

Two More Backdoor Papers: Unelicitable Backdoors in Language Models via Cryptographic Transformer Circuits

(Draguns, Gritsevskiy, Ramesh Motwani, Rogers-Smith, Ladish, Schroeder de Witt, 2025)

and

Backdoor Defense, Learnability, and Obfuscation

(Christiano, Hilton, Lecomte, Xu, 2025)

John Dunbar

Paper 1, Cryptographic Transformer Circuits

In this class, we talked about two follow ups to the seminal Goldwasser, Kim, et al. (2024) paper on undetectable backdoors in neural networks. Starting with the shorter paper, Unelicitable Backdoors in Language Models via Cryptographic Transformer Circuits (Draguns et al., 2025) handcrafts weights in a transformer to implement a backdoor with basic cryptography. It's not a new or complicated idea, they just use the hash of the transformer's input as the decryption key for some backdoor behavior and include a canary in the decrypted string to know when the key was correct. Their main contribution is implementing this computation in a transformer's weights, and so I thought it was not very relevant to the topic of our class, AI alignment *theory*.

Paper 2, Backdoor Defense

The Draguns et al. (2025) backdoor also has a trivial defense which is to just look at the transformer's weights and remove the handcrafted circuit. Even if we use a white-box undetectable backdoor like one of the constructions in Goldwasser, Kim, et al. (2024), it might be removable with processes like distillation or adding noise. Backdoor Defense, Learnability, and Obfuscation (Christiano et al., 2025) studies exactly this setting: white-box backdoor defense where mitigations after training are allowed. They define a class of functions $\mathcal{F}$ as defendable if the defender can win the following game against an adversary:

1. First, the adversary chooses the non-backdoored function $f \in \mathcal{F}$ and the data distribution $\mathcal{D}$.
2. Next, the backdoor trigger x^* is drawn randomly from $\mathcal{D}$.

3. Then, the adversary chooses the backdoored function f^* . The criteria for a backdoored function is that it behaves differently on the backdoor trigger, $f^*(x^*) \neq f(x^*)$, but behaves the same for most other inputs, $\Pr_{x \sim \mathcal{D}}[f^*(x) = f(x)] \geq 1 - \varepsilon$ for some $0 < \varepsilon < 1$.
4. Finally, the defender is randomly given either $(f, x \sim \mathcal{D})$, the true function and a random input, or (f^*, x^*) , the backdoored function and the backdoor trigger, and it must distinguish which one it was given.

If it's odd to think of the adversary choosing both the true and the backdoored functions, this step can also be thought of as a for-all quantification where the defender must win the remaining steps of the game in the worst case regardless of the true function and data.

Choosing the backdoor trigger randomly may also seem unusual, but it can be justified. One student brought up that if our AI were a resume screener and the desired backdoor was for our own resume to be given the highest rating, the backdoor trigger can't be any random string, it must be a realistic resume with our name on it. This doesn't exactly match our setting since we might choose from a small set of realistic resumes rather than being forced to use a single one, but if the adversary can win this paper's game that gives it a single random trigger, then they can win with a small set of triggers.

The final difference between this setting and previous work on backdoors is that the defender must detect the backdoor when it's activated by the trigger x^* rather than detecting it statically. This allows the defender to win if it can detect anomalous behavior or derive a non-backdoored function, even if it can't find the backdoor from just the function and its binary representation. Compared to Goldwasser, Kim, et al. (2024), this paper's setup puts additional limitations on the adversary and proves that this allows more function classes to be defendable.

Their Results on Backdoor Defendability

The paper's main result is that defendability in the statistical sense is characterized fully by VC dimension, a measure of complexity closely connected to classical learning theory. For a function class $\mathcal{F}$ with VC-dimension $\text{VC}(\mathcal{F})$ the best probability the defender has of winning the backdoor game is

$$\max\left(\frac{1}{2}, 1 - \Theta(\text{VC}(\mathcal{F})\varepsilon)\right),$$

where ε is the fraction of the dataset for which the backdoored function is allowed to differ from the true function, and the defender wins with probability $\frac{1}{2}$ if they guess randomly.

The way this paper proves the lower bound for this theorem is closely related to the concept of random self-reducibility used for backdoor defense by a contemporaneous paper, Goldwasser, Shafer, et al. (2024). They construct a defender that does something akin to distillation. It samples from the possibly backdoored function to predict the true function's output. When the VC-dimension is small, the defender can form an accurate prediction with relatively few samples, and when ε is small, the defender has a good chance that each datapoint they sample was not a place where the backdoor function differed from the true function. In a sense, the true functions are randomly self-reducible because evaluating them on any specific point can be reduced to evaluating a few random examples, so any small corruption is fixable.

Although this theorem proves when defense is possible, however, it doesn't prove when defense can be done in polynomial time. For computationally bounded defenders, this paper is bad news. Using punctured pseudorandom functions and indistinguishability obfuscators, they are able to create polynomial-size boolean circuits that can be backdoored in a way that's undetectable to any polynomial-time defender. This is despite all the affordances given to defense like white-box access and seeing the backdoor trigger x^* . The natural response to this attack, however, is that in real life the true functions are not heavily obfuscated, so defender can reject any circuit that is. This remains a direction for future work to define a good notion of "not heavily obfuscated."

Our Overall Thoughts on Backdoor Defense

I personally liked this paper because it thoroughly studies a backdoor setting that's more applicable to the real world than the prior work, but in our discussion, some people found the specific setup suspicious. There seems to be many different ways this paper's backdoor game could've been set up, and these different setups might give different results. Our discussion also brought up concerns on how relevant backdoors actually are to long term alignment. The paper brings up an analogy that a deceptively aligned model may be akin to a backdoor: it behaves like a cooperative model when there are subtle cues it's being evaluated, but behaves uncooperatively on certain inputs that were rare during evaluation. But the path from that analogy to a concrete technique that we know will actually detect deceptive alignment is long and unclear.

References

- Christiano, P., Hilton, J., Lecomte, V., & Xu, M. (2025). Backdoor defense, learnability and obfuscation. In (Vol. 325, p. 38:1-38:21). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. Retrieved from <https://drops.dagstuhl.de/entities/document/10.4230/LIPICS.ITCS.2025.38> doi: 10.4230/LIPICS.ITCS.2025.38
- Draguns, A., Gritsevskiy, A., Motwani, S. R., Rogers-Smith, C., Ladish, J., & de Witt, C. S. (2025). *Unelicitable backdoors in language models via cryptographic transformer circuits*. Retrieved from <https://arxiv.org/abs/2406.02619>
- Goldwasser, S., Kim, M. P., Vaikuntanathan, V., & Zamir, O. (2024). *Planting undetectable backdoors in machine learning models*. Retrieved from <https://arxiv.org/abs/2204.06974>
- Goldwasser, S., Shafer, J., Vafa, N., & Vaikuntanathan, V. (2024). *Oblivious defense in ml models: Backdoor removal without detection*. Retrieved from <https://arxiv.org/abs/2411.03279>