TRUSTED AUTONOMOUS SYSTEMS
DEFENCE CRC

Australia's AI Governance Frameworks & Pragmatic Tools Manage Ethical Risk

Dr S.Kate Devitt
Chief Scientist
Trusted Autonomous Systems

Australian Government
Department of Defence

NEXT GENERATION TECHNOLOGIES FUND

Queensland Government

1

Australian Values

2

Australian Values (Department of Home Affairs 2020)

Australian Government
Department of Home Affairs

- Respect for the freedom and dignity of the individual
- Freedom of religion (including the freedom not to follow a particular religion), freedom of speech and freedom of association
- Commitment to the rule of law, which means that all people are subject to the law and should obey it
- Parliamentary democracy whereby our laws are determined by parliaments elected by the people, those laws being paramount and overriding any other inconsistent religious or secular 'laws'
- Equality of opportunity for all people, regardless of their gender, sexual orientation, age, disability, race or national or ethnic origin
- A 'fair go' for all that embraces: mutual respect; tolerance; compassion for those in need; and equality of opportunity for all
- The English language as the national language, and as an important unifying element of Australian society

Department of Home Affairs. (2020). *Australian Values*. https://www.homeaffairs.gov.au/about-us/our-portfolios/social-cohesion/australian-values

3

Australia's AI Ethics Principles

Australian Government
Department of Industry, Science, Energy and Resources

1. **Human, societal, and environmental wellbeing**: AI systems should benefit individuals, society, and the environment.
2. **Human-centred values**: AI systems should respect human rights, diversity, and the autonomy of individuals.
3. **Fairness**: AI systems should be inclusive and accessible and should not involve or result in unfair discrimination against individuals, communities, or groups.
4. **Privacy protection and security**: AI systems should respect and uphold privacy rights and data protection and ensure the security of data.
5. **Reliability and safety**: AI systems should reliably operate in accordance with their intended purpose.
6. **Transparency and explainability**: There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI and can find out when an AI system is engaging with them.
7. **Contestability**: When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.
8. **Accountability**: People responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

Department of Industry Innovation and Science. (2019). *Australia's Artificial Intelligence Ethics Framework*. Retrieved 25 September from https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles

4

AI Action Plan for Australia (2021) commits to OECD AI principles :

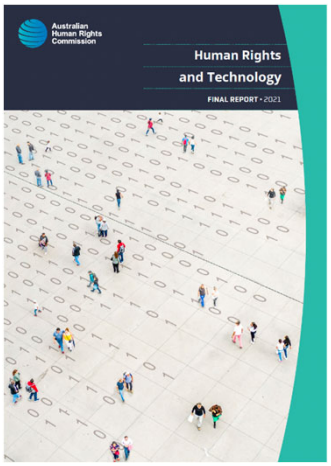
1. AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.
2. AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.
3. There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.
4. AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.
5. Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

Department of Industry Science Energy and Resources. (2021). *Australia's Artificial Intelligence Action Plan*. https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-action-plan

5

Human Rights

- require human rights impact assessments (HRIA) before any government department or agency uses an AI-informed decision-making system to make administrative decisions
- the government needs to make AI decision making transparent and explainable to affected individuals and give them recourse to challenge the decision
- Australian ethics principles should be used to encourage corporations and other non-government bodies to undertake a human rights impact assessment before using an AI-informed decision-making system.



Santow, E. (2021). *Human Rights and Technology Final Report*. Australian Human Rights Commission. https://tech.humanrights.gov.au/downloads

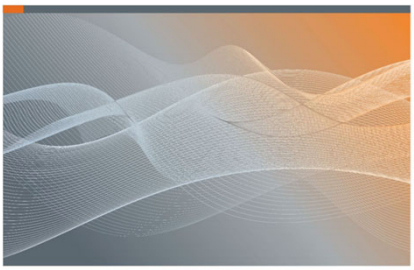
6

TRUSTED AUTONOMOUS SYSTEMS

OFFICIAL

Australian Government
Department of Defence

A Method for Ethical AI in Defence



Defence Science and Technology Group

DSTG-TR-3786

This is a report on the outcomes of a workshop only and does not represent an official position of Defence. It represents views expressed by participants and stakeholders of the workshop.

Defending Australia and its National Interests
www.defence.gov.au

OFFICIAL

7

TRUSTED AUTONOMOUS SYSTEMS

A method for Ethical AI in Defence

...while not a formally adopted view of the Australian government, the Method is the clearest articulation of ethical AI for defence among the Indo-Pacific allies as well one of the most concrete practices that U.S. allies have thus far developed for AI ethics implementation in defence (Lockman, 2021, pp.21 & 23)

Key references:

Devitt, S. K., Gan, M., Scholz, J., & Bolia, R. S. (2021). A Method for Ethical AI in Defence. **Defence Science & Technology Group** (DSTG-TR-3786). https://www.dst.defence.gov.au/publication/ethical-ai

Gaetjens, D., Devitt, S.K. & Shanahan, C. (2021). Ethical AI in Defence Case Study: Allied Impact. DST Technical Report. **Defence Science & Technology Group**

Lockman, Z. (2021). Responsible and Ethical Military AI Allies and Allied Perspectives: CSET Issue Brief. **Centre for Security and Emerging Technology, Georgetown University's Walsh School of Foreign Service**. https://cset.georgetown.edu/wp-content/uploads/CSET-Responsible-and-Ethical-Military-AI.pdf

Copeland, D., & Sanders, L. (2021, 8 October). Engaging with the industry: integrating IHL into new technologies in urban warfare. **Humanitarian Law and Policy ICRC Blog**. https://blogs.icrc.org/law-and-policy/2021/10/07/industry-ihl-new-technologies/

8

TRUSTED AUTONOMOUS SYSTEMS

Pragmatic tools for managing ethical risks in AI

https://youtube.com/playlist?list=PLaIRKAnA4EPXAqVOK537XQzpRto246vwX

9

TRUSTED AUTONOMOUS SYSTEMS

https://theodi.org/article/data-ethics-canvas/

10

TRUSTED AUTONOMOUS SYSTEMS

AI Ethics Checklist

A	Describe the military context the AI is for	E.g. Force Application, Force Protection, Force Sustainment, Situational Understanding, Personnel, Enterprise Logistics, Business Process Improv.
B	Explain the sort of decisions AI helps with	E.g. Is it a single decision-maker, multi-decision maker; once-off decisions vs. sequential decisions
C	Explain how the AI integrates with human operators to ensure effectiveness and ethical decision making in the anticipated context of use and countermeasures to protect against potential misuse	E.g. What are the human factors and system factors and what are your scenarios and T&E process?
D	Explain framework/s to be used	E.g. Method for Ethical AI in Defence, safety frameworks, human factors and legal frameworks suitable to the context etc…
E	Employ subject matter experts to guide AI development	E.g. If team lacks the expertise to undertake one or more of steps A-D, then they should onboard the skills gaps through hiring, consulting and oversight.
F	Employ appropriate verification and validation techniques to ensure compliance	E.g. Auditability, accountability, explainability and lawful abidance must be demonstrable

11

TRUSTED AUTONOMOUS SYSTEMS

AI Ethics Risk Matrix

AI Project management tool

Activity description	Ethical issue(s), principle(s)	Risk to the project objectives if ethical issue is not addressed	Actions/ Outcomes	Timeline	Person/s responsible	Status
Ensure operators act ethically under uncertainty	*Governance confidence*	*Operator misunderstands the accuracy and reliability of outputs or recommendations of the AI classifier*	*Experiments for implicit and explicit understanding of AI outputs by operator in ethical decision making*	*Q4 2020*	*Bob Cook*	*Pending*

12

Tool 3: Data Item Descriptor and Legal and Ethical Program Plan

- The draft Data Item Description (DID) provides guidance to contractors developing legal and ethical assurance programs for complex Defence AI systems. The DID will be distributed for review and comment by Defence and industry stakeholders before it is considered for Defence contracts.
- Where an ethical risk assessment is above a certain threshold, a Legal and Ethical Assurance Program Plan (LEAPP) should be supplied. This describes a contractor's plan for assuring software acquired under the contract meets Commonwealth legal and ethical assurance requirements. The LEAPP provides Defence with visibility into the contractor's legal and ethical planning, supports progress and risk assessment and provides input into Defence's internal planning, including weapons reviews under Article 36 of Additional Protocol 1.

13

TRUSTED AUTONOMOUS SYSTEMS

IWR
International Weapons Review

What is in a Legal and Ethical Assurance Program Plan (LEAPP)?

Project/ capability includes a description of:

- normal or expected use
- anticipated operating environment(s) and operational context(s)
- hardware/ software/ system type and design
- human/ machine integration design
- autonomous functions
- data sources and requirements

14

What is in a Legal and Ethical Assurance Program Plan (LEAPP)?

1. Description of AI decisions (planning/ executive)
2. Description of relevant rules and principles (direct/indirect) and their application to capabilities normal or expected use
3. Environmental/ operational factors relevant to AI functions/ risk
4. Ethical & Legal risk identification method/ frameworks
5. Test, evaluation and validation method(s)/ process(es)
6. Risk identification and mitigation report format
7. Stakeholder engagement plan
8. Ongoing risk analysis/ assurance/governance plan

15

IEEE STANDARDS

IEEE. (2021). IEEE 7000™-2021 - IEEE Standard Model Process for Addressing Ethical Concerns During System Design. https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html

Winfield, A. F. T., Booth, S., Dennis, L. A., Egawa, T., Hastie, H., Jacobs, N., Muttram, R. I., Olszewska, J. I., Rajabiyazdi, F., Theodorou, A., Underwood, M. A., Wortham, R. H., & Watson, E. (2021). IEEE P7001: A Proposed Standard on Transparency [Original Research]. Frontiers in Robotics and AI, 8(225). https://doi.org/10.3389/frobt.2021.665729

IEEE Std 7007™-2021 Ontological Standard for Ethically Driven Robotics and Automation System. https://site.ieee.org/sagroups-7007/

16

Summary: Reducing ethical risks of RAS-AI

- Describe the decisions the RAS-AI helps with
- Create scenarios with ethical risk
- Employ subject matter experts
- Engage stakeholders
- Test & evaluate in scenarios with ethical risk

17

TRUSTED AUTONOMOUS SYSTEMS

info@tasdcrc.com.au
www.tasdcrc.com.au
@TASDCRC
TASDCRC

Australian Government
Department of Defence

NEXT GENERATION TECHNOLOGIES FUND

Queensland Government

18