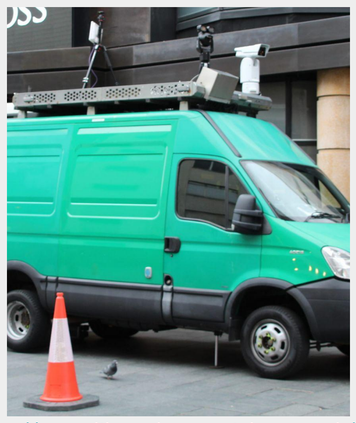
Monitoring AI Services for Misuse

Fourth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society
Seyyed Ahmad Javadi, Chris Norval, Richard Cloete, Jatinder Singh
Compliant & Accountable Systems Group, University of Cambridge
www.compacctsys.net

AI has much potential for misuse

Facial recognition cameras mounted on a vehicle



https://www.libertyhumanrights.org.uk/campaign/resist-facial-recognition/

PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

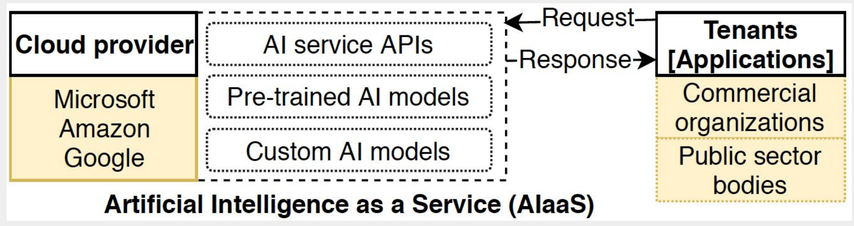
PHOTO: SIMON DAWSON/BLOOMBERG NEWS

"Criminals used artificial intelligence-based software to impersonate a chief executive's voice and demand a fraudulent transfer of €220,000 ($243,000) in March in what cybercrime experts described as an unusual case of artificial intelligence being used in hacking."

https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

AI Services can drive problematic applications

AI services provide access to pre-built models



Services include language, speech, vision, analytics, face and emotion detection, and more

Powerful AI capabilities, widely available at 'a few clicks' means AI services can easily be misused

This paper provides systematic ways forward on uncovering AI service misuse

Misuse indicators: mechanisms for discovering and alerting of certain AI service usage patterns warranting consideration or investigation

Taxonomy: supporting the formulation and implementation of misuse indicators

Domain	Dimension	Considerations
Source Information	Access level	Metadata
		Content
	Sensitivity	Sensitivity

Domain	Dimension	Considerations
Misuse analysis	Analysis type	Trait-based
		Discovery-based
	Scale	Overheads
		Tenant-specific
		Across-tenants
	Robustness	Efficacy
		Representativeness
		Circumvention

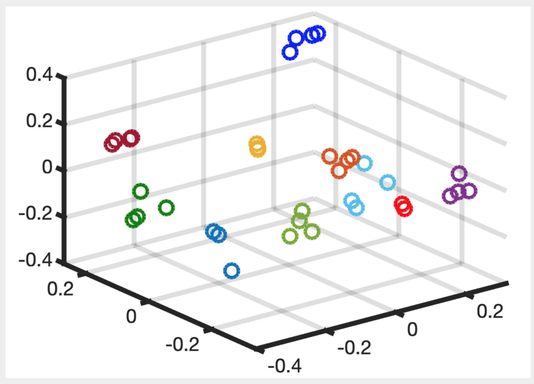
Domain	Dimension	Considerations
Record keeping	Duration	Temporary
		Permanent
	Sensitivity	Sensitivity

Supports a holistic assessment of an indicator's considerations, implications and consequences

Indicators in Practice

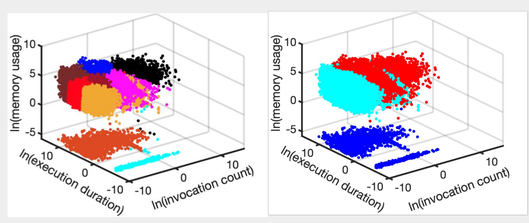
Exemplar: Surveillance through AI face services

- An AI face service that is used to detect many *different* faces over time could indicate that the service is being used to surveil
- We demonstrate that **clustering methods** show promise in identifying the number of unique faces processed by an AI face service



PCA plot of face encodings, showing each individual's (i.e. colour) encodings closely align

Exemplar: Landscaping patterns of use



Service usage patterns can be clustered to identify common and outlier behaviour

- Understanding general patterns of behaviour can *inform* as to where *attention and investigation is required*
- We explore the use of clustering methods to group customer usage patterns
- We show these can reveal **common patterns of behaviour** and those **anomalous**

The issues are contextual: indicators will vary and many implementations are possible