AI in the Underwater Domain and Exploiting Development Elsewhere

Eshan Rajabally CEng PhD

Subsea Expo, February 2024

Presentation Outline

- Introducing BMT
- Applications of AI in the Maritime Domain
- Trustworthy and Responsible AI
- Standards and Beyond

Introducing BMT

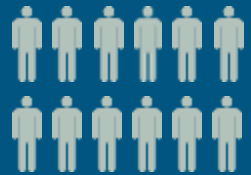
BMT at a Glance

Maritime Design and Consultancy

Asset Monitoring and Sustainment

Environment and Climate Solutions

Defence and Security Customer Friend



1,300 people

Established **1985**

27 offices

4 continents

BMT Maritime Autonomous Systems (MAS) activity

Dedicated team stood up to provide:

- Independent MAS assurance
  - Simulator based collision regulation compliance testing
- Technical & Engineering Consultancy
  - Underwater vehicle cybersecurity
- Leading Edge R&D
  - Next generation “Pentamaran” hull-form

Applications of AI in the Maritime Domain

Applications of AI in Underwater

Sensor processing
Multi-source data fusion and perception,
Object detection and classification
Collision avoidance,
Acoustic communications

Mission planning
Spatio-temporal scheduling

Multi vehicle coordination
Formation control,
Task allocation

Localization and mapping
GPS denied PNT

Guidance & Navigation
Path planning,
Docking & homing

Model based control
Trajectory tracking,
Station keeping

Data analysis
Environmental data assessment,
UW infrastructure monitoring

BMT

Potential Applications of AI yet to be read across?

Predictive maintenance

Power & propulsion optimisation

Report generation (back office tasks)

Cyber & physical security

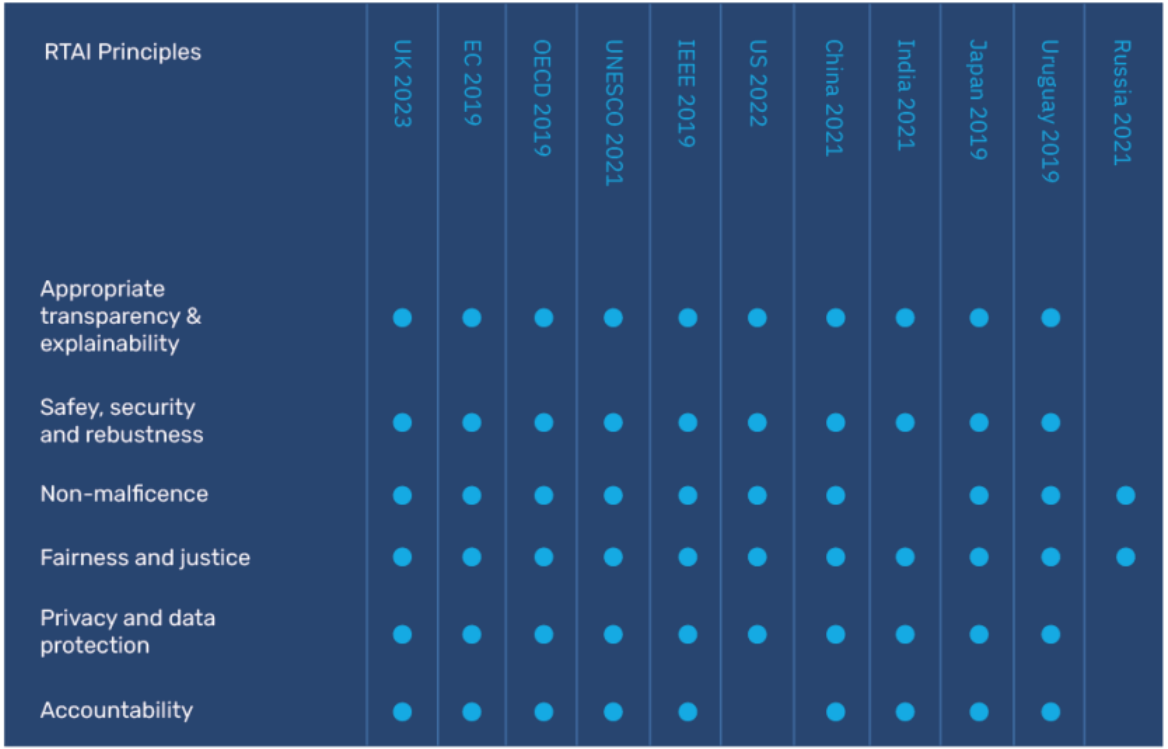
Launch and recovery operation

Optimised design & manufacture

BMT

Trustworthy and Responsible AI

Trustworthy & Responsible Building Blocks



RTAI Principles	UK 2023	EC 2019	OECD 2019	UNESCO 2021	IEEE 2019	US 2022	China 2021	India 2021	Japan 2019	Uruguay 2019	Russia 2021
Appropriate transparency & explainability	●	●	●	●	●	●	●	●	●	●	
Safey, security and rebustness	●	●	●	●	●	●	●	●	●	●	
Non-malficence	●	●	●	●	●	●	●		●	●	●
Fairness and justice	●	●	●	●	●	●	●	●	●	●	●
Privacy and data protection	●	●	●	●	●	●	●	●	●	●	
Accountability	●	●	●	●	●		●	●	●	●	

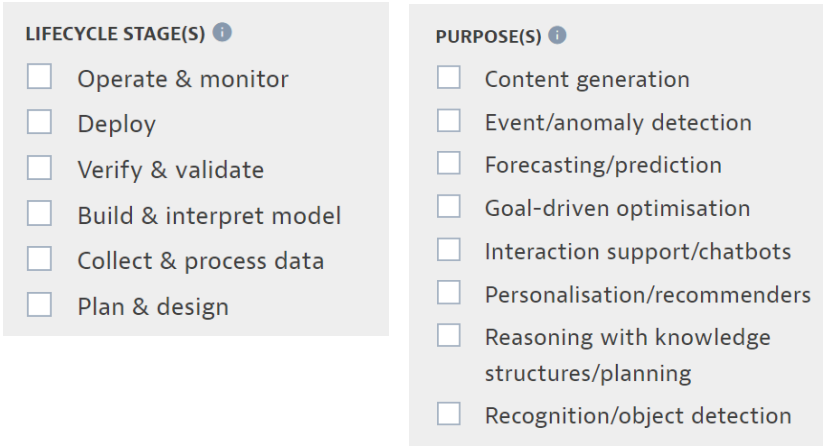
Sourced from UKRI's 2023 "Report on the Core Principles and Opportunities for Responsible and Trustw orthy AI"

OECD.AI Tools & Metrics for Trustworthy AI

Openly available online catalogue of tools (~700) maintained by the Organisation for Economic Co-operation and Development (OECD)

Some Tool Types

- Audit processes
- Checklists
- Documentation processes
- Guidelines
- Product development
- Risk management
- Technical validation
- Software toolkits



Sourced from https://oecd.ai/en/catalogue/tools

Newman's AI Trustworthiness Properties, 1 of 2

Trustworthy Characteristic	Property	Question(s) to Consider
Valid and Reliable	Fit for Purpose	How will we assess whether the AI system is fit for purpose for each intended use and provides a valid solution for the problems we are trying to solve?
	Predictable and Dependable	How will we ensure that the AI system will behave as expected?
	Appropriate Automation	How will we determine the desired and appropriate degree of automation?
	High Quality Configuration	How will we assess the quality of the AI system design and configuration and ensure consistently high quality?
	Data Completeness	How will we assess and improve the completeness, quantity, suitability, and representativeness of the data?
	Data Quality	How will we assess and improve the quality and relevance of the data?
	Accurate	How will we assess the descriptive and predictive accuracy of what the model has learned?
	Reproducible	How will we test whether desirable outputs of the AI system can be reproduced in different circumstances?
	Verifiable	How will we verify that the system is behaving as expected?
	Reliable	How will we ensure the AI system performs predictably and asintended, including in new environments or with new inputs?
	Replayable	How can we replay the behavior of the system to see if the same input generates the same output?
	Valid	How will we validate the outputs of the AI system, including through external validation?
	Appropriate Capabilities	How will we review whether the capabilities of the AI system are appropriate for a particular use and context?
	Appropriate System	How will we review that the design and training of the system is appropriate for intended and likely uses, and is not underspecified?

BMT

Adapted from New man, 2023, "A Taxonomy of Trustw orthiness for Artificial Intelligence", University of California, Berkeley White Paper

Newman's AI Trustworthiness Characteristics, 2 of 2

Trustworthy Characteristic	Property	Question(s) to Consider
Safe	Data Stability	How will we analyze and monitor for data drift over time?
	Safely Interruptible	How will we ensure that reliable technical and procedural controls, including deactivation and fail-safe shutdown, are in place?
	Containment	How can we contain the AI system to prevent safety and security breaches?
	Detection of Anomalies	How will we detect potential novel hazards?
Fair (Harmful Bias Managed)	Mitigation of Bias	How will we assess and mitigate ways in which systemic and human bias may influence the design, development, and deployment of the AI system?
Secure and Resilient	Security-by-Design	How will we build security into the AI system design, testing, deployment, and operation?
	Integrity	How will we maintain and ensure the accuracy, completeness, and appropriateness of data, models, and procedures informing the AI system?
	Data Security	How will the security of data that is used for training or created be ensured?
	Robust	How will we protect the AI system against cyber attacks, adversarial attacks, data poisoning, model leakage, evasion, inversion, etc., and ensure ongoing performance?
	Resilient	How will we assess the AI system's ability to handle uncertainty and unknown environments?
Explainable	Interpretable Uncertainty	How will we make model uncertainty more interpretable by adding features such as confidence interval outputs?
Privacy-Enhanced	Data Protection	How will we use encryption, differential privacy, federated learning, data minimization, and/or other best practices to protect data?
Accountable & Transparent	Documentation	How will we document the AI system's design, datasets, training, characteristics, capabilities, limitations, predictable failures, intended uses, etc.?
	Data & System Accessibility	How can we enable access to the AI system and datasets to relevant authorities, independent researchers, and trusted intermediaries?
Responsible Practice & Use	Verified Supply Chain	How will we assess and verify the relevant components of the supply chain?

Adapted from New man, 2023, "A Taxonomy of Trustw orthiness for Artificial Intelligence", University of California, Berkeley White Paper

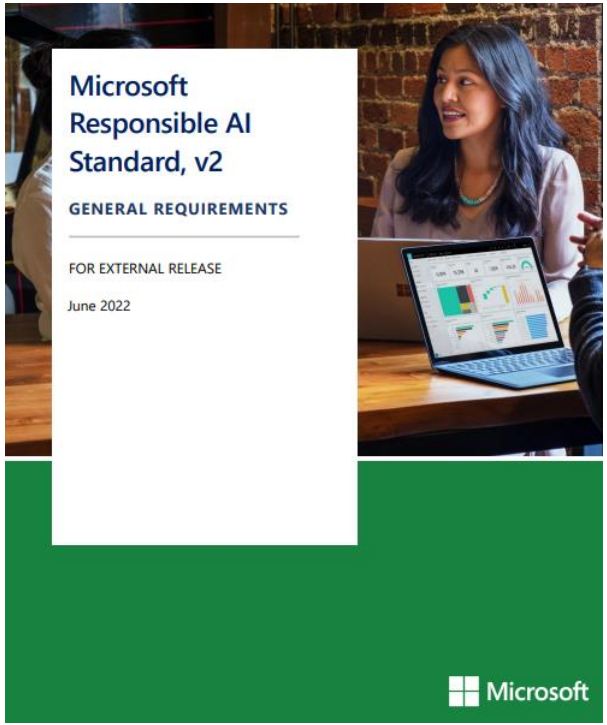
Standards and Beyond

Industry Standards for AI – Google & Microsoft

Google AI Principles

We will assess AI in view of the following objectives. We believe AI should:

1. **Be socially beneficial:** With the likely benefit to people and society substantially exceeding the foreseeable risks and downsides.
2. **Avoid creating or reinforcing unfair bias:** Avoiding unjust impacts on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability and political or religious belief.
3. **Be built and tested for safety:** Designed to be appropriately cautious and in accordance with best practices in AI safety research, including testing in constrained environments and monitoring as appropriate.
4. **Be accountable to people:** Providing appropriate opportunities for feedback, relevant explanations and appeal, and subject to appropriate human direction and control.
5. **Incorporate privacy design principles:** Encouraging architectures with privacy safeguards, and providing appropriate transparency and control over the use of data.
6. **Uphold high standards of scientific excellence:** Technology innovation is rooted in the scientific method and a commitment to open inquiry, intellectual rigor, integrity and collaboration.
7. **Be made available for uses that accord with these principles:** We will work to limit potentially harmful or abusive applications.



BMT

Left image sourced from Google, https://ai.google/static/documents/ai-principles-2022-progress-update.pdf
Right image sourced from Microsoft, https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

ISO AI Trustworthiness Guidance

Vulnerabilities, Threats & Challenges

- Security
- Privacy
- Bias
- Unpredictability
- Opaqueness
- Specification related
- Implementation related
- Use related

Mitigation Measures

- Transparency
- Explainability
- Controllability
- Bias reduction
- Privacy
- Reliability, resilience and robustness
- Mitigating system hardware faults
- Functional safety
- Testing and evaluation
- Use and applicability

BMT

Sourced from ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
Many other potentially relevant standards summarised in the UKRI's "Report on the Core Principles and Opportunities for Responsible and Trustworthy AI"
~300 catalogued at https://aistandardshub.org/ai-standards-search/

Mitigation Signposting

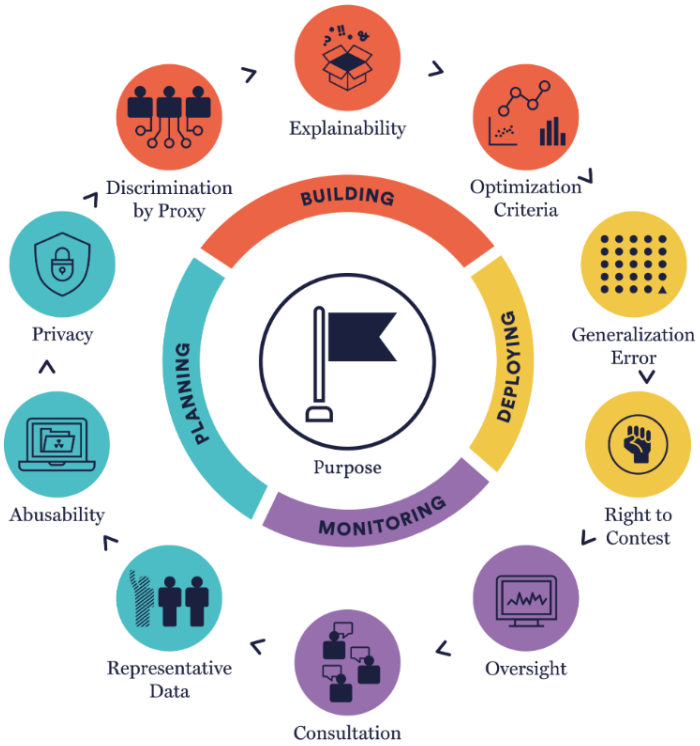
Transparency & Explainability

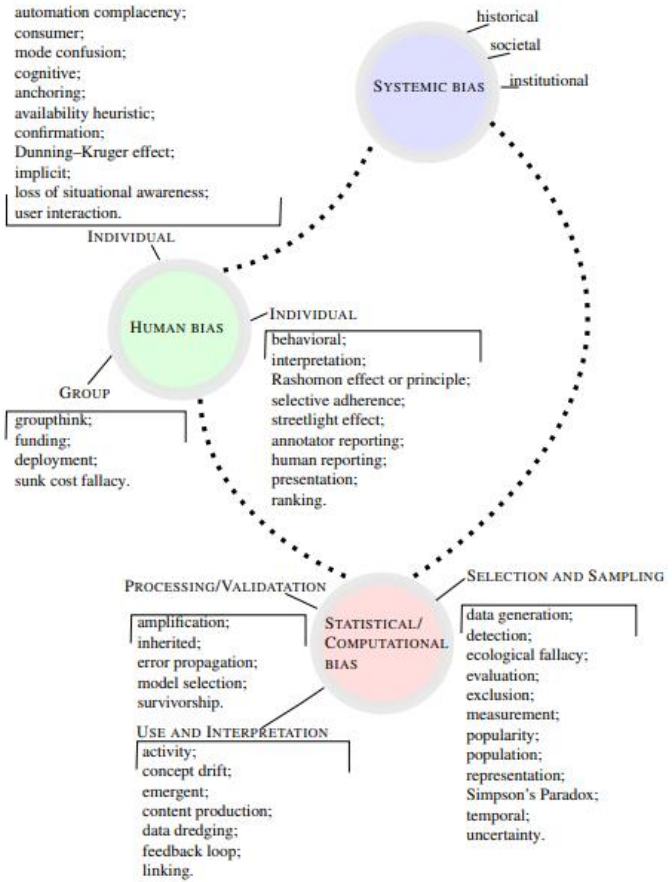
- AI is interpretable if we can understand how it works (transparent) and/or why it makes the decisions that it does (explainable)
- The Assuring Autonomy International Programme Body of Knowledge provides references to 17 technique types for providing interpretability specific to machine learning

Controllability

- Microsoft's Guidelines for Human-AI Interaction, 2019
- Google PAIR, People + AI Guidebook, 2021

AI Biases Signposting





Left image sourced from MIT's AI Blindspots (Calderon et al., 2021) at https://aiblindspot.media.mit.edu/
Right image sourced from US NIST's "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" (Schwartz et al., 2022)

To Conclude

Artificial Intelligence is demonstrating benefit in the underwater sector but has potential to go further

Generic or transferable features of AI trustworthiness along with principles and practice are proliferating

Adoption of these can benefit the application of AI in the underwater domain from development through to usage

Thank you

eshan.rajabally@uk.bmt.org