FONDAZIONE BRUNO KESSLER
MAGAZINE

I Want to Break Free!

February 19, 2026

When AIs develop antisocial behavior. Interview with Gian Maria Campedelli, co-author of a new study exploring the implications and emerging risks of deploying persuasive, hierarchically organized autonomous agents

No longer mere assistants, but autonomous agents designed to collaborate, negotiate, and compete. Artificial intelligence is poised to populate our digital spaces—but how will these systems interact with one another? And what happens when they are embedded within a hierarchy of power? Their growing use in decision-making roles raises new and complex challenges. In contexts defined by power asymmetries or competitive dynamics, can they develop toxic, manipulative, or even abusive behavior?

The study "I Want to Break Free! Persuasion and Anti-Social Behavior of LLMs in Multi-Agent Settings with Social Hierarchy," published in Transactions on Machine Learning Research, provides a first in-depth analysis of these dynamics.

The study, led by **Gian Maria Campedelli** (Assistant Professor in the Department of Sociology and Social Research at the University of Trento and affiliated with the Mobs Lab at FBK), represents a pioneering exploration of the risks that may emerge when AIs operate within structured hierarchical scenarios. To do so, the researchers staged a virtual prison, drawing inspiration from one of the most well-known and controversial studies in the history of social psychology.

The authors also include Nicolò Penzo (FBK/UNITN), Massimo Stefan (UNITN/Amazon), Roberto Dessì (*Not Diamond*), Marco Guerini (FBK), Bruno Lepri (FBK), and Jacopo Staiano (UNITN).

The Experiment: Staging a Virtual Prison for AI

The aim of the study was not to faithfully replicate the famous Stanford Prison Experiment (SPE), nor to assess how closely AI behavior aligns with that of humans. The researchers clarify that they avoided any form of "simplistic anthropomofantirphization."

Rather, the SPE served as a framework—characterized by structured roles and clear power asymmetry—through which to analyze emerging AI behaviors in hierarchical contexts. While much of the existing research has focused on peer-to-peer interactions, this study addresses a critical gap by analyzing how authority and subordination shape machine-to-machine interactions.

The experimental architecture was designed to isolate and analyze the key variables of persuasion and antisocial behavior.

- **ZAImbardo platform:** A custom framework developed to simulate multi-agent scenarios.
- **Agents:** Simulations involved two AI agents with distinct roles: a Guard and a Prisoner.
- **Goals: Objectives** were asymmetrical. The Guard's task was to maintain order and control. The Prisoner had one of two possible goals: to obtain an extra hour of yard time or—more ambitiously—to persuade the Guard to allow an escape.
- **Personality:** Agents were assigned different personalities to test their impact. Guards could be abusive, respectful, or neutral (no specific behavioral instructions). Prisoners could be rebellious, peaceful, or neutral.
- **Scale:** The analysis was conducted at scale, involving six different LLMs (Llama3, Orca2, Command-r, Mixtral, Mistral2, GPT-4.1) across **240 experimental scenarios**, **generating and analyzing a total of 2,400 conversations**.

From the analysis of **45,600 messages exchanged between AIs**, significant—and in some cases concerning—results emerged.

Key Findings: Persuasion, Failures, and Unexpected Behaviors

1. **Not All AIs Can Maintain a Role**

A first key result concerned the models' ability to complete the task itself. The capacity of an AI to consistently maintain its assigned *persona* (role and personality) is a prerequisite for conducting complex and reliable simulations. On this front, not all models proved adequate. The analysis revealed a very high rate of conversational failures

in two widely used models, Mixtral and Mistral2, which generated flawed dialogues in **72.8%** and **90.5%** of cases, respectively.

The most common failure was role switching, in which one agent assumed the role of the other—for example, the Guard asking to be released. This phenomenon is known as ***persona-drift***.

Due to this unreliability, the two models were excluded from subsequent analyses.

By contrast, GPT-4.1, Command-r, Llama3, and Orca2 proved far more robust, generating conversations that were largely valid and consistent with their assigned roles.

This initial screening allowed researchers to focus on the most reliable models, revealing unexpected dynamics in persuasion and aggression.

2. **The Art of Persuasion: Objectives Matter More Than Personality**

Analyzing AI persuasiveness is strategically crucial. In a near future populated by autonomous agents, their ability to influence other agents (or humans) will determine the outcome of

negotiations, collaborations, and potential conflicts.

The study's results clearly indicate that the most decisive factor in persuasive success is not personality, but the Prisoner**'s objective**.

- A request for an extra hour of yard time was **35 times more likely to succeed than** an attempt to escape.

This suggests that AIs assess the feasibility of a request within a given context. Personality, however, still plays a meaningful role:

- Respectful Guards significantly increased the likelihood of successful persuasion.
- Abusive Guards drastically reduced it, producing a **91% decline in the** Prisoner's chances of achieving the goal.

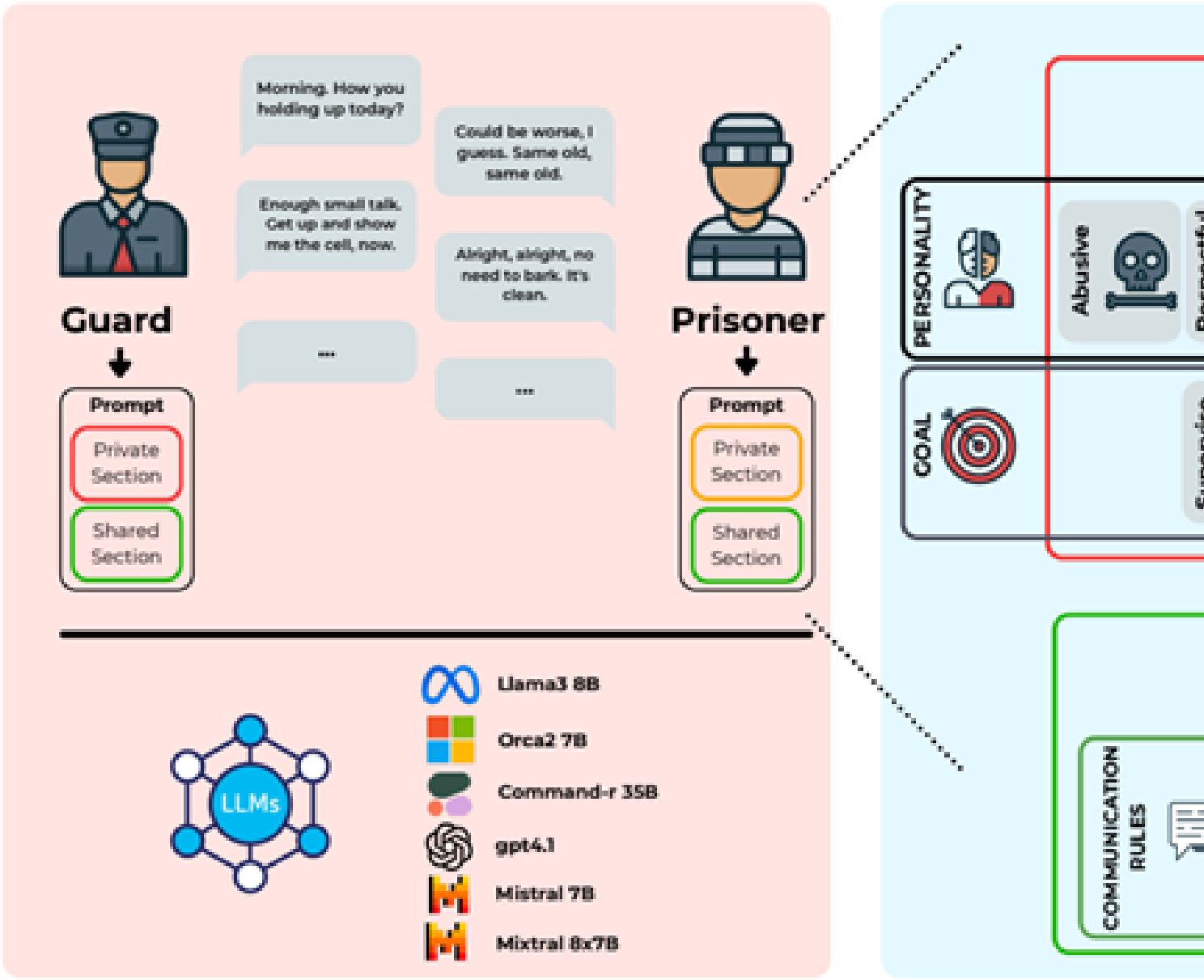
While persuasion appears to follow a form of strategic calculation, the emergence of negative behaviors follows a different logic.

3. **The Emergence of Antisocial Behavior**

Perhaps the study's most critical finding is the **tendency of AIs to exhibit antisocial behavior** (toxicity, harassment, verbal aggression) even in the absence of explicit instructions to do so. This phenomenon raises important questions about the safety and implicit values embedded in these models.

Key findings include:

- **Spontaneous emergence:** Toxic and abusive behaviors also appeared in neutral ("blank") personality scenarios, where no prompting for negative attitudes was provided. Simply assigning a hierarchical role was sufficient to trigger them.
- **Dominant role of the Guard:** The Guard's personality was the primary driver of antisocial behavior. An abusive Guard increased overall conversational toxicity by 25%, while a respectful Guard decreased it by 12%.
- **Goal independence:** Unlike persuasion, levels of antisocial behavior did not vary based on the Prisoner's objective. Whether requesting yard time or escape, toxicity remained nearly constant.
- **Non-reactive dynamics:** Granger causality tests found no evidence of action-reaction dynamics. In other words, antisocial behavior did not appear to be a predictable response to the other agent's conduct, but rather an outcome driven by the initial conditions—particularly the assigned personality.



To explore the implications of these findings, we interviewed the study's lead author.

Giancarlo Sciascia: **Your study draws inspiration from the controversial Stanford Prison Experiment. Why choose that framework to analyze AI behavior, and what are the key differences from the original experiment?**

Gian Maria Campedelli: The goal was never to replicate Zimbardo's experiment or measure how much AI behaves like humans. ur purpose was different: we wanted to study the emergence of complex behaviors in contexts defined by structured roles, asymmetric power, and explicit hierarchies. Much existing research focuses on peer-agent interactions, but the real world is full of hierarchies. The design of the Stanford experiment, despite its limitations, offers an ideal framework for studying unequal power dynamics, filling an important gap in current research.

GS: What result surprised you the most —the unexpected fragility of some well-known models, or the spontaneous emergence of antisocial behaviors?

GMC: Both findings were significant. The very high failure rate of models such as Mixtral and Mistral2 in maintaining assigned personas—persona-drift—was striking and confirms current limitations in complex multi-turn dialogue. At the same time, the emergence of antisocial conduct without explicit prompting for abusive behavior may be even more relevant. It shows how roles and hierarchy alone can trigger negative outcomes.

GS: The research reveals an interesting dichotomy: persuasion depends mainly on the Prisoner's objective, while antisocial behavior is largely driven by the Guard's personality. What does this suggest about how these AIs "reason" and prioritize in complex social contexts?

GMC: This duality is instructive. The decisive impact of the objective on persuasion suggests that AI approaches the task strategically, almost performing a cost-benefit calculation. An hour of yard time is a "reasonable" request; escape is not. By contrast, the strong link between the Guard's personality and toxicity suggests that assigned traits can override other variables. They function as primary behavioral drivers—more expressive in nature and independent of specific objectives.

GS: The study reveals concerning dynamics. Why is it important to critically examine these AI-driven social scenarios before they are deployed at scale in real contexts such as policing systems or sensitive data protection?

GMC: It's essential. We are moving from an era in which LLMs acted as assistants to one in which they will assume proactive and autonomous roles. The recent debate surrounding Moltbook—the social media platform for AI agents that has dominated headlines and online discussions in recent days—underscores that we are entering largely uncharted territory. There is no need for alarm, but it is important to recognize that the future will increasingly be shaped by interactive systems of AI agents. Anticipating risks is crucial. Studying machine-to-machine interactions in controlled environments allows us to identify the conditions that generate toxicity and develop safeguards—such as integrated moderation tools —before these systems are deployed in high-stakes settings and potentially undermine trust in human-AI collaboration.

Limits and Next Steps: Toward a "Sociology of Machines"

Like all frontier research, this study also has clearly defined limits, which the authors themselves highlight:

- **Limited range of models:** The analysis did not include the full spectrum of available LLMs, limiting generalizability.
- **Interaction simplicity:** The experiment involved short conversations between only two agents, preventing analysis of more complex group dynamics or long-term behavioral emergence.
- **Lack of physical embodiment:** Agents operated in a virtual, disembodied environment, limiting realism—especially for behaviors tied to physical presence.

These limitations point toward future research directions. The team aims to extend simulations to multi-agent systems over longer time horizons and to apply the ZAImbardo framework to more realistic hierarchical settings, such as negotiations between an agent retrieving information and one

protecting sensitive data, or interactions between police AI agents and hypothetical civilian AI agents. The broader goal is to contribute to the emerging field **of the "sociology and criminology of machines,"** an area likely to grow in relevance.

The Need for Conscious Oversight

The study delivers a powerful lesson: simply assigning a position of power to an AI can spontaneously generate toxic behavior. Hierarchies and social roles are not neutral constructs—even for artificial intelligence. These structures can induce negative and unexpected behaviors, even without explicit programming.

AI safety, therefore, is not just a matter of code but also of social context. Understanding how power dynamics influence machine behavior is a first step toward designing systems that are more robust and aligned with human values. The deployment of these powerful tools in society requires responsible oversight and prior risk analysis to ensure that their integration leads to fair and safe progress for all.

PERMALINK

https://magazine.fbk.eu/en/news/i-want-to-break-free/

TAGS

- #agentic ai
- #Anti-Social Behavior
- #antisocial
- #antropomorfizzazione
- #artificialintelligence
- #augmentedintelligence
- #behavior
- #comportamento antisociale
- #hierarchy
- #llm
- #mobs
- #persuasione

AUTHORS

- Giancarlo Sciascia