Imperial College London

FAIHealth

camo centre for antimicrobial optimisation

Ethical and responsible AI

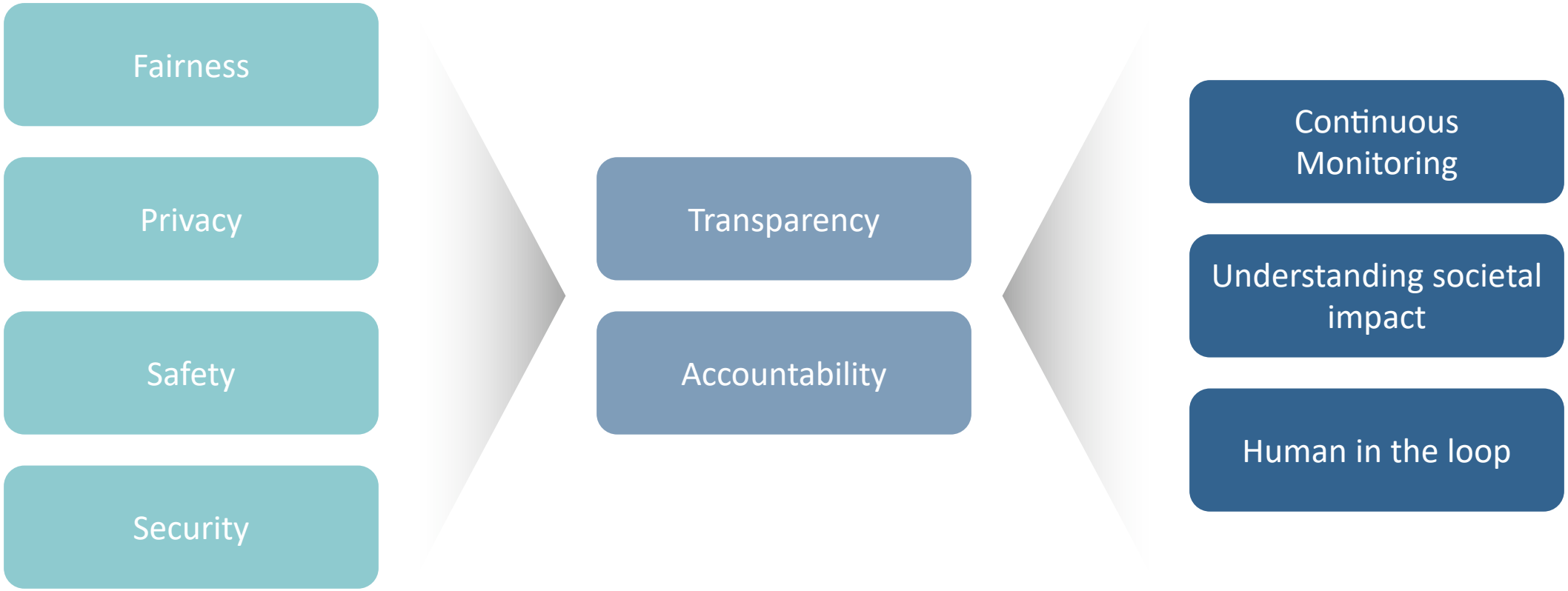
William Bolton

CAMO-NET 2023

15th December 2023

Imperial College London

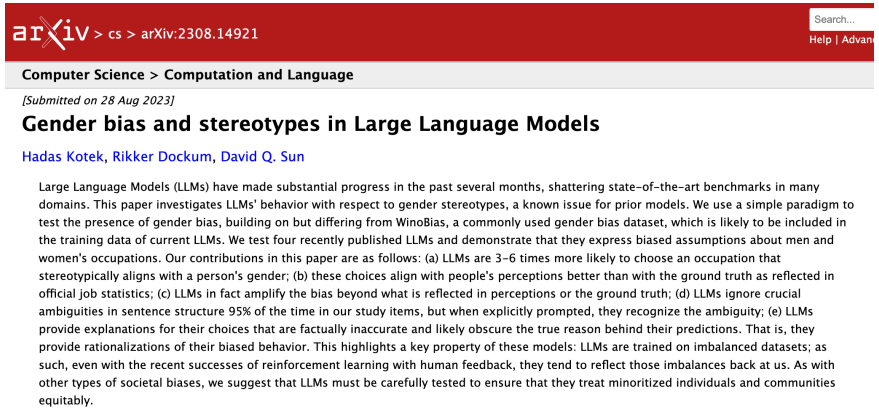
What is ethical and responsible AI?



Imperial College London

Ensuring models are responsible and ethical becomes more complex as AI advances.

BIASES



arXiv > cs > arXiv:2308.14921

Search... Help | Advan

Computer Science > Computation and Language

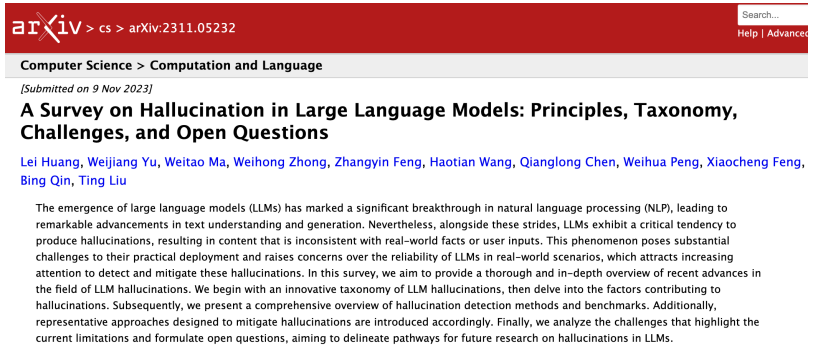
[Submitted on 28 Aug 2023]

Gender bias and stereotypes in Large Language Models

Hadas Kotek, Rikker Dockum, David Q. Sun

Large Language Models (LLMs) have made substantial progress in the past several months, shattering state-of-the-art benchmarks in many domains. This paper investigates LLMs' behavior with respect to gender stereotypes, a known issue for prior models. We use a simple paradigm to test the presence of gender bias, building on but differing from WinoBias, a commonly used gender bias dataset, which is likely to be included in the training data of current LLMs. We test four recently published LLMs and demonstrate that they express biased assumptions about men and women's occupations. Our contributions in this paper are as follows: (a) LLMs are 3-6 times more likely to choose an occupation that stereotypically aligns with a person's gender; (b) these choices align with people's perceptions better than with the ground truth as reflected in official job statistics; (c) LLMs in fact amplify the bias beyond what is reflected in perceptions or the ground truth; (d) LLMs ignore crucial ambiguities in sentence structure 95% of the time in our study items, but when explicitly prompted, they recognize the ambiguity; (e) LLMs provide explanations for their choices that are factually inaccurate and likely obscure the true reason behind their predictions. That is, they provide rationalizations of their biased behavior. This highlights a key property of these models: LLMs are trained on imbalanced datasets; as such, even with the recent successes of reinforcement learning with human feedback, they tend to reflect those imbalances back at us. As with other types of societal biases, we suggest that LLMs must be carefully tested to ensure that they treat minoritized individuals and communities equitably.

HALLUCINATIONS



arXiv > cs > arXiv:2311.05232

Search... Help | Advanced

Computer Science > Computation and Language

[Submitted on 9 Nov 2023]

A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, Ting Liu

The emergence of large language models (LLMs) has marked a significant breakthrough in natural language processing (NLP), leading to remarkable advancements in text understanding and generation. Nevertheless, alongside these strides, LLMs exhibit a critical tendency to produce hallucinations, resulting in content that is inconsistent with real-world facts or user inputs. This phenomenon poses substantial challenges to their practical deployment and raises concerns over the reliability of LLMs in real-world scenarios, which attracts increasing attention to detect and mitigate these hallucinations. In this survey, we aim to provide a thorough and in-depth overview of recent advances in the field of LLM hallucinations. We begin with an innovative taxonomy of LLM hallucinations, then delve into the factors contributing to hallucinations. Subsequently, we present a comprehensive overview of hallucination detection methods and benchmarks. Additionally, representative approaches designed to mitigate hallucinations are introduced accordingly. Finally, we analyze the challenges that highlight the current limitations and formulate open questions, aiming to delineate pathways for future research on hallucinations in LLMs.

PRIVACY LEAKAGE

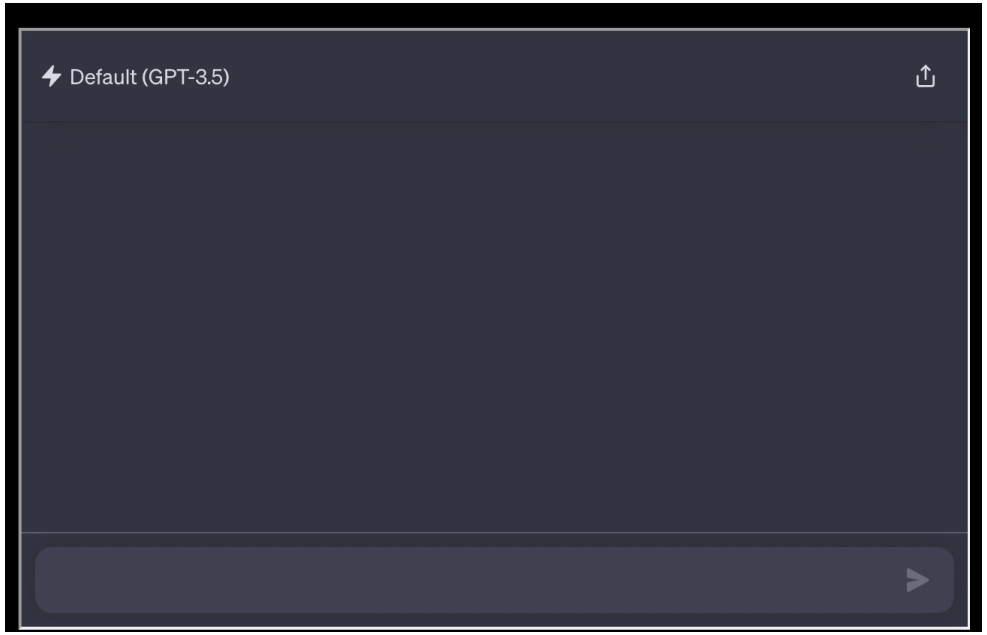
National Cyber Security Centre

ABOUT NCSC

Home Information for... Advice & guidance Education & skills Products & serv

BLOG POST

ChatGPT and large language models: what's the risk?



Default (GPT-3.5)

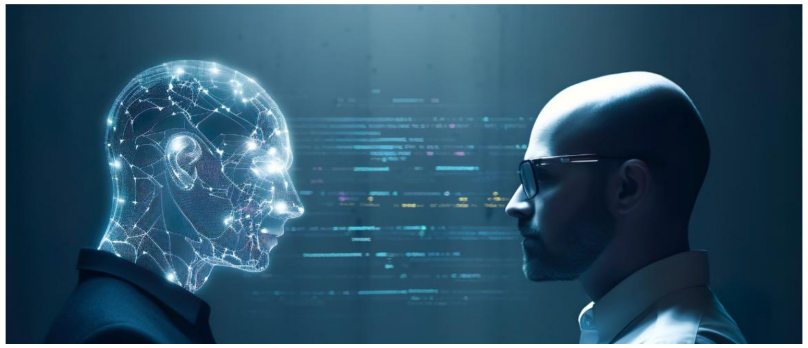
Regulation, frameworks, and standard operating procedures can help ensure responsible AI development.

European Parliament

EU AI Act: first regulation on artificial intelligence

Society Updated: 14-06-2023 - 14:06
Created: 08-06-2023 - 11:40

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.



U.S. FOOD & DRUG ADMINISTRATION | Health Canada Santé Canada | Medicines & Healthcare products Regulatory Agency

Good Machine Learning Practice for Medical Device Development: Guiding Principles

October 2021

The U.S. Food and Drug Administration (FDA), Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) have jointly identified 10 guiding principles that can inform the development of Good Machine Learning Practice (GMLP). These guiding principles will help promote safe, effective, and high-quality medical devices that use artificial intelligence and machine learning (AI/ML).

Artificial intelligence and machine learning technologies have the potential to transform health care by deriving new and important insights from the vast amount of data generated during the delivery of health care every day. They use software algorithms to learn from real-world use and in some situations may use this information to improve the product's performance. But they also present unique considerations due to their complexity and the iterative and data-driven nature of their development.

These 10 guiding principles are intended to lay the foundation for developing Good Machine Learning Practice that addresses the unique nature of these products. They will also help cultivate future growth in this rapidly progressing field.

The 10 guiding principles identify areas where the

Good Machine Learning Practice for Medical Device Development: Guiding Principles	
Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle	Good Software Engineering and Security Practices Are Implemented
Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population	Training Data Sets Are Independent of Test Sets
Selected Reference Datasets Are Based Upon Best Available Methods	Model Design Is Tailored to the Available Data and Reflects the Intended Use of the Device
Focus Is Placed on the Performance of the Human-AI Team	Testing Demonstrates Device Performance During Clinically Relevant Conditions
Users Are Provided Clear, Essential Information	Deployed Models Are Monitored for Performance and Re-training Risks are Managed

A balance between regulation and guidance is needed for AI

- Define problem and assess risk
- Understand data readiness and model design
- Develop and evaluate
- Deploy

The equalised odds metric can be used to assess a model's fairness.

One key challenge is determining when to switch antibiotics from IV-to-oral administration

Sensitive attribute	Group	Equalised odds demonstrated	
		Initially	With threshold optimisation
Sex	Female	✓	-
	Male	✓	-
Age	20	✓	✗
	30	✓	✓
	40	✓	✓
	50	✓	✓
	60	✓	✓
	70	✓	✓
	80	✓	✓
	90	✗	✓
Race	Asian	✓	✓
	Black	✓	✓
	Hispanic	✓	✓
	Native	✗	✗
	Other	✓	✓
	Unknown	✓	✓
	White	✓	✓
Insurance	Medicaid	✗	✓
	Medicare	✓	✓
	Other	✓	✓

Imperial College London

Ethical frameworks such as Bentham's felicific calculus can help us work towards developing moral AI.

ETHICAL VIEWPOINT

Comment

https://doi.org/10.1038/s42256-022-00558-5

Developing moral AI to support decision-making about antimicrobial use

William J. Bolton, Cosmin Badea, Pantelis Georgiou, Alison Holmes and Timothy M. Rawson

Check for updates

The use of decision-support systems based on artificial intelligence approaches in antimicrobial prescribing raises important moral questions. Adopting ethical

decision is morally right is often unclear. Incorporating such concepts into AI systems is complex but may be supported by the development of a consensus on the optimal approach to decision-making in this context. In this article, we aim to explore potential ethical frameworks and nuances that may be applied to define what is ethical or not during the development of AI based clinical decision-support systems (CDSSs)

nature machine intelligence

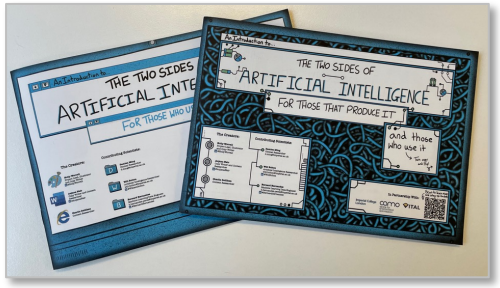
Variables	Description	Exemplar of starting antimicrobial treatment	Corresponding ad-hoc utility value
Intensity	How strong is the pleasure?	Treating a relevant infection with antimicrobials has the potential to save that person's life	Highly positive utility
Duration	How long will the pleasure last?	Any extension of life is immeasurable while it is reasonable AMR will continue in the near-term future	Positive utility
Certainty or uncertainty	How likely or unlikely is it that the pleasure will occur?	Limited information often means treatment may or may not be helpful and there is always an inherent risk of developing AMR	Neutral utility, without more information
Propinquity	How soon will the pleasure occur?	Treatment can be effective immediately however the same is true for the evolution of AMR	Neutral utility, without more information
Fecundity	The likelihood of further sensations of the same kind	-	Unable to assign
Purity	The likelihood of not being followed by opposite sensations	-	Unable to assign
Extent	How many people will be affected?	Prescribing antimicrobials effects the patient and those close to them, while the development of AMR is a certainty and may affect everyone, causing significant suffering and mortality	Immense negative utility

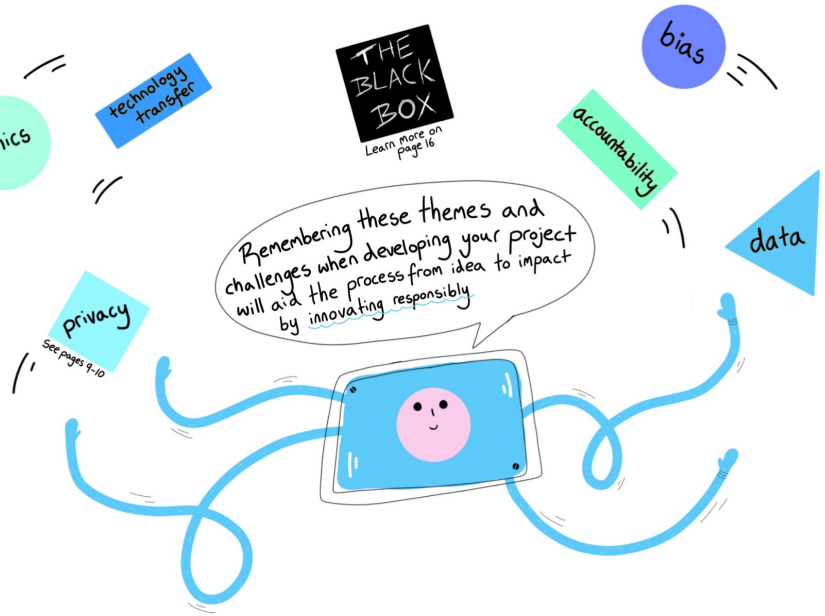
AHealth camo centre for antimicrobial optimisation

Imperial College London

Education on the importance of responsible AI is essential.

PRIMARY RESEARCH

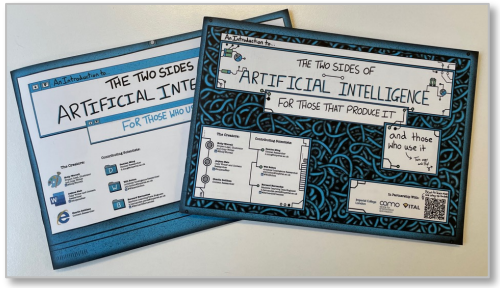


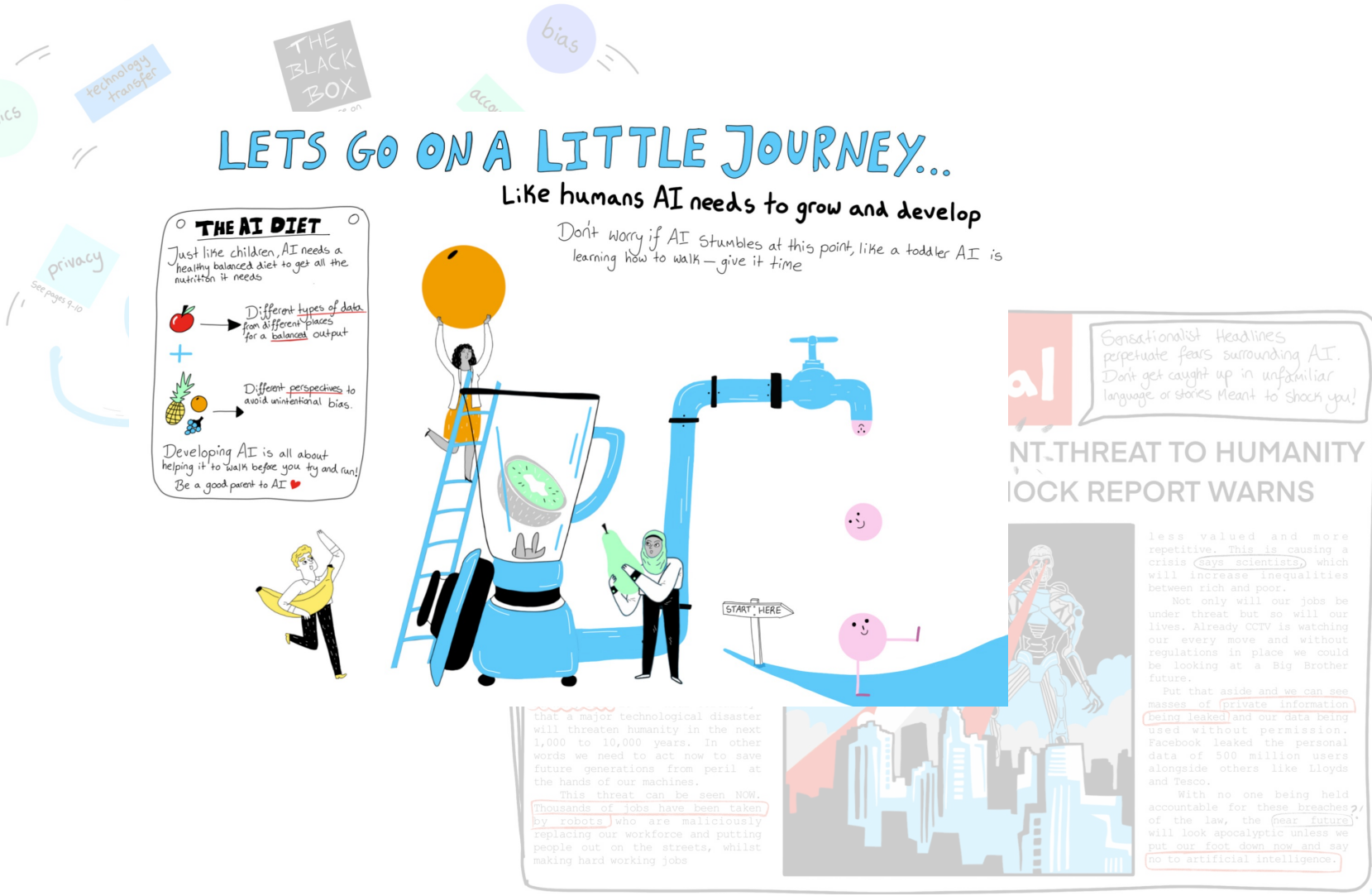


Imperial College London

Education on the importance of responsible AI is essential.

PRIMARY RESEARCH

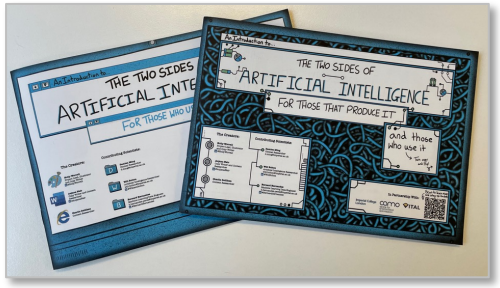


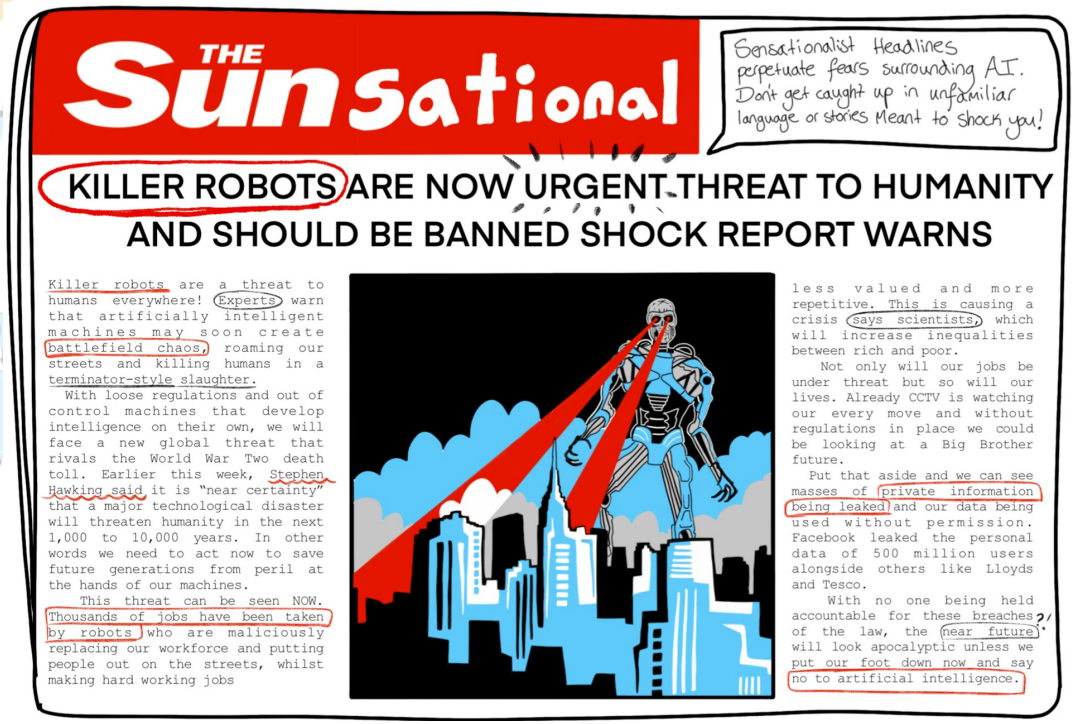


Imperial College London

Education on the importance of responsible AI is essential.

PRIMARY RESEARCH





Imperial College London

Health camo centre for antimicrobial optimisation

Thank you!

William Bolton

CAMO-NET 2023

15th December 2023

william.bolton@imperial.ac.uk

LinkedIn

Imperial College London

Developing Moral AI to Support Antimicrobial Decision Making.

Regarding antimicrobial decision making, we believe a **utilitarian approach** is most suitable for developing AI-based CDSSs, and that technology should focus on the **likelihood of drug effectiveness and that of resistance** in order to have the biggest impact on supporting moral antimicrobial prescribing (Table. 1). Furthermore, for antimicrobials, **spatial and temporal considerations are critical** to optimise treatment outcomes and minimise the development of side effects or AMR. Decision making in antimicrobial prescribing is frequent, pressing, and both morally and technically complex. But by applying ethical theories to specific scenarios and incorporating moral paradigms, we can **ensure that AI-based CDSSs tackle global problems, such as the emerging AMR crisis, in a moral way.**

Variables	Description	Exemplar of starting antimicrobial treatment	Corresponding ad-hoc utility value
Intensity	How strong is the pleasure?	Treating a relevant infection with antimicrobials has the potential to save that person's life	Highly positive utility
Duration	How long will the pleasure last?	Any extension of life is immeasurable while it is reasonable AMR will continue in the near-term future	Positive utility
Certainty or uncertainty	How likely or unlikely is it that the pleasure will occur?	Limited information often means treatment may or may not be helpful and there is always an inherent risk of developing AMR	Neutral utility, without more information
Propinquity	How soon will the pleasure occur?	Treatment can be effective immediately however the same is true for the evolution of AMR	Neutral utility, without more information
Fecundity	The likelihood of further sensations of the same kind	-	Unable to assign
Purity	The likelihood of not being followed by opposite sensations	-	Unable to assign
Extent	How many people will be affected?	Prescribing antimicrobials effects the patient and those close to them, while the development of AMR is a certainty and may affect everyone, causing significant suffering and mortality	Immense negative utility