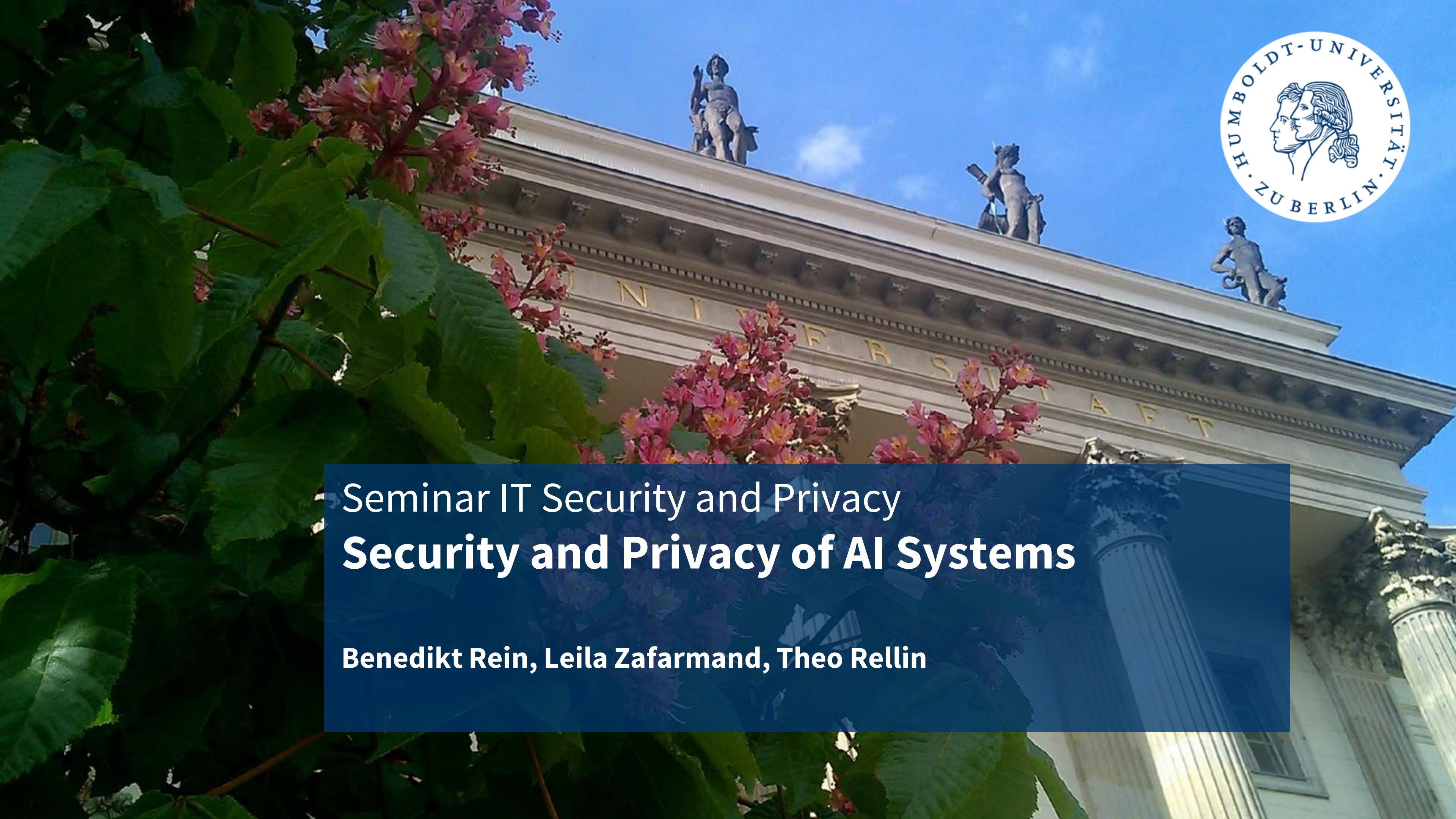


HUMBOLDT-UNIVERSITÄT ZU BERLIN

Seminar IT Security and Privacy

Security and Privacy of AI Systems

Benedikt Rein, Leila Zafarmand, Theo Rellin

Outline

1. Definitions
    a. Frameworks
    b. EU AI Act
    c. Scope and Research Question
2. Systematic Mapping Process
3. Conducted Search
4. Results & Analysis
5. Next Steps

AI Security Framework Overview

- Value based frameworks
- Lifecycle based frameworks

- Auditing and compliance frameworks
- Risk frameworks

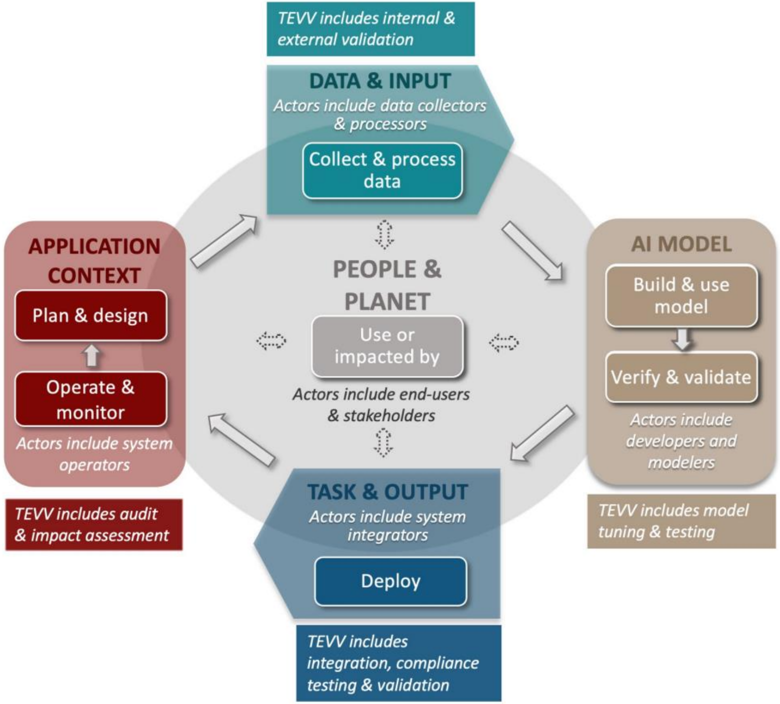
- Regulatory and standardization efforts

Specific Frameworks

- Google’s Secure AI Framework
- Microsoft Responsible AI
- National Institute of Standards and Technology (US Gov)
  - AI Risk Management Framework

- IEEE - Ethically Aligned Design
- Journal of Physics - An AI Security Framework

Frameworks - General Concept

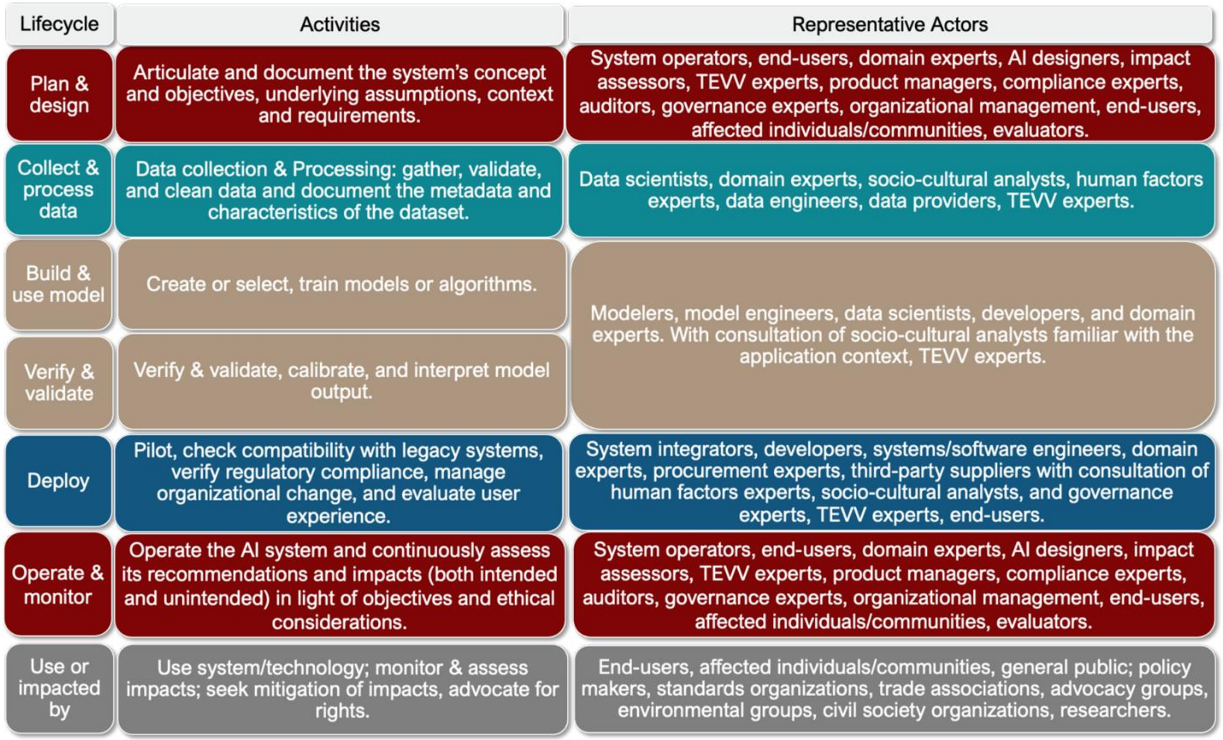


TEVV includes internal & external validation
DATA & INPUT
Actors include data collectors & processors
Collect & process data
AI MODEL
Build & use model
Verify & validate
Actors include developers and modelers
TEVV includes model tuning & testing
PEOPLE & PLANET
Use or impacted by
Actors include end-users & stakeholders
TASK & OUTPUT
Actors include system integrators
Deploy
TEVV includes integration, compliance testing & validation
APPLICATION CONTEXT
Plan & design
Operate & monitor
Actors include system operators
TEVV includes audit & impact assessment

Figure 1: Lifecycle and Key Dimensions of an AI System. Modified from OECD (2022) OECD Framework for the Classification of AI systems | OECD Digital Economy Papers. Risk management should be continuous, timely, and performed throughout the AI system lifecycle, starting with the plan & design function in the application context.

Source: AI Risk Management Framework: Second Draft (National Institute of Standards and Technology - U.S. Department of Commerce)

AI Lifecycle Approach - NIST (US Gov.)

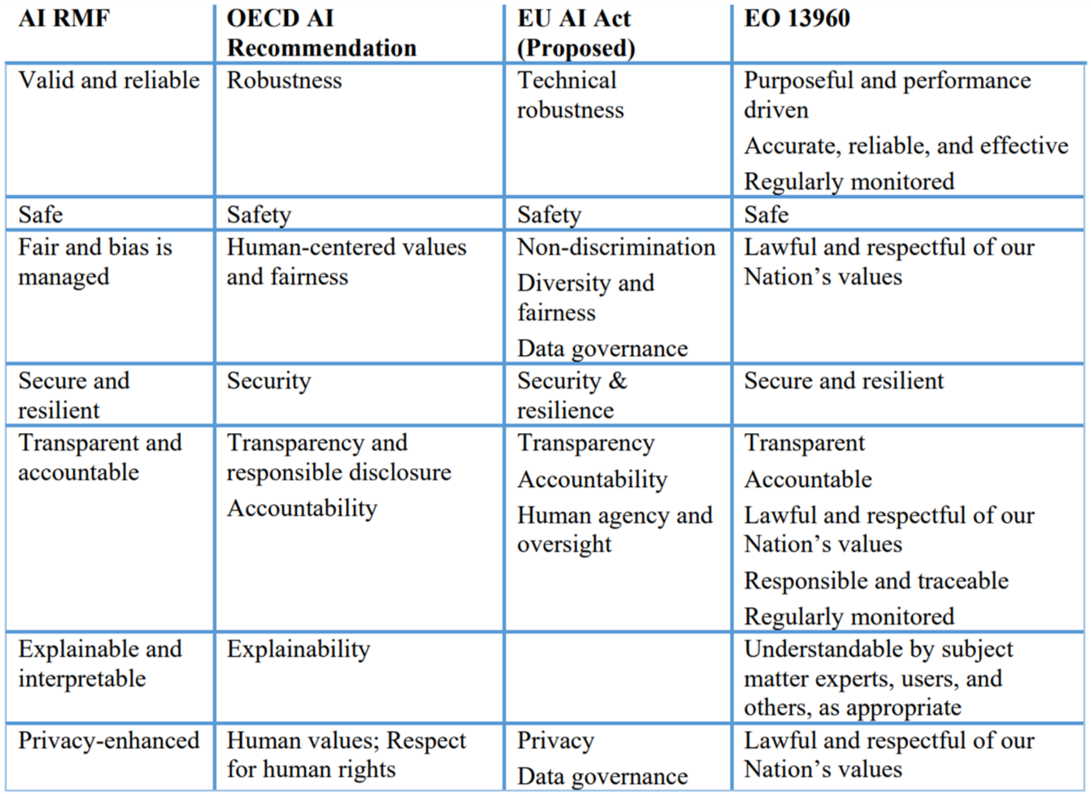


Lifecycle	Activities	Representative Actors
Plan & design	Articulate and document the system's concept and objectives, underlying assumptions, context and requirements.	System operators, end-users, domain experts, AI designers, impact assessors, TEVV experts, product managers, compliance experts, auditors, governance experts, organizational management, end-users, affected individuals/communities, evaluators.
Collect & process data	Data collection & Processing: gather, validate, and clean data and document the metadata and characteristics of the dataset.	Data scientists, domain experts, socio-cultural analysts, human factors experts, data engineers, data providers, TEVV experts.
Build & use model	Create or select, train models or algorithms.	Modelers, model engineers, data scientists, developers, and domain experts. With consultation of socio-cultural analysts familiar with the application context, TEVV experts.
Verify & validate	Verify & validate, calibrate, and interpret model output.	
Deploy	Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.	System integrators, developers, systems/software engineers, domain experts, procurement experts, third-party suppliers with consultation of human factors experts, socio-cultural analysts, and governance experts, TEVV experts, end-users.
Operate & monitor	Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives and ethical considerations.	System operators, end-users, domain experts, AI designers, impact assessors, TEVV experts, product managers, compliance experts, auditors, governance experts, organizational management, end-users, affected individuals/communities, evaluators.
Use or impacted by	Use system/technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights.	End-users, affected individuals/communities, general public; policy makers, standards organizations, trade associations, advocacy groups, environmental groups, civil society organizations, researchers.

Figure 2: AI actors across the AI lifecycle.

Source: AI Risk Management Framework: Second Draft (National Institute of Standards and Technology - U.S. Department of Commerce)

Frameworks - Value Based Approach



AI RMF	OECD AI Recommendation	EU AI Act (Proposed)	EO 13960
Valid and reliable	Robustness	Technical robustness	Purposeful and performance driven Accurate, reliable, and effective Regularly monitored
Safe	Safety	Safety	Safe
Fair and bias is managed	Human-centered values and fairness	Non-discrimination Diversity and fairness Data governance	Lawful and respectful of our Nation's values
Secure and resilient	Security	Security & resilience	Secure and resilient
Transparent and accountable	Transparency and responsible disclosure Accountability	Transparency Accountability Human agency and oversight	Transparent Accountable Lawful and respectful of our Nation's values Responsible and traceable Regularly monitored
Explainable and interpretable	Explainability		Understandable by subject matter experts, users, and others, as appropriate
Privacy-enhanced	Human values; Respect for human rights	Privacy Data governance	Lawful and respectful of our Nation's values

Source: AI Risk Management Framework: Second Draft (National Institute of Standards and Technology - U.S. Department of Commerce)

Microsoft Values and Compliance Goals

- Fairness
  - **Quality of service**
  - Allocation of resources
  - Minimization of stereotyping
- Reliability & Safety
  - General guidance
  - Failures and Remedification
  - **Ongoing monitoring and evaluation**
- Privacy & Security
  - Privacy standard compliance
  - **Security policy compliance**

- Inclusiveness
  - Accessibility standard compliance
- Transparency
  - Traceability of decision making
  - Communication with stakeholders
  - Disclosure of AI interaction
- Accountability
  - **Impact Assessment**
  - **Oversight of adverse impacts**
  - **Fit for purpose**
  - Data governance

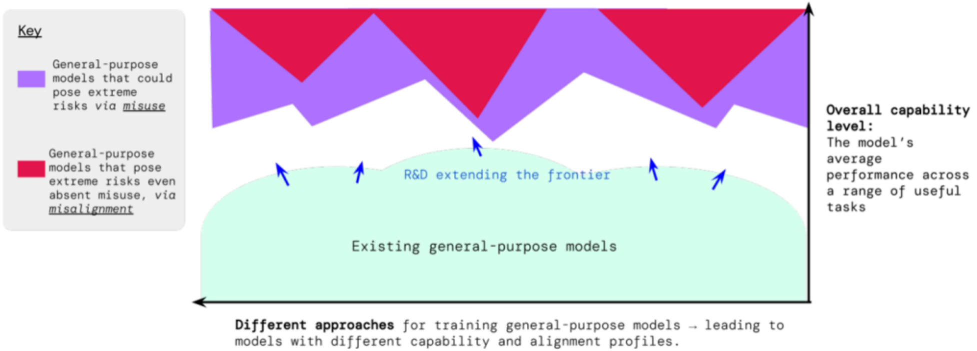
Source: Microsoft Responsible AI Standard, v2 GENERAL REQUIREMENTS

Additional Microsoft Frameworks

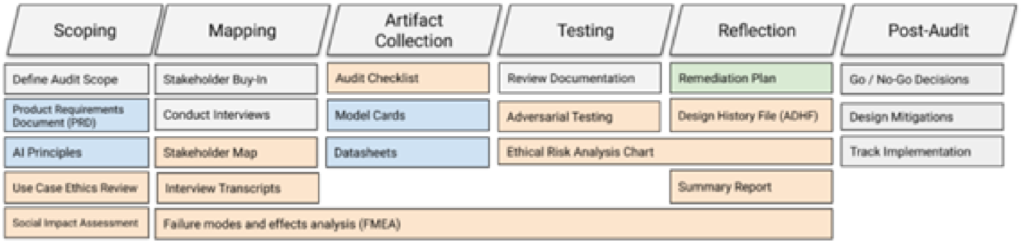
- Management Level
  - Aether oversight board and consulting
- Engineers and Practitioners
  - Responsible AI Strategy in Engineering (RAISE)
- **Impact assessment guide + templates**
  - Microsoft: all documents are WiP

Google Publications

Source: Model evaluation for extreme risk


Key
General-purpose models that could pose extreme risks via misuse
General-purpose models that pose extreme risks even absent misuse, via misalignment
R&D extending the frontier
Existing general-purpose models
Overall capability level: The model's average performance across a range of useful tasks
Different approaches for training general-purpose models → leading to models with different capability and alignment profiles.

Source: Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing



Scoping	Mapping	Artifact Collection	Testing	Reflection	Post-Audit
Define Audit Scope	Stakeholder Buy-In	Audit Checklist	Review Documentation	Remediation Plan	Go / No-Go Decisions
Product Requirements Document (PRD)	Conduct Interviews	Model Cards	Adversarial Testing	Design History File (ADHF)	Design Mitigations
AI Principles	Stakeholder Map	Datasheets	Ethical Risk Analysis Chart		Track Implementation
Use Case Ethics Review	Interview Transcripts			Summary Report	
Social Impact Assessment	Failure modes and effects analysis (FMEA)				

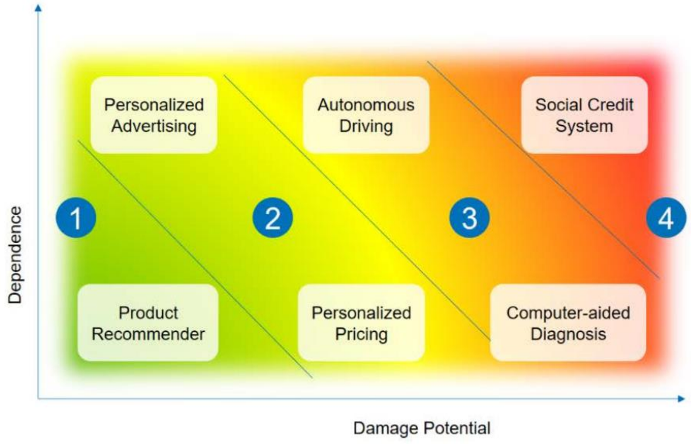
Figure 2: Overview of Internal Audit Framework. Gray indicates a process, and the colored sections represent documents. Documents in orange are produced by the auditors, blue documents are produced by the engineering and product teams and green outputs are jointly developed.

Google’s Secure AI Framework (SAIF), Based on NIST

1. Build and Expand existing security foundations
2. Extend detection and response
3. Automate defenses
4. Harmonize platform level controls
5. Adapt controls for AI deployment
6. Contextualize AI system risks

EU AI Act

- Unacceptable risk
- High risk
- Limited risk
- Generative AI
- General AI



Source: Gutachten der Datenethikkommission Kurzfassung

Preliminary enforcement ideas:

- Fines up to €30 million or 6% of global entity income
- Submitting false or misleading documentation to regulators can result in fines

EU Research and Frameworks

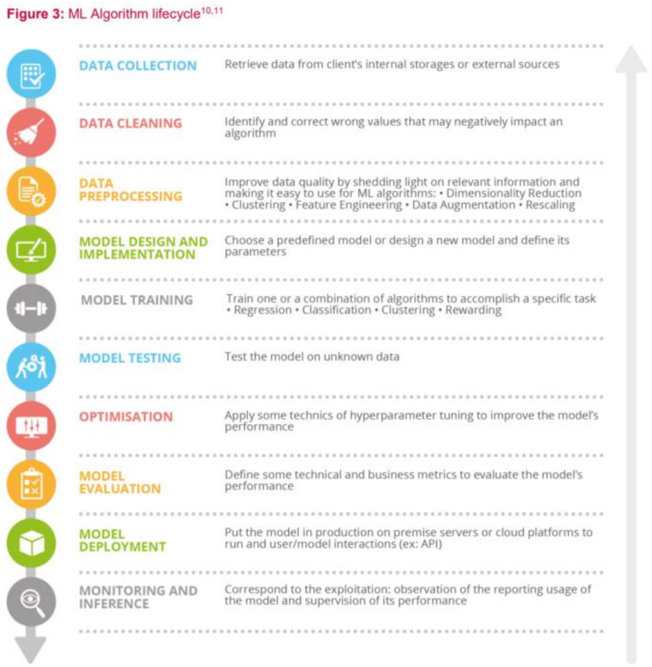
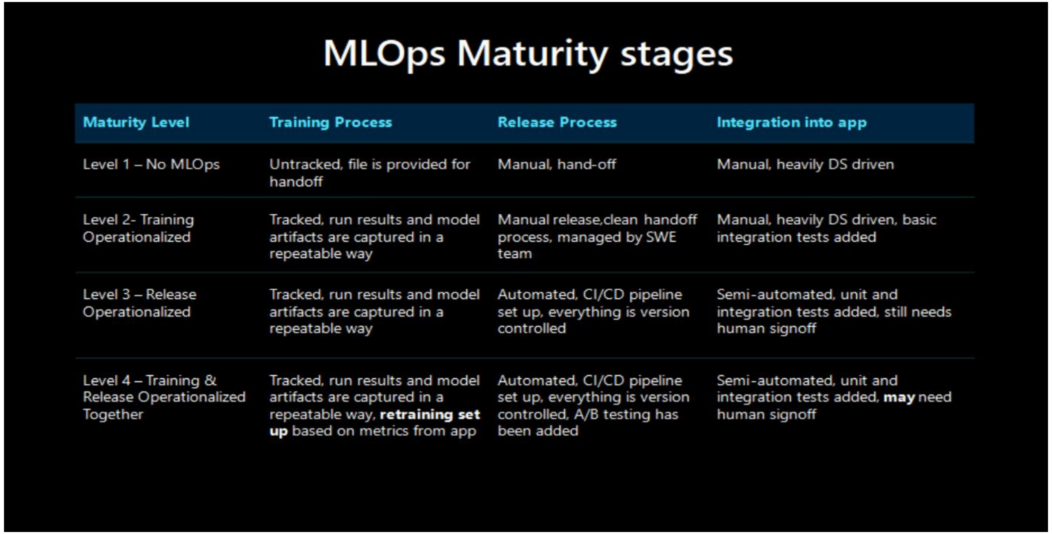


Figure 3: ML Algorithm lifecycle[10,11]

DATA COLLECTION	Retrieve data from client's internal storages or external sources
DATA CLEANING	Identify and correct wrong values that may negatively impact an algorithm
DATA PREPROCESSING	Improve data quality by shedding light on relevant information and making it easy to use for ML algorithms: • Dimensionality Reduction • Clustering • Feature Engineering • Data Augmentation • Rescaling
MODEL DESIGN AND IMPLEMENTATION	Choose a predefined model or design a new model and define its parameters
MODEL TRAINING	Train one or a combination of algorithms to accomplish a specific task • Regression • Classification • Clustering • Rewarding
MODEL TESTING	Test the model on unknown data
OPTIMISATION	Apply some technics of hyperparameter tuning to improve the model's performance
MODEL EVALUATION	Define some technical and business metrics to evaluate the model's performance
MODEL DEPLOYMENT	Put the model in production on premise servers or cloud platforms to run and user/model interactions (ex: API)
MONITORING AND INFERENCE	Correspond to the exploitation: observation of the reporting usage of the model and supervision of its performance



MLOps Maturity stages

Maturity Level	Training Process	Release Process	Integration into app
Level 1 – No MLOps	Untracked, file is provided for handoff	Manual, hand-off	Manual, heavily DS driven
Level 2- Training Operationalized	Tracked, run results and model artifacts are captured in a repeatable way	Manual release,clean handoff process, managed by SWE team	Manual, heavily DS driven, basic integration tests added
Level 3 – Release Operationalized	Tracked, run results and model artifacts are captured in a repeatable way	Automated, CI/CD pipeline set up, everything is version controlled	Semi-automated, unit and integration tests added, still needs human signoff
Level 4 – Training & Release Operationalized Together	Tracked, run results and model artifacts are captured in a repeatable way, **retraining set up** based on metrics from app	Automated, CI/CD pipeline set up, everything is version controlled, A/B testing has been added	Semi-automated, unit and integration tests added, **may** need human signoff

Source: European Union Agency for Cybersecurity: Securing ML Algorithms

Regulatory and Standardization Efforts

- Google
  - **Fostering industry support for SAIF** with the announcement of key partners and contributors in the coming months and continued industry engagement to help develop the NIST AI Risk Management Framework and ISO/IEC 42001 AI Management System Standard (the industry's first AI certification standard).

- Facebook
  - Open Loop
- NIST - defining US standards
- ISO - defining global standards

- OpenAI - public communication, information and opinion shaping
  - Or just advertisement?!

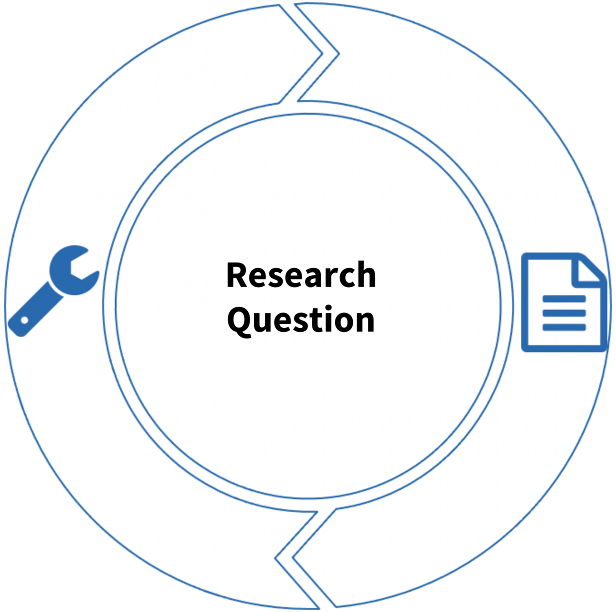
Research Question

Facebook

Google

Microsoft

Industry



Research Question

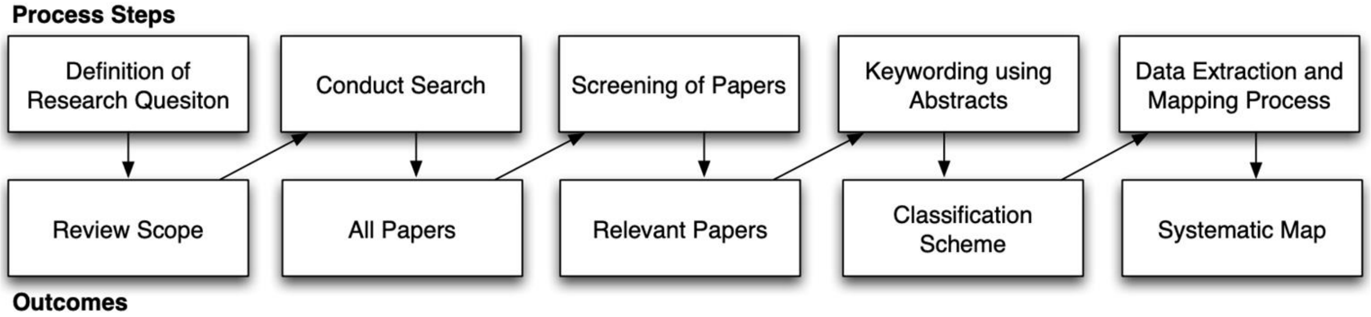
Research

Research Question

What are the key frameworks used in industry to evaluate the security and privacy implications of Artificial Intelligence systems and are they in line with research findings?

Systematic Mapping Process

1. Database Choices
2. Search Strings & Query
3. Inclusion & Exclusion Criteria
4. Final Results & Reference Management
5. Data Extraction and Visuals

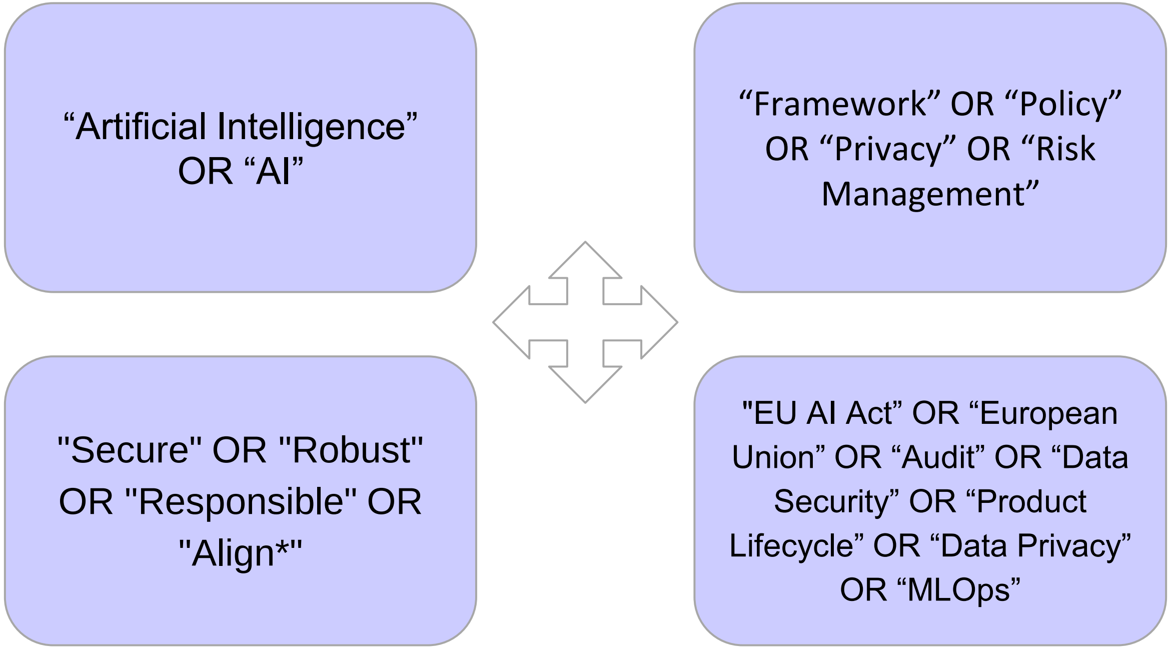


Source: Petersen et al. (2008)

Database Choices

Database	Reasons
IEEE	- high-quality papers for computer science - wide range of research papers, Journals as well as Conference Papers and Early Access Papers
ACM	- cutting-edge research - interdisciplinary perspective from various fields - global influence in this field
Ebsco Host	- provides access to seven out of eight journals in the Basket of Eight, compiled by the Association for Information Systems and containing leading information systems journals

Search Strings & Query



“Artificial Intelligence” OR “AI”

“Framework” OR “Policy” OR “Privacy” OR “Risk Management”

"Secure" OR "Robust" OR "Responsible" OR "Align*"

"EU AI Act” OR “European Union” OR “Audit” OR “Data Security” OR “Product Lifecycle” OR “Data Privacy” OR “MLOps”

Inclusion - Exclusion Criteria

Inclusion	Exclusion
● Relevance	● Language
● Recency (one paper from 1994 excluded)	● Duplicates
● Forum type (Journals, Magazines, Books, Early Access)	● Cybersecurity specific
● From Computer Science or Software Engineering Domain	● Forum type (News Articles)

Final Results & Reference Management

- iterative process, refining query, apply exclusion criteria
- from >10.000 papers to 91 final results
- used Mendeley as Reference Management-Tool
- started to tag all papers for the classification framework

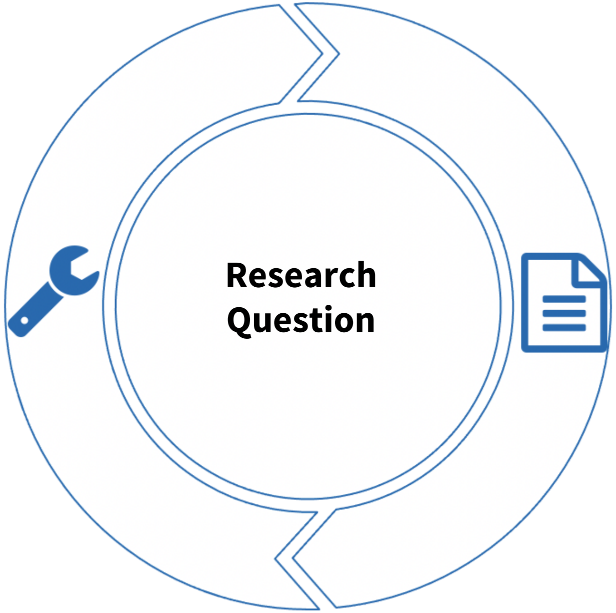
Research Question

Facebook

Google

Microsoft

Industry



Research Question

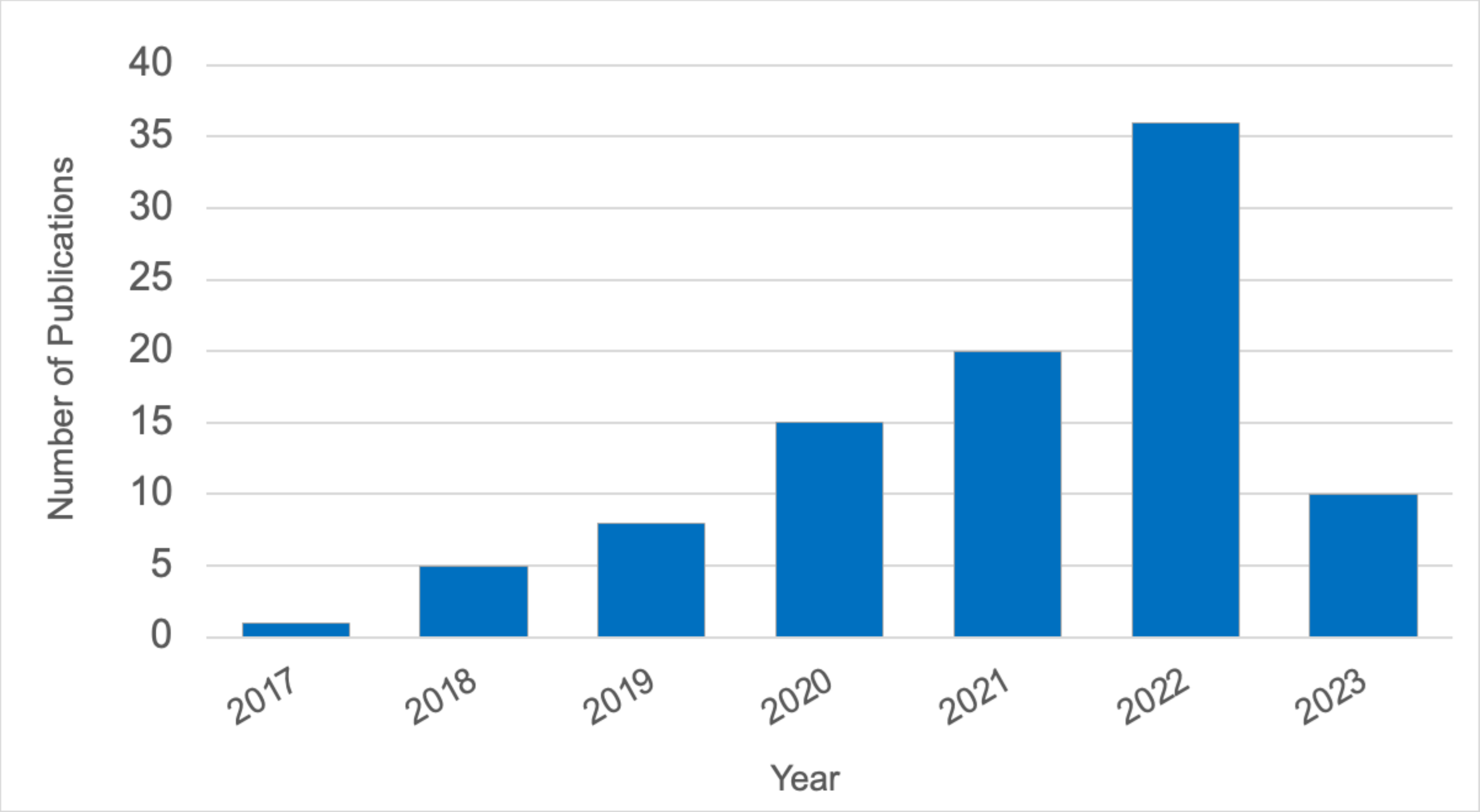
Research

IEEE

ACM

EbscoHost

Publications in Each Year

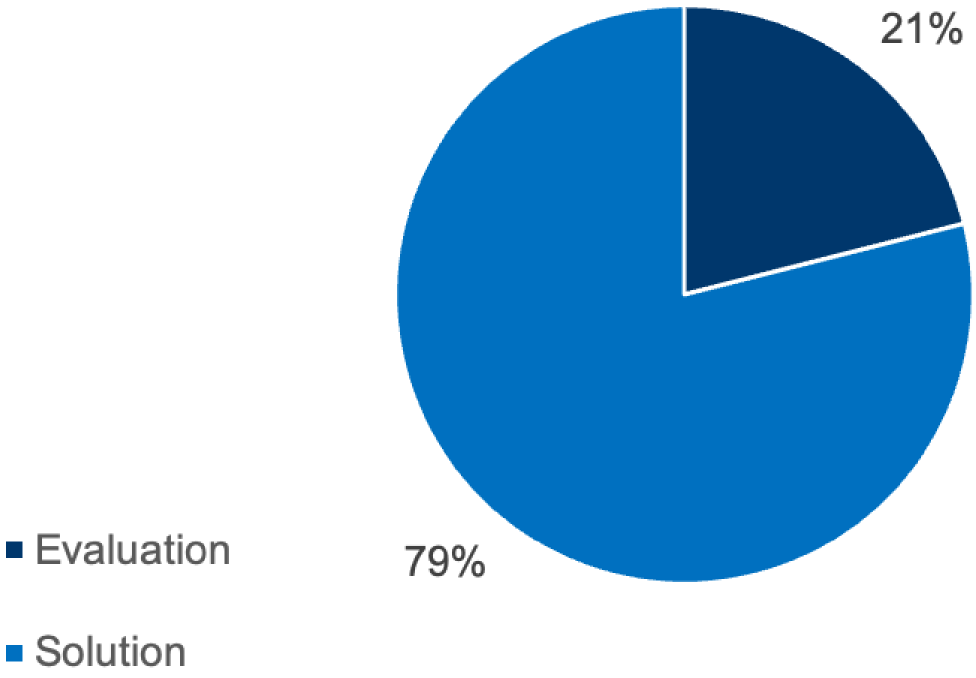


Contribution Type

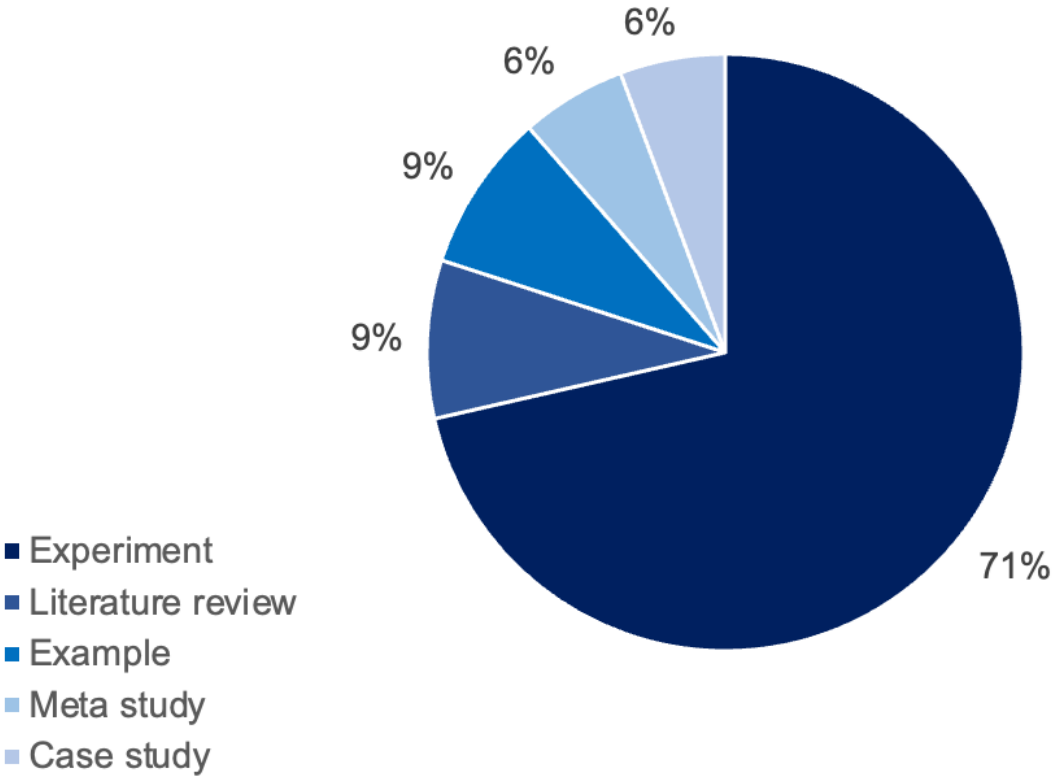
Validation Research	Investigates novel techniques that have not been implemented in practice. It employs rigorous methods to study the properties of the proposed solution.
Evaluation Research	Examines the implementation and consequences of a technique. Assesses benefits, drawbacks and impact of implemented technique.
Solution Proposal	Propose a novel or significantly improved technique with supporting arguments, without a full blown validation.
Philosophical Papers	Presents a new perspective or conceptual framework to understand existing phenomena. It offers a different way of looking at things.
Opinion Papers	The paper contain the author's opinion about what is wrong or good about something, how we should do something, without relying on related work and research methodologies.
Experience Papers	Experience papers explain on what and how something has been done in practice. It has to be the personal experience of the author.

Petersen et al. (2008)

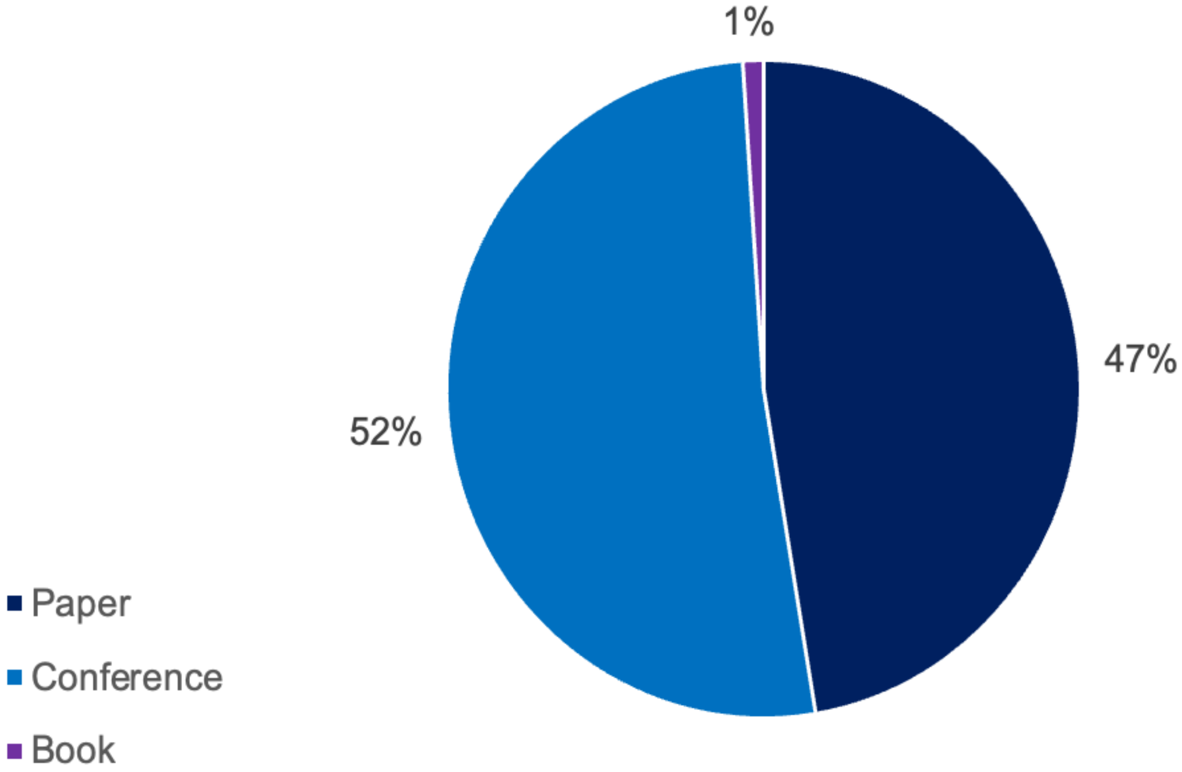
Classification Scheme - Contribution Type



Classification Scheme - Evaluation Method



Types of Publication



Next Steps

- Classification Framework: Mapping values/lifecycle to abstracts
- Systematic Map based on Classification Framework
- More detailed work with papers for discussion

Thank you for your attention!