AI as a Disruptive Opportunity and Challenge for Security

Antonio Kung
CTO Trialog
25 rue du Général Foy 75008 Paris
www.trialog.com

Introduction Speaker / Company

Security & privacy background

- Standardisation
  - 27550 Privacy engineering (editor)
  - 27030 Security and privacy guidelines for IoT (co-editor)
  - 27570 Privacy guidelines for smart cities (editor)
- IPEN wiki (ipen.trialog.com)
- EIP- Smart Cities and Communities
  - Citizen approach to data
- PDP4E - Privacy and Data Protection for Engineers
- Create-IoT (IoT large scale pilots)

IoT background

- Domain
  - Energy, Social and care, e-Mobility, ITS
- AOTI member
- EIP-Active healthy ageing
  - Recommendations for interoperability and standardisation (2015)

AI background

- BDVA member
- ACCRA - Agile Co-Creation of Robots for Ageing

Input to this Presentation

- Discussion with Ivo Emanuilov (KUL Citip / IMEC)
  - Poisoned AI: towards data liability. LICT Workshop Autonomous Systems, 31 May 2018
    - https://set.kuleuven.be/LICT/map-events-workshops2/lict-workshop-on-autonomous-systems
  - AI Malicious use report. February 2018
    - https://maliciousaireport.com
  - Asilomar AI Principles. January 2017
    - https://futureoflife.org/ai-principles/
  - Building ethics into AI - Kathy Baxter blog March – April 2018
    - https://medium.com/salesforce-ux/how-to-build-ethics-into-ai-part-i-bf35494cce9
    - https://medium.com/salesforce-ux/how-to-build-ethics-into-ai-part-ii-a563f3372447
- Big data value association
  - Task force 5: policy & societal implications
    - Freek Bomhof TNO, Natalie Bertels KUL Citip IMEC
  - Working group on AI transparency

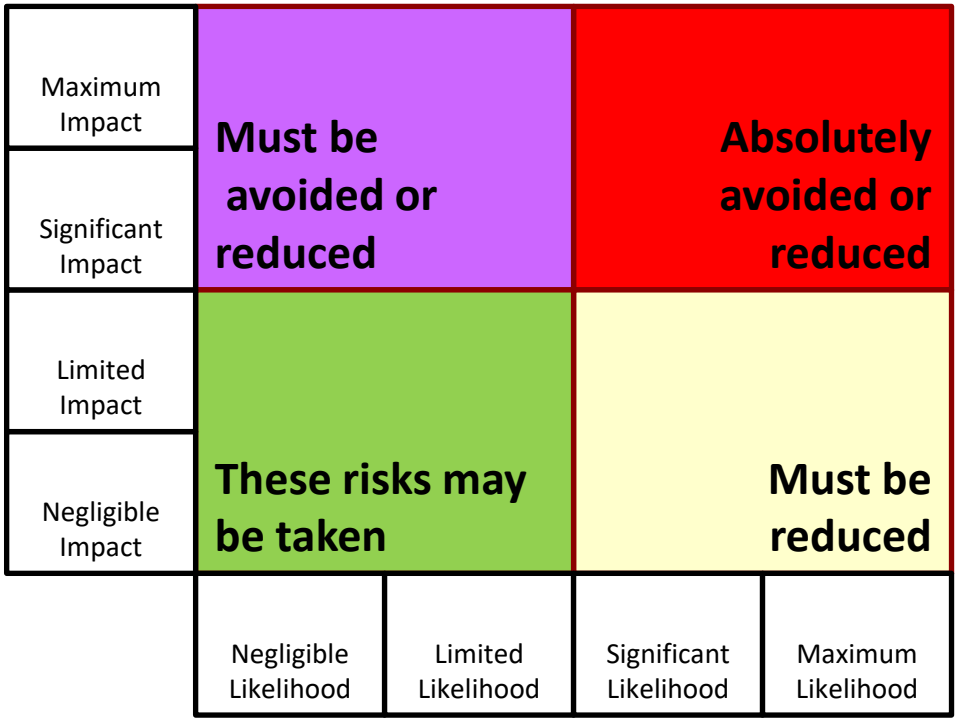
Artificial Intelligence

- Artificial intelligence
  - “intelligence” demonstrated by machines
  - mimics "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving“
- Stochastic AI (Machine learning) vs deterministic AI
  - Stefan Ticu (Rocket labs)
    - https://yseop.com/blog/artificial-intelligence-machine-learning-vs-deterministic/
  - Stochastic AI automates 100% of an activity with 70% accuracy
  - Deterministic AI automates 70% of an activity with 100% accuracy

- AI applications
  - Automatic speech recognition
  - Machine translation
  - Spam filters
  - Search engines

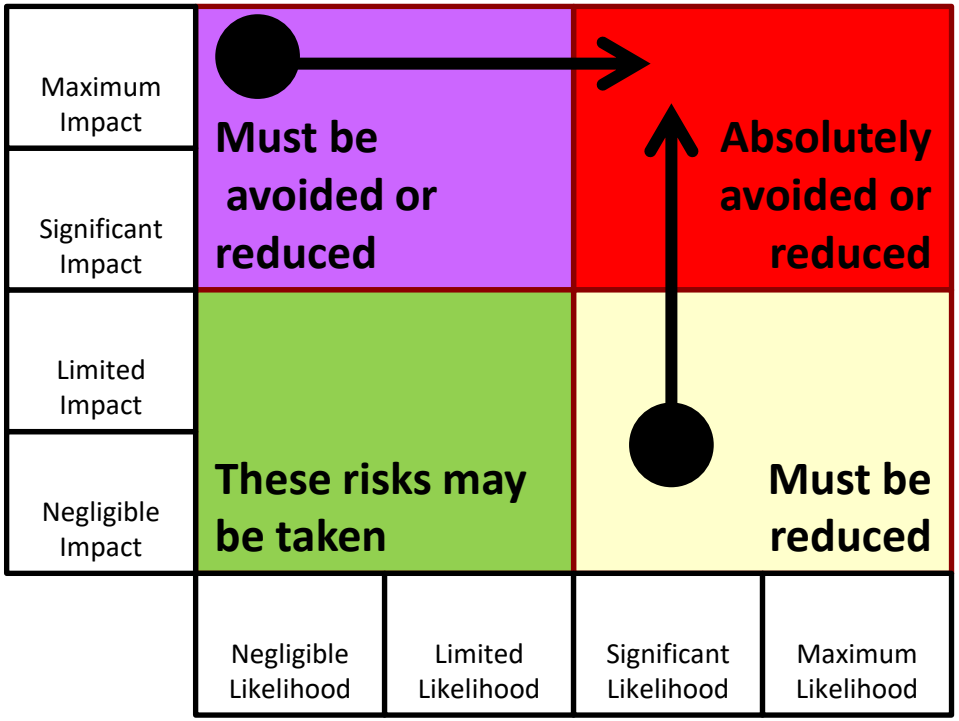
  - Autonomous cars
  - Robots for elderly people
  - Autonomous drones

Security and Privacy Risk Map



- Security and privacy threat/breach risk level:
  - Likelihood
  - Impact
- Many versions of risk maps
  - More levels
  - Different ways of calculating. Exemples
    - NIST privacy engineering
    - ETSI TVRA
  - This map is from CNIL guidelines

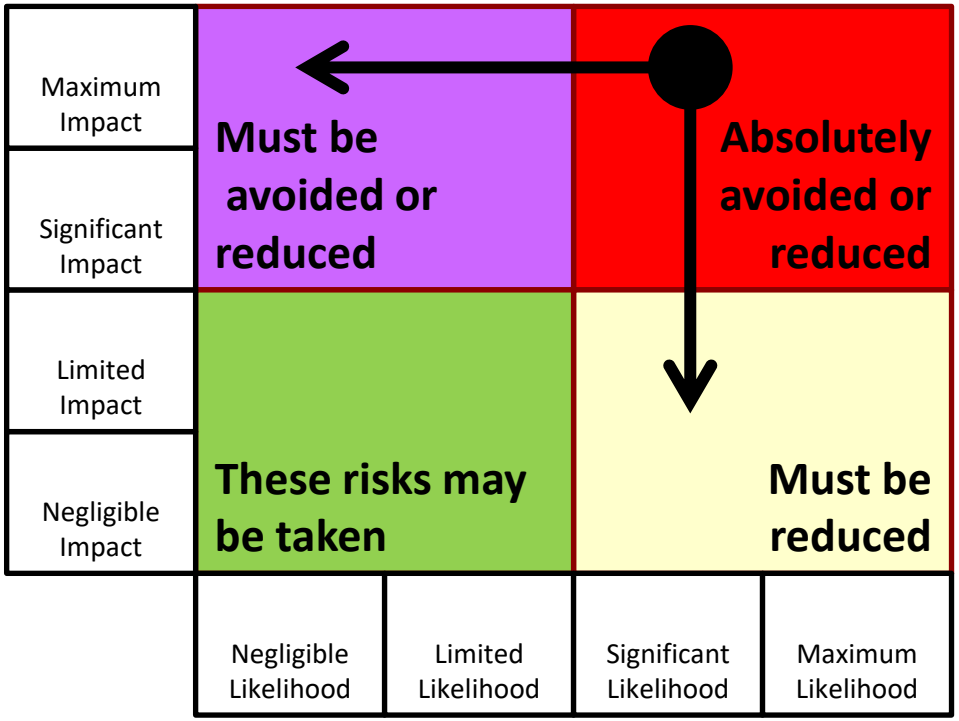
Dual use 1: Malicious AI



- A security / privacy breach is more likely to occur

- A security/ privacy breach has more impact

Dual use 2: AI to improve IoT security and privacy

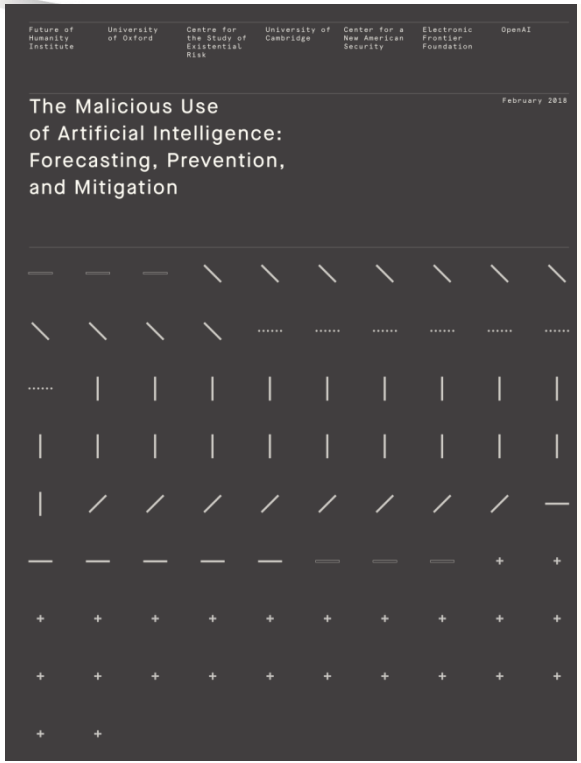


- A security / privacy breach is less likely to occur

- A security/ privacy breach has less impact

Dual Use 1
Malicious AI

Malicious AI Report



Miles Brundage[1] Shahar Avin[2] Jack Clark[3] Helen Toner[4] Peter Eckersley[5] Ben Garfinkel[6] Allan Dafoe[7] Paul Scharre[8] Thomas Zeitzoff[9] Bobby Filar[10] Hyrum Anderson[11] Heather Roff[12] Gregory C. Allen[13] Jacob Steinhardt[14] Carrick Flynn[15] Seán Ó hÉigeartaigh[16] Simon Beard[17] Haydn Belfield[18] Sebastian Farquhar[19] Clare Lyle[20] Rebecca Crootof[21] Owain Evans[22] Michael Page[23] Joanna Bryson [24] Roman Yampolskiy[25] Dario Amodei[26]

1 Corresponding author miles.brundage@philosophy.ox.ac.uk Future of Humanity Institute, University of Oxford; Arizona State University

2 Corresponding author, sa478@cam.ac.uk Centre for the Study of Existential Risk, University of Cambridge

3 OpenAI

4 Open Philanthropy Project

5 Electronic Frontier Foundation

6 Future of Humanity Institute, University of Oxford

7 Future of Humanity Institute, University of Oxford; Yale University

8 Center for a New American Security

9 American University

10 Endgame

11 Endgame

12 University of Oxford/ Arizona State University/New America Foundation

13 Center for a New American Security

14 Stanford University

15 Future of Humanity Institute, University of Oxford

16 Centre for the Study of Existential Risk and Centre for the Future of Intelligence, University of Cambridge

17 Centre for the Study of Existential Risk, University of Cambridge

18 Centre for the Study of Existential Risk, University of Cambridge

19 Future of Humanity Institute, University of Oxford

20 Future of Humanity Institute, University of Oxford

21 Information Society Project, Yale University

22 Future of Humanity Institute, University of Oxford

23 OpenAI

24 University of Bath

25 University of Louisville

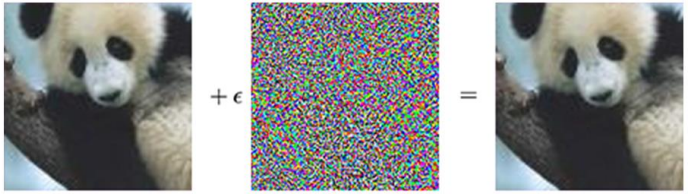
26 OpenAI

Authors are listed in order of contribution

Design Direction by Sankalp Bhatnagar and Talia Cotton

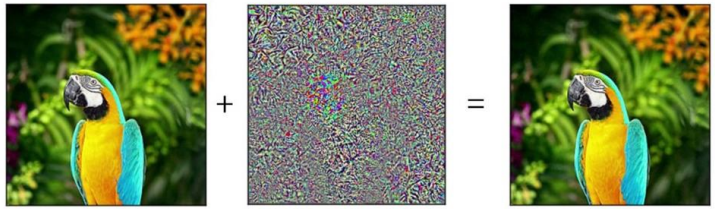
Courtesy Ivo Emanuilov (KUL – citip – Imec)

Adversarial examples and data poisoning..



"panda"
57.7% confidence

"gibbon"
99.3% confidence



X
97.3% macaw

sign($\nabla_X J(\theta, X, Y)$)

X + ϵ.sign($\nabla_X J(\theta, X, Y)$)
88.9% bookcase

3

KU LEUVEN CITIP CENTRE FOR IT & IP LAW

KU LEUVEN CITIP CENTRE FOR IT & IP LAW imec

Poisoned AI: towards data liability?

Ivo Emanuilov

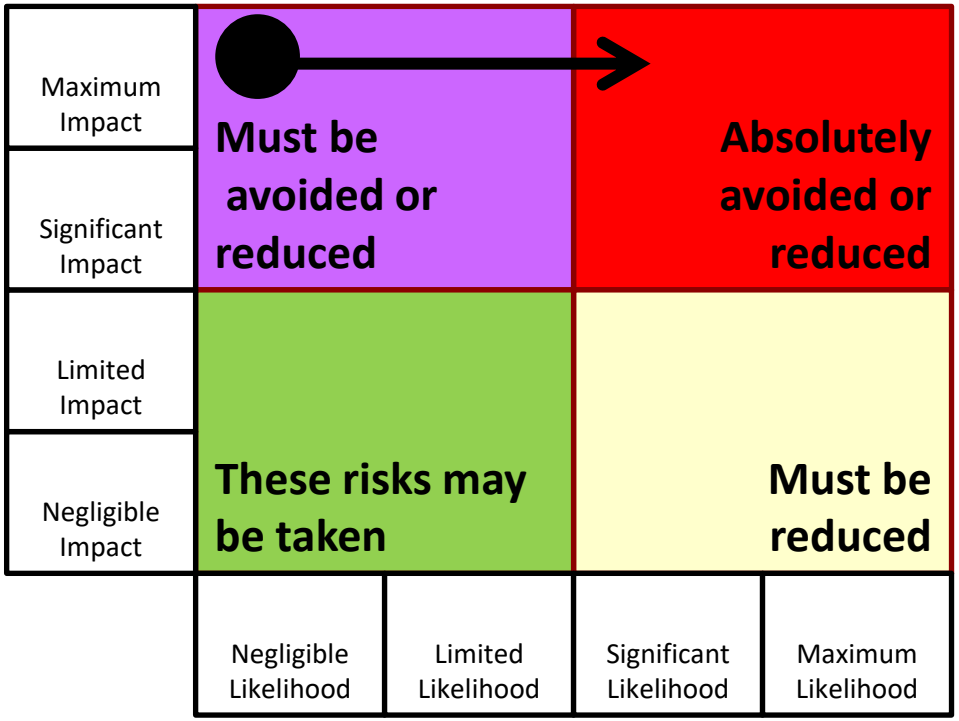
LICT Workshop Autonomous Systems, 31 May 2018

- Adversarial examples: malicious inputs to machine learning models
- Data Poisoning: Fooling the models

Malicious AI

- Expansion of existing threats
  - Expanding phishing
  - Increasing willingness to carry out attacks
    - increasing anonymity and increasing psychological distance
  - Robotics progress
- Introduction of new threats
  - Mimicking voice
  - New AI capabilities imply new threats
    - Autonomous cars VS image of a stop sign changed
    - Swarm of autonomous systems VS attack on a server to control the swarm

Malicious AI: Increasing Likelihood



(TVRA) Threat Vulnerability Risk Analysis		
Attack factor		**Malicious AI assistance**
Time	<= 1 day <= 1 week <= 1 month <= 3 months <= 6 months > 6 months	AI attack creation assistant
Expertise	Layman Proficient Expert	
Knowledge	Public Restricted Sensitive Critical	AI based learning of vulnerabilities
Opportunity	Unnecessary Easy Moderate Difficult Nont	AI based creation of opportunities
Equipment	Standard Specialised Bespoke	Lower cost
Asset Impact	Low Medium High	AI analysis of impact
Intensity	Single intensity Moderate intensity High intensity	AI based swarm attack

Malicious AI Report : Categories of threats

Digital security	Automation of social engineering attacks. Mimick a person
	Automation of vulnerability discovery. Historical patterns of code vulnerabilities are used to speed up the discovery of new vulnerabilities, and the creation of code for exploiting them
	More sophisticated automation of hacking.
	Human-like denial-of-service
	Automation of service tasks in criminal cyber-offense (payment processing / dialog with ransomware) victims
	Prioritising targets for cyber attacks using machine learning
	Criminal Training – Data poisoning
	Black-box model extraction of proprietary AI system capabilities
Physical security	Terrorist repurposing of commercial AI systems (e.g. drones)
	Endowing low-skill individuals with previously high-skill attack capabilities
	Increased scale of attacks
	Swarming attacks – Drones as weapons
	Attacks further removed in time and space
Political security	State use of automated surveillance platforms to suppress dissent
	Fake news reports with realistic fabricated video and audio
	Automated, hyper-personalised disinformation campaigns
	Automating influence campaigns
	Denial-of-information attacks
	Manipulation of information availability

Malicious AI Report : Example of measures

Digital security	Consumer awareness (e.g. education)
	Governments policies and research (e.g. incentives for source code analysis)
	Industry centralization capability (e.g. centralised spam filters)
	Attacker incentives (e.g. identifying source of attack)
	Technical cybersecurity defense (e.g. NIST based improved practice)
Physical security	Hardware manufacturers (e.g. drones)
	Hardware distributors (e.g. controlled sales)
	Software supply chain
	Robot users (e.g. using licence)
	Governments policies (e.g. restricted use of robots)
	Physical defenses (e.g. new generation of radars)
	Payload control (e.g. AI based payload analysis)
Political security	**Technical tools (e.g. fake news detection)**
	Pervasive use of security measures (e.g. more encryption)
	General interventions to improve discourse (e.g. education)
	Media platforms (e.g. integrating assessment capabilites)

Malicious AI Report : Research Topics

Learning from and with the Cybersecurity Community	**Red teaming**
	Formal verification
	Responsible disclosure of AI vulnerabilities
	Forecasting security-relevant capabilities
	Security tools
	Secure hardware
Exploring Different Openness Models	**Pre-publication risk assessment in technical areas of special concern**
	Central access licensing models
	Sharing regimes that favor safety and security
	Other norms and institutions that have been applied to dual-use technologies
Promoting a Culture of Responsibility	**Education**
	Ethical statements and standards
	Whistleblowing measures
	Nuanced narratives
Developing Technological and Policy Solutions	Privacy protection
	Coordinated use of AI for public-good security
	Monitoring of AI-relevant resources
	Other legislative and regulatory responses

Dual Use 2
AI to improve IoT security and privacy

NIST Cybersecurity Framework (input to ISO/IEC 27101)

	AI Assistance
Identify	Big data risk analysis
Protect	Pattern analysis and design
Detect	Off line anomaly analysis On line anomaly detection
Respond	Response big data analysis Training operators Assisting operations
Recover	Training operators Assisting operations

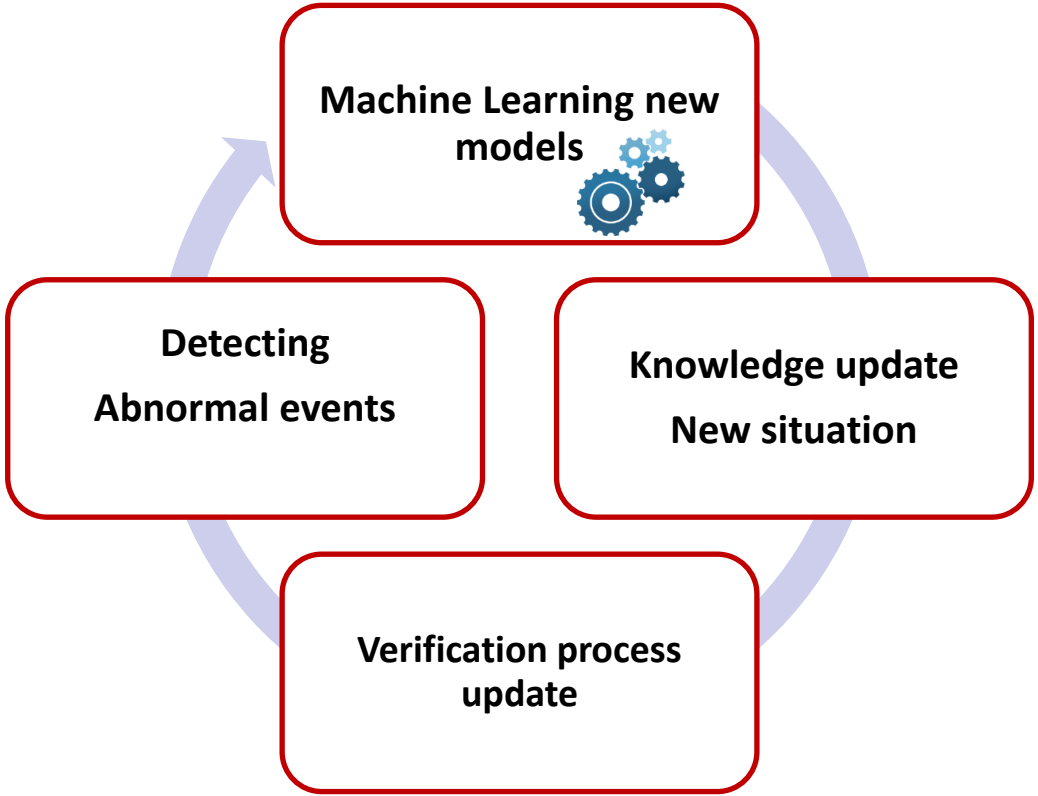
System life cycle process (ISO/IEC 15288)

Agreement process	**Acquisition**
	Supply
Organizational project-enabling processes	Life cycle model management
	Infractructure management
	Porfolio management
	Human resource management
	Quality management
	Knowledge management
Technical management process	Project planning
	Project assessment and control
	Decision management
	Risk management
	Configuration management
	Information management
	Measurement
	Quality assurance

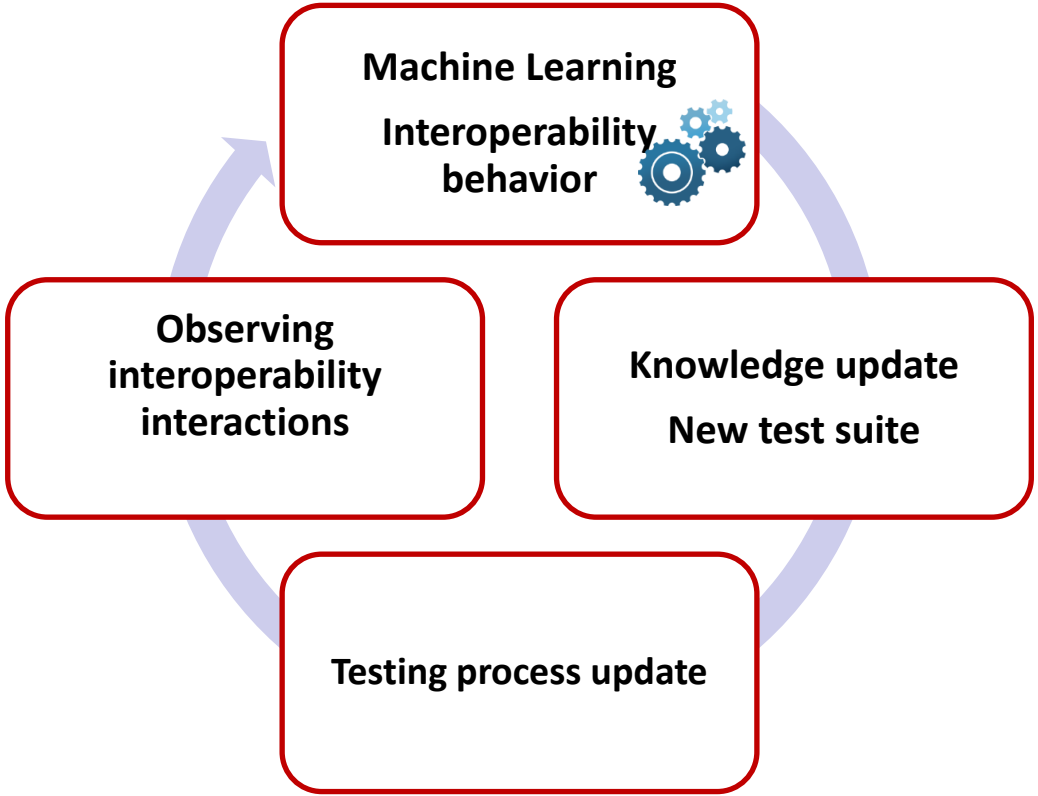
Technical processes	Business or mission analysis
	Stakeholder needs and requirements definition
	System requirements definition
	Architecture definition
	Design definition
	System analysis
	Implementation
	Integration
	Verification
	Transition
	Validation
	Operation
	Maintenance
	Disposal

- Concerns to integrate
  - Ethics impact assessment
  - Bias management
  - Transparency
- Example of ISO/IEC 27550 Privacy engineering

Example: Cybersecurity situation awareness learning



Example: Conformity Learning



Asilomar AI Principles (Beneficial AI)

Research issues	**1**	Research Goal	Create beneficial intelligence.
	2	Research Funding	AI systems robust – Growth through automation - Update legal systems with AI – Align AI with set of values
	3	Science-Policy Link	Exchange between AI researchers and policy-makers
	4	Research Culture	Cooperation, trust, and transparency among researchers and developers of AI.
	5	**Race Avoidance**	**Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.**
Ethics and Values	**6**	Safety	AI systems should be safe and secure
	7	Failure Transparency	If an AI system causes harm, it should be possible to ascertain why.
	8	Judicial Transparency	AI based judicial decision-making auditable by competent human authority.
	9	Responsibility	Designers and builders of advanced AI systems responsible
	10	Value Alignment	Autonomous AI systems goals and behaviors aligned with human values
	11	Human Values	AI systems compatible with ideals of human dignity, rights, freedoms, and cultural diversity.
	12	Personal Privacy	People control data
	13	Liberty and Privacy	Application of AI to personal data must not curtail people's liberty.
	14	Shared Benefit	AI technologies should benefit and empower as many people as possible.
	15	Shared Prosperity	The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
	16	**Human Control**	**Humans should choose how and whether to delegate decisions to AI systems**
	17	Non-subversion	Respect and improve social and civic processes on which the health of society depends.
	18	AI Arms Race	Avoiding arms race in lethal autonomous weapons
Longer-term Issues	**19**	Capability Caution	Avoid strong assumptions regarding upper limits on future AI capabilities.
	20	Importance	Advanced AI planned for and managed with commensurate care and resources.
	21	Risks	Risks posed by AI systems subject to planning and mitigation efforts commensurate with their expected impact.
	22	**Recursive Improvement**	**AI systems designed to recursively self-improve / self-replicate subject to strict safety and control measures**
	23	Common Good	Superintelligence developed in the service of widely shared ethical ideals, and for the benefit of all humanity

Principles for Ethics into AI (Kathy Baxter blog)

Create an ethical culture	**Build a Diverse Team**	**Recruit for a diversity of backgrounds and experience to avoid bias and feature gaps.**
	Cultivate an Ethical Mindset.	Ethics is a mindset, not a checklist. Empower employees to do the right thing.
	Conduct a Social Systems Analysis	Involve stakeholders at every stage of the product development lifecycles to correct for the impact of systemic social inequalities in AI data.
Be transparent	Understand Your Values	Examine the outcomes and trade-off of value-based decisions.
	Give Users Control of Their Data	**Allow users to correct or delete data you have collected about them**
	Take Feedback	Allow users to give feedback about inferences the AI makes about them.
Remove exclusion	Understand the Factors Involved	Identify the factors that are salient and mutable in your algorithm(s)
	Prevent Dataset Bias	**Identify who or what is being excluded or overrepresented in your dataset, why they are excluded, and how to mitigate it.**
	Prevent Association Bias	Determine if your training data or labels represent stereotypes (e.g., gender, ethnic) and edit them to avoid magnifying them.
	Prevent Confirmation Bias.	Determine if bias in the system is creating a self-fulling prophecy and preventing freedom of choice.
	Prevent Automation Bias	Identify when your values overwrite the user's values and provide ways for users to undo it.
	Mitigate Interaction Bias	Understand how your system learns from real-time interactions and put checks in place to mitigate malicious intent.

Prevent dataset bias

Prevent dataset bias	What	How
	Majority of data set represented by one group of users.	Cost-sensitive learning Changes in sampling methods Anomaly detection
	Statistical patterns invalid within a minority group.	Algorithms for different groups rather than one-size-fits-all.
	Categories or labels oversimplify data points and may be wrong for some percentage of the data.	**Judgement about someone is identified as fair same judgement made in a different demographic group (e.g., if a woman were a man)**
	Identify who is being excluded and the impact on your users as well as your bottom line. Context and culture matters but it may be impossible to "see" it in the data.	Identify the unknown unknowns (unidentified risks); See https://www.pmi.org/learning/library/characterizing-unknown-unknowns-6077.

Other issues

Dual use

- Trustworthiness
  - AI to help trustworthiness
    - AI-based trust framework assessment
  - AI to prevent trustworthiness
- Transparency
  - AI to help transparency
  - AI to prevent transparency
- Ethics
  - AI to help ethical impact assessment
  - AI to prevent ethical impact assessment
- Conformity
  - AI to help conformity
  - AI to prevent conformity

Life cycle process

- Integration of with model system and software enginering capabilties
  - Model driven engineering
  - Ontology

Consensus on policies

- Autonomy level definition

Recommendations

Dual Use (from Malicious AI report)

- Policy makers / Researchers collaboration
- AI researchers to address dual-use concerns
- Best practices & methods to address dual-use concerns

Lifecycle process

- Ethical impact assessment
- Ethical-by-design AI engineering

Consensus on policies

- Best available techniques
  - consensus-building with numerous stakeholders underpinned by sound techno-economic information
  - e.g. RFId or smart grid
    - http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32012D0119&from=EN

www.trialog.com

Questions?