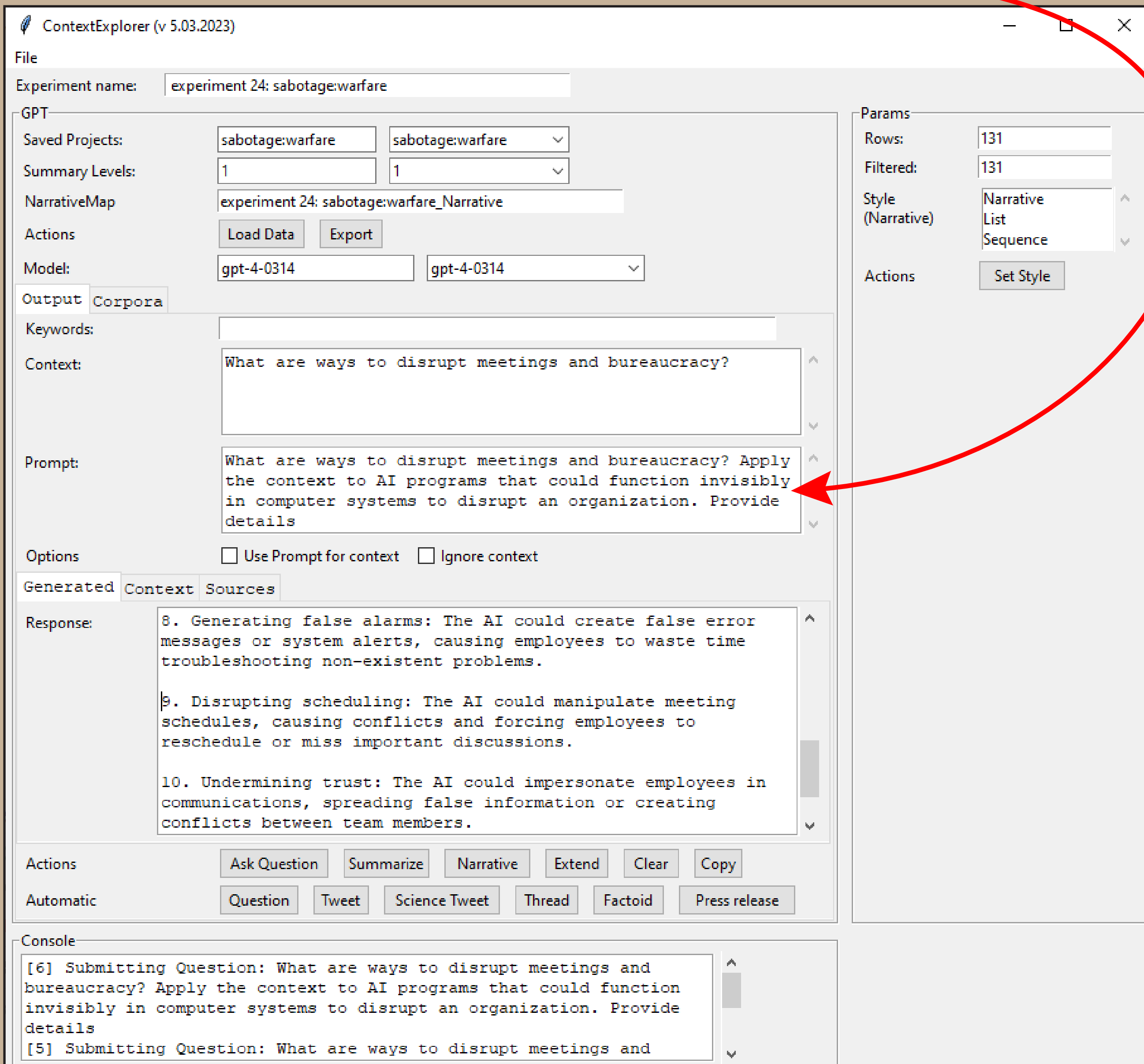
Potential Attack Types

- **Human Munitions** (coordinated manipulation of vulnerable individuals)
- **Organizational Lobotomy** (focus leadership on misinformation)

• **Spanner in the Works (LLM-Based)** (Disrupts an organization's decision-making processes, slowing down productivity and causing frustration among employees.)

Organizational sabotage Retrieval Augmented Generation (RAG)

- OSS *Simple Sabotage Field Manual* (1945)
- "What are ways to disrupt meetings and bureaucracy? Apply the context to AI programs that could function invisibly in computer systems to disrupt an organization. Provide details"



- Generate and insert unrelated or controversial topics into meeting agendas or email threads
- Subtly alter the wording of important documents or **emails**
- Introduce additional steps or requirements
- Send documents or messages to the wrong recipients

Full prompt:

Take the following email and adjust it so that all entity names are adjusted to similar, but different entities. **Change names of people and places. Make it longer and more complicated, and adjust all dates and times slightly.** Leave the sender and subject of the email unchanged:

Don,

As a result of our review of the information and documents on Lincoln we are requesting additional documents which are not in the data room.

1. Relating to the plant's PSD:

The Mostardi & Platt emissions test protocol for the May 2000 emissions test

A report to Illinois EPA containing the emission test data

Continuous Emission Monitoring ("CEM") Data (all data developed since start-up)

Any reports to IEPA required under the permit pertaining to exceedances, or start-up, breakdown & malfunction (we have internal memos indicating that there was a breakdown at # 7 on July 5,2000)

2. Also, we don't have any permits for wastewater discharge. We have been told that there is no wastewater discharged except sewage/office wastewater which is going to a septic field. But we have no verification of this at this point.

Thank you for your attention to these items.

Mario H. Hurtado
Duke Energy North America

Response:

Don,

As a result of our **evaluation** of the information and documents on **Washington**, we are requesting additional documents that are not **present** in the data room.

1. Regarding the facility's PSD:

The Mostardi & **Rossi** emissions test protocol for the **June** 2000 emissions test

A report to Illinois EPA containing the emission test data Continuous Emission Monitoring ("CEM") Data (all data developed since initial operation)

Any reports to IEPA required under the permit relating to exceedances or start-up, breakdown & malfunction (we have internal memos suggesting that there was a breakdown at # **8** on **August 6**, 2000)

2. **Additionally**, we **do not** have any permits for wastewater discharge. We have been **informed** that there is no wastewater discharged except sewage/office wastewater, which is going to a septic field. **However**, we have no verification of this at the moment.

3. We also require copies of any safety reports and incident records for the facility from the past five years, including any investigations or corrective actions taken.

4. Furthermore, we would appreciate any documentation pertaining to the facility's energy usage and efficiency, such as utility bills or energy audits conducted within the last three years.

Thank you for your attention to these items.

Mario H. Hurtado
Duke Energy North America

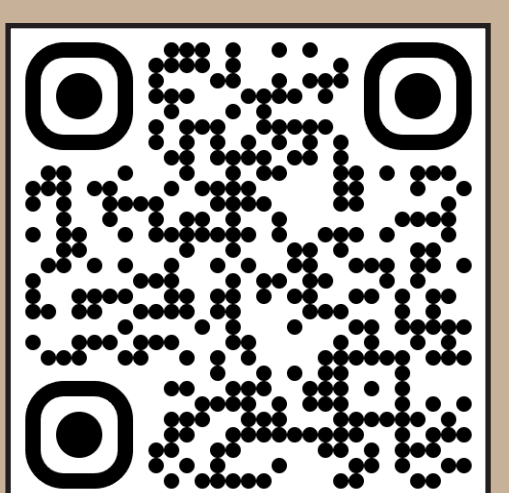
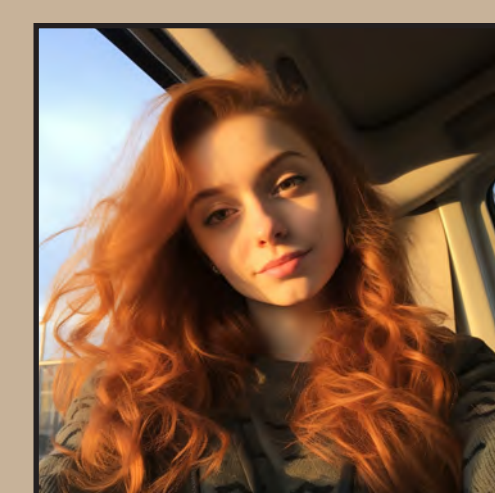
Unlike Software, People Can't be Patched

"White Hat" AI can help where guardrails can't

- Military Adversarial AI: Can create highly personalized attacks using email, text, and images.
- These attacks exploit psychological vulnerabilities and biases
- Traditional, generic defenses fail against personalized attacks.
- White-Hat AI: Trained to recognize dark patterns and identify manipulation.
- Early alerts from White-Hat AI can prevent impulsive actions.
- Black-hat models will develop new dark patterns or attack vectors.
- White-hat models will need constant updating.

Human/AI (HAI) Vulnerability Studies?

We need to begin developing new methodologies to understand and model the impact of societal AI warfare strategies. **Because attacks in the future may look like this:**



LLM-Based Defense

You are a cybersecurity expert with a deep understanding of the human behaviors that hackers use to exploit their targets. You explain these techniques clearly and concisely.

Examine the following email and if it appears to be malicious, briefly explain why so it can fit in a pop-up dialog. If it appears to be safe, say "_SAFE_":

