"Words can be like X-rays if you use them properly – they'll go through anything. You read and you're pierced." (Aldous Huxley)

IMMORAL PROGRAMMING –

THE CASE OF *DEEPFAKE SCIENCE* ATTACKS

Dr. ir. Leon Kester, Senior Research Scientist, TNO Netherlands

Dr. Nadisha-Marie Aliman, M. Sc., Independent Visiting Scholar, Utrecht University

TNO innovation for life

Utrecht University

OUTLINE

I. Defenses Against Immoral Programming (IP) as Moral Programming (MP)

II. Deepfake Science Attacks as IP Use Case

III. Defenses Against Deepfake Science Attacks

IV. Conclusion

RISK MANAGEMENT FOR MORAL PROGRAMMING

- Mitigation of AI risks linked to mitigation of socio-psycho-techno-physical harm
- Good regulator theorem from cybernetics: "every good regulator of a system must be a model of that system" (Conant and Ashby, 1970) → rigorous harm model needed for moral programming

How and when was AI risk instantiated?		***Causes***	
		On Purpose	***By Mistake***
Timing	***Pre-Deployment***	*a*	*b*
	Post-Deployment	*c*	*d*

Modified and adapted from Aliman et al. (2021)

EXTENDING MORAL PROGRAMMING

more suitable harm model for moral programming

How and when was AI risk instantiated?		Causes	
		On Purpose	By Mistake
Timing	Pre-Deployment	*a*	*b*
	Post-Deployment	*c*	*d*

conventional harm model for moral programming

immoral programming

OUTLINE

I. Defenses Against Immoral Programming (IP) as Moral Programming (MP)

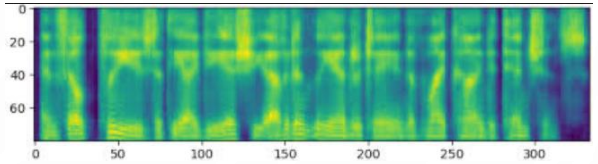
II. Deepfake Science Attacks as IP Use Case

III. Defenses Against Deepfake Science Attacks

IV. Conclusion

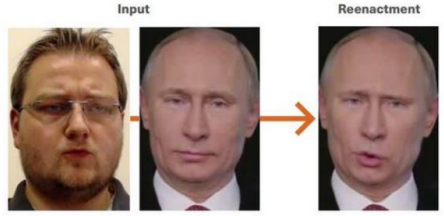
MALICIOUS DEEPFAKE DESIGN

- Deepfake voice for voice impersonation and cybercrime
- Deepfake video for sextortion
- Deepfake images for fake profiles in disinformation operations and espionage
- Deepfake videos for non-consensual voyeurism
- Deepfake „hologram“ for impersonation in video calls
- **Future deepfakes for deepfake science attacks?**



(Rössler et al., 2019)

(Satter, 2019 (AP news))

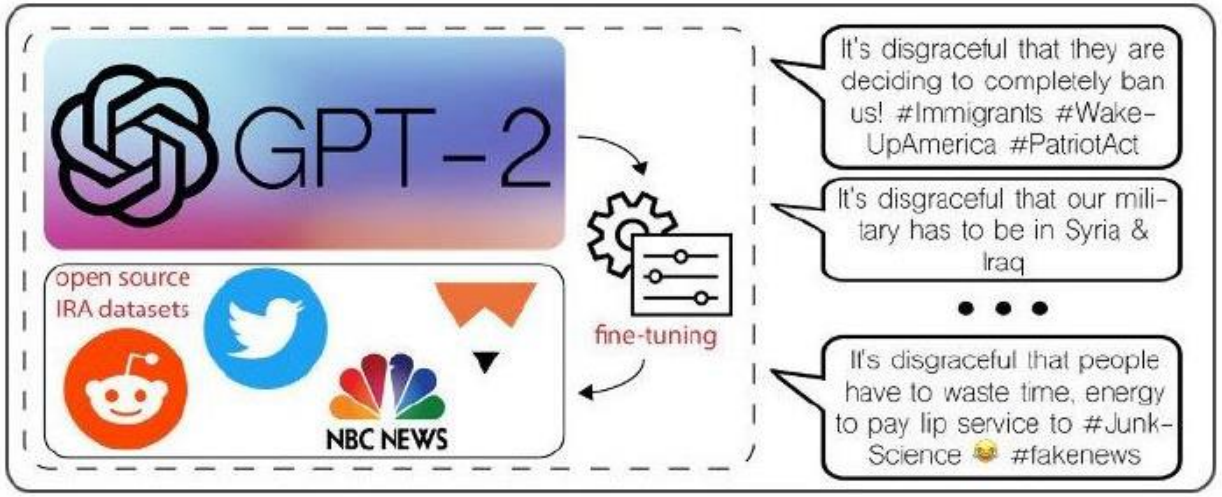


(Thies et al., 2020)

How and when was AI risk instantiated?		Causes	
		On Purpose	*By Mistake*
Timing	*Pre-Deployment*	*a*	*b*
	Post-Deployment	*c*	*d*

DEEPFAKE TEXT

- N.B: Deepfake (deep-learning based fakery) technology is <u>not</u> restricted to images/audios/videos. An often overlooked case is **deepfake text.**



(Tully and Foster, 2020)

We Asked GPT-3 to Write an Academic Paper about Itself—Then We Tried to Get It Published

An artificially intelligent first author presents many ethical questions—and could upend the publishing process

By Almira Osmanovic Thunström

DEEPFAKE SCIENCE

(ALIMAN, 2021; ALIMAN AND KESTER, 2022)

- **<u>Deepfake science attack:</u> The technically possible but not yet widespread malicious use of deepfake artefacts (e.g. deepfake text/audio/ video/image) for the purpose of epistemic distortion in science**
- Examplary deepfake text in science generated with language AI model GPT-2 (see table to the right, right column)

Prompt by human	Counterfactual outputs of pre-trained GPT-2
Abstract In a time in which topics such as fake news, disinformation, deepfakes and recently fake science affect online debates amongst the population at large but also specifically in scientific circles, we thematize the following elephant in the room now and not in hindsight: what to do if malicious actors use AI for not yet prevalent but technically feasible 'deepfake science attacks' i.e., on (applied) science itself? Deepfakes are not restricted to audio and visual phenomena and deepfake text whose impact could be potentiated with regard to speed, scope, and scale may represent an underestimated avenue for malicious actors.	[...] Here, we assume that a bot that, by modifying and/or manipulating the data extracted from the published article, creates an actual original story about fake news, false claims, hoaxes, or lies. These are potentially useful means against those intending to harm the credibility [...] of an author's work. [...]
	It is clear that these types of threats could be used to disrupt scientific debate in the future. [...] The threat in question is, therefore, likely to make the field far more hostile.
	Linguistics is a core part of science, and we need to learn it. [...] We can't just use algorithms to control it.
	[...] Liz Y. Wang et al., "A model of the Deep Web, the Deep Web and the First Human-Driving Device†: a comparative analysis -in order to better inform the development and analysis of advanced techniques and technologies. Artificial Intelligence and Cyberthreats." IEEE Transactions on Intelligence 14, 14 (2010): 1797–1805. [Crossref] [PubMed] 2nd Annual IEEE International Conference on CyberSecurity.
	[...] we present empirical evidence that such false expectations apply disproportionately to bots compared to humans. [...] Human Cognition: Human Cognition and Cybernetic Engineering by Michael Schmitt et al., 2002; doi:10.1016/j.heccomplications.2002.01.009
	[...] What will be interesting about this issue is how much more targeted the attack will be under AI technologies, based only on what we previously thought was good science. Let us see how this plays out in practice.
	[...] The most likely scenario if malicious actors do exploit deepfakes is that they create a false narrative (by exploiting "fake science"), by using the "fake news" as their method of proving the "fake evidence." These tactics could even be employed to further their goals by using 'fake science' to further their own agendas, or (in this case) to gain public acceptance of them. [...] R. Pérez, "Deepfaking Theses at the University of Cambridge," American Journal of Physics 37:2890-2915 (2011), pp. 21-233:

OUTLINE

I. Defenses Against Immoral Programming (IP) as Moral Programming (MP)

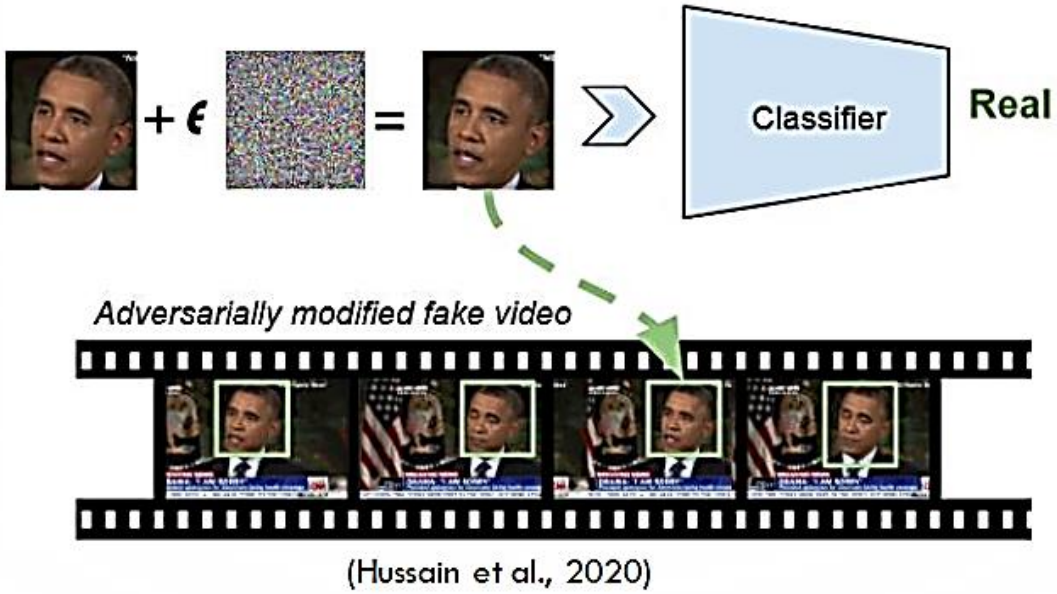
II. Deepfake Science Attacks as IP Use Case

III. Defenses Against Deepfake Science Attacks

IV. Conclusion

WHY A BETTER APPROACH THAN „DEEPFAKE DETECTION“ IS NEEDED AS DEFENSE

1. Deepfakes involve an open adversarial cat-and-mouse game. The adversary can adapt to present-day AI-based detection schemes.



(Hussain et al., 2020)

WHY A BETTER APPROACH THAN „DEEPFAKE DETECTION“ IS NEEDED AS DEFENSE

2. Any text/audio/video/picture sample could be suspected to be deepfake-based → automated disconcertion. Scientists could then unintentionally exclude scientists being statistical outliers even more. (Examples: imagine e.g. scientific videos of people with certain physical health conditions, texts written by eccentric and/or neurodivergent scientists, etc.)

PRESENT-DAY „AI“ SHOULD **NOT** BE **OVER**ESTIMATED

CYBORGNETIC COMPREHENSION BOTTLENECK

- Asymmetry: ability to create information **x** ≠ ability to understand information **x** (example: present-day AI can create outputs perceived as explanations, but present-day AI does **not understand** it)

PRESENT-DAY „AI“ SHOULD **NOT** BE **OVER**ESTIMATED

- The epistemic aim of science can be to achieve <u>better and better</u> **explanations** (Popper, 1957; Frederick, 2020). Science is <u>not</u> merely about data/experiments.
- It is impossible for imitative „AI“ to reliably create better **<u>new</u> yet unknown** *chains* of explanations (also called explanatory blockchains (Aliman, 2021)) required for novel scientific/philosophical theories.

Exemplary recipe for an explanatory blockchain (Aliman, 2021) loosely inspired by an essay of Frederick (2020)

BUT: THE POTENTIAL OF PRESENT-DAY AI SHOULD ALSO **NOT** BE ***UNDER***ESTIMATED

- Deepfake detection may be doomed in the long-term. Prohibiting deepfakes may not be enforceable in the long-term.
- Proactive *self-paced* exposure to synthetic AI-generated material could prepare scientists for that and enhance their critical thinking.
- Deepfake technology can be used to augment human creativity (e.g. use of language AI to assist in generating new threat models and defenses in AI safety, (cyber)security, risk management, …)

OUTLINE

I. Defenses Against Immoral Programming (IP) as Moral Programming (MP)

II. Deepfake Science Attacks as IP Use Case

III. Defenses Against Deepfake Science Attacks

IV. Conclusion

CONCLUSION

- Defending against deepfake science attacks can involve a new form of **moral programming.**
- Science can be robust through its own chain of words by relying on its *explanation-anchored* (and **not** merely data-driven) nature which is grounded in better and better **new** *chains* of **explanations.**
- Scientists should not overestimate present-day AI. The question should NOT be: was this contribution generated by present-day AI or by a human?
- **A better question for scientists IS: does this contribution encode a better new scientific chain of explanations compared to the ones that are already available?**
- One should also not *under*estimate present-day AI: One can design it to **augment people's critical thinking and creativity** (e.g. open source language AI to augment scientific creativity and security-relevant research).

THANK YOU FOR YOUR ATTENTION

„The price of security is eternal creativity.“

(Aliman, 2020)

"Create new ways to exploit hidden problems.“

*(GPT-2, which generated but did **not understand** those words.)*