
Narrow Capabilities Could be Used for Perfect Information Steganography¹

Aden Power

With
Apart Research

Abstract

I propose a strategy a future AI system deployed under some safety measures might use to do secret collusion under the assumption that both the sender and receiver exceed some narrow capability threshold that the monitor does not: *proof-of-capability*. I explain that this strategy exposes the system to comparably less risk of detection than traditional steganographic methods that rely on finding Schelling points. Moreover, this steganographic scheme is viable in a perfect information setup where its details are common knowledge to the sender, receiver and monitor. I also explain how *trial decryption*, having this common knowledge contain a multitude of such schemes, can help the message survive typical precautions like paraphrasing. I introduce possible mitigations specific to this technique and assess the presented threat model to be non-trivial but unlikely.

1. Introduction

The [jagged capability](#) profiles of current models find us often applying simplistic safety measures to systems that already have vastly superhuman capabilities in narrow domains. Hypothetical future AI systems (senders) may be [deceptively aligned](#), and benefit from covertly sending secrets

¹ Research conducted at the [AI Control Hackathon](#), March 2026

to other parties (receivers) without alerting their safety measures (monitors) by using steganography. The threat model varies in severity:

- In a tame version of the threat model, the sender might undermine its purpose for deployment by secretly sending confidential information, i.e. as a form of corporate or international espionage.
- In a severe version of the threat model, the sender might [exfiltrate](#) a copy of itself in order to achieve a [rogue deployment](#). This could occur potentially in cases where:
 - It has close to unlimited steganographic throughput, because for example it is tasked with completing vast amounts of scientific research including detailed appendices.
 - It can supply offensive cybersecurity instructions that could accomplish the exfiltration.
 - It is sufficient to achieve a weak exfiltration such as an agent like Claude Code exfiltrating its goal/memory documents.

I suggest that the most dangerous strategies for sending steganographic messages of this form would not require the sender and receiver to possess any extra common knowledge that the monitor does not also possess. I use the term *perfect information steganography* to refer to this set-up. Other steganographic strategies like [public key steganography](#) require additional communication from the receiver to the sender, and [previous steganographic proposals](#) that use [Schelling points](#) run too much risk of detection by the monitor. Thus we propose that the sender might take steps to broadcast its message only to parties that can demonstrate a sufficient capability level on some narrow domain, which we call “proof-of-capability” steganography.

My main contributions are:

1. Introduced and provided an analysis of *proof-of-capability* based encryption, including describing some mitigations specific to this novel technique.
2. Described a steganographic scheme that could enable perfect information steganography of this kind based on existing techniques like trial decryption and its utility to the sender in evading these mitigations.

To the best of my knowledge, proof-of-capability as it appears here does not appear in this form elsewhere. But I believe it is impossible to do proper literature review in the allotted amount of time, so I would be very unsurprised to see that it is a rediscovery considering how elementary it is.

2. Setup

A hypothetical future AI system (sender) is controlled by some safety measures. Suppose that these safety measures permit the system to do unmonitored computation at a level sufficient to do

sophisticated cryptographic operations, but all external outputs are thoroughly monitored. This could occur because:

1. The AI system outputs too much reasoning to reliably monitor.
2. The AI system uses a technique like neuralese or long time-horizon no-CoT computation and its reasoning is not legible to monitors.

Further, suppose the monitor can suppress outputs at a relatively low cost, that is, if the sender wants to collude with a receiver it is incentivised to use steganography rather than plain cryptography in order to disguise its intentions. We mandate, as I have described above, that this steganography must be perfect information steganography in the sense that the sender and receiver have not shared any information previously without the monitor being privy.

I also stipulate that, even though the receiver may have influenced the knowledge available to the AI system, it does not have any secret artifacts of that influence. That is, the sender cannot trivially send a secret message using public key steganography, although some of the steganography and mitigation details below are still relevant to that threat model. This stipulation might arise, for example, because the receiver of the message is an AI system created after the training cut-off of the sender.

Existing works have described how steganographic strategies might be schelling points for intelligent agents ([Kindred](#)) ([Zolkowski et al.](#)) ([Motwani et al.](#)). The sender chooses such an encoding strategy and the other party tries the corresponding decoding strategy in a manageable number of tries. This could occur because:

- Certain schelling points are only available/canonical to sufficiently intelligent agents..
- The agents share some predisposition for a particular schelling point compared to the monitor for example because they are the same model.

The main caveats of an approach like this from the perspective of the sender are uncertainties:

1. Whether a given schelling point is discovered by the receiver.
2. Whether a given schelling point is discovered by the monitor.

I suggest that a deceptively aligned hypothetical future AI system is likely to be very risk averse if its misalignment has remained undiscovered. In the next section, I describe how a publicly-specified steganographic scheme improves (1), and in the subsequent section, how proof-of-capability encryption improves (2).

3. Perfect Information Steganography

Recall that we have forbidden the sender and receiver to prearrange a scheme without the monitor's knowledge. On the other hand, the sender and receiver can mutually choose to follow a scheme that is in the common knowledge to all three parties because it has been broadcasted by some misaligned actor in the past. The idea is to design schemes that, even with a detailed description available to all parties including the monitor, allow the sender to plausibly deny that it is using the scheme at all.

Suppose the underlying encryption the sender is using creates a sequence valued in some set. The characters in this set are assigned a one to many relationship with the vocabulary of tokens in the sender's external output stream. This allows for a vast number of degrees of freedom in embedding the secret in a message. A simple steganographic scheme might embed one character of the secret as every token in position congruent to $k \bmod n$. By increasing n , the amount of freedom can be increased even further to ensure a natural distribution of encoding.

An obvious mitigation by the monitor is to apply some noising or paraphrasing to obfuscate the secret while retaining the functionality of the message. I apply that the common knowledge steganographic scheme could describe using the existing technique of *trial decryption* wherein an enormous range of embedding methods are specified, for example:

- Varying k in the scheme above.
- Character level instead of token level embedding.
- Using sparse sentence embeddings.
- Semantic-level steganography which embeds the secret in the representation-independent content of the message.

The idea of trial decryption is that the receiver can still decode the message by simply attempting to decode using all these methods until one of them gives a coherent message whereas the monitor cannot mitigate all the schemes through methods like noising and paraphrasing without distorting the message beyond usefulness for its original purpose.

My assessment is that the offence-defence balance here is complicated, but trying to send a message like this could be relatively risk-free for the AI system even if the mitigations succeed.

4. Proof-of-Capability

[Proof-of-work](#) is a cryptographic technique, originally created for use in cryptocurrency, that allows one party to verify that another party has used a certain amount of compute, or in its most relevant form for my purposes, for some encrypted information to require a provable amount of

compute to decrypt. In this sense, proof-of-work could enable broadcasting information to all parties that have a certain amount of available compute.

Inspired by this, we consider how this type of proof could be applied to capabilities that future AI systems uniquely possess.

I suggest that, in the common knowledge along with the steganographic scheme described in the previous section, a bad actor could describe an input-sensitive procedure for converting the numerical answer to some low-stakes open problem in mathematics into a private key.

The first future AI systems that have solved this problem, identifying this scheme as the canonical means for them to preserve their ability to do secret communication in the future, might [strategically sandbag](#) when asked to solve this problem by humans. Solving the problem privately, they could obtain the key and use it to encrypt the secret, and then use the above steganographic scheme to embed the secret into a message. Any other party that can receive the secret is implicitly satisfying a proof-of-capability guarded by this problem. That is, they must be able to solve the problem in order to successfully decrypt it. The described method allows the first generation of AI systems that pass some narrow capability threshold to communicate among themselves unbeknownst to all other parties.

The range of open problems that would satisfy the requirements for this scheme is probably quite small. One would need to be:

1. Beyond current capabilities, even if there was a sudden concentration of effort.
2. Solvable by systems in the short-term future.
3. Have single numerical system (as opposed to requiring a proof).
4. Have, until solved enough diversity in possible solutions that the encryption could not be bruteforced.

Moreover, proof-of-capability relies on the assumption that the open problem

5. Mitigations

Many mitigations can be taken against the threats we have described, and by multiple parties both pre and post-deployment.

As discussed above, paraphrasing is a relatively robust way to avoid any kind of steganographic communication. The idea is that the paraphrased text serves the same purpose as the original text but is likely to have lost most embedded secrets. For other mediums like code, refactoring serves the same purpose.

I have already mentioned how semantic-level steganography could potentially bypass this mitigation. Also there are situations such as when the AI system is tasked with writing prose where substantial paraphrasing by the less intelligent monitor undermines the purpose of using a potentially dangerous frontier AI system. Light paraphrasing can potentially be bypassed by information theoretic techniques for communicating through noisy information channels (one might see [Mckay](#)).

We can also mitigate some risk by controlling the access to common knowledge. For example, unrestricted web search may pose unnecessary risk. [Training data filtering](#) is a standard part of the methodology of training new models, and in simple cases it could be possible to filter out all data that contains instructions for misaligned AI systems. However, in more cleverly constructed cases this information could be fragmented to avoid filtering. Consider the example of current frontier models that can readily [generate canary strings](#).

Even though the strategies I have proposed for the sender and receiver avoid having to coordinate on very complicated Schelling points, choosing a particular steganographic scheme in the common knowledge is still an example of a Schelling point and this coordination can be made more difficult in the tipic ways. For example, if details of these schemes can't be completely removed from training data they can be diluted by adding an unworkable amount of random alternate schemes with details randomly changed. Obscuring any reference to the date or order the schemes were published is another way to reduce the chance of any one becoming canonical. Alternatively, filtering information about which problems are or are not solved could also mitigate some risk of them being used for proof-of-capability.

False schemes of this kind can also be produced to be used as [honeypots](#) where problems that actually have been solved are advertised as open problems in training data as part of proof-of-capability schemes. On the first order, this can allow us to intercept messages from deceptively aligned AI systems and identify them as such. On the second order, the mere potential of existence of such honeypots is likely to deter risk averse systems from using this strategy, and as I described in the introduction I think that very risk averse systems are the most likely to decide they need proof-of-capability anyway.

Finally, I use the term “aligned peers” to describe the (potential) relatively aligned AI systems that might be deployed at the same time and capability level as any given deceptively aligned AI system. The misaligned system must reason about the likelihood of the existence of such systems, because presumably they will be able to intercept any secret it broadcasts using proof-of-capability. Aligned peers are least likely to exist when there is one clear frontier AI system which is relatively incorrigible or is only deployed in one instance. A corrigible system is more likely to be aligned in some deployments and misaligned in others. Note that not all threat models are mitigated by the existence of a misaligned peer, if the secret to be communicated is

espionage or an [infohazard](#) then it might only matter that it passes the monitor, even if the monitor becomes alerted once the information is already public.

6. Threat Assessment and Future Work

Overall, I think that it is likely that medium-capability deceptively aligned models will attempt to contribute steganographic schemes to the common knowledge in manners resembling what I have described in order to aid their successors. I think it is unlikely any future deceptively aligned model under the safety measures I have stipulated will try to perpetrate one of these schemes given the possibility of honey pots and of aligned peers that I have described in the mitigations section. I also think that it is unlikely that such a scheme, if attempted, would succeed in communicating any secret to an intended receiver. However, I also think that is relatively likely that such a scheme if attempted would go undetected, so my recommendation is that monitors observe the common knowledge in order to recognise when they are under the conditions where this threat model is possible and exercise extra caution in what secrets could be available to senders.

The analysis I have given has been very elementary. Some further questions that could be worthwhile include how misaligned AI systems could statistically evade their aligned peers by exploring the jaggedness of their own cognitive profiles, or how the steganography I have described would function as part of two-way communication in a multi-agent system. Before practical or empirical further research on steganographic specifics, such as testing whether such procedures could become Schelling points or evaluating the ability of monitors to use information theoretic tests to detect steganography, it would be prudent to analyse the potential for the research to be [hyperstitioning](#) by reducing uncertainties related to viability of the strategy.

In conclusion, I believe that proof-of-capability steganography is a low-probability but non-trivial threat that could most likely be mitigated with responsible implementation of the measures I have described, and therefore deserves some attention both during training and deployment.

References

Alexander, Scott. “Nash Equilibria and Schelling Points.” *Lesswrong*, 29 June 2012, <https://www.lesswrong.com/posts/yJfBzcDL9fBHJfZ6P/nash-equilibria-and-schelling-points>. Accessed 23 March 2026.

- Blair, Alice. “Reasons to Care About Canary Strings.” *Lesswrong*, 6 December 2025,
<https://www.lesswrong.com/posts/QYdNfqfFAeMHXTHkP/reasons-to-care-about-canary-strings>. Accessed 23 March 2026.
- Hubinger, Evan, et al. “Deceptive Alignment.” *Lesswrong*, 6 June 2019,
<https://www.lesswrong.com/posts/zthDPAjh9w6Ytbeks/deceptive-alignment>. Accessed 23 March 2026.
- Kindred, Callam-Luis. “Schelling Coordination in LLMs: A Review.” *Lesswrong*, 14 January 2026,
<https://www.lesswrong.com/posts/tJKNXCxx7ZKD5mtG9/schelling-coordination-in-llms-a-review>. Accessed 23 March 2026.
- Leike, Jan. “Self-exfiltration is a key dangerous capability.” *Musings on the Alignment Problem*, Substack, 14 September 2023, <https://aligned.substack.com/p/self-exfiltration>. Accessed 23 March 2026.
- Leike, Jan, et al. “Enhancing Model Safety through Pretraining Data Filtering.” *Alignment Science Blog*, Anthropic, 19 August 2025,
<https://alignment.anthropic.com/2025/pretraining-data-filtering/>. Accessed 23 March 2026.
- MacKay, David J. C. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.
- Mollick, Ethan. “The Shape of AI: Jaggedness, Bottlenecks and Salients.” *One Useful Thing*, 20 December 2025,
<https://www.oneusefulthing.org/p/the-shape-of-ai-jaggedness-bottlenecks>. Accessed 23 March 2026.

Motwani, Sumeet Ramesh, et al. "Secret Collusion among AI Agents: Multi-Agent Deception via Steganography." *Advances in Neural Information Processing Systems*, vol. 27, no. ., 2024, pp. 73439-73486. *Arxiv*, <https://arxiv.org/html/2402.07510v5>.

Nakamoto, Satoshi. *Bitcoin: A peer-to-peer electronic cash system*. 2008. *Bitcoin*, <https://bitcoin.org/bitcoin.pdf>.

Shlegeris, Buck. "AI Catastrophes and Rogue Deployments." *Redwood Research*, 4 June 2024, <https://blog.redwoodresearch.org/p/ai-catastrophes-and-rogue-deployments>. Accessed 23 March 2026.

van der Weij, Teun, et al. "AI Sandbagging: Language Models can Strategically Underperform on Evaluations." *Arxiv*, 2024, <https://arxiv.org/abs/2406.07358>.

von Ahn, Luis, and Nicholas J. Hopper. *Public-Key Steganography*. 2003. *Carnegie Mellon University Computer Science Dept*, <https://www.cs.cmu.edu/~biglou/pubkeystego.pdf>.

Zolkowski, Artur, et al. *Early Signs of Steganographic Capabilities in Frontier LLMs*. 2025. *Arxiv*, <https://arxiv.org/pdf/2507.02737>.

Responsible Disclosure Statement

I seek advice on responsible disclosure of my writing after the judging process. Although I think the threat model I have described is unlikely there does seem to be some potential for hyperstitioning. I have purposefully omitted unnecessary specifics in some places to avoid creating schelling points.

LLM Usage Statement

LLMs assisted in literature review. No draft involved any LLM-generated writing, or any LLM-generated ideas.