



# Trusting AI for Cybersecurity and Defence

## A Double-Edged Sword

Mariarosaria Taddeo

Digital Ethics Lab - Oxford Internet Institute, University of Oxford

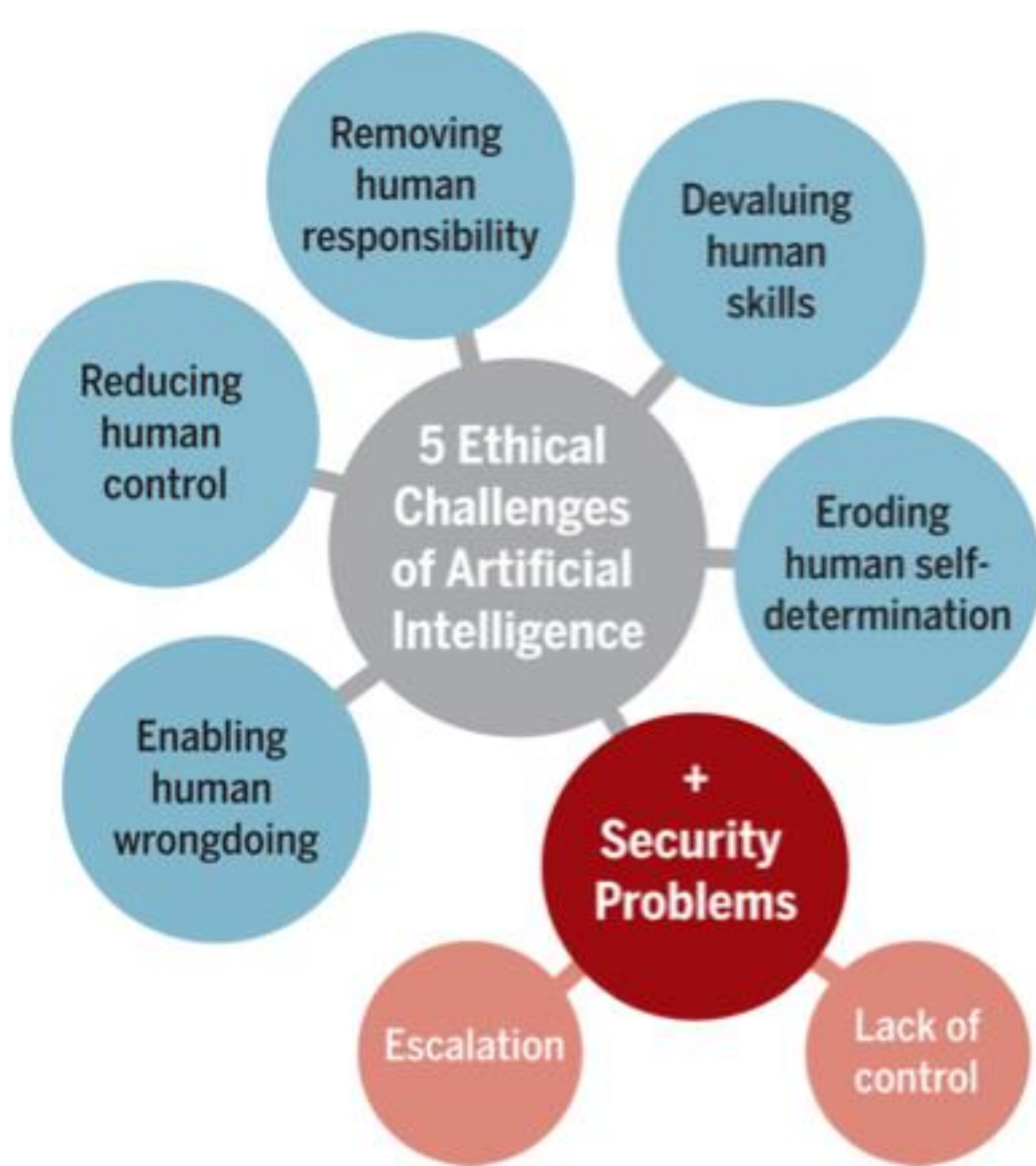
Alan Turing Institute, London

@RosariaTaddeo

AI is a growing resource of interactive,  
autonomous, and self-learning agency, which  
can be used to perform tasks that would  
otherwise require human intelligence to be  
executed successfully

AI is a growing resource of interactive,  
autonomous, and self-learning agency, which  
can be used to perform tasks that **would**  
**otherwise require human intelligence to be**  
**executed successfully**

AI is a growing resource of **interactive,**  
**autonomous, and self-learning agency,** which  
can be used to perform tasks that would  
otherwise require human intelligence to be  
executed successfully



# Ethical challenges of AI





Trust, but no control or predictability







AI failures remain  
human  
responsibilities



Pilots need to  
keep landing  
airplanes, so they  
can if AI won't



## ATTACK ORIGINS

#	Country
2367	United States
1656	China
154	Hong Kong
142	Netherlands
107	Mil/Gov
97	Canada
78	Russia
64	Taiwan
61	Thailand
56	Germany

## ATTACK TARGETS

#	Country
4505	United States
157	Hong Kong
85	Thailand
72	Canada
65	Iceland
57	Singapore
52	Portugal
50	Australia
37	Bulgaria
36	France

## ATTACKS

Timestamp	Attacker		Target		Type	
	Organization	Location	IP	Location	Service	Port
2014-06-21 08:56:27.08	Webhosting.Net	Miami, United States	67.215.180.162	Miami, United States	nethios-ns	137
2014-06-21 08:56:27.63	Hurricane Electric	Stanford, United States	184.105.139.67	Seattle, United States	snmp	161
2014-06-21 08:56:27.66	CHINANET-HN Hengyang	Changsha, China	218.77.79.43	Kirkville, United States	ftp	21
2014-06-21 08:56:27.69	Virtual Line	Lodz, Poland	217.76.117.131	Kirkville, United States	ms-term-services	3389
2014-06-21 08:56:29.09	Webhosting.Net	Miami, United States	67.215.180.162	Miami, United States	CrazyNet	17500
2014-06-21 08:56:29.13	CariNet	San Diego, United States	66.240.192.138	Kirkville, United States	unknown	7071
2014-06-21 08:56:29.16	Virtual Line	Lodz, Poland	217.76.117.131	Kirkville, United States	ms-term-services	3389

## ATTACK TYPES

#	Service	Port
1496	ssh	22
371	ms-sql-s	1433
314	http-alt	8080
307	http	80
256	domain	53
235	CrazyNet	17500
169	microsoft-ds	445

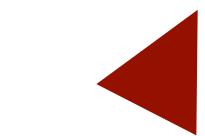


The 2019 WEF Global Risks Report ranks cyber attacks among the **top 5 sources** of severe global risks

Gemalto reports that in the first half of 2018 cyber attacks **compromised 4.5 billion records**, almost twice the amount of records (2.7 billion) compromised during the entire 2017

A Microsoft study shows **that 60% of the attacks in 2018 lasted less than an hour** and relied on new forms of malware

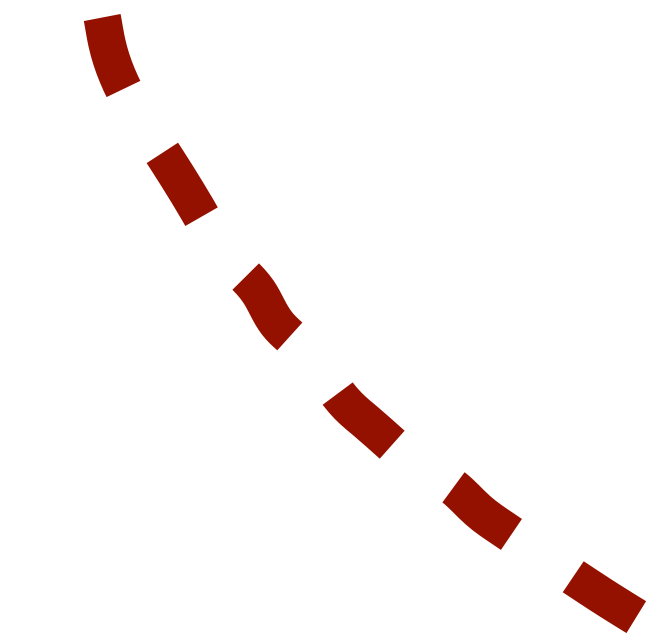
Non-kinetic cyber  
attacks easier to  
launch



Offence-  
persistent  
environment

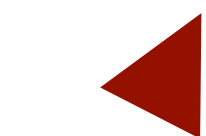


Weaponisation  
of cyberspace





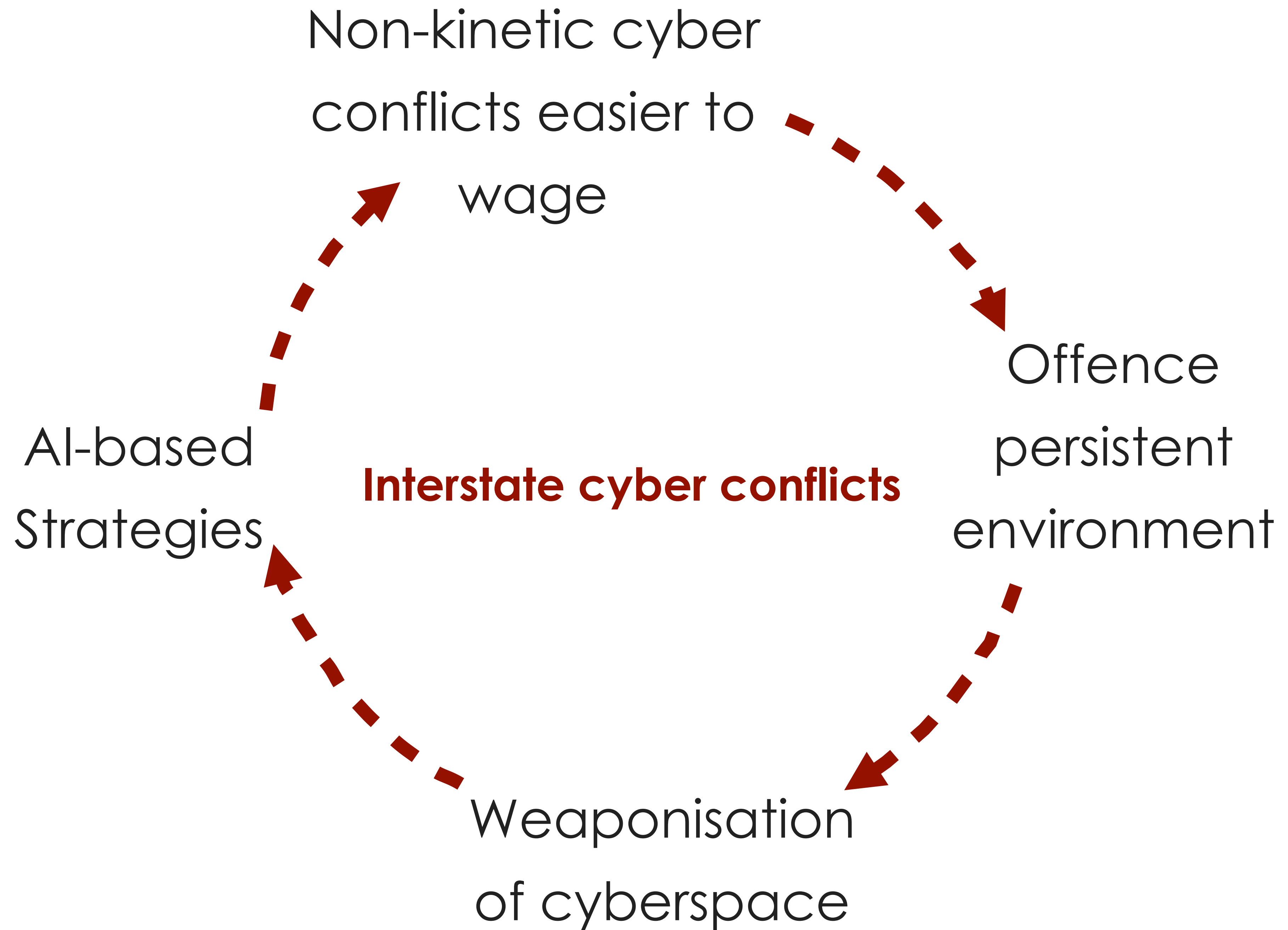
Non-kinetic cyber  
attacks easier to  
launch

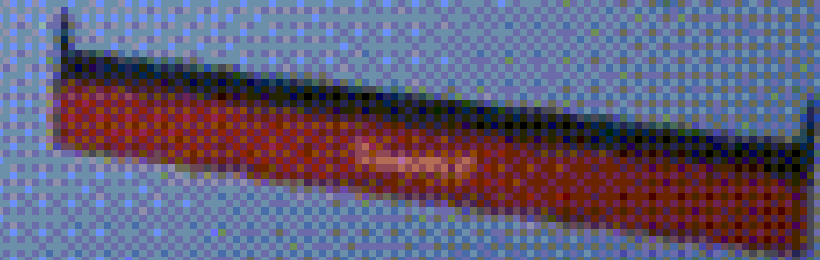
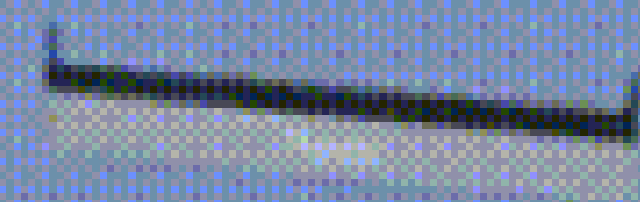
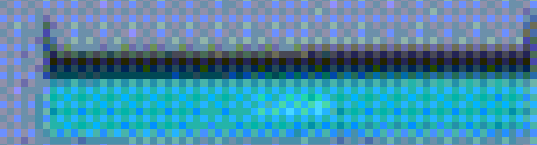
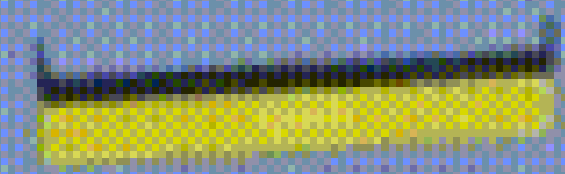
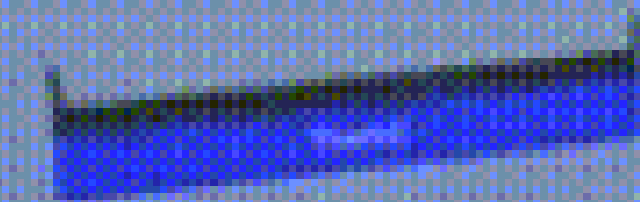
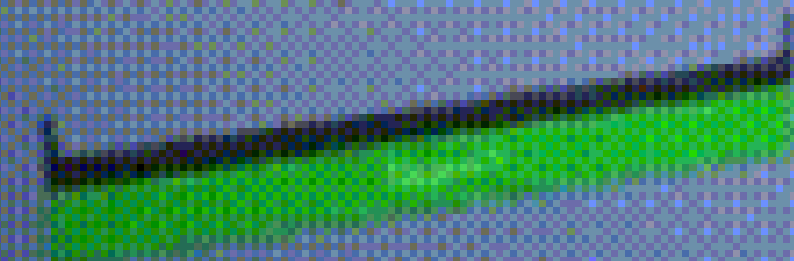
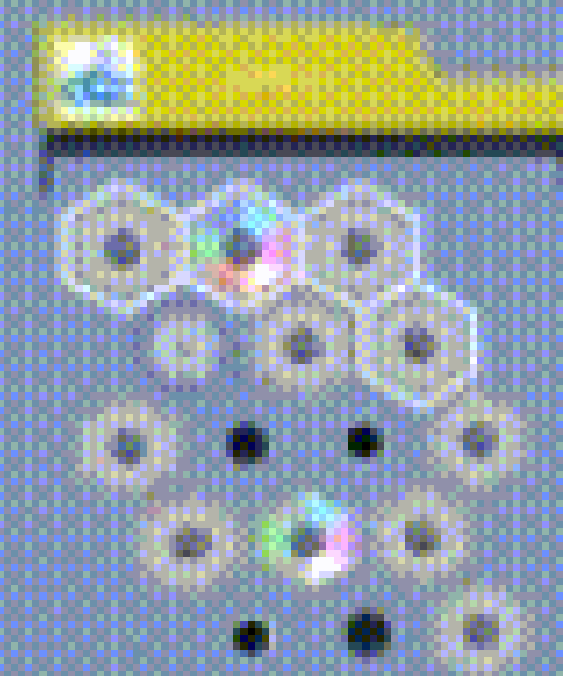
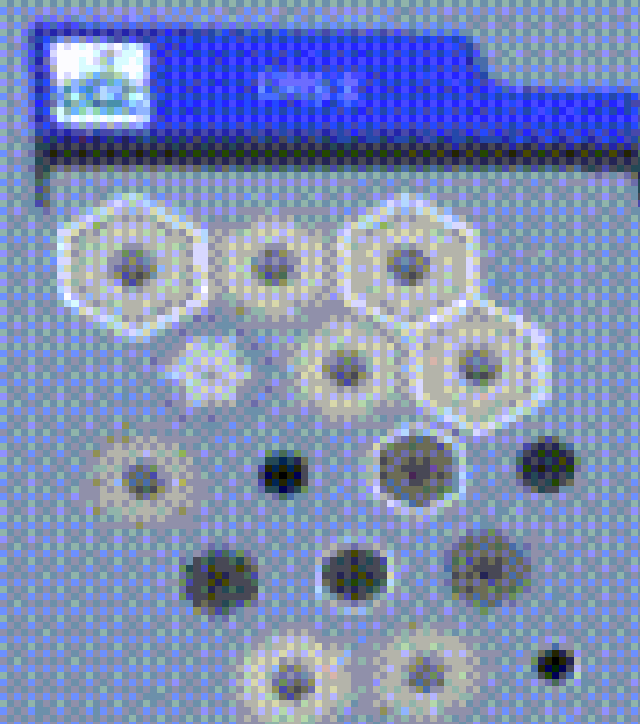


**Interstate cyber conflicts**

Offence  
persistent  
environment

Weaponisation  
of cyberspace









System **R**obustness

System **R**esponse

System **R**esilience



## Trust in AI

- US executive order on AI
- EU Commission Cybersecurity Act
- Commission's Guidelines for AI
- 2017 IEEE report on the development of standards for AI in cybersecurity

Trust is a form of delegation  
with no control

It is based on the assessment of the  
trustworthiness of the trustee

## Trust in AI

Taddeo 2010

**Trustworthiness** is a measure:

- of the predictability of the  
behaviour of the trustee
- of the risk run by the trustor, should  
the trustee behave differently



Trust is a form of delegation  
with no control

It is based on the assessment of the  
trustworthiness of the trustee

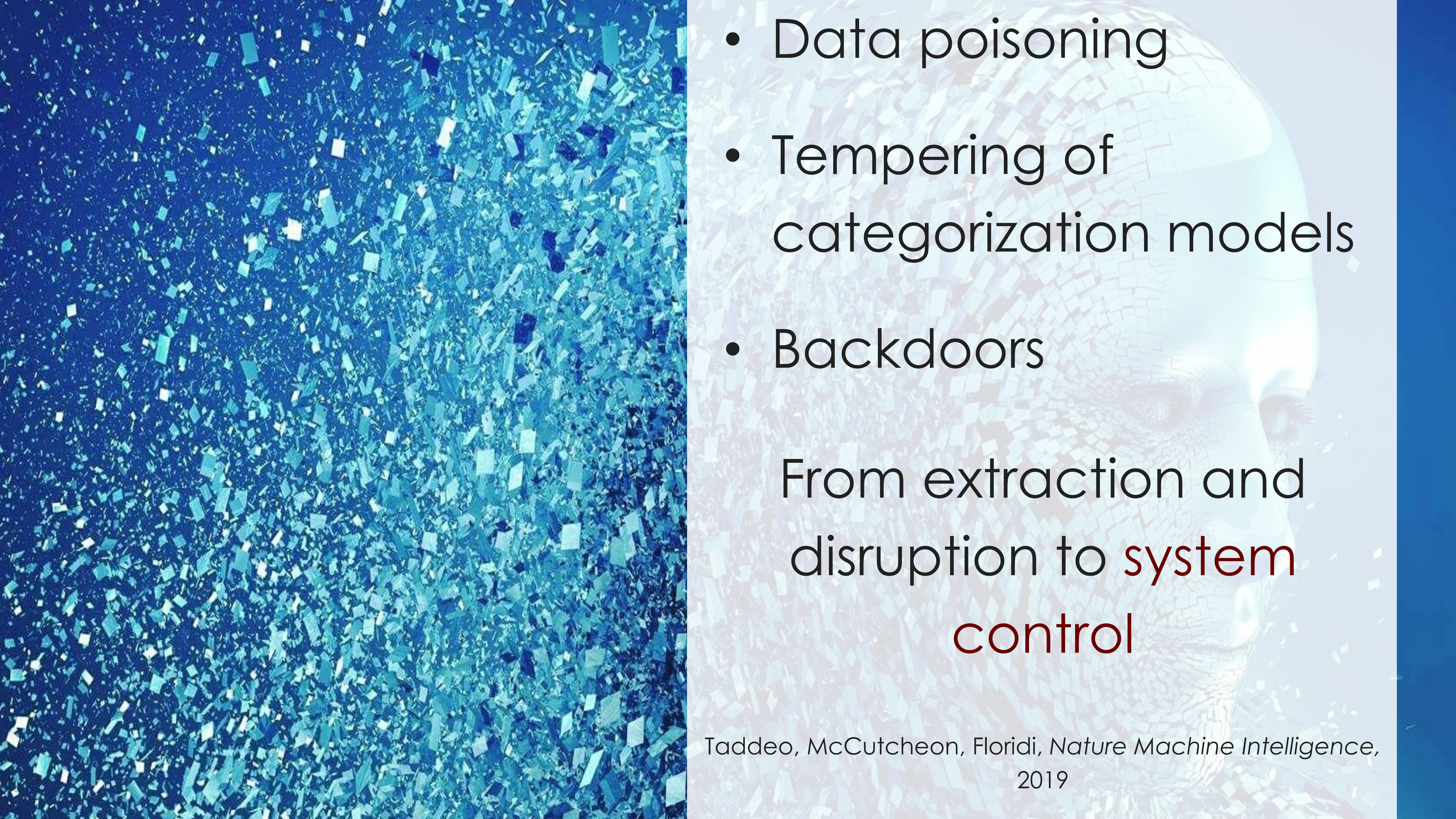
## Trust in AI

Taddeo 2010

Trustworthiness is a measure:

- of the predictability of the  
behaviour of the trustee
- of the risk run by the trustor, should  
the trustee behave differently



- 
- Data poisoning
  - Tempering of categorization models
  - Backdoors

From extraction and  
disruption to **system  
control**



# Robust AI

- ISO standard (ISO/IEC NP TR 24029-1) on robustness of neural networks
- 2019 DARPA's GARD
- US executive order on AI mandated standards for reliable, robust, and trustworthy AI systems
- China Electronics Standardization Institute established three working groups on AI



Attacks on AI can be **deceptive** and **be deceived**

E.g. a backdoor in a neural network

AI systems are **not transparent**: it is hard to understand what exactly determines a given outcome

Trusting AI in cybersecurity is conceptually misleading,  
and operationally dangerous



An aerial view of a city at sunset, with a blue network overlay of glowing nodes and lines. The text is contained within a white rounded rectangle with an orange border.

In-house development

Adversarial training

Parallel and dynamic  
monitoring



“AI is a growing resource of interactive, autonomous,  
and self-learning agency ...”

To help with the 3Rs AI must be **reliable** (**delegation + control**) rather than trustworthy

# Digital Ethics Lab



Every Bit as Good





Thank you

Mariarosaria Taddeo

Digital Ethics Lab - Oxford Internet Institute, University of Oxford

Alan Turing Institute, London

@RosariaTaddeo