Deployment of Watermarking?

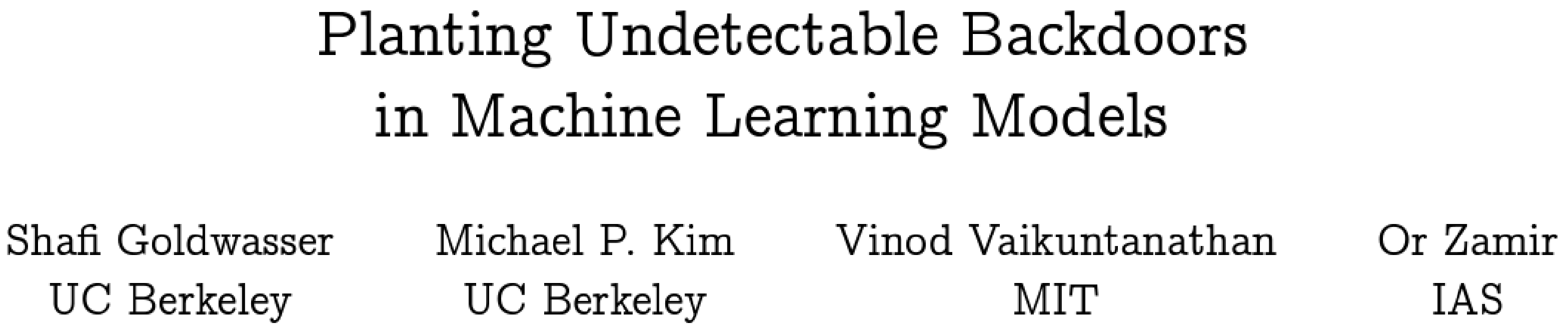
Prototype of my scheme implemented by OpenAI's Hendrik Kirchner

It seems to work, even with a few hundred tokens!

Questions that have come up:

- Will customers rebel and switch to a competing language model?
- Can we coordinate? (Cf. White House commitment)
- Is watermarking unfair to, e.g., ESL students?
- Who gets access to the detection tool? Everyone? Only Canvas, TurnItIn.com, journalists?

Cryptographic Backdoors in ML



Planting Undetectable Backdoors
in Machine Learning Models

Shafi Goldwasser
UC Berkeley

Michael P. Kim
UC Berkeley

Vinod Vaikuntanathan
MIT

Or Zamir
IAS

Result: Suppose you control the training data of a depth-2 ReLU network with random initial weights. Then you can cause there to be a secret input on which the network "goes crazy," which is polynomial-time undetectable assuming a standard cryptographic conjecture (hardness of Planted Clique)

(Generalizing to realistic neural nets is a major problem)

Lemons to Lemonade!

Insight: An undetectable backdoor could be **great** for AI alignment! A means to identify, control, and shut down a powerful AI that only its human creators knew about

Cf. Adi et al. 2018

"The Off-Switch Problem"

But Goldwasser et al. won't immediately work: we need an **unremovable** backdoor, not an **undetectable** one

What does "unremovable" even mean here?

Two Obvious Attacks by the Super-AI

(1) It trains a second AI that lacks the backdoor

Except then it faces its own version of the alignment problem!

(2) It encases itself in code saying "If shutdown command is output, overwrite by 'stab humans harder' "

But what if the AI **wants** to shut itself down in some cases?

Challenge: Give a formal definition of "unremovable backdoor" that's not immediately killed by these attacks

Idea: Assume the model has "backdoor-like" behaviors that it **likes**. Then can we also insert a backdoor that it **doesn't** like?

Summary

Whether or not it can stop the robot uprising, cryptography can clearly play a role in mitigating near-term harms from generative AI, from cheating to fraud to theft to privacy violations

Watermarking, backdoors, privacy-preserving ML, CAPTCHAs, and much more

Typically, can’t just layer existing crypto protocols on top of ML—not only because of efficiency, but because the goals are conceptually new

“Neurocryptography” is at any rate a better name for this subject than “deep crypto”…