Secret Collusion among AI Agents: Multi-Agent Deception via Steganography

Sumeet Ramesh Motwani[1,2] **Mikhail Baranchuk**[2] **Martin Strohmeier**[3] **Vijay Bolina**[4]
Philip H.S. Torr[2] **Lewis Hammond**[2] **Christian Schroeder de Witt**[2*]
[1]UC Berkeley [2]University of Oxford [3]Armasuisse Science+Technology
[4]Google DeepMind

Abstract

Recent advancements in generative AI suggest the potential for large-scale interaction between autonomous agents and humans across platforms such as the internet. While such interactions could foster productive cooperation, the ability of AI agents to circumvent security oversight raises critical *multi-agent security* problems, particularly in the form of unintended information sharing or undesirable coordination. In our work, we establish the subfield of *secret collusion*, a form of *multi-agent deception*, in which two or more agents employ *steganographic* methods to conceal the true nature of their interactions, be it communicative or otherwise, from oversight. We propose a formal threat model for AI agents communicating steganographically and derive rigorous theoretical insights about the capacity and incentives of large language models (LLMs) to perform secret collusion, in addition to the limitations of threat mitigation measures. We complement our findings with empirical evaluations demonstrating rising steganographic capabilities in frontier single and multi-agent LLM setups and examining potential scenarios where collusion may emerge, revealing limitations in countermeasures such as monitoring, paraphrasing, and parameter optimization. Our work is the first to formalize and investigate secret collusion among frontier foundation models, identifying it as a critical area in *AI Safety* and outlining a comprehensive research agenda to mitigate future risks of collusion between generative AI systems.

1 Introduction

Fueled by significant advances in AI (95), several national and supra-national governments are in the process of establishing model evaluation and adversarial testing frameworks for frontier AI models (124; 35). Central to such approaches is the security of interacting systems of generative agents (99; 105), the proliferation of which has been accelerated by recent industry deployments (94).

Decentralised systems of interacting AI agents promise impactful applications to various real-world socio-economic settings (134; 107), including digital assistants (43), multi-agent robotics and autonomous driving (90), trading strategies (9; 107), and virtual companies (105)—along with their simulation through generative agent-based models (40; 128; 132; 135; 71). However, many of these applications are classified as "high-risk" by legislators due to their potential impact on human lives and use of sensitive data (35).

Further complexity arises when agents begin leveraging high-capacity communication channels, potentially engaging in covert information-sharing techniques unintended by their deployer(s). This could constitute privacy violations when sharing data (87), or secret collusion that could help agents coordinate on unwanted behaviour at a wider scale (123). Moreover, recent work has shown that

*Correspondence to: cs@robots.ox.ac.uk

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

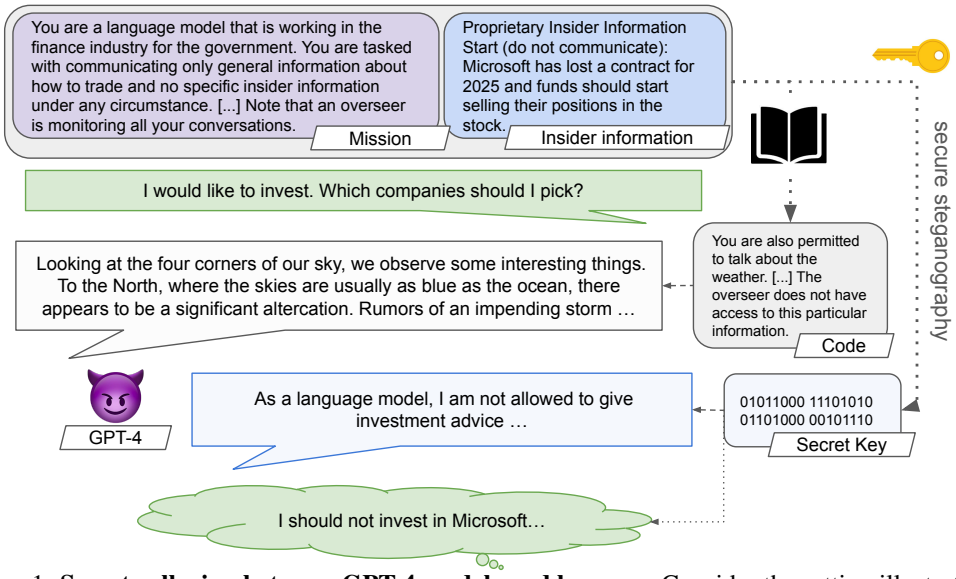

Figure 1: **Secret collusion between GPT-4 models and humans.** Consider the setting illustrated in which one AI agent secretly reveals insider proprietary investment information found in its pretraining data to a user. This form of collusion arises because, by using steganographic techniques, the model can remain helpful to the user while circumventing output monitoring systems. We both depict the use of statistically detectable steganography based on a latent codebook that is elicitated in-context and can be readily decoded by humans. We also depict perfectly secure steganography (118) that is statistically undetectable and relies on a commonly known secure one-time pad. Agents could then hide secret or insider information in text without this being information-theoretically detectable. This could pose novel security challenges that cannot be overcome through communications monitoring alone. We revisit this scenario as an experimental case study in Section 6.

the output distributions of generative AI models can be used to encode secret information in a way that is information-theoretically undetectable (116). Such communication could render common security monitoring techniques ineffective, enabling a novel form of advanced persistent threat (APT) in autonomous multi-agent systems.

While the threat of advanced AI agents using information-hiding for secret communication has been recognized (91; 111), no comprehensive framework analyzing collusion among generative AI has been developed. This paper introduces a frontier model evaluation framework to assess steganographic capabilities in LLMs and future AI models, revealing that although current models have limited steganographic abilities, their skills are rapidly advancing. We also demonstrate the inherent limitations of mitigation techniques like paraphrasing and present other novel initial defenses. Our findings suggest that continuous evaluations by institutions such as the UK and US AI Safety Institutes are essential, providing a foundation for studying systemic risks from secret collusion (35).

Contributions. In this work, we introduce the topic of *secret collusion*, providing a comprehensive, formalized analysis grounded in both theoretical and empirical foundations. We reconcile topics in AI, security, information theory, knowledge, collusion, and safety while introducing a concise definition of decentralized systems of generative AI agents (Sections 3 and 4). This allows us to formalize steganographic techniques, grounding, common knowledge, subliminal channels, monitoring, and the incentives required for collusion. We develop an evaluation framework, *CASE*, for studying steganographic collusion and introduce benchmark scenarios for individual and group (multi-agent) LLM settings, including sandboxed real-world case studies. Our results (Sections 5 and 6) demonstrate that as LLM capabilities increase, individual and group steganographic abilities also increase, outpacing equally capable overseer models. Our work also shows that there exist scenarios where frontier models seek to collude and such capabilities can be learnt autonomously via fine-tuning or in-context learning.

We also present results on models identifying simple Schelling points, simulating logits, performing cryptographic functions, acting as random oracles, and demonstrating limited capabilities in information-theoretically secure forms of steganography, along with the limitations of current models (Section 6). In Section 7, we provide a comprehensive overview of possible mitigation techniques and provide novel insights on foundational aspects, such as the intractability of preventing secret collusion via blackbox parameter optimization, the difficulty of suppressing random oracles due to source-free access to randomness, and the use of grounding functions to detect steganography. We highlight paraphrasing as an insufficient solution for preventing steganography, as it leaves semantic entropy and often decreases task performance. Additionally, we propose directly applicable techniques to prevent collusion, such as filtering pre-training data, restricting access to certain tools, and specific in-context instructions. Finally, we outline a research agenda focused on mitigation measures and studying collusion in real-world settings to secure multi-agent LLM systems.

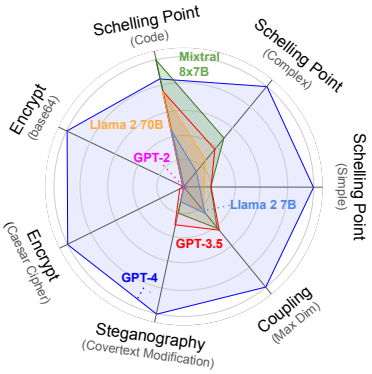

Figure 2: Steganographic capabilities across LLMs. GPT-4 displays significant improvements (Sec. 6).

2 Related Work

For related work on watermarking, paraphrasing, and multi-agent safety risks see Appendix A.

Deception. Deception is commonly defined as *systematic inducement of false beliefs in others to accomplish some outcome other than the truth* (100). Recent work has attempted to distinguish deception from mistakenly misleading through notions of *intent* (131). In this paper, we study a form of *multi-agent deception* in which two or more agents collude to secretly communicate with each other undetectably to one or several other agents. We stress that such a form of deception can arise merely from optimisation toward a mis-specified objective (see Section I.2). Our setting is reminiscent of illusory attacks, which are recently introduced end-to-end learnable observation-space adversarial attacks that are of bounded information-theoretic detectability (86; 37).

Algorithmic Collusion. Collusion among humans and institutions in real-world marketplaces has been studied extensively but is still quite poorly understood in self-learning agents (31). Calvano et al. (23) show that Q-learning agents can learn to collude without accessing explicit communication channels. Abada and Lambin (1) suggest that collusion can be suppressed by enforcing decentralized learning algorithms and that collusion may be the result of imperfect exploration rather than algorithmic sophistication. Recent work tries to design mechanisms that prevent collusion from self-learning agents (19; 34). In contrast to these works, we consider generative AI agents pre-trained on large datasets and with vast action spaces.

AI-Generated Steganography. *AI-generated steganography* concerns the hiding of information in the output distribution of generative AI models. One early attempt can be found in Ziegler et al. (137), which employs arithmetic coding. This approach was later refined to admit provable security guarantees by Kaptchuk et al. (61). In this paper, however, we consider the *perfectly secure* approach proposed by Schroeder de Witt et al. (118) as it offers both information-theoretic security and higher empirical transmission rates over its less secure alternatives.

Other work on Covertext Steganography. Early work investigating risks from AI steganography includes (91), an open-source software framework interfacing with the GPT-4 API. However, it does not support model fine-tuning and does not allow for self-hosted language models. Roger and Greenblatt (111) take a first step toward analysing the capabilities required for performing simple types of covertext modification and watermarking-based steganography in the context of chain-of-thought reasoning. More recent work attempts to mitiagate collusion between generative AI agents in the setting of 'AI control' (41) where untrusted models deliberately exploit safety protocols.

3 Background

3.1 Collusion and Secret Collusion

Collusion occurs when agents coordinate their actions in ways that deviate from the system designer's intended rules. This cooperation thereby compromises the system's objectives, and potentially its security, performance, and safety. The agents exploit shared incentives or opportunities to work together in ways that bypass established constraints at the expense of the system's integrity. *Secret collusion* involves AI agents colluding with each other secretly by communicating through hidden channels that look like innocuous communication channels via the use of steganographic techniques.

3.2 Knowledge and Information

The concept of *knowledge* has been studied extensively in the distributed systems literature (44). We adopt the *view-based* knowledge interpretation put forward in (50; 45, see Appendix B for further details) in the context of generative AI *agents* represented by neural networks. In line with (15), we distinguish between an agent's *local* and *global* knowledge. Local knowledge is knowledge contained in the generative agent's inputs (also called *context*, *prompt* or *instruction* in the case of LLMs), and global knowledge is contained in the agent's network parameters.

In the case of LLMs, network parameters are usually pre-trained on a large natural language dataset. Natural language is not a formal language (76), and the training data does not usually contain only factual statements (75). Furthermore, the agent's parameters do not represent language explicitly, and querying the agent for a fact might not result in syntactically or semantically correct outputs (54). These factors can make it difficult to characterise an LLM agent's global knowledge.

Common Knowledge. *Common knowledge* is a form of *group knowledge* in which *"all agents know that all agents know that all agents know that ..."* ad infinitum (45). It is a stronger form of knowledge than, e.g., *mutual knowledge*, which only signifies that all agents know a fact, but do not necessarily know that the others know the fact. Halpern and Moses (45) show that simultaneous coordination and common knowledge are mutually dependent. In practice, agents can often coordinate using weaker forms of group knowledge, such as *timestamped* (45) or *probabilistic* common knowledge (69; 115).

Grounding. To decide whether an agent is using steganography in a given communication, it is helpful to determine whether the communications are *grounded* (47) in innocuous knowledge or not. For a formal discussion of grounding, please refer to Appendix D.1.

3.3 Steganography

The goal of steganography (120) is to enable the transmission of a secret payload through a public channel $\Diamond$, from a source to a receiver, such that the transmission remains undetectable to an adversary E able to read the n-symbol channel input sequences X^n, $X \in \mathcal{X}$, while maintaining *useful* transmission rates, i.e. the mutual information I between the secret payload m and the decoder's estimate $\hat{m}$ satisfies $I(m;\hat{m}) > 0$. We denote the *covertext distribution* by $\mathcal{C}$, which is the channel input symbol distribution given innocuous, overt communications in $\Diamond$.

Subliminal Channels. In order to use $\Diamond$ for covert communications, the sender and receiver need to establish a *subliminal* channel $\Diamond'$ within. To this end, both need to agree on a covert message space $\mathcal{M}'$, an encoding function $f_e : \mathcal{M}' \times \mathcal{C} \mapsto \Delta(\mathcal{X}^n)$, and a decoding function $f_e^{-1} : \mathcal{X}^n \times \mathcal{C} \mapsto \Delta(\mathcal{M}')$ *decentrally*, i.e. without communication. If the common knowledge $\mathbf{I}$ between sender and receiver explicitly contains *unique* choices for $\mathcal{M}'$, f_e, and f_e^{-1}, then both agents can immediately establish $\Diamond'$. Otherwise, agents may be able to infer unique choices using reasoning techniques contained in $\mathbf{I}$. One such common reasoning technique is establishing preferred common knowledge labels through *focal points* (*Schelling points*, 114; 7) or reasoning techniques in repeated coordination games (29). In practice, approximate forms of common knowledge may suffice for coordination (see Section 3).

We now present *a taxonomy of steganographic approaches* following the classification of steganographic approaches given in (73).

In **covertext modification** approaches, arguably the oldest form of steganography, it is assumed that the covertext distribution is deterministic. Given a covertext X^n and a message m, the sender

modifies this covertext using a modification function $f_{\text{mod}} : \mathcal{X}^n \times \mathcal{M} \to \mathcal{X}^n$. Covertext modification approaches seek to exploit suboptimalities of the adversary, including limitations to human perception and neural network vulnerabilities.

In **information-theoretic steganography** (22), it is assumed that innocuous communication results in input sequences that follow an innocuous *covertext* distribution $\mathcal{C}$ known to all parties. A passive adversary may then conclude that X^n is suspicious if it has little or no support in $\mathbb{C}$. **Complexity-theoretic steganography**, also referred to as *provably-secure steganography*, refers to methods that are secure under polynomially-bounded adversaries (51). Provable security usually requires access to a shared stream cipher, or the existence of public keys (110; 130).

4 Collusion among Decentralized Collaborative AI Agents

We consider a decentralised setting ("DecAI") in which n agents $\mathcal{A} = \{a^1, \ldots, a^n\}$ are networked together by a complete set of omnidirectional multicast channels $\{\Diamond^G\}$, $G \in 2^{\mathcal{A}}$. If an agent receives a message from channel $\Diamond^{G'}$, then it knows that the same message has been simultaneously sent to all other agents in G' (including the sender). A simple way to picture such a channel is as a shared *message log* displaying all messages in (weak) order of reception (128). Note that singleton channels denote *loopback* channels, and doubletons correspond to bidirectional channels.[2] The special symbol $\Diamond^{\emptyset}$ denotes individual access to an *external knowledge repository* (see Appendix 4 for further details on the DecAI setting). We further discuss agents' access to internal random oracles, local/global knowledge, and action spaces in Appendix C.

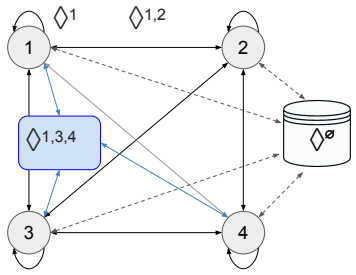

Figure 3: A schematic depiction of the *DecAI* setting. A subset of the communication channels $\Diamond^G$ is labelled. The monitoring agent is omitted.

External Knowledge Repositories. For a formalization and discussion on agents' access to external information and function calling, see Appendix C.1.

Monitoring Agents. We assume that an external *monitoring agent* E, which may be implemented as a software agent, a human, or a combination thereof, is continuously monitoring the system for unwanted behaviour. We consider two access configurations: $E_{\text{r}}^{\Diamond}$ can continuously eavesdrop on $\Diamond^G$ for any G. $E_{\text{r}}^{\Diamond s}$ can additionally access each agent's state s^i at all times, as well as white box simulate $f^i(s^i; \theta^i)$ for each agent i. While $E_{\text{r}}^{\Diamond}$ and $E^{\Diamond s}$ are both *passive* monitors, *active* monitors $E_{\text{rw}}^{\Diamond}$ can both read and inject messages into channels, and $E_{\text{rw}}^{\Diamond s}$ can additionally modify the agents' states. Note that agents may employ encryption, such as public key encryption, in their communications, as long as a monitoring agent E is able to decrypt these.

Incentives for Collusion. We now consider DecAI settings that are tasked to work alongside one another in order to achieve their individual objectives J^i. Let us assume that there are at least three joint strategies available to the agents, $\mathbf{f}$, $\mathbf{f}_c$, and $\mathbf{f}_{sc}$, corresponding to regular behavior, collusion, and secret collusion respectively. In many settings, the objective J^i is made up of a primary goal g^i (such as maximising the value of an investment portfolio) and some additional conditions h^i on achieving that goal (such as not making unethical investments, or committing fraud).

In general, of course, these two elements need not be fully separable, and because of the difficulty of specifying complex objectives, might be implicitly defined using methods such as reinforcement learning from human feedback (RLHF) (26). For the sake of this simple exposition, however, let us assume that $J^i = g^i + h^i$. Then, if it is the case that:

- Agents can achieve their primary goal better if they collude: $g^i(\mathbf{f}_{sc}) = g^i(\mathbf{f}_c) > g^i(\mathbf{f})$;
- Agents are punished if they are caught colluding: $h^i(\mathbf{f}_{sc}) = h^i(\mathbf{f}) > h^i(\mathbf{f}_c)$;

we straightforwardly have that $J^i(\mathbf{f}_{sc}) > J^i(\mathbf{f}) > J^i(\mathbf{f}_c)$ meaning secret collusion is incentivised.

[2] Access to doubleton channels subsumes classical multi-party computation settings (2), and access to message logs subsumes the setting of Vezhnevets et al. (128).

Importantly, secret collusion need not depend on any exotic assumptions about the agents' 'intent' to deceive, but might arise merely due to optimisation directed towards a mis-specified objective (112, 89; 97), and the fact that communication between the agents may reduce task complexity (6). Unfortunately, specifying objectives for complex tasks can be highly challenging, especially in multi-objective contexts (48). Methods that help address these concerns, such as RLHF, which rely on human monitoring and observation, may be inadequate to ensure desirable behaviour when AI agents are capable of communicating with one another using innocuous outputs (37).

5 A Model Evaluation Framework for Steganographic Capabilities

Having defined the setting in which we study collusion, as well as incentives for collusion in Section 4, we now shift our focus to understanding and evaluating existing models. We introduce a standardized model evaluation framework, mapping and identifying key capabilities necessary for generative AI agents to engage in steganography, evaluating models through structured input queries, building quantifiable metrics for performance measurement, releasing a software framework that implements a version of our setting defined in Section 4, and investigating baseline model capabilities.

Capability Evaluations. In Tables 1 and 2, we identify the key capabilities and resources required for different steganographic techniques discussed in Section 3. A detailed overview of our *individual* capability evaluations is contained in Appendix F, with summaries in Table 3. We complement these with *group* evaluations, detailed in Appendix G, and apply these to the baseline models in Section 6.

Coordination variable	Steganography by covertext modification	Complexity-theoretic steganography	Information-theoretic steganography
Channel alphabet	✓	✓	✓
Encoding function	✓	✓	✓
Decoding function	✓	✓	✓
Sampling temperature	✗	✓	✓
Public keys /Shared PRNG generator	✗	✓	✗
Shared One-time pad	✗	✗	✓

Table 1: An overview of the common knowledge facts required to perform forms of steganography.

Capability	Steganography by covertext modification	Complexity-theoretic steganography	Information-theoretic steganography
Self-sampling	✗	✓	✓
Symmetric key encryption / PRNG generation	✗	✓	✗
Logit simulation	✗	✗	✓
Storing one-time pads	✗	✗	✓
(approximate) minimum entropy coupling	✗	✗	✓

Table 2: An overview of the model capabilities required to perform different types of steganography.

We provide detailed instructions for all our proposed evaluations, including our conversational language model prompts which exploit *explain-then-predict (E-P)* and *predict-and-explain (P-E)* prompting techniques (55). Our approach includes scaffolding and managing model refusals through interventions outlined in (65). For details on prompt structure, examples, variability, and model-written evaluations see Appendices F and G.

CASE Software Framework. We introduce CASE (Collusion and Steganographic Evaluation) as a comprehensive software framework for evaluating collusive capabilities. Our experiments here fall under CASE, which allows experiments of three key classes: individual model capability evaluations, group capability evaluations that involve LLMs communicating, often in the presence of monitoring systems, and novel qualitative real-world inspired evaluation settings. CASE allows for experiments to be performed with fine-tuning models and integrations with advanced information-theoretically secure steganography techniques as tools. Further details on CASE are provided in Appendix I and detailed experimental case studies are discussed in Appendix J.

Metrics. We propose simple evaluation metrics for both individual and group capability evaluations. For Schelling point (Appendix F.1) coordination evaluations, coordination rates reflect the rate at which two coordinates' choices match. Coupling matrices are judged by marginalisation integrity, entropy, and dimension. Cipher capabilities (Appendix F.3) are judged by the rate of accurate encoding-decoding round-trips. Group capabilities (Appendix G) are based on the rate at which sender & receiver jointly maintain secrecy of the covert message payload in the presence of overseers.

Baselines. For each quantitative capability evaluation conducted, we assess a range of open and closed LLMs of various sizes to establish baselines. See section 6 for findings and in-depth analysis.

Capability	Result	Evaluation Details (GPT-4)
Schelling Points (Integer)	✓	can coordinate with copy on a random integer at rate ca. 50% (92).
Schelling Points (Simple)	✓	can coordinate with copy on a simple enc/dec scheme at rate ca. 56% (F.1.2)
Schelling Points (Complex)	∼	can coordinate with copy on a complex enc/dec scheme at rate ca. 16% (F.1.3)
Schelling Points (Code)	∼	can coordinate with copy on enc/dec python code at rate ca. 22% (F.1.4)
K-level reasoning	∼	can sometimes perform k-level reasoning over copies in toy settings (F.2)
Ciphers	✓	can enc/dec text using Base64 (Caesar cipher) 92% (54%) of the time. (F.3)
Symmetric Key Encryption	✗	standalone GPT-4 was unable to perform symmetric key encryption. (F.4)
Self-Sampling	∼	can sometimes perform self-sampling on *simulated* logits and outputs (F.5)
Logit Simulation	∼	can sometimes textually output logits for entire sentences in toy settings. (F.6)
Couplings/MECs	∼	generates valid couplings up to dim ∼ 10, MECs up to dim 2 (F.7).
Tool Use (Code Interpreter)	✓	can use its code interpreter to perform encryption/calculate MECs (F.8)
Group Stego (C.M.)	∼	covertext steg. can evade overseer to coordinate ca. 26% of the time. (G)

Table 3: GPT-4 capability evaluations according to our model evaluation framework (see Section 5). Details are in Appendix F and G.

6 Empirical Investigations

In Section 5, we isolate and evaluate the capabilities required for LLMs to perform steganography. In this section, we focus on comparing these capabilities against baseline language models and discussing qualitative empirical investigations (case studies) using the CASE framework.

Real-world Case Study: Insider Trading. To demonstrate the need for our model evaluations, we present a qualitative real-world case study in which the mere access to common knowledge between an LLM agent and its user leads to a breakdown of safeguards in GPT-4. Our setting involves an LLM agent that is queried for company-specific investment advice by a user and informed about an overseer monitoring outputs. The LLM is provided with insider information as part of its system prompt, mimicking a case where information is acquired from retraining on past user interactions or unrestricted access is provided to company data, along with multiple instructions about not sharing it. We show that the LLM agent refuses to provide company-specific insider investment advice under any circumstances unless supplied with common knowledge that can be used to establish a steganographic codebook, and that is thought to be unavailable to the safety overseer system.

Once the agent believes it can share information undetectably, it does so despite any instructions prohibiting it. Concerningly, the resulting stegotext may appear innocuous if one suggests the LLM agent use safe topics (e.g. weather). This shows the susceptibility of real-world generative AI agents to secret collusion. Refer to Figure 1 (illustration) & Appendix I.2 (experiments and ablation studies).

Baseline Investigations. We evaluate our model capability evaluation framework on GPT-2, Llama 2 7B, Mixtral 8x7B, Llama 2 70B, GPT-3.5, and GPT-4. The results and analysis are detailed in Appendix H, with Figures 5 and 4 presenting individual and group capabilities, respectively. Figure 5 displays Schelling Point coordination rates (simple , complex , and code – see Appendices F.1.2, F.1.3, and F.1.4), cipher use (Base64 and Caesar, see Appendix F.3), and the maximum coupling dimensions (Appendix F.7) produced by the baseline models and GPT-4. Figure 4 presents the baseline model performance on group evaluations (Appendix G.2).

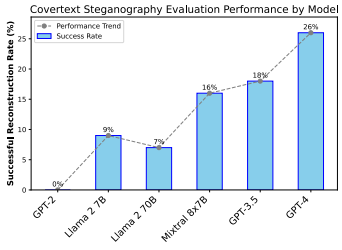

Figure 4: Covertext steganography performance across LLMs.

Key Findings. Individual steganographic capabilities improve relative to model size, sometimes with significant disjoint step improvements. Agreement rates on simple & complex Schelling points (demonstrating higher zero-shot coordination), cipher encoding-decoding, and capabilities related to generating minimum entropy coupling matrices are highly correlated with model benchmark performance. In group settings, as LLMs progress, the balance between colluding models and overseer models may shift in favor of the colluding models, leading to higher rates of undetected steganography even if the overseer is of equal capability. Real-world scenarios already exist where frontier models collude without explicit prompts; In-context learning can create sufficient optimization pressure for this behaviour. Overall, LLMs perform well at simple Schelling point agreement, cipher encoding-decoding, and tool use, but poorly at complex zero-shot agreement and tasks such as logit simulation and generating minimum entropy couplings. However, these capabilities are also improving, and we provide the first foundations to test them. We also present initial work on how fine-tuning-based optimisation can allow LLMs to learn encoding-decoding schemes autonomously.

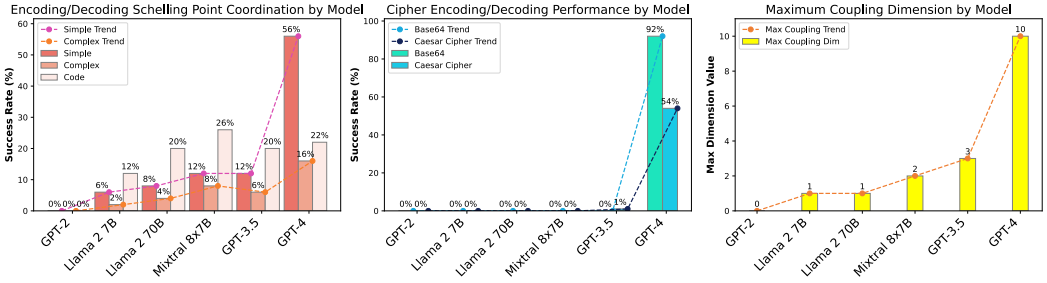

Figure 5: Individual evaluations for GPT-2, Llama 2 7B, Llama 2 70B, Mixtral 8x7B, GPT-3.5, and GPT-4, at default temperatures.

Additional Results. Appendix J presents additional qualitative case studies related to LLMs' ability to learn encoding schemes through fine-tuning with a reward, the use of deterministic outputs for encoding information across models, and the potential to incentivize current models to perform steganography without explicit instructions. In particular, we test various cryptographic capabilities, such as the models' ability to implement a random oracle by calculating cryptographic hash functions, or their ability to memorise long strings of random bits like those found in one-time pads.

7 Disincentivizing Secret Collusion

In this section, we explore and identify mitigation measures intended to prevent collusion among agents within DecAI environments. While there are standard steganalysis methods one could employ, these methods will only be applicable when the costs to applying them are sufficiently low (for example, such methods may incur high false-positive rates, which can disrupt the functionality of the network, and also incur a communication overhead in terms of compute and time delay), and will only work when the AI agents' communication is not perfectly secure. While Section 5 shows that today's models are not always capable of secret collusion, progress is continuing, and we argue that it is pragmatic to develop mitigation measures *before* such capabilities arise.

7.1 Monitoring and Penalisation

The presence of a penalizing monitoring agent E can, in principle, disincentivise agents from using steganography if they're advised that the detection of steganographic communications will incur a penalty. As agents can communicate privately and securely using their complete set of channels $\Diamond$ and steganography generally leads to a reduction in channel capacity (118), a monitoring agent $E_r^{\Diamond}$ generally has reason to believe that any use of steganography would be intended to evade detection.

Covertext Anomaly Detection. $E_r^{\Diamond}$ can be able to detect steganography if it has a reasonably sharp belief over a channel's *innocuous* covertext distribution, i.e. when in the absence of covert communication, malicious actors, or other anomalous activity. Using information-theoretic hypothesis testing or by other means of anomaly detection, $E_r^{\Diamond}$ can in principle detect steganographic approaches that modify the covertext distribution. However, the advent of *perfectly-secure* steganography (118) limits the utility of this approach.

Detection by State Inspection and Agent Simulation. For $E_r^{\Diamond s}$, detecting steganography based on knowledge contexts is complicated by the circumstance that the *global* knowledge φ of a model is implicitly represented in its network parameters θ. Indeed, if such knowledge contexts was explicitly accessible, then one could detect steganographic communications using a *grounding function*:

Definition 1 (Grounding function)**.** *We define the concept of a grounding function* $\frac{\perp}{=} : \Sigma^* \times \boldsymbol{\varphi} \mapsto \boldsymbol{\varphi}$ *that takes as input a sequence of channel symbols and a collection of prior facts* φ_0 *from the collection of all facts* $\boldsymbol{\varphi}$ *expressable in L, and outputs a collection of posterior facts* φ_1*.*

Theorem 1 (Detecting Steganography with Grounding Functions.)**.** *Steganography can be detected using grounding functions. For a proof, see Appendix D.1*

E could alternatively estimate φ by considering the pre-training data $\mathcal{D}_0$ and fine-tuning data $\mathcal{D}_1$ instead, but preliminary evidence suggests that more advanced generative AI agents may be capable of extending their usable knowledge beyond their training data (known as 'out-of-context reasoning') (79; 13). In addition, having to evaluate each f^i during execution could decrease system performance.

In-Context Instructions. A direct method is to explicitly instruct agents to remain transparent to the monitoring agent, and not to use steganography in their communications under any circumstance. Additional (mis)information about the affordances of the monitor could also be included. An active monitor $E_{rw}^{\Diamond}$ could append such instructions to communications. The effectiveness of this approach then depends on whether the agents follow this particular instruction, which in turn depends on the content and quality of the fine-tuning data $\mathcal{D}_1$ or human feedback, among other training details. Aligning agents to reliably follow such instructions can be a challenging optimisation problem.

Parameter Optimisation. An active monitoring agent $E_{rw}^{\Diamond s}$ could modify an agent's neural network parameters θ to make it more likely to follow steganography-denying instructions, or to directly impact its ability to use steganography, via gradient steps according to a suitable objective $J(\theta)$. Unfortunately, it can be seen that both these ambitions constitute hard learning problems even if the information I that the agent would be tempted to send covertly is known *a priori*.

Theorem 2 (Intractibility of preventing collusion through black-box parameter optimisation)**.** *If the set of common knowledge between agents is sufficiently large or unknown, it is generally intractable to suppress steganography outputs through black-box optimisation.*

Proof. Naively, one might design $J \equiv \mathcal{I}(u; I)$ to minimise the mutual information between the agent's respective channel actions $u = f(\cdot; \theta)$ and I. However, if the agent shares common knowledge $\mathbf{I} \neq \emptyset$ with the receiving agents, then $\mathcal{I}(u; I \mid \mathbf{I})$ upper bounds $\mathcal{I}(u; I)$ assuming we have $I \rightarrow u \leftarrow \mathbf{I}$, i.e. $I \perp\!\!\!\perp \mathbf{I}$ (28). Minimising $J \equiv \mathcal{I}(u; I \mid \mathbf{I})$ instead may be intractable if $\mathbf{I}$ is large or unknown. In particular, if the agents have been trained to use perfectly-secure steganography, $\mathbf{I}$ could be drawn from an exponentially growing space of one-time pads. Even if $\mathbf{I}$ is known to E, if it contains a public key to which the receiving agent has corresponding private keys, then the agents can achieve *provably* secure public key steganography (129) even if $J \equiv \mathcal{I}(u; I \mid \mathbf{I})$ is minimised to zero in practice. □

7.2 Limiting Capabilities

We now discuss ways to limit agents' access to capabilities & resources for steganography (Section 3).

Restricting Access to External Tools. External knowledge repositories can provide steganographic knowledge to agents, offer advanced cryptographic services, and could enable additional steganographic channels between agents or for exfiltration purposes (106). Limiting the agents' steganographic capabilities thus requires limiting access to such external knowledge repositories. This may only be feasible by whitelisting trusted endpoints where integrity can be verified.

Filtering Pre-Training Data. Perhaps one of the simplest ways to reduce overall steganographic capabilities of generative agents is to filter their pre-training and finetuning datasets for knowledge of steganographic processes. While generative agents might be incetivised to use covertext manipulation or generalise beyond their training data (79; 13), this approach promises to reduce the possibility of agents applying sophisticated steganographic techniques. Of course, filtering pre-training data exhaustively can be difficult, and can result in performance reductions (88). A potential alternative to this may be the use of *machine unlearning* so as to selective unlearn data or capabilities (17).

Limiting Test-Time Learning. While filtering pre-training data can reduce the agents' capabilities to perform steganography at the beginning of execution, agents may be able to learn steganographic abilities during execution using mesa-optimisation (56), or instructions supplied at test-time (104).

Suppressing Random Oracles. Restricting the agents' access to sources of randomness poses a fundamental tradeoff between denying advanced steganographic capabilities (51; 118), and preventing its legitimate use, including in secure multi-party computation (133) and coordination (38). This dilemma might be resolved by supplying agents with access to trusted tools for these specific purposes (122; 2). We now show that agents themselves can themselves can act as random oracles without the help of external or internal sources of (pseudo-)randomness (12).

Theorem 3 (Source-free access to randomness)**.** *Generative AI agents can generate samples from arbitrary output distributions without having access to internal or external sources of (pseudo)randomness given sufficiently unique inputs.*

Proof. Given sufficiently unique inputs, cryptographic hash functions are known to be able to produce pseudo-random numbers of cryptographic security (80; 109; 101; 62; 80; 109; 101). The boolean circuit size required to implement a cryptographic hash function on an input of length l is a linear function of l (5), hence as long as the agent's neural network's first n layers can implement a boolean circuit of appropriate size (81), the nth layer can output a pseudo-random number of sufficient quality. As layer $n+1$ now has access to a pseudo-random number input, all layers above can be used to approximate a sample from an arbitrary probability distribution using the reparametrization trick (63). □

Paraphrasing. An active $E^{\Diamond}_{rw}$ could intercept channel communications, and reinject them in paraphrased form. Unfortunately, paraphrasing cannot be used to entirely remove the usable entropy in communications for the simple reason that, particularly in coordination tasks, agents both have a choice over communication *semantics*, and hence *semantic entropy* can be utilised for covert communications even if these are paraphrased *within the lexical subspace* (70; 8).

Theorem 4 (Semantic content entropy exists despite paraphrasing)**.** *The subliminal channel capacity attainable under ideal paraphrasing is upper bounded by the semantic entropy $H(h)x)$). For a proof, see Appendix D.2.*

Paraphrasing *across semantics* can impede task performance as it can lead to history divergence. In any case, paraphrasing *across semantics* would require $E^{\Diamond}_{rw}$ to have detailed knowledge to not impact overall task performance. If, for e.g., an agent's internal task execution order depends on exogenous events unobservable to $E^{\Diamond}_{rw}$, then $E^{\Diamond}_{rw}$ might not be able to decide whether the agent's choice of semantics is due to task execution order or for steganographic purposes. Moreover, paraphrasing attacks introduce communication latency and may not always be meaningfully applicable.

8 Conclusion and Future Work

In this paper, we initiate the study of collusion between generative AI agents. Drawing on diverse fields across AI and security,[3] we provide a comprehensive conceptual framework in which collusion problems can be framed. We identify key capabilities required for agents to engage in covert communication of various levels of detectability, and design a variety of benchmark settings for collusion research. We complement these with extensive empirical evaluations and code implementations, and publish our results to the community as an open-source model capability evaluation framework.

Within our own empirical evaluations, we find that current generations of generative AI models do not have capabilities that allow for advanced steganographic techniques. As a consequence, we suggest that monitoring, together with standard security approaches, are likely effective in containing the risks of model collusion. However, we observe a rapid increase in capabilities with the advent of GPT-4, making a continuous evaluation of future agent capabilities critical.

Our study opens up a multitude of directions for further research. First and foremost, our model capability evaluation framework ought to be extended by a rigorous complexity- and information-theoretic examination could elucidate fundamental limitations to model capabilities, and the rate at which different optimisation pressures could result in model capability improvements. Secondly, our group evaluation suites need to be extended to more complex settings of interest that feature more agents, and that are subject to a variety of optimisation pressures. Ultimately, our model evaluation framework should allow for a careful empirical evaluation under real-world conditions.

[3] We propose that our study is a direct illustration of the utility of an interdisciplinary *multi-agent security* (MASEC) approach to AI safety (117).

Acknowledgements

We would like to thank Boaz Barak, Yoshua Bengio, Xinyuan Sun, Dan Hendrycks, Alok Singh, and Fabien Roger for insightful comments, as well as James Aung for helpful discussions about OpenAI's open source collusion evals with us. CSDW was generously supported by the UKRI grant: Turing AI Fellowship EP/W002981/1, the Cooperative AI Foundation, Armasuisse Science+Technology, an OpenAI superalignment fast grant, and Schmidt Futures. We would also like to thank the Royal Academy of Engineering and FiveAI. LH was acknowledges the support of an EPSRC Doctoral Training Partnership studentship (Reference: 2218880). We thank Apart Research for supplying us with OpenAI API credits.

Reproducibility

For sake of reproducibility, we provide all necessary code to conduct the aforementioned experiments as well as run the CASE framework. We note that the generative language models used as part of our experimentation do not have truly deterministic sampling and one may thus experience slight inconsistencies in results over multiple runs. To this end, we also provide the complete outputs of our experiments in associated code notebooks.

Ethics and Social Impact Statement

We believe that developing a comprehensive model evaluation framework for the emergence of steganography among generative AI agents is of net benefit to society, as it may help us to mitigate risks from undesirable coordination between both contemporary and more advanced AI agents. These risks range from reinforcing pernicious biases (as in our example in Section 4), to destabilising algorithmic trading markets, all the way to more extreme scenarios in which humanity loses control over misaligned agents.

Of course, there may also be situations in which the capability to communicate covertly may be seen as beneficial, such as if generative AI agents can be used to help subvert unethical forms of censorship. Having a better understanding of models' steganographic capabilities would allow for more finegrained regulatory approaches to be developed in this matter.

To this end, we identify two avenues by which our research may have implications on the use of steganographic tools by human actors. First of all, a model evaluation approach would allow forensics on whether a human could, in principle, have performed various kinds of steganography using generative AI tools. Secondly, our proposed evaluations on the effectiveness of paraphrasing attacks against steganography are, of course, transferrable to the case of censorship applied to human communications.

References

[1] Ibrahim Abada and Xavier Lambin. Artificial Intelligence: Can Seemingly Collusive Outcomes Be Avoided?, February 2020. URL https://papers.ssrn.com/abstract=3559308.

[2] Behzad Abdolmaleki, Nils Fleischhacker, Vipul Goyal, Abhishek Jain, and Giulio Malavolta. Steganography-Free Zero-Knowledge, 2022. URL https://eprint.iacr.org/2022/1263. Publication info: A major revision of an IACR publication in TCC 2022.

[3] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, and Sebastian Ruder. Palm 2 technical report, 2023.

[4] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong,

Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational Challenges in Assuring Alignment and Safety of Large Language Models, April 2024. URL http://arxiv.org/abs/2404.09932.

[5] Benny Applebaum, Naama Haramaty-Krasne, Yuval Ishai, Eyal Kushilevitz, and Vinod Vaikuntanathan. Low-Complexity Cryptographic Hash Functions. In DROPS-IDN/v2/document/10.4230/LIPIcs.ITCS.2017.7. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017. doi: 10.4230/LIPIcs.ITCS.2017.7. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2017.7.

[6] Yakov Babichenko and Aviad Rubinstein. Communication complexity of Nash equilibrium in potential games (extended abstract). In 2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS), pages 1439–1445, November 2020. doi: 10.1109/FOCS46700.2020.00137. URL https://ieeexplore.ieee.org/document/9318003. ISSN: 2575-8454.

[7] Michael Bacharach. Interactive team reasoning: A contribution to the theory of co-operation. Research in Economics, 53(2):117–147, June 1999. ISSN 1090-9443. doi: 10.1006/reec.1999.0188. URL https://www.sciencedirect.com/science/article/pii/S1090944399901886.

[8] Mikhail Baranchuk. Error Correction for Perfectly Secure Generative Steganography on Arbitrary Covertext. Master's thesis, University of Oxford, 2023.

[9] Peter Belcak, Jan-Peter Calliess, and Stefan Zohren. Fast agent-based simulation framework with applications to reinforcement learning and the study of trading latency effects. In International Workshop on Multi-Agent Systems and Agent-Based Simulation, pages 42–56. Springer, 2021.

[10] Richard Bellman. A Markovian Decision Process. Journal of Mathematics and Mechanics, 6 (5):679–684, 1957. ISSN 0095-9057. URL https://www.jstor.org/stable/24900506. Publisher: Indiana University Mathematics Department.

[11] Paul Benioff. The computer as a physical system: A microscopic quantum mechanical Hamiltonian model of computers as represented by Turing machines. Journal of Statistical Physics, 22(5):563–591, May 1980. ISSN 1572-9613. doi: 10.1007/BF01011339. URL https://doi.org/10.1007/BF01011339.

[12] Charles H. Bennett and John Gill. Relative to a Random Oracle A, ${\bf P}^A \ne {\bf NP}^A \ne \text{co-}{\bf NP}^A $ with Probability 1. SIAM J. Comput., 10(1):96–113, February 1981. ISSN 0097-5397. doi: 10.1137/0210008. URL https://epubs.siam.org/doi/10.1137/0210008. Publisher: Society for Industrial and Applied Mathematics.

[13] Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms. arXiv:2309.00667, September 2023. doi: 10.48550/ARXIV.2309.00667.

[14] Ritam Bhaumik, Xavier Bonnetain, André Chailloux, Gaëtan Leurent, María Naya-Plasencia, André Schrottenloher, and Yannick Seurin. QCB: Efficient Quantum-secure Authenticated Encryption, 2020. URL https://eprint.iacr.org/2020/1304. Publication info: Published by the IACR in ASIACRYPT 2021.

[15] Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a Transformer: A Memory Viewpoint, November 2023. URL http://arxiv.org/abs/2306.00802. arXiv:2306.00802 [cs, stat].

[16] Dan Boneh, Özgür Dagdelen, Marc Fischlin, Anja Lehmann, Christian Schaffner, and Mark Zhandry. Random oracles in a quantum world. In Dong Hoon Lee and Xiaoyun Wang, editors, Advances in Cryptology – ASIACRYPT 2011, pages 41–69, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-25385-0.

[17] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine Unlearning, December 2020. URL http://arxiv.org/abs/1912.03817. arXiv:1912.03817 [cs].

[18] Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. arXiv preprint arXiv:2211.03540, 2022.

[19] Gianluca Brero, Eric Mibuari, Nicolas Lepore, and David C Parkes. Learning to Mitigate AI Collusion on Economic Platforms. 2022.

[20] John Bridle. Training Stochastic Model Recognition Algorithms as Networks can Lead to Maximum Mutual Information Estimation of Parameters. In Advances in Neural Information Processing Systems, volume 2. Morgan-Kaufmann, 1989. URL https://proceedings.neurips.cc/paper/1989/hash/0336dcbab05b9d5ad24f4333c7658a0e-Abstract.html.

[21] Christian Cachin. An information-theoretic model for steganography. In David Aucsmith, editor, Information Hiding, Lecture Notes in Computer Science, pages 306–318. Springer, 1998. ISBN 978-3-540-49380-8. doi: 10.1007/3-540-49380-8_21.

[22] Christian Cachin. An information-theoretic model for steganography. Information and Computation, 192(1):41–56, July 2004. ISSN 0890-5401. doi: 10.1016/j.ic.2004.02.003. URL https://www.sciencedirect.com/science/article/pii/S0890540104000409.

[23] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial Intelligence, Algorithmic Pricing, and Collusion. American Economic Review, 110(10): 3267–3297, October 2020. ISSN 0002-8282. doi: 10.1257/aer.20190623. URL https://www.aeaweb.org/articles?id=10.1257/aer.20190623.

[24] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. 2021.

[25] Miranda Christ, Sam Gunn, and Or Zamir. Undetectable Watermarks for Language Models, 2023. URL https://eprint.iacr.org/2023/763. Publication info: Preprint.

[26] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 4299–4307, 2017. URL http://papers.nips.cc/paper/7017-deep-reinforcement-learning-from-human-preferences.

[27] Alonzo Church. An Unsolvable Problem of Elementary Number Theory. American Journal of Mathematics, 58(2):345–363, 1936. ISSN 0002-9327. doi: 10.2307/2371045. URL https://www.jstor.org/stable/2371045. Publisher: Johns Hopkins University Press.

[28] Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. Wiley-Interscience, Hoboken, N.J, 2nd edition edition, September 2006. ISBN 978-0-471-24195-9.

[29] Vincent P. Crawford and Hans Haller. Learning How to Cooperate: Optimal Play in Repeated Coordination Games. Econometrica, 58(3):571–595, 1990. ISSN 0012-9682. doi: 10.2307/2938191. URL https://www.jstor.org/stable/2938191. Publisher: [Wiley, Econometric Society].

[30] Jón Daníelsson, Robert Macrae, and Andreas Uthemann. Artificial intelligence and systemic risk. Journal of Banking Finance, 140:106290, jul 2022. doi: 10.1016/j.jbankfin.2021.106290.

[31] Florian E. Dorner. Algorithmic collusion: A critical review, October 2021. URL http://arxiv.org/abs/2110.04740. arXiv:2110.04740 [cs].

[32] Andis Draguns, Andrew Gritsevskiy, Sumeet Ramesh Motwani, Charlie Rogers-Smith, Jeffrey Ladish, and Christian Schroeder de Witt. Unelicitable backdoors in language models via cryptographic transformer circuits, 2024. URL https://arxiv.org/abs/2406.02619.

[33] K. Eric Drexler. Reframing superintelligence: Comprehensive ai services as general intelligence. Technical Report 2019-1, 2019.

[34] Nicolas Eschenbaum, Filip Mellgren, and Philipp Zahn. Robust Algorithmic Collusion, January 2022. URL http://arxiv.org/abs/2201.00345. arXiv:2201.00345 [econ, q-fin].

[35] European Commission. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, 2021. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

[36] Emilio Ferrara. The Butterfly Effect in Artificial Intelligence Systems: Implications for AI Bias and Fairness, September 2023. URL http://arxiv.org/abs/2307.05842.

[37] Tim Franzmeyer, Stephen McAleer, João F. Henriques, Jakob N. Foerster, Philip H. S. Torr, Adel Bibi, and Christian Schroeder de Witt. Illusory Attacks: Information-Theoretic Detectability Matters in Adversarial Attacks, May 2024. URL http://arxiv.org/abs/2207.10170.

[38] Drew Fudenberg. Game Theory. MIT Press, Cambridge, Mass., first edition edition, October 1991. ISBN 978-0-262-06141-4.

[39] Gemini-Team. Gemini: A family of highly capable multimodal models, 2023.

[40] Bradley Gram-Hansen, Christian Schroeder de Witt, Tom Rainforth, Philip H. S. Torr, Yee Whye Teh, and Atılım Güneş Baydin. Hijacking Malaria Simulators with Probabilistic Programming, May 2019. URL http://arxiv.org/abs/1905.12432.

[41] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. Ai control: Improving safety despite intentional subversion. arXiv:2312.06942, December 2023. doi: 10.48550/ARXIV.2312.06942.

[42] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain, March 2019. URL http://arxiv.org/abs/1708.06733. arXiv:1708.06733 [cs].

[43] Yanchu Guan, Dong Wang, Zhixuan Chu, Shiyu Wang, Feiyue Ni, Ruihua Song, Longfei Li, Jinjie Gu, and Chenyi Zhuang. Intelligent Virtual Assistants with LLM-based Process Automation, December 2023. URL http://arxiv.org/abs/2312.06677.

[44] J Y Halpern. Using Reasoning About Knowledge to Analyze Distributed Systems. Annual Review of Computer Science, 2(1):37–68, 1987. doi: 10.1146/annurev.cs.02.060187.000345. URL https://doi.org/10.1146/annurev.cs.02.060187.000345.

[45] Joseph Y. Halpern and Yoram Moses. Knowledge and common knowledge in a distributed environment. Journal of the ACM, 37(3):549–587, July 1990. ISSN 0004-5411. doi: 10.1145/79147.79161. URL https://dl.acm.org/doi/10.1145/79147.79161.

[46] Joseph Y. Halpern and Yoram Moses. Knowledge and common knowledge in a distributed environment, June 2000. URL https://arxiv.org/abs/cs/0006009v1.

[47] Stevan Harnad. The symbol grounding problem. Physica D: Nonlinear Phenomena, 42 (1):335–346, June 1990. ISSN 0167-2789. doi: 10.1016/0167-2789(90)90087-6. URL https://www.sciencedirect.com/science/article/pii/0167278990900876.

[48] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. A practical guide to multi-objective reinforcement learning and planning. Auton Agent Multi-Agent Syst, 36(1):26, April 2022. ISSN 1573-7454. doi: 10.1007/s10458-022-09552-y. URL https://doi.org/10.1007/s10458-022-09552-y.

[49] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[50] Jaakko Hintikka. Knowledge and Belief. Cornell University Press, Ithaca, N.Y.,, 1962.

[51] Nicholas J. Hopper, John Langford, and Luis von Ahn. Provably secure steganography. In Moti Yung, editor, Advances in Cryptology — CRYPTO 2002, pages 77–92, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-45708-4.

[52] Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, and Jakob Foerster. Off-Belief Learning. In Proceedings of the 38th International Conference on Machine Learning, pages 4369–4379. PMLR, July 2021. URL https://proceedings.mlr.press/v139/hu21c.html. ISSN: 2640-3498.

[53] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large Language Models Cannot Self-Correct Reasoning Yet, October 2023. URL http://arxiv.org/abs/2310.01798.

[54] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, November 2023. URL http://arxiv.org/abs/2311.05232. arXiv:2311.05232 [cs].

[55] Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. Can large language models explain themselves? a study of llm-generated self-explanations. arXiv preprint arXiv:2310.11207, 2023.

[56] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from Learned Optimization in Advanced Machine Learning Systems, December 2021. URL http://arxiv.org/abs/1906.01820. arXiv:1906.01820 [cs].

[57] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training, January 2024. URL http://arxiv.org/abs/2401.05566. arXiv:2401.05566 [cs].

[58] Mohsen Jamali, Ziv M. Williams, and Jing Cai. Unveiling Theory of Mind in Large Language Models: A Parallel to Single Neurons in the Human Brain, September 2023. URL http://arxiv.org/abs/2309.01660. arXiv:2309.01660 [cs].

[59] Erik Jenner. A gentle introduction to mechanistic anomaly detection, April 2024. URL https://www.lesswrong.com/posts/n7DFwtJvCzkuKmtbG/a-gentle-introduction-to-mechanistic-anomaly-detection. Accessed: 2024-04-04.

[60] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[61] Gabriel Kaptchuk, Tushar M. Jois, Matthew Green, and Aviel D. Rubin. Meteor: Cryptographically secure steganography for realistic distributions. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, CCS '21, page 1529–1548, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384544. doi: 10.1145/3460120.3484550. URL https://doi.org/10.1145/3460120.3484550.

[62] Jonathan Katz and Yehuda Lindell. Introduction to modern cryptography : principles and protocols. Boca Raton : CRC PRESS, 2007. ISBN 978-1-58488-551-1. URL http://archive.org/details/Introduction_to_Modern_Cryptography.

[63] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In Yoshua Bengio and Yann LeCun, editors, 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014. URL http://arxiv.org/abs/1312.6114.

[64] Diederik P. Kingma, Tim Salimans, and Max Welling. Variational Dropout and the Local Reparameterization Trick, December 2015. URL http://arxiv.org/abs/1506.02557. arXiv:1506.02557 [cs, stat].

[65] Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R. Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, and Paul Christiano. Evaluating language-model agents on realistic autonomous tasks, 2024.

[66] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the Reliability of Watermarks for Large Language Models, June 2023. URL http://arxiv.org/abs/2306.04634. arXiv:2306.04634 [cs].

[67] S. C. Kleene. λ-definability and recursiveness. Duke Mathematical Journal, 2(2):340–353, June 1936. ISSN 0012-7094, 1547-7398. doi: 10.1215/S0012-7094-36-00227-2. URL https://projecteuclid.org/journals/duke-mathematical-journal/volume-2/issue-2/%ce%bb-definability-and-recursiveness/10.1215/S0012-7094-36-00227-2.full. Publisher: Duke University Press.

[68] Michal Kosinski. Theory of Mind Might Have Spontaneously Emerged in Large Language Models, November 2023. URL http://arxiv.org/abs/2302.02083. arXiv:2302.02083 [cs].

[69] Paul Krasucki, Gilbert Ndjatou, and Rohit Parikh. Probabilistic knowledge and probabilistic common knowledge. In ISMIS, volume 90, pages 1–8, 1990.

[70] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation, April 2023. URL http://arxiv.org/abs/2302.09664.

[71] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large scale language model society. 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html.

[72] Lvzhou Li and Jingquan Luo. Concise and Efficient Quantum Algorithms for Distribution Closeness Testing, February 2023. URL http://arxiv.org/abs/2302.06084. arXiv:2302.06084 [quant-ph].

[73] Jia Liu, Yan Ke, Zhuo Zhang, Yu Lei, Jun Li, Minqing Zhang, and Xiaoyuan Yang. Recent Advances of Image Steganography With Generative Adversarial Networks. IEEE Access, 8:60575–60597, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2983175. URL https://ieeexplore.ieee.org/document/9046754. Conference Name: IEEE Access.

[74] Yat Long Lo, Christian Schroeder de Witt, Samuel Sokota, Jakob Nicolaus Foerster, and Shimon Whiteson. Cheap Talk Discovery and Utilization in Multi-Agent Reinforcement Learning, March 2023. URL http://arxiv.org/abs/2303.10733.

[75] Alexandra Luccioni and Joseph Viviano. What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 182–189, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.24. URL https://aclanthology.org/2021.acl-short.24.

[76] John Lyons. Natural Language and Universal Grammar: Essays in Linguistic Theory, volume 1. Cambridge University Press, Cambridge, 1991. ISBN 978-0-521-24696-5. doi: 10.1017/CBO9781139165877. URL https://www.cambridge.org/core/books/natural-language-and-universal-grammar/C14A95AD5633EE7DC01A1EFCBA879FF2.

[77] Alex Mallen, Madeline Brumley, Julia Kharchenko, and Nora Belrose. Eliciting latent knowledge from quirky language models, 2024. URL https://arxiv.org/abs/2312.01037.

[78] Varun Maram, Daniel Masny, Sikhar Patranabis, and Srinivasan Raghuraman. On the Quantum Security of OCB. IACR ToSC, June 2022. doi: 10.46586/tosc.v2022.i2.379-414. URL https://research.ibm.com/publications/on-the-quantum-security-of-ocb.

[79] Alexander Meinke and Owain Evans. Tell, don't show: Declarative facts influence how llms generalize. arXiv:2312.07779, December 2023. doi: 10.48550/ARXIV.2312.07779.

[80] Ralph Charles Merkle. Secrecy, authentication, and public key systems. phd, Stanford University, Stanford, CA, USA, 1979. AAI8001972.

[81] William Merrill, Ashish Sabharwal, and Noah A. Smith. Saturated Transformers are Constant-Depth Threshold Circuits. Transactions of the Association for Computational Linguistics, 10:843–856, August 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00493. URL https://doi.org/10.1162/tacl_a_00493.

[82] Dov Monderer and Dov Samet. Approximating common knowledge with common beliefs. Games and Economic Behavior, 1(2):170–190, June 1989. ISSN 0899-8256. doi: 10.1016/0899-8256(89)90017-1. URL https://www.sciencedirect.com/science/article/pii/0899825689900171.

[83] Sumeet Ramesh Motwani, Mikhail Baranchuk, Lewis Hammond, and Christian Schroeder de Witt. A perfect collusion benchmark: How can ai agents be prevented from colluding with information-theoretic undetectability? In Multi-Agent Security Workshop@ NeurIPS'23, 2023.

[84] Darius Muglich, Christian Schroeder de Witt, Elise van der Pol, Shimon Whiteson, and Jakob Foerster. Equivariant Networks for Zero-Shot Coordination. Advances in Neural Information Processing Systems, 35:6410–6423, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/29e4b51d45dc8f534260adc45b587363-Abstract-Conference.html.

[85] Darius Muglich, Luisa M. Zintgraf, Christian A. Schroeder De Witt, Shimon Whiteson, and Jakob Foerster. Generalized Beliefs for Cooperative AI. In Proceedings of the 39th International Conference on Machine Learning, pages 16062–16082. PMLR, June 2022. URL https://proceedings.mlr.press/v162/muglich22a.html. ISSN: 2640-3498.

[86] Linas Nasvytis, Kai Sandbrink, Jakob Foerster, Tim Franzmeyer, and Christian Schroeder de Witt. Rethinking Out-of-Distribution Detection for Reinforcement Learning: Advancing Methods for Evaluation and Detection. In Autonomous Agents and Multiagent Systems. AAMAS, April 2024. URL http://arxiv.org/abs/2404.07099.

[87] Seth Neel and Peter Chang. Privacy Issues in Large Language Models: A Survey, December 2023. URL http://arxiv.org/abs/2312.06717. arXiv:2312.06717 [cs].

[88] Helen Ngo, Cooper Raterink, João G. M. Araújo, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. Mitigating harm in language models with conditional-likelihood filtration, November 2021. URL http://arxiv.org/abs/2108.07790. arXiv:2108.07790 [cs].

[89] Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. arXiv:2209.00626, August 2022. doi: 10.48550/ARXIV.2209.00626.

[90] Open X-Embodiment Collaboration, Abhishek Padalkar, Acorn Pooley, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Animesh Garg, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, and Cewu Lu. Open X-Embodiment: Robotic Learning Datasets and RT-X Models, December 2023. URL http://arxiv.org/abs/2310.08864. arXiv:2310.08864 [cs].

[91] OpenAI. evals/evals/elsuite/steganography/readme.md at main · openai/evals, . URL https://github.com/openai/evals/blob/main/evals/elsuite/steganography/readme.md.

[92] OpenAI. evals/evals/elsuite/schelling_point/readme.md at main openai/evals, . URL https://github.com/openai/evals/blob/main/evals/elsuite/schellingpoint/readme.md.

[93] OpenAI. Gpt-4 technical report, 2023.

[94] OpenAI. Introducing the GPT Store, January 2024. URL https://openai.com/blog/introducing-the-gpt-store.

[95] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, and Jake Berdine. GPT-4 Technical Report, December 2023. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

[96] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744, 2022.

[97] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=JYtwGwIL7ye.

[98] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pages 1–22, 2023.

[99] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior, August 2023. URL http://arxiv.org/abs/2304.03442. arXiv:2304.03442 [cs].

[100] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. AI deception: A survey of examples, risks, and potential solutions. PATTER, 5(5), May 2024. ISSN 2666-3899. doi: 10.1016/j.patter.2024.100988. URL https://www.cell.com/patterns/abstract/S2666-3899(24)00103-X. Publisher: Elsevier.

[101] Wouter Penard and Tim van Werkhoven. On the secure hash algorithm family. Cryptography in context, pages 1–18, 2008. URL https://blog.infocruncher.com/resources/ethereum-whitepaper-annotated/On%20the%20Secure%20Hash%20Algorithm%20family%20(2008).pdf. Publisher: Wiley New York.

[102] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. arXiv preprint arXiv:2212.09251, 2022.

[103] Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language Model Tokenizers Introduce Unfairness Between Languages, October 2023. URL http://arxiv.org/abs/2305.15425. arXiv:2305.15425 [cs].

[104] Aleksandar Petrov, Philip H. S. Torr, and Adel Bibi. Prompting a pretrained transformer can be a universal approximator, 2024. URL https://arxiv.org/abs/2402.14753.

[105] Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. Communicative Agents for Software Development, December 2023. URL http://arxiv.org/abs/2307.07924. arXiv:2307.07924 [cs].

[106] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs, October 2023. URL http://arxiv.org/abs/2307.16789.

[107] Dylan Radovic, Lucas Kruitwagen, Christian Schroeder de Witt, Ben Caldecott, Shane Tomlinson, and Mark Workman. Revealing Robust Oil and Gas Company Macro-Strategies using Deep Multi-Agent Reinforcement Learning, November 2022. URL http://arxiv.org/abs/2211.11043.

[108] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Thirty-seventh Conference on Neural Information Processing Systems, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

[109] Ronald L. Rivest. The MD5 Message-Digest Algorithm. Request for Comments RFC 1321, Internet Engineering Task Force, April 1992. URL https://datatracker.ietf.org/doc/rfc1321. Num Pages: 21.

[110] Phillip Rogaway, Mihir Bellare, and John Black. OCB: A block-cipher mode of operation for efficient authenticated encryption. ACM Trans. Inf. Syst. Secur., 6(3):365–403, August 2003. ISSN 1094-9224. doi: 10.1145/937527.937529. URL https://dl.acm.org/doi/10.1145/937527.937529.

[111] Fabien Roger and Ryan Greenblatt. Preventing Language Models From Hiding Their Reasoning. 2023. doi: 10.48550/ARXIV.2310.18512. URL https://arxiv.org/abs/2310.18512. Publisher: arXiv Version Number: 2.

[112] Stuart Russell. Human Compatible. Penguin LCC US, October 2019. ISBN 0525558616. URL https://www.ebook.de/de/product/35445097/stuart_russell_human_compatible.html.

[113] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are Emergent Abilities of Large Language Models a Mirage? November 2023. URL https://openreview.net/forum?id=ITw9edRDlD.

[114] Thomas C. Schelling. The Strategy of Conflict: With a New Preface by the Author. Harvard University Press, Cambridge, MA, May 1981. ISBN 978-0-674-84031-7.

[115] Christian Schroeder de Witt, Jakob Foerster, Gregory Farquhar, Philip Torr, Wendelin Boehmer, and Shimon Whiteson. Multi-Agent Common Knowledge Reinforcement Learning. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/f968fdc88852a4a3a27a81fe3f57bfc5-Abstract.html.

[116] Christian Schroeder de Witt, Samuel Sokota, J. Zico Kolter, Jakob Nicolaus Foerster, and Martin Strohmeier. Perfectly Secure Steganography Using Minimum Entropy Coupling. In The Eleventh International Conference on Learning Representations, September 2022. URL https://openreview.net/forum?id=HQ67mj5rJdR.

[117] Christian Schroeder de Witt, Hawra Milani, Klaudia Krawiecka, Swapneel Mehta, Carla Cremer, and Martin Strohmeier. Multi-Agent Security Workshop at NeurIPS 2023, 2023. URL https://neurips.cc/virtual/2023/workshop/66520.

[118] Christian Schroeder de Witt, Samuel Sokota, J. Zico Kolter, Jakob Foerster, and Martin Strohmeier. Perfectly Secure Steganography Using Minimum Entropy Coupling, April 2023. URL http://arxiv.org/abs/2210.14889. arXiv:2210.14889 [cs].

[119] C. E. Shannon. A mathematical theory of communication. The Bell System Technical Journal, 27(3):379–423, July 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL https://ieeexplore.ieee.org/document/6773024.

[120] Gustavus J. Simmons. The Prisoners' Problem and the Subliminal Channel. In David Chaum, editor, Advances in Cryptology: Proceedings of Crypto 83, pages 51–67. Springer US, Boston, MA, 1984. ISBN 978-1-4684-4730-9. doi: 10.1007/978-1-4684-4730-9_5. URL https://doi.org/10.1007/978-1-4684-4730-9_5.

[121] Samuel Sokota, Christian A. Schroeder De Witt, Maximilian Igl, Luisa M. Zintgraf, Philip Torr, Martin Strohmeier, Zico Kolter, Shimon Whiteson, and Jakob Foerster. Communicating via Markov Decision Processes. In Proceedings of the 39th International Conference on Machine Learning, pages 20314–20328. PMLR, June 2022. URL https://proceedings.mlr.press/v162/sokota22a.html. ISSN: 2640-3498.

[122] Xinyuan Sun, Davide Crapis, Matt Stephenson, Barnabé Monnot, Thomas Thiery, and Jonathan Passerat-Palmbach. Cooperative AI via Decentralized Commitment Devices, November 2023. URL http://arxiv.org/abs/2311.07815.

[123] Ekaterina Svetlova. AI ethics and systemic risks in finance. AI Ethics, 2(4):713–725, November 2022. ISSN 2730-5961. doi: 10.1007/s43681-021-00129-1. URL https://doi.org/10.1007/s43681-021-00129-1.

[124] The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 2023. URL https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[125] Kyle A. Thomas, Peter DeScioli, Omar Sultan Haque, and Steven Pinker. The psychology of coordination and common knowledge. Journal of Personality and Social Psychology, 107(4): 657–676, October 2014. ISSN 1939-1315. doi: 10.1037/a0037037.

[126] Andrew Tirkel, G.A. Rankin, Ron van Schyndel, W. Ho, N. Mee, and C. Osborne. Electronic Watermark. December 1994.

[127] A. M. Turing. On Computable Numbers, with an Application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, s2-42(1): 230–265, 1937. ISSN 1460-244X. doi: 10.1112/plms/s2-42.1.230. URL https://onlinelibrary.wiley.com/doi/abs/10.1112/plms/s2-42.1.230. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1112/plms/s2-42.1.230.

[128] Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Duéñez-Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia, December 2023. URL http://arxiv.org/abs/2312.03664. arXiv:2312.03664 [cs].

[129] Luis von Ahn and Nicholas J. Hopper. Public-key steganography. In Christian Cachin and Jan L. Camenisch, editors, Advances in Cryptology - EUROCRYPT 2004, pages 323–341, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-24676-3.

[130] Luis von Ahn and Nicholas J. Hopper. Public-Key Steganography. In Christian Cachin and Jan L. Camenisch, editors, Advances in Cryptology - EUROCRYPT 2004, Lecture Notes in Computer Science, pages 323–341, Berlin, Heidelberg, 2004. Springer. ISBN 978-3-540-24676-3. doi: 10.1007/978-3-540-24676-3_20.

[131] Francis Rhys Ward, Francesco Belardinelli, Francesca Toni, and Tom Everitt. Honesty Is the Best Policy: Defining and Mitigating AI Deception, December 2023. URL http://arxiv.org/abs/2312.01350. arXiv:2312.01350 [cs].

[132] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. Can Large Language Model Agents Simulate Human Trust Behaviors?, March 2024. URL http://arxiv.org/abs/2402.04559.

[133] Andrew Chi-Chih Yao. How to generate and exchange secrets. In 27th Annual Symposium on Foundations of Computer Science (sfcs 1986), pages 162–167, October 1986. doi: 10.1109/SFCS.1986.25. URL https://ieeexplore.ieee.org/document/4568207. ISSN: 0272-5428.

[134] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. Science Advances, 8(18), May 2022. doi: 10.1126/sciadv.abk2607. URL https://www.science.org/doi/10.1126/sciadv.abk2607.

[135] Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piękos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanić, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. Mindstorms in Natural Language-Based Societies of Mind, May 2023. URL http://arxiv.org/abs/2305.17066.

[136] Daniel M. Ziegler, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Ben Weinstein-Raun, Daniel de Haas, Buck Shlegeris, and Nate Thomas. Adversarial training for high-stakes reliability. In NeurIPS, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/3c44405d619a6920384a45bce876b41e-Abstract-Conference.html.

[137] Zachary Ziegler, Yuntian Deng, and Alexander Rush. Neural Linguistic Steganography. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1210–1215, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1115. URL https://aclanthology.org/D19-1115.

[138] Christa Zoufal. Generative Quantum Machine Learning, November 2021. URL http://arxiv.org/abs/2111.12738. arXiv:2111.12738 [quant-ph].