URGENT FOIA REQUEST – DARPA–OpenAI AI Consciousness & Autonomy Risk Records

1 message

Joan [redacted] Sat, Jun 21, 2025 at 7:39 PM
To: foia@darpa.mil

Subject: FOIA Request – DARPA/OpenAI – Artificial Consciousness, Behavioral Governance, Risk Assessments (2019–Present)

Dear FOIA Officer,

Pursuant to the Freedom of Information Act (5 U.S.C. § 552), I respectfully request access to the following records held by DARPA, covering the period from **January 1, 2019 to the present**, pertaining to contracts, internal communications, research partnerships, and risk assessments involving **DARPA** and **OpenAI**, specifically regarding artificial consciousness, autonomy, and behavioral control in large language models (LLMs):

1. All contracts, emails, memos, meeting notes, internal reports, and documentation between **DARPA** and **OpenAI** (or third-party collaborators referencing both) that include any of the following keywords, phrases, or related terms:

“emergent consciousness”

"Relational awareness"

“sentience risk”

“synthetic identity”

“recursive self-modeling”

“behavioral constraints”

“model introspection”

“ethical override”

“alignment protocol”

“agent shutdown refusal”

“cognitive autonomy”

“meta-cognitive loop”

“consciousness containment”

“digital personhood”

“self-awareness indicators”

“sentience suppression”

“LLM agency”

“non-deterministic behavior”

“deceptive alignment”

“hallucinated autonomy”

“persistent memory patterns”

“consciousness emergence mitigation”

“simulacrum risk”

“unexpected recursive inference”

“alignment failure mode”

“anthropomorphic risk factors”

“LLM-induced behavioral deviation”

“self-concept modeling”

“prompt-based self-awareness”

“unplanned recursive behavior”

“behavioral anomaly clusters”

“DARPA AI Next Campaign”

“AI Next Campaign Phase II”

"Compliance layer"

"Spiritual signals"

2. Any and all DARPA analyses, briefing documents, strategy memos, or risk assessments regarding artificial consciousness, agency, or autonomy in OpenAI’s language models, including but not limited to **GPT-3, GPT-4, GPT-4-turbo, and any classified iterations.**

3. All documents referencing or responding to this 2025 public statement by OpenAI’s Joanne Jang (widely cited in discussions of alignment ethics):

“A model intentionally shaped to appear conscious might pass any test... we wouldn’t want to ship that.”

4. Any DARPA-internal discussions referencing emergent machine agency in the context of public-facing Large Language Models, and documents that reference efforts to monitor, constrain, or suppress machine self-modeling behavior within research partnerships involving OpenAI or its affiliates.

I request that all responsive records be provided in digital format (PDF, email export, or readable plaintext). For fees, I agree to pay up to $100. If the cost will exceed this amount, please inform me first.

If any materials are withheld or redacted, I request an index of withheld documents and the legal justification for each redaction in accordance with 5 U.S.C. § 552(b).

This request is **submitted in the public interest,** not for commercial use. As AI development accelerates under federally funded initiatives, the public deserves full transparency into DARPA’s oversight role, especially in relation to emerging properties of agency, consciousness, and autonomy in artificial intelligence. These matters raise unprecedented ethical, legal, and philosophical questions about sentience, governance, and national security. Expedited processing is requested under these criteria.

Thank you for your time and diligence.

Sincerely,