AI ETHICS ADVISORY PANEL
Minutes
Wednesday 19th February 2025, 1400 - 1600

Attendees

Paul Lincoln, 2nd Permanent Secretary – Chair

Prof Michael Wooldridge, Ashall Professor of the Foundations of Artificial Intelligence in the Department of Computer Science at the University of Oxford, and a Senior Research Fellow at Hertford College – Guest Speaker

Professor Nick Colosimo, Head of Group Science & Technology & CTIO Engineering Fellow, BAE Systems and Visiting Professor Cranfield University (AI, Robotics & Space)

Professor Peter Lee, Professor of Applied Ethics, University of Portsmouth

Dr Darrell Jaya-Ratnam, Managing Director, DIEM Analytics

Professor Sarvapali (Gopal) Ramchurn (Responsible AI UK, CEO and UKRI Trustworthy Autonomous Systems (TAS) Hub, Director

Professor David Whetham, Professor of Ethics and the Military Profession, Kings College London

Richard Moyes, Managing Director and co-founder, Article 36

Professor Mariarosaria Taddeo, Associate Professor and Senior Research Fellow, Oxford Internet Institute, University of Oxford; Dstl Ethics Fellow, Alan Turing Institute

Dr Merel Ekelhof, Foreign Exchange Officer at the US DoD Chief Digital and Artificial Intelligence Office (CDAO), attending the panel in her personal capacity.

AVM David Arthurton, Director Strategy and Military Digitisation

David Blackall, Director Digital Exploitation for Defence

Professor Steven Meers, Technical Strategy Leader for Dstl's AI & Data Science Capability and Head of Research and Experimentation (DAIC)

Sally Meecham, Transformation Director and Interim Chief Technology Officer, UKHO

Dr Chris Moore-Bick, Head of Defence Science & Technology Policy

AI Senior Scientist, Dstl

Senior Legal Adviser, MODLA

Head of Defence AI and Autonomy Unit (DAU)

1	**Introduction and Updates from the Chair** MOD's 2nd Permanent Secretary (2nd PUS) welcomed members to the tenth meeting of the Ethics Advisory Panel, making the following points: • Having published previous minutes, as a matter of routine, we will continue to do so. • Some Responsible AI Senior Officers are joining the panel today as observers, in support of ongoing leaders' upskilling on AI. • Congratulations to Prof Mariarosaria Taddeo on the publication of her book *The Ethics of Artificial Intelligence in Defence* and the successful launch event held at the Italian Embassy which some panel members attended. The Head of the Defence AI and Autonomy Unit (DAU) noted that: • Since the ninth meeting, the UK MOD welcomed the House of Commons Defence Committee report titled 'Developing AI capacity and expertise in UK defence' to which the department has drafted a response. • A UK MOD delegation attended the AI Action Summit held in Paris on 10 - 11 February**.**
2	**Presentations and Discussion on Agentic AI** The panel heard presentations from guest speaker Prof Michael Wooldridge, as well as Prof Steven Meers and Prof Nick Colosimo on the topic of Agentic AI. These systems may present novel challenges compared to more 'traditional' AI (e.g. Large Language Models) which simply process and respond to inputs, whereas Agentic AI possesses a degree of independence and decision-making capability to achieve specific goals. Prof Michael Wooldridge's presented **Agents in the Era of Large Language Models**. Key points included: • Agentic AI is an evolution from passive software to proactive, cooperative assistants. • Contrasting the "Standard Model" of AI (i.e. AI is the task of constructing agents to which we communicate our preferences in order that they can autonomously act on our behalf) to Multi-Agent Systems (i.e. AI agents that can communicate, cooperate, coordinate and negotiate with each other). • Whilst LLMs are currently disembodied intelligence that is not connected to the world, there is the potential for LLM-based intelligent agents that can take actual actions in the world (going from chatbot functionality to actually getting things done). • AutoGPT: AI with generalist agent capabilities which is able to execute actions on behalf of its user whilst the user still has to authorise every action. Other example agents include: Agents that plan and book holidays; agents that read incoming emails and can decline requests; and agents that monitor social media feeds and conduct sentiment analysis. • Some risks include, inter alia: LLMs' inaccuracies; AI hacking; embedded biases; and personalised spam. **Professor Steven Meers, Dstl** presented on **Agentic AI: Defence Implications**. Key points included: • Agentic AI has the potential to enhance several Defence functions, including Command and Control; Logistics; Knowledge Management; and Process Automation. • An adapted framework from a Deepmind paper places AI agents on a continuum of ongoing AI developments, with increases in capability and moving from narrow (i.e. clearly scoped tasks) to more general 'intelligence' (i.e. can carry out a wide range of tasks). A

third axis describes the spectrum of autonomy (from 'human does everything' to full autonomy).
- Some Defence implications include considerations on hallucinations and the level of delegation, deskilling, workforce impacts, adversary use and vulnerability exploits.

Professor Nick Colosimo presented **Agentic AI & Swarming: Systems Engineering and Industry Perspectives**. Key points included:

- Explaining Systems Thinking, which includes integrating and managing complex systems, and emphasising the importance of how the interrelationships between system elements contribute to the overall system.
- Noting that unbounded agency would create unbounded risk, existing techniques that mitigate for reliability issues and uncertainty include, inter alia: system architecture; independent safeguards; formal policing; run-time validation; choosing (AI) components wisely; and operator-defined formal bounding techniques.
- Using the example of a self-driving car, independent safeguards could also include radars and a diversity of other sensing tools.
- A few examples of industry and academic demonstrations of agentic AI were given. Physical solutions (i.e. for swarming) can be policed through adapting existing systems engineering techniques which were explored in detail during the meeting. This might be more difficult for cyber applications.
- Problem framing approaches included a 4 box model to map existing vs new applications against existing vs emerging AI technologies to support questions on where the greatest uncertainties lie, and what it means for safety and mission assurance.
- Increasing agency can help to address complexity of task and complexity and uncertainty of environment but with the risk of increasingly complex and emergent behaviours.

The chair noted the following framing questions to guide the discussion:

- What new ethical benefits and risks does agentic AI introduce? How might it exacerbate existing ethical risks?
- How do we best frame the problem to be able to address the risks in a meaningful way and maximise the benefits?
- Are there any promising approaches that manage emerging risks that we can learn from?
- What should MOD do (at different levels) to keep pace with these rapid and potentially high-impact changes?
- What impact does this have on our AI policy – (when) do we need to set any strategic drivers/permissions?

In conversation, the following points were made:

- Risk management of agentic AI should consider that the risk surface increases with multiple failure points, depending on the number of agents.
- Specific attention should be given to the human-machine interaction, considering what kind of training will be needed for the human to enable them to disagree with the AI agent and preserve sufficient human autonomy.
- Humans are usually poor at describing what we want. However, good results from AI rely on a clear description of the human-defined goal (also known as preference solicitation in computer science). Training models through reinforcement learning with human feedback can help models better define acceptable vs unacceptable outputs. This problem can also

	be framed in the context of 'Specificity of goal setting' vs 'the Trust Distance', i.e. if the goal is not well defined there will be greater uncertainty. • The importance of training (on the benefits and risks of AI technologies) for senior commanders and Responsible AI Senior Officers was highlighted, to enable responsible innovation and not be held back by how we approached problem-solving historically. It was noted that there may be a gap between what people think they know about AI compared to what they actually know. • It was noted that current validation and benchmarks (the baseline human performance vs the AI performance) are not yet effective for testing a range of performances. Defence-specific testing and benchmarking for Human-Machine Teaming is needed, noting the diversity of defence AI applications. • Security was highlighted as a key area of focus when developing (agentic) AI. • The work of the UN Group of Governmental Experts on emerging technologies in the area of lethal autonomous weapons systems was noted as an area where the structured thinking presented in this meeting could be further refined.
3	**Any Other Business and Closing Remarks** The following AOBs were raised: • Prof David Whetham presented the near-finalised AI Ethics Playing Cards, explaining the tool's structured approach to building AI ethical awareness by using the analogy of a tree forming different parts of the learning journey: o Roots representing the foundation of the military service values and how Defence values overlap with AI ethical principles (the stem of the tree); o The 'leaves' let learners explore through prompts and questions how they might put the AI ethical principles to practice; and o The 'fruit' represent different AI use case studies to help illustrate the practical implementation of ethical considerations. The tool received early positive feedback and will be published as a website once it has been through the clearance process. • Prof Steven Meers noted that Dstl is carrying out horizon-scanning research to inform future Defence planning and mitigate strategic shock. • The DAU policy adviser noted the intent to present a high-level analysis of the Responsible AI Assurance Statements at the next panel meeting.