INTERNATIONAL
DIGITAL
SECURITY
FORUM
VIENNA

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

AIT AUSTRIAN INSTITUTE OF TECHNOLOGY GMBH

Trustworthy and Socially Responsible AI

IDSF 2023 – Panel – 2023-09-19

Dr. Ross King

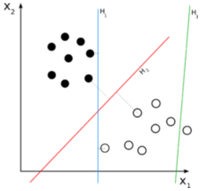
Head of Competence Unit

Data Science & Artificial Intelligence

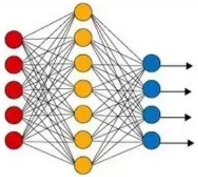
AIT Austrian Institute of Technology

ARTIFICIAL INTELLIGENCE DRIVERS

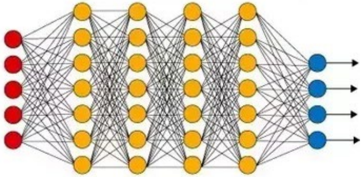
Machine Learning



Classifiers



Neural Networks



Deep Learning

HOW DID WE GET HERE? A TIMELINE

2015: OpenAI founded

2017: Google releases the seminal paper "Attention Is All You Need" *

2018: OpenAI releases GPT-1

2019: OpenAI releases GPT-2

2020: OpenAI releases GPT-3

2022: BigScience collaboration releases BLOOM

2022: 30. November: OpenAI releases ChatGPT

2023: The world of LLMs explodes…

… LLaMA, Alpaca, PaLM2, Claude, Falcon, Llama2, …

* Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

LARGE LANGUAGE MODELS – THREATS

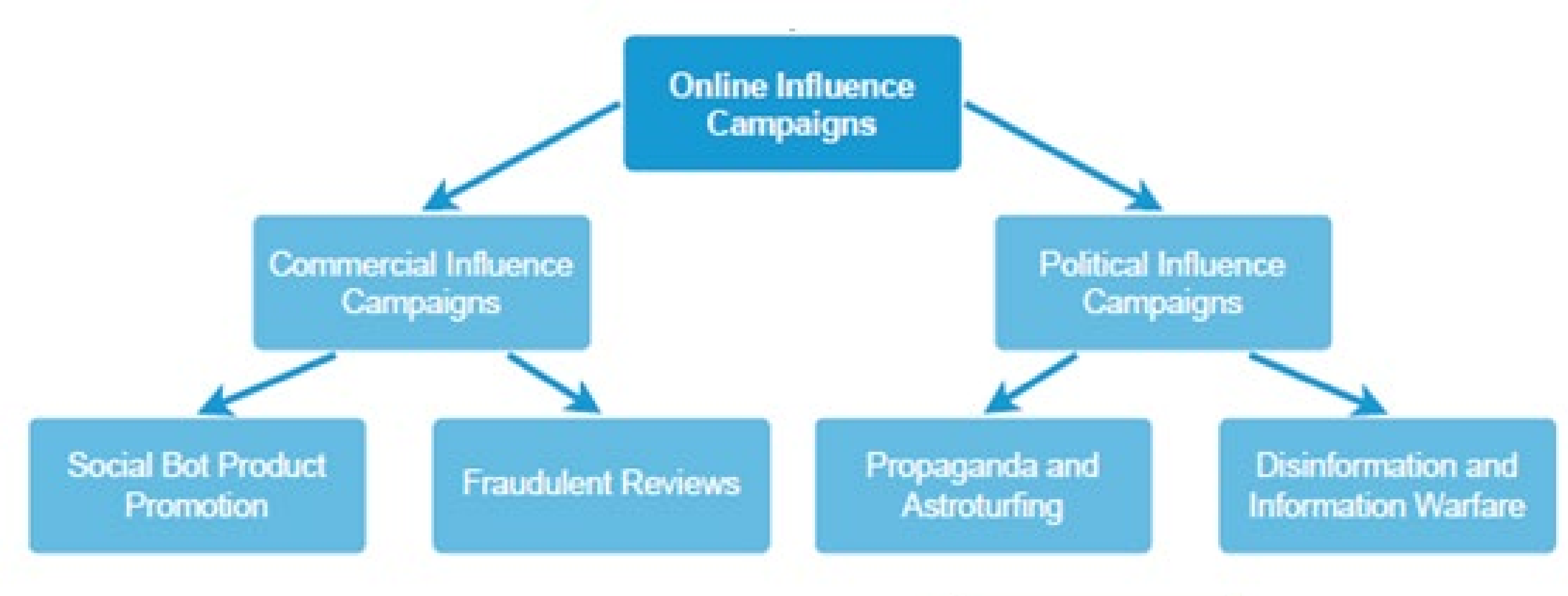

Online Influence Campaigns
Commercial Influence Campaigns
Political Influence Campaigns
Social Bot Product Promotion
Fraudulent Reviews
Propaganda and Astroturfing
Disinformation and Information Warfare

Figure 2 from: Machine Generated Text: A Comprehensive Survey of Threat Models and Detection Methods
EVAN CROTHERS, NATHALIE JAPKOWICZ, and HERNA VIKTOR
arXiv:2210.07321v2 [cs.CL] 19 Nov 2022

THE CHALLENGE: TRUSTWORTHY AND SOCIALLY RESPONSIBLE AI

AUSTRIAN INSTITUTE OF TECHNOLOGY

How do we ensure the development and deployment of AI systems that prioritize ethics, transparency, fairness, and accountability, while at the same time fostering trust and benefiting society?

- AI Act?
- Trustworthiness by design?
- …?

FAIRe-AI

Fostering Austria's Innovative strength and Research excellence in Artificial Intelligence

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

THANK YOU

Ross King

ross.king@ait.ac.at