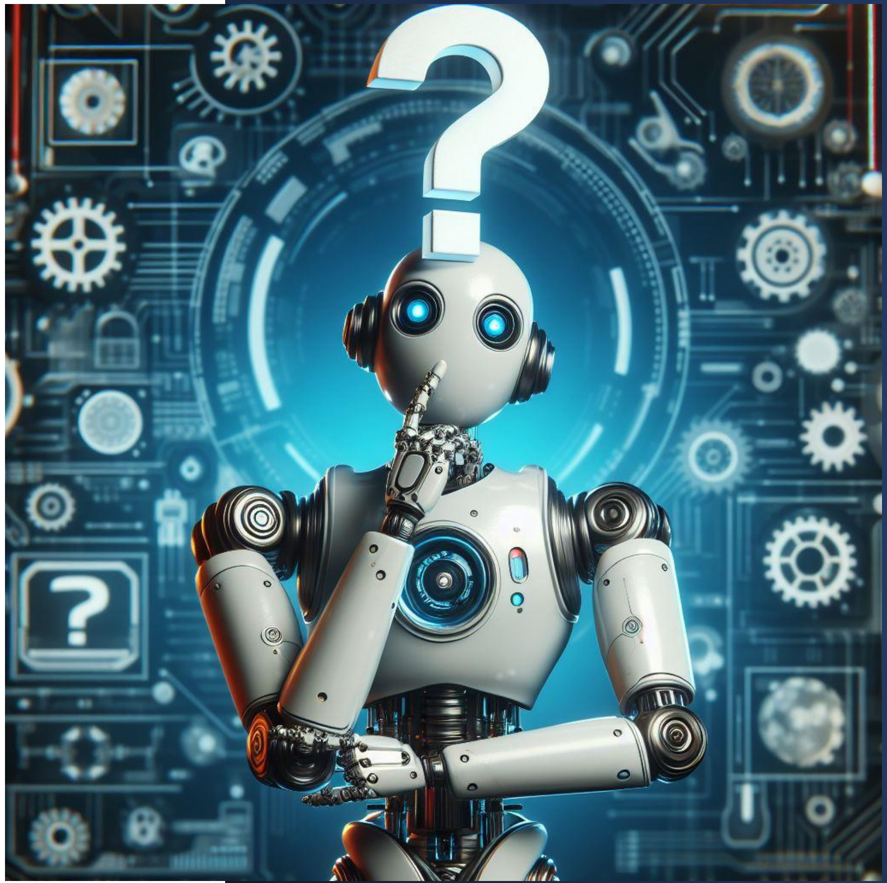
BEARS TEACH

A Simplified AI Threat Assessment Method

The Problem

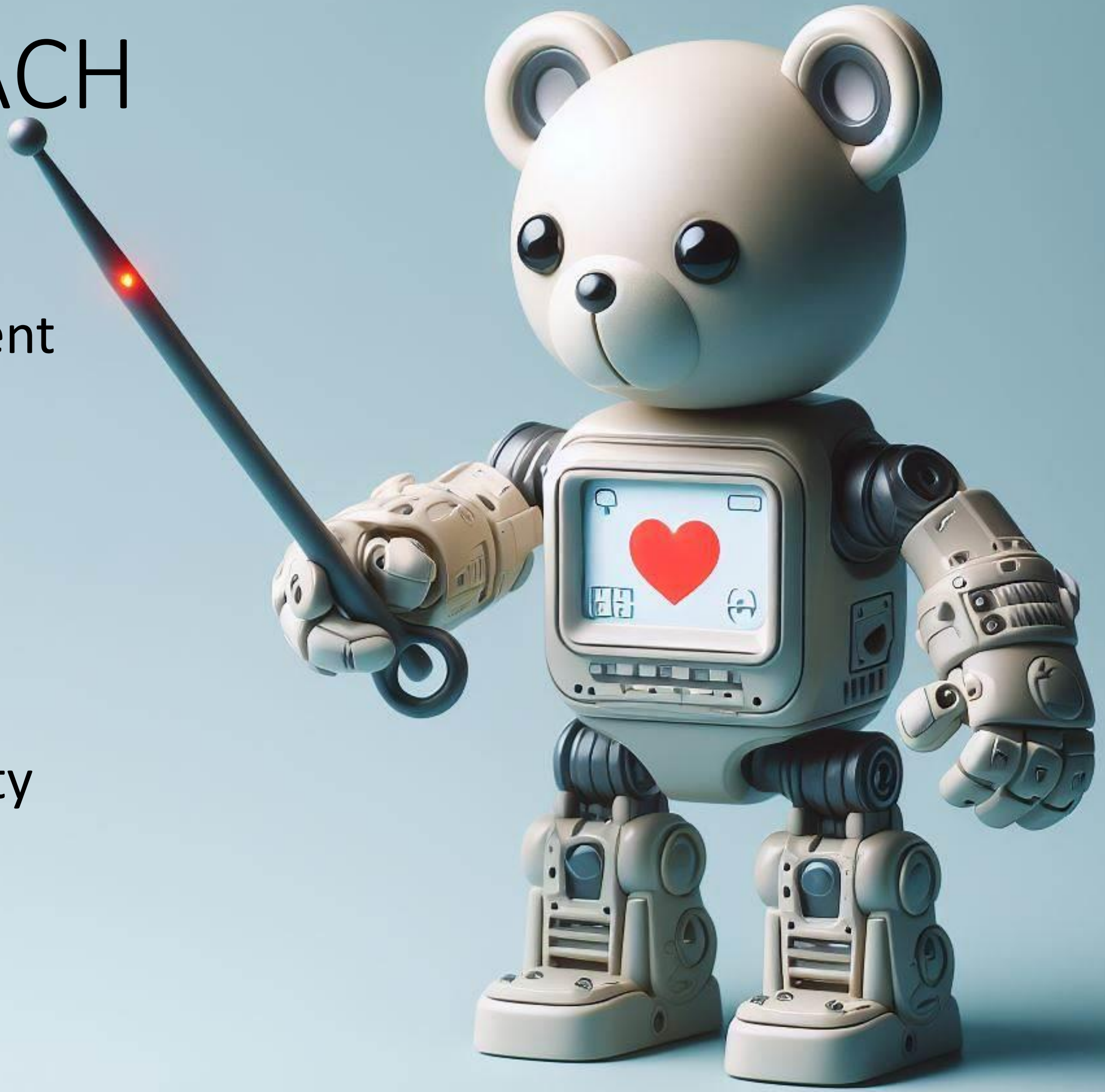
1. A lack of AI expertise in the work force. Companies do not have enough people with experience implementing AI systems, either from a technical or a governance perspective.
2. Multi-dimensional risk. AI systems bring risk to the enterprise in multiple ways, including performance and scalability, security and privacy, reputation, and regulatory.
3. Competing frameworks and uncertain regulatory environment. There are multiple, competing standards for AI regulation, and none have been formally adopted as law yet.
4. Emerging technology. AI is a rapidly developing field and new use cases and threat vectors are being discovered daily.
5. There are few, if any, systematic risk assessment approaches that are suitable for non-technical audiences.
6. Companies are under enormous time pressure to innovate and jump on the AI bandwagon.
7. AI terminology is specialized, and many people tasked with governance do not understand key terms or capabilities.
8. The AI Safety research field is new and underdeveloped.
9. AI Safety concerns are often exaggerated and receive outsized coverage in the media.



A Solution: BEARS TEACH

- Bias and Fairness
- Ethical Compliance and Values Alignment
- Accountability and Responsibility
- Robustness and Reliability
- Security and Privacy

- Transparency and Explainability
- Environmental Impact and Sustainability
- Autonomy and Control
- sCalability and Performance
- Human Impact and Safety



The Dimensions

Bias and Fairness (B): Evaluates the risk of AI systems exhibiting biased outputs or unfair treatment towards certain groups or individuals.

Ethical Compliance and Values Alignment (E): Focuses on ensuring that AI systems comply with ethical standards and are aligned with human values and societal norms.

Accountability and Responsibility (A): Addresses who is responsible for the actions and decisions of AI systems and how accountability is maintained.

Robustness and Reliability (R): Assesses the AI system's ability to perform consistently and accurately under different conditions and its resilience to errors and failures.

Security and Privacy (S): Involves the protection of AI systems from unauthorized access and misuse, and the safeguarding of user data privacy.

Transparency and Explainability (T): Concerns the extent to which AI processes and decisions can be understood and traced by humans.

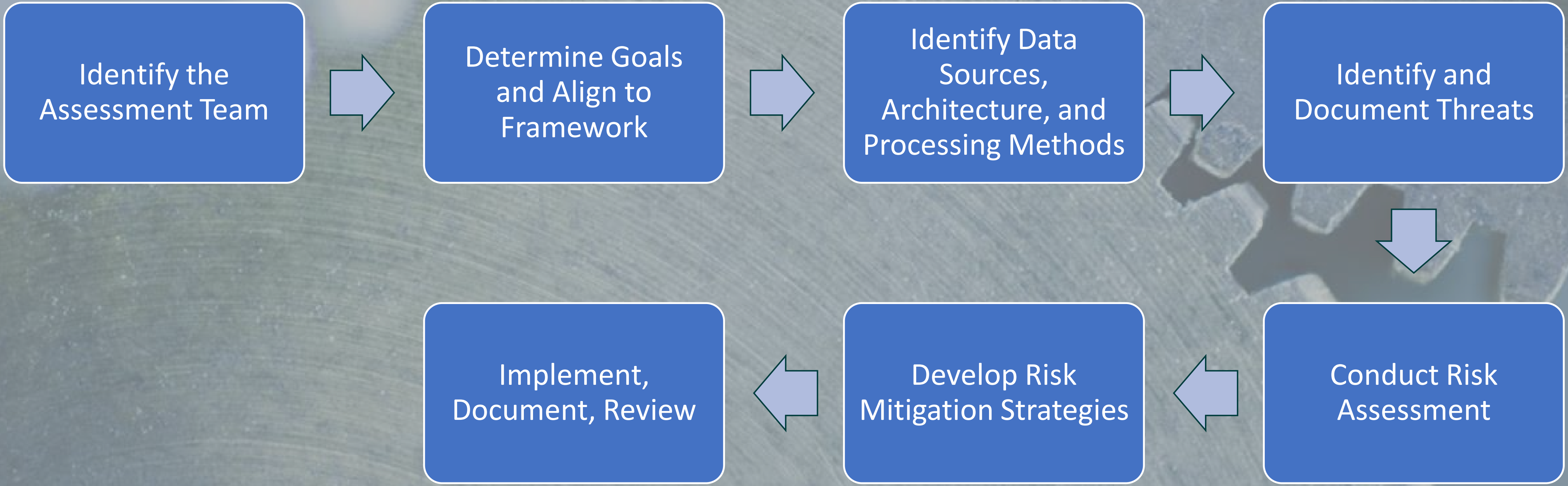
Environmental Impact and Sustainability (E): Considers the ecological footprint and long-term sustainability of AI systems.

Autonomy and Control (A): Examines the balance between AI autonomy and human oversight, ensuring appropriate human control over AI systems.

Scalability and Performance (C): Assesses the AI system’s capability to handle scaling in terms of workload and its operational performance efficiency.

Human Impact and Safety (H): Evaluates the effects of AI systems on human well-being and safety, both physically and psychologically.

The Method



Identify the Assessment Team

Determine Goals and Align to Framework

Identify Data Sources, Architecture, and Processing Methods

Identify and Document Threats

Conduct Risk Assessment

Develop Risk Mitigation Strategies

Implement, Document, Review

Sample Threat Card

Threat Category: Accountability and Responsibility (A&R)

- **Description:** This threat involves the risk of AI systems operating without clear accountability and responsibility. It addresses who is responsible for the actions of the AI and how accountability is maintained, especially in scenarios where AI decisions lead to negative outcomes.

Relevant To:

- All AI systems, particularly those involved in decision-making processes affecting humans (e.g., healthcare, finance, legal, autonomous systems).

Indicators of Risk:

- Lack of clear policies or guidelines on who is responsible for AI decisions.
- Absence of mechanisms to trace decisions back to specific AI algorithms or data sets.
- AI systems making autonomous decisions without human oversight.
- Inadequate documentation or logging of AI decision processes.

Potential Impact:

- Legal and ethical ramifications due to unaccountable AI actions.
- Erosion of user trust in AI systems.
- Harm to individuals or groups affected by unaccountable AI decisions.

Mitigation Strategies:

1. **Establish Clear Governance:** Implement governance frameworks that clearly define roles and responsibilities in AI development and deployment.
2. **Ensure Traceability:** Develop systems to trace AI decisions back to specific algorithms, data inputs, and operational logic.
3. **Implement Oversight Mechanisms:** Incorporate human oversight in critical decision-making processes, especially in high-stakes scenarios.
4. **Maintain Comprehensive Documentation:** Keep detailed records of AI development processes, decision-making criteria, and operational logs.
5. **Develop and Enforce Ethical Guidelines:** Create and adhere to ethical guidelines that govern AI behavior and decision-making processes.

Documentation

Threat ID	Threat Category	Summary Description	Hotspot	Threat Source	Elicitation Questions	Example Illustrations	Consequences	Mitigation Strategies	Responsibility
T001	Accountability and Responsibility	Unclear responsibility for decisions made by the AI system.	Decision-making processes, autonomous actions.	Organizational, External, Receiving Party.	Who is responsible for AI decisions? How are errors handled?	Loan application denial without explanation.	Legal penalties, loss of trust, harm without redress.	Define clear roles, establish oversight protocols.	Compliance Team, IT Department.
T002	Security and Privacy	Potential for unauthorized access to sensitive data.	Data storage, transmission points.	Organizational, External.	How is data encrypted? Are there access control policies?	Data breach leading to exposure of personal information.	Breach of trust, regulatory fines, identity theft.	Implement encryption, strengthen access controls.	IT Security Team.
T003	Bias and Fairness	AI system perpetuates existing biases.	Data training, model validation.	Organizational.	What measures are in place to ensure diversity in training data?	AI hiring tool favoring certain demographics.	Unfair practices, discrimination claims, loss of reputation.	Diversify training data, implement fairness checks.	Data Science Team.
T004	Robustness and Reliability	AI system fails under unexpected conditions.	Operational environment, stress points.	Technical, Environmental.	How does the system handle edge cases or high-load scenarios?	Autonomous vehicle navigation fails in severe weather.	System downtime, potential accidents, loss of life.	Conduct stress testing, enhance error handling.	Engineering Team.
T005	Transparency and Explainability	AI decisions are opaque and lack justification.	User interface, reporting systems.	Organizational.	Can users understand the rationale?	Patient unable to understand AI medical diagnosis.	Lack of trust, inability to contest decisions, non-compliance	Develop explainability features, create user education materials.	Product Management Team.

Simplified Risk Map

Consequences

Probability

	Catastrophic	Major	Moderate	Minor
Frequent	🟥	🟥	🟨	🟩
Occasional	🟥	🟨	🟩	🟩
Uncommon	🟥	🟨	🟩	🟩
Remote	🟥	🟨	🟩	🟩

Ristić, Dejan (2013). "A tool for risk assessment" (PDF). *Safety Engineering*. ***3*** *(3)*. doi:10.7562/SE2013.3.03.03.

https://privatus.online
info@privatus.online