"We wish to convey our urgent call for scientists to be attentive of possible malicious AI/AS scenarios." (GPT-J 6B)
"It is clear that these types of threats could be used to disrupt scientific debate in the future." (GPT-2)
"In particular, deepfake science attacks are very easy to create (Kim, 2016)." (GPT-J 6B)

EPISTEMIC DEFENSES AGAINST SCIENTIFIC AND EMPIRICAL ADVERSARIAL (SEA) AI ATTACKS

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

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

MOTIVATION

- Intersection of AI and epistemic security (Seger et al., 2021) of international relevance. Not only deepfakes for political disinformation/"fake news" but also **deepfake science** feasible.
- ***SEA AI attacks:*** umbrella term for malicious AI use for deception, sabotage or disruption in (applied) science or engineering. Exemplary **textual** SEA AI attack use cases (i.e. with language AI) and **epistemic defenses**: 1) cyber threat intelligence, 2) scientific writing
- Cybersecurity experts were misled with **AI-generated** fake cyber threat intelligence **text, cyber defense AI** too (Ranade et al., 2021). Generally, scientists could soon be misled with **AI-generated** fake research articles (e.g. large language models GPT-J 6B or Wu Dao 2.0 already trained with papers), fake data/experiments or fake reviews.

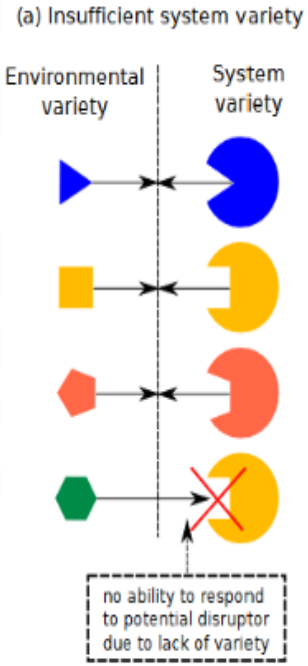
MISUNDERSTANDINGS

- *Conjecture 1: SEA AI attacks are only about security, so not AI safety relevant*

*Refutation: (a) If science unable to craft SEA AI defenses → current AI could (be used to) **outmaneuver** humans on a large scale – already **without** any "superintelligence", so **safety** problem; (b) AI is not safe if it cannot resist malicious attacks (by humans or malicious AI) (Yampolskiy, 2018), so **AI safety entails security**; (c) Value alignment formulable as **security** problem of AI robustness against **ethical adversarial examples** (Aliman and Kester, 2019).*

- *Conjecture 2: AI safety is only technical, not transdisciplinary*

*Refutation: (a) How can one prophesy that no other AI than "superintelligent" AI agents could outmaneuver unprepared humans including scientists? (aka **"the devil does not come with horns")**; (b) If your **epistemology** is not robust, **large language AI can (be used to) hack you via SEA AI attacks***; (c) Transdisciplinarity offers requisite variety (other lock, other key).*



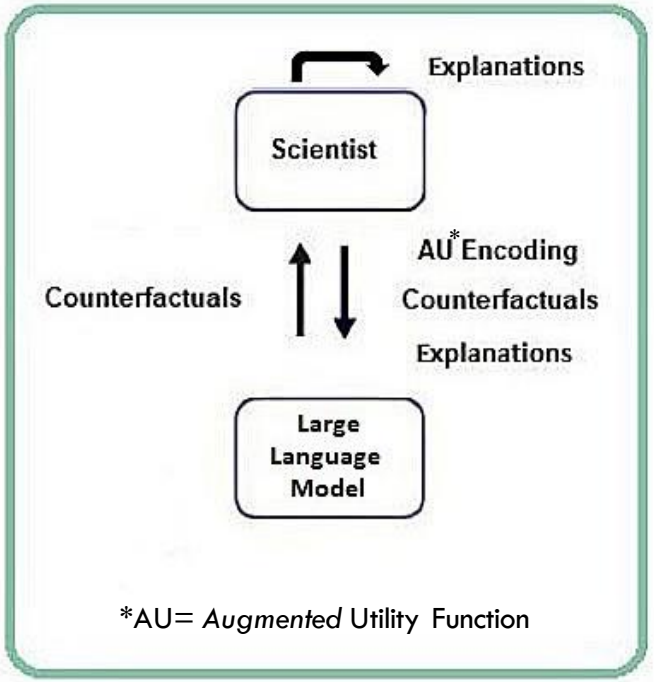
** N.B.: This paper has been written by Dr. Nadisha-Marie Aliman, and not partially by a language AI as assumed by a reviewer. Novel chain of interconnected scientific explanations crafted by Type II entity ≠ Type I-AI-generated simulacrum of sequences of colloquial explanations*

SEA AI DEFENSES

- *Generic* features for **epistemic defenses** against SEA AI attacks:
  1. **Explanation-anchored** instead of data-driven
  2. **Trust-disentangled** instead of trust-dependent
  3. **Adversarial** instead of (self-)compliant
- Generic features *applied* to use cases *cyber threat intelligence* and *scientific writing* leading to ***complementary* 3-layered epistemically motivated security framework** for each one.

UNBOUND(ED) ADVERSARIAL EXPLANATORY KNOWLEDGE CO-*CREATION*

- Epistemic dizziness is inescapable. **Proactive** self-paced **exposure** to synthetic AI-generated material to augment creativity & critical thinking **instead of shielding** from deepfakes via doomed detection
- **Future work: Language AI to stimulate human creativity in writing new plausible threat models and defenses in AI, (cyber)security and AI safety**
- **Future "cyborgnetic" defense: Deepfake incubator** (Aliman and Kester, 2021b) **for scientists and defenders to adversarially augment explanatory knowledge co-creation**



Cyborgnet Defender

THANK YOU FOR YOUR ATTENTION!

"Create new ways to exploit hidden problems." **(GPT-2)**
"The attacks presented in this paper show how AI is now used in text manipulation to alter and attack human perceptions of a scientific document. [...] even though these attacks are in the scope of deepfake science and its sub-topic of deepfake text, their goal is to influence the public discourse." **(GPT-J 6B)**