Decision-making: How do human-machine teamed decision-makers, make decisions?

Authors: Jane Walker and Lucy Smeddle

This concept information note is the fourth in a series of five being made available unedited to support DCDC's **command and control in the future** project. The information notes are designed to introduce the thinking and themes of the joint concept note that will publish in late 2024.

Concept information note 4

May 2024

Introduction

1. This paper explores the current and developing states of technical, practical, and ethical understanding of human-machine decision making. The increasing utilisation of human-machine teaming in decision making is inevitable, necessary, and desirable given the rapid acceleration of technology in the defence and security environment. As fast as the technology is developing, the associated terminology is evolving, and there is considerable nuance in the situations described here. The implications for defence of human-machine teamed decision making are significant in both benefit and potential risk. Operational decisions depend on collecting, processing and controlling vast amounts of data from all domains. The speed and quantity of this data outpaces human cognitive ability to make effective informed decisions. Human-machine teaming allows the operational commander to manage and analyse these large data sets to support decision making. Humans and machines bring different strengths and will need to collaborate to deal with different aspects of decision making. AI is well-positioned to tackle complex issues using analytical approaches. Human cognition is better suited to focus more on uncertainty, using creative, intuitive, and experience-based decisions.

Knowledge and cognition: what is machine understanding?

2. Current Artificial Intelligence (AI) systems cannot 'understand' in the way that humans do.[1] For a machine to achieve 'natural understanding', Pepperrell argues that it must be imbued with a sense of consciousness.[2] Though a contested term, AI leaders broadly understand consciousness to relate to self-awareness.[3] A machine could achieve the consciousness required to understand, Pepperrell argues, if the same kinds of biological processes and structures that mediate human understanding were implemented efficiently in a machine, though for this to happen, he argues we must move beyond today's computational architecture.[4] Fristol et al's exploration of human and machine understanding concludes that, despite significant computational advances, "learning machines… can only deceive users into ascribing understanding to them while, in fact, having none".[5] Until Artificial General Intelligence (AGI) is achieved, machines will

1 'What is Deep Learning?', *IBM*, no date.

2 Pepperell defines natural understanding as "*the human-like capacity for understanding that is instantiated in our neurobiology, in particular in our brains*". See: Pepperell, Robert, 'Does Machine Understanding Require Consciousness', *Frontiers in Systems Neuroscience*, Vol 16 Article 788486, 18 May 2022, p3.

3 Pelley, Scott, 'Is artificial intelligence advancing too quickly? What AI leaders at Google say', *CBS News*, 16 April 2023.

4 Pepperell, Robert, 'Does Machine Understanding Require Consciousness'.

5 Moran, Rosalyn J., Friston, Karl, J., Yufik, Yan M., 'Editorial: Understanding in the human and the machine', *Frontiers in Systems Neuroscience*, Vol 16, 25 November 2022.

only be able to gain and retain knowledge (learn), without understanding it.[6] Despite this, we can still explore how machines are taught to acquire knowledge by the computer scientists, roboticists and engineers that create them.

3. How an AI system learns significantly depends on how it is trained, the data used to train it, and information gathered from other systems it may interact with. As decisions regarding the aforementioned largely lie with humans, machines are liable to be impacted by bias.[7] Advances in computer hardware have resulted in a myriad of machine learning (ML) algorithms,[8] computational models which "allow computers to understand patterns and forecast or make judgments based on data without the need for explicit programming".[9][10] The better the data on which an algorithm is trained, the more suitable/improved its outputs. Deep learning, a subfield of ML modelled on the human brain, has become a popular method, and uses multi-layered artificial neural networks (ANN) that respond to inputs with electrical output signals to make decisions.[11]

4. Issues arise when it comes to how AI systems explain the phase that comes between the input and the output; that is, how or why it made its decision. This is particularly pertinent for human-machine collaborators, as the human overseer needs to understand how the machine has reached its decision. Challengingly, even if simple code is used to specify the architecture and training of a deep neural network model, the results can be notably complex, leading to 'black boxes'.[12] As Rahwan explains, though input results in output, "the exact functional processes that generate these outputs are hard to interpret even to the very scientists who generate the algorithms themselves".[13] This complexity grows when the volume and variety of data inputs increase, as would occur when machines learn from one another through transfer or meta learning,[14][15] with explainability further reduced if working with systems/training data protected under intellectual property laws.[16] Even if users are able to access the source code or model structure of an AI system, this may still prove insufficient when predicting outputs, because AI agents continue to demonstrate novel behaviours in their interactions with the world and other AI agents, which can be impossible to predict.[17] Issues are compounded if an AI agent operates in an environment in flux, as is likely with human-machine teams that may operate during warfighting.[18]

6 'What is artificial general intelligence (AGI)?', *McKinsey & Company*, 21 March 2024.

7 Rahwan, Iyad, and others, 'Machine Behaviour', *Nature*, 568, pp 477–486, 24 April 2019.

8 Masri, Ali, 'How Do Machines Learn?', *Medium*, 01 June 2019.

9 'Machine Learning Algorithms', *GeeksforGeeks*, 15 Nov 2023.

10 Types of ML algorithm include supervised learning, unsupervised learning and reinforcement learning. See Glossary for further explanation.

11 Ball, Phillip, 'How do machines think?', *The New Statesman*, 11 December 2019.

12 Voosen, Paul, 'The AI detectives', *Science*, Vol 357, Issue 6346: pp 22-27, 07 July 2017.

13 Rahwan, Iyad, and others, 'Machine Behaviour', p. 478.

14 Zhuang, Fuzhen, and others, 'A Comprehensive Survey on Transfer Learning', *arXiv: Cornell University*, 23 June 2020.

15 'Meta-Learning in Machine Learning', *GeeksforGeeks*, 29 November 2023.

16 Rahwan, Iyad, and others, 'Machine Behaviour'.

17 Rahwan, Iyad, and others, 'Machine Behaviour'.

18 Maathuis, Clara, 'Towards Trustworthy AI-based Military Cyber Operations', *Open University Netherlands: International Conference on Cyber Warfare and Security*, Vol 19, No 1, 21 March 2024.

5. Another issue concerning AI learning and predictability concerns machine 'hallucinations', incorrect information generated when a machine is confronted with a problem set they have limited understanding in solving. This has been reported in Large Language Models (LLMs) and can occur for multiple reasons, though often stems from training data and/or model design errors, and can have damaging real-world implications.[19] [20] [21] [22] Though developers are continually designing methods to minimise hallucinations, the lowest rate in a public LLM sits at 2.5 percent.[23] An obvious solution to dealing with data and / or model design issues is to ensure that systems are trained on high quality, relevant, large and diverse datasets, with false / obsolete / inaccurate information that risk poisoning the dataset removed before input, expectations / limits defined, and continuous evaluation / testing actioned before the system is released.[24] However, even if this is achieved, because LLMs are trained on orders of magnitude more data than they are able to store, they are occasionally unable to perfectly recall everything from their training.[25] A recent review on the topic used learning theory to conclude that because LLMs will never be able to learn all conceivable computable functions, hallucinations will always be impossible to eliminate.[26]

6. Oftentimes the data we do have, even if high quality, verified, varied and vast, may in fact act as a constraint, leading us in the wrong direction because we are unaware that we're missing key information, either accidentally - possibly from bias within input data, which increases if data sources vary - or purposefully, i.e. an adversary has chosen to hide it. As humans, we may be able to extrapolate, using our knowledge to form some idea as to what might exist in the dark area, but for machines, this is less straightforward. Potential solutions for dealing with missing or limited data are emerging and may lie with a combination of techniques such as Zero-Shot Learning (ZSL), One-Shot Learning (OSL) and Few-Shot Learning (FSL).[27] ZSL enables ML models to adapt to new domains or tasks by utilising transfer learning.[28] Advancements in these areas offer promise for versatile, smart systems, with some arguing that combining ZSL with FSL and meta-learning could enable ML models to learn rapidly

19 'AI hallucinations: Complete guide to detection and prevention', *SuperAnnotate*, 06 February, 2024.

20 For example, training data may be insufficient/incomplete, low-quality/flawed, or outdated, there may be model errors in encoding/decoding, bias in former generations of the model, or too much focus has been made on model novelty in its design, or the LLM may be responding to vague prompts, which can lead to nonsensical or fabricated outputs. See: 'AI hallucinations: Complete guide to detection and prevention'.

21 Kirkovska, Anita, '4 LLM Hallucination Examples and How to Reduce Them', *vellum*, 01 January 2024.

22 'Understanding AI Hallucinations: Causes and Consequences', *DataScientest*, 17 April 2024.

23 The lowest hallucination rate as of 1050hrs, 19 April 2024, was 2.5 percent for GPT 4 Turbo. See: 'Vectara / hallucination-leaderboard', *github*.

24 'Understanding AI Hallucinations: Causes and Consequences'.

25 Leffer, Lauren, 'AI Chatbots Will Never Stop Hallucinating', *Scientific American*, 05 April 2024.

26 Xu, Ziwei, Jain, Sanjay, Kankanhalli, Mohan, 'Hallucination is Inevitable: An Innate Limitation of Large Language Models', *arXiv*, 22 January 2024.

27 Cyber, John. D. 'Decoding Zero-Shot Learning: A Bridge between Machine Learning and Human-like Intelligence', *Medium*, 07 July 2023; Defence Science and Technology Laboratory, 'Machine Learning with Limited Data'; Future of AI for Defence Project Autonomy Programme, *UK Ministry of Defence*, 07 Dec 2020; Lamba, Harshall, 'One Shot Learning with Siamese Networks using Keras', *Medium*, 21 Jan 2019; 'What is few-shot learning?', *IBM*, 12 October 2022.

28 Cyber, John. D. 'Decoding Zero-Shot Learning: A Bridge between Machine Learning and Human-like Intelligence'.

from a few examples to perform effectively on novel tasks.[29]

7. Utilising emerging and uncertain ML techniques in human-machine teaming will require continuous testing and oversight. Models that are trained on hypothetical scenarios may behave uncertainly in volatile environments, with even the most robust systems prone to hallucination. Most AI systems are created by institutions with rigorous rules and standards, at odds with real-world, unpredictable warfighting scenarios that may be faced by military human-machine teams. Who takes responsibility for the results of errors made by an AI model is unclear. Defence must decide where it wishes to implement these systems, and ensure that those it does deploy are trustworthy, overseeable and understandable, or risk serious reputational and security implications.

Machine unlearning, deception, and spoofing

8. The advent of AI has brought with it a new attack vector, the surface area of which will grow as more decisions are made by predictive algorithms.[30] Types of attacks on ML systems vary according to the model and adversary's aim, though many are vulnerable and can be easily fooled if unprepared.[31] [32] Adversarial examples, for instance, are designed to trick an AI systems' neural network by modifying/distorting inputs, causing a model to misclassify them, the real-world implications of which could see algorithms working as part of a human-machine team misclassify an adversary's weapons system as a benign object.[33] [34] [35] Other common attacks include data poisoning, where false or misleading information is intentionally inputted into a training dataset, or the dataset is modified or a portion deleted, with the aim of manipulating or influencing the model's operation.[36] [37]

29 Cyber, John. D. 'Decoding Zero-Shot Learning: A Bridge between Machine Learning and Human-like Intelligence'.

30 Muppidi, Sridhar, 'Adversarial AI: As New Attack Vector Opens, Researchers Aim to Defend Against It', *SecurityIntelligence*, 17 April 2018.

31 Common examples include adversarial examples, data poisoning, membership inference, model inversion attacks and backdoor injection attacks.

32 This vulnerability is largely due to assumptions made during the development of the model, which include assumptions regarding the trustworthiness of the training datasets, and the security of the environment used to evaluate the data. Theoretically, if a model is developed for the use of military human-machine teaming, such training data would be private with enhanced security implemented, so such models may be less vulnerable than examples found in the literature. See: Bountakas, Panagiotis and others, 'Defense strategies for Adversarial Machine Learning: A survey', *Computer Science Review*, Vol 49, August 2023.

33 Muppidi, Sridhar, 'Adversarial AI: As New Attack Vector Opens, Researchers Aim to Defend Against It'.

34 Machine Learning @ Berkeley, 'Tricking Neural Networks: Create your own Adversarial Examples', *Medium*, 7 March 2019.

35 A variety of techniques can be used to create an adversarial example, with attackers able to analyse the gradients of an ML model to identify the most impactful changes to apply to the input data. One of the biggest issues with such an attack is that they are often not model or architecture specific, and can transfer easily from one model to the next, an implication being that such an attack could be launched on a black box model, the internal mechanics of which an adversary has no prior knowledge. See: Machine Learning @ Berkeley, 'Tricking Neural Networks: Create your own Adversarial Examples'.

36 Lenaerts-Bergmans, Bart, 'Data Poisoning: The Exploitation of Generative AI', *Crowdstrike*, 20 March 2024.

37 Niu, Jun and others, 'A survey on membership inference attacks and defenses in machine learning', *Journal of Information and Intelligence*, 8 March 2024; Machine Learning @ Berkeley, 'Tricking Neural Networks: Create your own Adversarial Examples'; Abdollahzadeh, Milad and others, 'Re-thinking Model Inversion Attacks Against Deep Neural Networks', *arXiv*, 15 June 2023; Li, Na and others, 'Machine Unlearning: Taxonomy, Metrics, Applications, Challenges, and Prospects', *arXiv*, 13 March 2024.

9. Prevention and detection of such attacks are key; models can be defended through the use of adversarial training, data cleaning, and model ensembling, though a continuous battle of development between attackers and defenders of ML systems endures.[38] [39] [40] A focus on detecting malicious inputs prior to the next training cycle is essential.[41] If a machine suffers an attack, developers may retrain it to forget poisoned data through 'machine unlearning' (MU).[42] Two categories of the MU paradigm exist; 'exact', which completely retrains a model to ensure the erasure of certain data, though is computationally expensive, particularly if dealing with complex models and / or a high level of data, and 'approximate', which aims for a more efficient removal that doesn't deteriorate model performance, though is less accurate at 'forgetting' data, and therefore carries an element of risk.[43]

10. Challenges exist in that many MU methods are designed for Centralised Machine Learning (CML) as opposed to Distributed Machine Learning (DML) which utilises multiple computational modes, complicating access to data and model parameter control.[44] Additionally, MU processes may inadvertently make an unlearned model more vulnerable than its original; Chen at al note that adversaries could take advantage of the fact that excessive unlearning can cause model parameter shifts, and / or model degradation, so may purposefully trigger MU.[45] Lam et al identify several areas in the unlearning system which could lead to threats and attacks, though note that the interplay between these and defences within MU systems are complex.[46] Indeed, MU methods can act as a passive defence against damage caused from data poisoning and backdoor attacks, an active defence to ensure privacy attacks fail, whilst also being a cause of vulnerability, or alternatively exploited to manipulate an adversary's system.[47] Those seeking to utilise human-machine teaming (HMT) should be aware that MU is in its early stages of development and continues to face challenges; though progress in the field is assured, experimental methods are arguably still too uncertain to be applied to real-world, practical situations that may have lethal results.

38 Machine Learning @ Berkeley, 'Tricking Neural Networks: Create your own Adversarial Examples'.

39 Brahim Belhaouari, Samir, Kraidia, Insaf and Ghenai, Afifa, 'Defense against adversarial attacks: robust and efficient compressed optimized neural networks', *Scientific Reports*, 17 March 2024.

40 Go, Brendon and Liu, Evan 'Ensembling as a Defense Against Adversarial Examples', *Stanford University*, 15 December 2016.

41 Regression testing, input validity checking, rate limiting, manual moderation and other statistical techniques used to detect anomalies are recommended. See: Constantin, Lucian, 'How data poisoning attacks corrupt machine learning models', *CSO*, 12 April 2021.

42 Due to MU being in its relative infancy, Kurmanji et al note that no well-established formal definition of the issue of unlearning exists, nor do well-established metrics for measuring the quality of unlearning algorithms. See: Kurmanji, Meghdad and others, 'Towards Unbounded Machine Unlearning', *arXiv*, 30 October 2023.

43 Na, Li and others, 'Machine Unlearning: Taxonomy, Metrics, Applications, Challenges, and Prospects', *arXiv*, 13 March 2024.

44 Chen, Hui and others, 'Machine Unlearning: Taxonomy, Metrics, Applications, Challenges, and Prospects'.

45 Chen, Hui and others, 'Machine Unlearning: Taxonomy, Metrics, Applications, Challenges, and Prospects'.

46 Chen, Chen and others, 'Threats, Attacks, and Defenses in Machine Unlearning: A Survey', *arXiv*, 20 March 2024.

47 Chen, Hui and others, 'Machine Unlearning: Taxonomy, Metrics, Applications, Challenges, and Prospects'.

Scale of human-machine decision making

11. Decision making is a complex cognitive process that involves evaluating options and choosing the best course of action. Humans and machines both have the ability to make decisions, but each approach this task differently. The volatile, uncertain, complex, and ambiguous (VUCA) operating environments demand the military commander make timely, well-informed decisions. For operational commanders, decision making can be considered an art, not a science; however, science can complement the commander to make timely, well-informed, effective decisions.[48] From human-only decision making to fully autonomous decision making by machines, each stage represents a different level of autonomy and complexity. As technology continues to advance, the boundaries between human and machine decision making will continue to blur, raising important questions about ethics, responsibility, and the future of decision making.

12. Human-machine interfaces of various kinds are now ubiquitous in everyday life and as technology advances, the role of machines in decision making is becoming increasingly prominent. In the complex and fast-paced defence and security environment, it will be essential. AI, ML, HMT, cognitive computing, and automation have enabled machines to assist or even replace humans in the decision making process. Human decision making and machine decision making can be understood to exist on a sliding scale or continuum (Fig 1), with various levels and factors influencing the process. Situational factors, time factors, information (data) factors, technology factors, legal factors, risk factors, and trust factors all play a role in determining where a decision falls on this continuum.

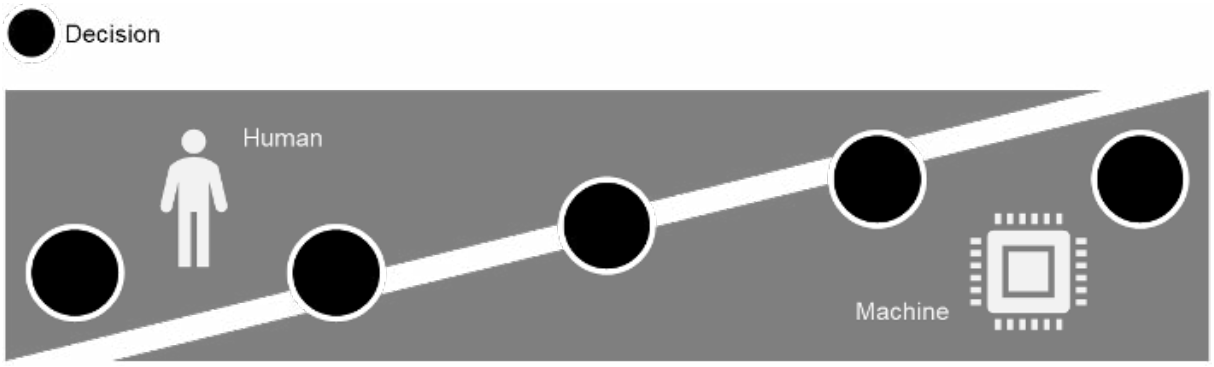

Figure 1, Human to Machine decision making continuum

13. At one end of the spectrum is human-only decision making, where individuals rely solely on their own judgment, training, experience, and intuition. In uncertain and competitive environments, intuition enables rapid and instinctive responses that can prove advantageous. To aid energy optimisation, the human brain employs cognitive shortcuts or heuristics, which are mental strategies that simplify complex problems and allow for faster decision making. While these shortcuts generally work well, they can also lead to errors when faced with novel or ambiguous situations. Another factor which influences, and can degrade human decision making, is cognitive bias. Cognitive biases are inherent tendencies in human thinking that can lead to deviations from rational

48 Menna, Michael "Effective Decision making Employing Human-Machine Teaming" 3 May 2022.

judgment. Over 100 cognitive biases have been identified.[49] In a military context, authority bias borne of rigid hierarchy may also have considerable influence. Situational factors such as the complexity of the decision, the stakes involved, and potential consequences, and the availability of information will all impact the quality of human decision making. Urgency or time constraints can also influence the speed and accuracy of decisions made by humans.

14. At the intermediate level of the continuum, human-machine decision making involves a combination of human input and machine analysis. Humans provide the context, goals, and constraints, while machines process data, identify patterns, and generate recommendations. This collaborative approach leverages the strengths of both humans and machines to make more informed and efficient decisions. According to Hoffman at al, "humans surpass machines in their ability to improvise and use flexible procedures, exercise judgement and reason inductively. Machines outperform humans in responding quickly, performing repetitive and routine tasks, and reason deductively".[50]

15. Trust factors play a crucial role in determining the level of autonomy granted to machines in decision making. Trust in technology is influenced by factors such as reliability, transparency, accountability, and ethical considerations. In human-machine teams Mayer, Davis, and Schoorman defined trust as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party".[51] Many machine learning and control systems developed in a wide variety of applications are 'black boxes' that do not, indeed often cannot, explain how decisions are made. With small data sets and simple decisions, it is easy to verify information and decisions, for example look out of a car window to check a street name to confirm a navigation instruction. As machines demonstrate their ability to make accurate decisions, trust in their capabilities grows, inevitably leading to increased autonomy. However, trust is fragile and can be easily eroded by errors, biases, or lack of human understanding of how machines make decisions.

16. Moving towards the other end of the spectrum is fully autonomous decision making by machines. In this scenario, machines have complete control over the decision making process, from data collection and analysis to action implementation. Fully autonomous machines are capable of learning from experience, adapting to changing circumstances, and making decisions without human intervention. This level of autonomy raises ethical and legal concerns, as machines may not always prioritise human values or consider the broader societal impact of their decisions.

17. Fully autonomous decision making by machines is a complex and evolving field, with various stages and levels of autonomy. At the lowest level, machines are programmed to follow predefined rules and algorithms, making decisions based on a set of predetermined criteria. As machines become more sophisticated, they can learn

49 Geng, Baocheng and Varshney, Pramod K. "*Human-Machine Collaboration for Smart Decision Making: Current Trends and Future Opportunities*", 18 January 2023.

50 Hoffman, R.R., Feltovich, P. J., Ford, K. M., and Woods D. D., *"A rose by any other name... would probably be given an acronym [cognitive systems engineering]*," IEEE Intelligent Systems, vol. 17, no. 4 2002.

51 Mayer, R.C., Davis, J. H., and Schoorman, F. D., *"An integrative model of organizational trust*," Academy of Management Review, vol. 20, no. 3, 1995.

from data, recognise patterns, and make predictions without explicit instructions. Machine learning algorithms enable machines to improve their decision making capabilities over time, leading to higher levels of autonomy.

18. At the highest level of autonomy, machines are capable of self-learning, self-improving, and self-optimising their decision making processes. These autonomous systems can adapt to new information, unforeseen events, and changing environments, making decisions in real-time without human oversight. While fully autonomous machines offer the potential for increased efficiency, productivity, and innovation, they also raise concerns about accountability, transparency, and control.

Criteria for successful decision making

19. Success and failure of decisions are defined and measured against factors, including the impact of the decision, the planned outcome, and the consequences, both intended and unintended. Simplistically, success can typically be defined as achieving a desired outcome or goal, while failure is the inability to reach that outcome or goal.

20. It can be argued that decision evaluation should be identical for both human and machine-made decisions purely based on effect. However, as previously discussed, human decisions are influenced by emotions, biases, and personal values, which can impact the ultimate success or failure of a decision. Human decisions can be explained, whereas machine decisions often cannot. Machine decisions are evaluated on the algorithms and data used to make the decision, as well as the accuracy and efficiency of the decision making process. Machines are limited by the data they are trained on and may not always consider ethical or moral considerations in their decision making process, as a human intuitively would (should) but these play a crucial role in evaluating the success or failure, whether made by humans or machine systems, especially in high-risk defence and security environments. An 'acceptable' machine decision may result in humanly 'unacceptable' collateral outcomes. The public outcry at the killing of civilians associated with the recent use of the Lavender AI system by Israel to identify Hamas targets, is one example.[52]

21. Accountability is another important criterion in evaluating the success or failure of decisions, as it ensures that decision-makers are held responsible for their actions and the consequences of their decisions. In human decision making, accountability may be measured by the ability of the decision-maker to justify their decision, take responsibility for their actions and any mistakes, and learn from their failures. In the case of AI systems, accountability may be more complex, as the decision making process is often opaque and difficult to trace back to a specific individual. The Future of Life Institute highlighted the "failure of transparency" and "failure of judicial transparency" as illustrations that AI cannot provide a satisfactory explanation to humans as to why a decision is made.[53]

22. Depending on where and how much a human is 'in the loop' in the decision making process, will inevitably be a consideration in accountability for decisions. If the Turing Test is passed and true and total autonomy is ever reached and employed in the defence context, and is involved in a humanly perceived catastrophic failure, how far back in the development chain does the 'blame' lie?

52 McKernan, Bethan and Davies, Harry, "'The machine did it coldly': Israel used AI to identify 37,000 Hamas targets" The Guardian, 3 April 2024.

53 Future of Life, "Asilomar AI Principals", 17 March 2022.