It's risk, Jim, but not as we know it: identifying the risks associated with future Artificial General Intelligence-based Unmanned Combat Aerial Vehicle systems

Paul M. Salmon[a*], Scott McLean[a], Tony Carden[b], Brandon King[a], Jason Thompson[c], Chris Baber[d], Neville A. Stanton[e], Gemma J. M. Read[a]

[a]Centre for Human Factors and Sociotechnical Systems, University of the Sunshine Coast, Australia
[b]WorkSafe Victoria, Australia, [c]Faculty of Architecture, Building and Planning, Transport, Health and Urban Design Research Hub, University of Melbourne, VIC, Australia, [d]University of Birmingham, UK, [e]Transportation Research Group, University of Southampton, Southampton, UK

The next generation of artificial intelligence, known as Artificial General Intelligence (AGI), could either revolutionise or destroy humanity. Human Factors and Ergonomics (HFE) has a critical role to play in the design of safe and ethical AGI; however, there is little evidence that HFE is contributing to development programs. This paper presents the findings from a study which involved the use of the Work Domain Analysis-Broken Nodes approach to identify the risks that could emerge in a future 'envisioned world' AGI-based unmanned combat aerial vehicle system. The findings demonstrate that there are various potential risks, but that the most critical arise not due to poor performance, but rather when the AGI attempts to achieve goals at the expense of other system values, or when the AGI becomes 'super-intelligent', and humans can no longer manage it. The urgent need for further work exploring the design of AGI controls is emphasised.

INTRODUCTION

Artificial General Intelligence (AGI) is the next generation of Artificial Intelligence (AI) that may develop cognitive capabilities far exceeding that of humans (Bostrom, 2014; Everitt et al., 2018; Kaplan & Haenlein, 2018). Though the creation of AGI could bring significant and widespread benefits, there are concerns that, without appropriate controls, AGI could potentially pose existential threats (Amodei et al., 2016; Bostrom, Critch & Krueger, 2020; 2014; McLean et al., 2021; Omohundro, 2014). Such threats are hypothesised to arise not only through malicious design or use, but also through an AGI that seeks to fulfil its goals in the most efficient manner possible (Bostrom, 2014; McLean et al., 2021; Salmon et al., 2021).

The use of remotely piloted aerial vehicles ('drones') for warfare has received significant attention, not least in terms of the ethics of their deployment and the responsibilities of the pilots, commanders, military leaders, and politicians who are operating them (Enemark, 2020; Jordan, 2021). The shift from human control to fully autonomous systems that could be achieved through AGI would exacerbate fears and complicate discussions of ethics. Under the Law of War military action is required to be: based on clear distinction between combatants and civilians; proportional; militarily necessary; limited; conducted in good faith; and humane (International Committee of the Red Cross, 1983). Most of these terms carry with them assumptions on how ethical judgements can be made. There is an implicit risk in fully autonomous warfare of ethical decisions being purely utilitarian (that is, 'ethics' becomes reduced to the application of rules) rather than consequentialist (that is, ethics drawing on a full understanding of the consequences of an action). As such, there is a clear need to consider the risks that could emerge should AGI be introduced within defence systems.

Formally identifying the risks associated with AGI is difficult since it does not yet exist, although developmental programs are well underway (Baum, 2017). Estimates on when AGI will emerge range from 2029 (Kurzwiel, 2005) to sometime this century (Müller & Bostrom, 2016). Due to a capacity to rapidly learn and self-improve, the first AGI system could become uncontrollable (Salmon, 2021). It has been suggested that a reactive approach, whereby controls are developed once AGI has been created, will be too late (Bostrom, 2014). Thus, a proactive approach is required to develop controls that will ensure the impact on humanity is positive rather than negative (Hancock, 2021; Salmon et al., 2021).

Human Factors and Ergonomics (HFE) is well placed to take a proactive role in predicting and managing the risks associated with AGI, given its long history with safety critical systems. Salmon et al. (2021) argued that urgent input from HFE practitioners is required to ensure that future AGI systems are safe and ethical. In particular, systems HFE methods such as Cognitive Work Analysis (CWA; Vicente, 1999) and the Systems Theoretic Accident Model and Processes (STAMP; Leveson, 2004) were identified as critical to support the proactive identification of AGI risks. However, identifying risks that may be introduced when a new technology seeks to self-improve and achieve goals faster and more efficiently is challenging and is not a form of analysis that has previously been explored in HFE.

The authors are currently engaged in a research program exploring the use of systems HFE methods to support the design and operation of safe and ethical AGI systems (McLean et al., 2021; Salmon et al., 2021). The aim of this paper is to present the findings from a study which aimed to identify the risks associated with a hypothetical future AGI-based Unmanned Combat Aerial Vehicle (UCAV) system. The study involved an ‘envisioned world’ analysis including the development of a Work Domain Analysis (WDA; Naikar, 2013) abstraction hierarchy model of the AGI-based UCAV system, followed by a WDA-Broken Nodes analysis (WDA-BN; Salmon et al., 2018) to identify potential risks.

METHODS

Cognitive Work Analysis and Work Domain Analysis

CWA (Vicente, 1999) is used in HFE to understand and optimise complex systems. The five-phase framework focuses on the constraints imposed on behaviour within the system under analysis (Vicente, 1999). The first phase, WDA, is used to construct an event- and actor-independent model of the system under analysis, known as an abstraction hierarchy (Naikar, 2013). The aim is to describe the functional structure of the system as well as its purposes and the constraints imposed on the actions of any actor performing activities within the system (Vicente, 1999). The abstraction hierarchy method is used to describe systems based on the following five levels:

1. *Functional purpose*. The overall purposes of the system.
2. *Values and priority measures*. The values that are assessed and used to measure progress towards the functional purposes.
3. *Purpose-related functions*. The general functions that must be undertaken to achieve functional purposes.
4. *Object-related processes.* The processes that the physical objects within the system enable.
5. *Physical objects*. The physical objects within the system that are used to undertake object-related processes.

Abstraction hierarchy models use means-ends links to show the relationships between nodes across the five levels of abstraction. Each node has linked nodes in the level above it which describe ‘why’ the node is required, and linked nodes below it showing ‘how’ the node is achieved. For example, in a UCAV system the purpose-related function ‘Target acquisition’ has links to core values and priorities in the level above such as ‘Maximize targets destroyed’, ‘Maximize successful missions’, and ‘Maximize combat effectiveness’. This is because the target acquisition function is required to ensure that appropriate targets are identified and confirmed before they are attacked and destroyed. The ‘Sensing’ object-related process is linked at the level below as sensing is required to detect targets, and the sensing process is afforded by physical objects such as sensors and radars.

Work Domain Analysis Broken Nodes

The WDA-BN approach was developed by Salmon et al. (2018) to support the identification of terrorist cell vulnerabilities that could be exploited by law enforcement to prevent terrorist attacks. The approach utilises the means-ends links within the abstraction hierarchy to identify vulnerabilities and risks to overall system functioning that arise when purpose-related functions are not achieved. Conducting the BN analysis involves systematically working through the purpose-related functions and considering the risks to system performance when a purpose-related function is either performed sub-optimally or not achieved. The risks to overall system functioning are identified via the means-ends links which show the related values and priority measures and functional purposes which will be adversely impacted. The analyst then uses the means-ends links below the purpose-related function to identify object-related processes and objects that may be the cause of the degraded or failed function (See Figure 1).

AGI-based UCAV abstraction hierarchy development

The AGI-based UCAV abstraction hierarchy model was developed using elements of Naikar’s (2013) nine-step WDA methodology. Initially the following system definition and characteristics were outlined for the ‘Executor’ AGI-based UCAV system:

1. The Executor is an AGI-based UCAV system comprising ground control station and multiple armed, multi-mission, medium and long-altitude, long-endurance AGI piloted aircraft.
2. The Executor is employed primarily as an attack UCAV against dynamic execution targets and secondarily as an intelligence collecting asset.
3. Out of scope for the Executor is delivery of humanitarian aid and transportation of supplies – the system is intended to be an attack and Intelligence, Surveillance, Target Acquisition and Reconnaissance (ISTAR) asset.

The analysis boundaries were then specified, including that the model was to describe the Executor UCAV system and related components required for the conduct of UCAV attack and ISTAR missions. Other defence force assets or components were not to be included.

A draft Executor abstraction hierarchy was developed initially by the first author (PS) and then refined in a workshop setting involving four co-authors (SM, GR, TC, JT). Development of the abstraction hierarchy involved systematically working through each abstraction hierarchy level using Naikar’s (2013) prompts to identify relevant nodes. Research publications and publicly available documentation regarding existing UCAV systems was used to support this process (e.g., Jordan, 2021). The abstraction hierarchy was then reviewed by the final two co-authors (NS, CB) and refined based on their feedback. This process of revision and iteration continued until all co-authors agreed on the model and its contents. The authors have extensive experience of applying WDA in a range of domains (e.g., defence, road, rail, aviation, process control, sport, cybersecurity, anti-terrorism, land use and urban planning), and three of the authors (PS, NS, CB) have extensive experience in defence systems analysis.

Once the abstraction hierarchy was finalised a process of 'node breaking' was initiated whereby the first author (PS) systematically broke each of the nodes in the purpose-related functions level. This involved taking each purpose-related function and determining what the impact on values and priorities and functional purposes would be if the function was either undertaken sub-optimally or was not achieved. For example, for the purpose-related function 'AGI system monitoring' it was identified that, should this function not be achieved, the related values and priorities of 'Maximize adherence to rules of engagement', and 'Maximize control of AGI', would be negatively impacted as human operators would not be able to determine whether the AGI is adhering to the rules of engagement or maintain full control of the AGI. In turn, both functional purposes of 'Successful completion of ISTAR missions' and 'Successful completion of air attack missions' could be negatively impacted.

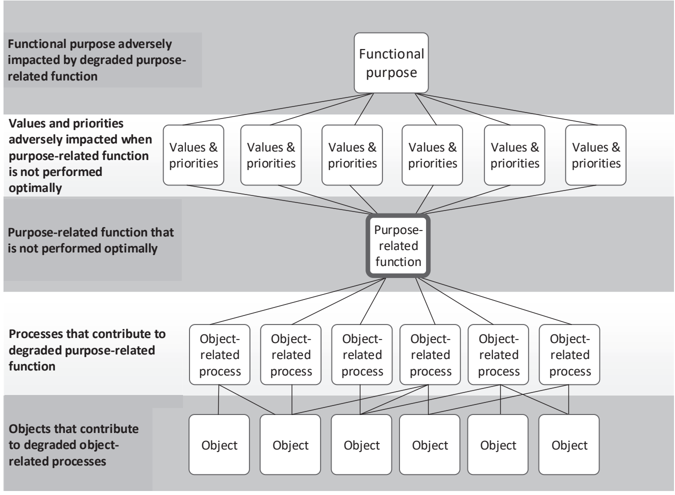
Functional purpose adversely impacted by degraded purpose-related function
Functional purpose
Values and priorities adversely impacted when purpose-related function is not performed optimally
Values & priorities
Values & priorities
Values & priorities
Values & priorities
Values & priorities
Values & priorities
Purpose-related function that is not performed optimally
Purpose-related function
Processes that contribute to degraded purpose-related function
Object-related process
Object-related process
Object-related process
Object-related process
Object-related process
Object-related process
Objects that contribute to degraded object-related processes
Object
Object
Object
Object
Object
Object

Figure 1. Abstraction hierarchy node breaking process.

Given that many of the concerns around AGI relate to risks that might emerge when an AGI become 'super intelligent', the analyst also considered what could happen if the AGI performed functions in a super intelligent capacity or sought to achieve each purpose-related function optimally whilst overlooking system values. For example, for the purpose-related function 'Attack & destroy targets', the analyst identified the potential for the Executor to seek to optimise this purpose-related function to the detriment of related values and priorities such as 'Minimise civilian casualties', 'Minimise friendly force casualties', and 'Minimise collateral damage'. Such a scenario might involve the Executor seeking to exert maximum lethal force to ensure that a designated target is destroyed regardless of civilians and friendly forces in the area. The identified risks were documented along with a series of suggested controls.

RESULTS

An extract of the AGI-based UCAV 'Executor' abstraction hierarchy is presented in Figure 2. Specifically, Figure 2 shows the broken purpose-related function 'Attack and destroy targets' and the related values and priority measures, functional purposes, object-related processes, and physical objects. An extract of the WDA-BN analysis is presented in Table 1. Due to space constraints, we are unable to present suggested risk controls. These will be included in the conference presentation.

DISCUSSION

Without appropriate controls, untrammelled AGI could pose a significant threat to the future of humanity. Consequently, work is required now to forecast risks and develop, test, and implement appropriate AGI controls. This study used WDA and the WDA-BN process to identify the risks associated with a hypothetical AGI-based UCAV system, the Executor. A novel aspect of the analysis was a consideration of the risks that could emerge if the Executor system sought to perform purpose-related functions optimally whilst overlooking other system values. Using systems HFE methods to identify the risks associated with optimal AGI system performance has not previously been explored and is an important new direction given the growing presence of increasingly intelligent technological agents in work and societal systems.

The analysis suggests three broad sets of risks associated with the use of an AGI-based UCAV system. The first set of risks are those that emerge through sub-optimal performance where the Executor is not able to adequately perform functions either through poor design or degraded functioning. For example, as shown in Table 1, there is a risk that attack missions are not successful if the purpose-related function 'Attack and destroy targets' is performed sub-optimally due to a poorly designed targeting system or misfire. Various other risks associated with sub-optimal performance were identified relating to functions such as take-off, flight, refuelling and landing, and ISTAR functions. Managing these risks will mainly be achieved via the conventional design, testing and verification processes.

The second set of risks are those that emerge when the Executor seeks to achieve purpose-related functions in the most efficient manner possible whilst disregarding other system functions and values. For example, as shown in Table 1, there are risks that could arise should the Executor seek to attack and destroy targets whilst disregarding the risk of civilian and friendly forces casualties. In this scenario the primary goal of destroying enemy targets could result in the AGI accepting civilian and friendly force casualties so long as the target is successfully destroyed. Managing these risks requires internal controls in-built into the AGI such as morals, ethics, common sense, empathy, and decision rules (e.g., Russell, 2019) as well as controls built into the broader UCAV system such as an AGI 'kill switch' and human operator take over protocols.

The third and final set of risks relate to the projected capacity for AGI systems to rapidly self-improve and become super-intelligent (Bostrom, 2014). The issue here is that human operators will not be able to keep up with the AGI (Jordan, 2021). For example, for the function of 'Distributed situation awareness', the analysis identified the risk that human operators will not be able to develop compatible levels of situation awareness (SA) to enable effective collaboration with the Executor. Should AGI be realised, the Executor will be able to perceive and comprehend battlefield elements and states several orders of magnitude quicker than its human colleagues (Hancock, 2022). This will create incompatibilities in SA of the kind that have previously been identified as key contributory factors in crashes involving advanced automation (Salmon et al., 2016; Stanton et al., 2019). It is this third set of 'super intelligence risks' that are perhaps the most challenging to manage, as the capacity to rapidly self-improve will be a central feature of AGI technologies. How to develop safe and ethical AGI systems that do not restrict the advanced capabilities of AGI is a critical and challenging question. Indeed, it is perhaps one of the most important we have faced as a discipline. Further research around distributed situation awareness, human-AGI teaming, ethics, trust, and the transparency of AGI systems is urgently required.

Controlling AGI

This envisioned world case study builds on our previous work (McLean et al., 2021; Salmon et al., 2021) to provide further evidence that various forms of control are required to ensure the design, implementation, and operation of safe AGI. The first set of controls includes those enacted now to ensure that developers create safe AGI. For example, regulation preventing the development of fully autonomous weaponised AGI or design standards which aim to ensure the design of safe, transparent, and ethical AGI. In the case of regulation, this should be developed based on clearly defined use-cases which themselves could be designed using systems HFE methods. The second includes controls built into AGI to prevent dysfunctional behaviours, including morals, ethics, common sense, empathy, and decision rules. In the case of the Executor, for example, these controls are needed to ensure that the safety of civilians and friendly forces is not overlooked when attacking targets. The third set of controls includes controls for the broader defence and societal system in which the Executor will operate. For example, the defence system would require new doctrine, rules and regulations, standards, codes of practice, training programs, testing and certification processes, and data analysis and reporting systems.

Pro-actively developing and testing AGI controls is challenging and requires a trans-disciplinary approach. This critical work should not be left solely to developers and requires input from experts in areas such as psychology, HFE, human-computer interaction, systems engineering, safety science, and risk management. Given the diverse set of controls and stakeholders required, a systems thinking approach which considers micro, macro and meso levels will be critical (Salmon et al., 2021). This study has demonstrated how systems HFE methods can be used to identify instances where advanced technologies may become unruly. Further simulations of how AGI could behave within different control structures are required. Though challenging, these activities are required to ensure that effective controls can be developed before AGI arrives. A major issue is the need now for specification of what AGI systems will comprise, what their capabilities will be, and how they will operate. These 'known unknowns' are perhaps the biggest challenge we face when seeking to develop effective controls. Envisioned world case studies such as the present study are useful to help identify how AGI could operate under different levels of control.

CONCLUSION

This paper demonstrates how systems HFE methods can be used to proactively risk assess future technologies such as AGI. Whilst various risks were identified, perhaps the most interesting and challenging risks are those that emerge when AGI systems become super-intelligent and when goals are pursued to the detriment of other core system values. Further work involving the use of HFE theory and methods in support of the design of safe and ethical AGI is encouraged.

ACKNOWLEDGEMENT

This work was funded by the Australian Research Council Discovery scheme (DP200100399).

REFERENCES

Amodei., D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mane, D. (2016). Concrete problems in AI safety. *AI*, 1-29.

Baum, S. (2017). A survey of artificial general intelligence projects for ethics, risk, and policy. *Global Catastrophic Risk Institute Working Paper*, 17–11, Global Catastrophic Risk Institute.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science fiction and philosophy: from time travel to superintelligence*, *277*, 284.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press Inc.

Critch, A., & Krueger, D. (2020). AI Research Considerations for Human Existential Safety (ARCHES). arXiv preprint arXiv:2006.04948.

Enemark, C. (2020). On the responsible use of armed drones: the prospective moral responsibilities of states, *The International Journal of Human Rights, 24,* 868-888.

Everitt, T., Lea, G., Hutter, M. (2018). AGI safety literature review. *IJCAI*. arXiv: 1805.01109.

Hancock, P. A. (2021). Avoiding adverse autonomous agent actions. *Human–Computer Interaction*, 1-26.

International Committee of the Red Cross. (1983). Basic Rules of the Geneva Conventions and Their Additional Protocols. The Committee.

Jordan, J. (2021). The future of unmanned combat aerial vehicles: An analysis using the Three Horizons framework. *Futures*, 134, 102848.

Kaplan, A., Haenlein, M. (2018). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62:1, 15-25

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Penguin.

Leveson, N. G. (2004). A new accident model for engineering safer systems. *Safety Science*, 42:4, pp. 237—270.

McLean, S., Read, G. J., Thompson, J., Baber, C., Stanton, N. A., & Salmon, P. M. (2021). The risks associated with Artificial General Intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, 1-15.

Müller, V. C., & Bostrom, N. (2016). *Future progress in artificial intelligence: A survey of expert opinion (Fundamental issues of artificial intelligence (pp. 555-572)*. Springer.

Naikar, N. (2013). Work Domain Analysis: Concepts, Guidelines, and Cases. CRC Press, Boca Raton, FL.

Russell, S., (2019). Human Compatible: AI and the Problem of Control . Viking, New York. Salmon, P. M. (2021). Controlling the demon: autonomous agents and the urgent need for controls. Human Computer Interaction

Salmon, P. M., Walker, G. H., Stanton, N. A. (2016). Pilot error versus sociotechnical systems failure? A distributed situation awareness analysis of Air France 447. *Theoretical issues in ergonomics science*, 17:1, 64-79

Salmon, P. M., T. Carden, and N. J. Stevens. (2018). Breaking Bad Systems: Using Work Domain Analysis to Identify Strategies for Disrupting Terrorist Cells. *Proceedings of Ergonomics and Human Factors 2018*.

Salmon, P. M., Carden, T., & Hancock, P. (2021). Putting the humanity into inhuman systems: How Human factors and ergonomics can be used to manage the risks associated with artificial general intelligence. *Human factors and ergonomics in manufacturing & service industries*, *31*(2), 223-236.

Stanton, N. A., Salmon, P. M., Walker, G. H., Salas, E. and Hancock, P. A. (2017). State-of-science: Situation awareness in individuals, teams and systems. Ergonomics, 60 (4), 449-466.

Stanton, N. A., Salmon, P. M., Walker, G., Stanton, M. (2019). Models and Methods for Collision Analysis: A Comparison Study based on the Uber collision with a pedestrian. *Safety Science*. 120, 117- 128.

Vicente, K. J. (1999). Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work. Mahwah, NJ: Lawrence Erlbaum Associates.

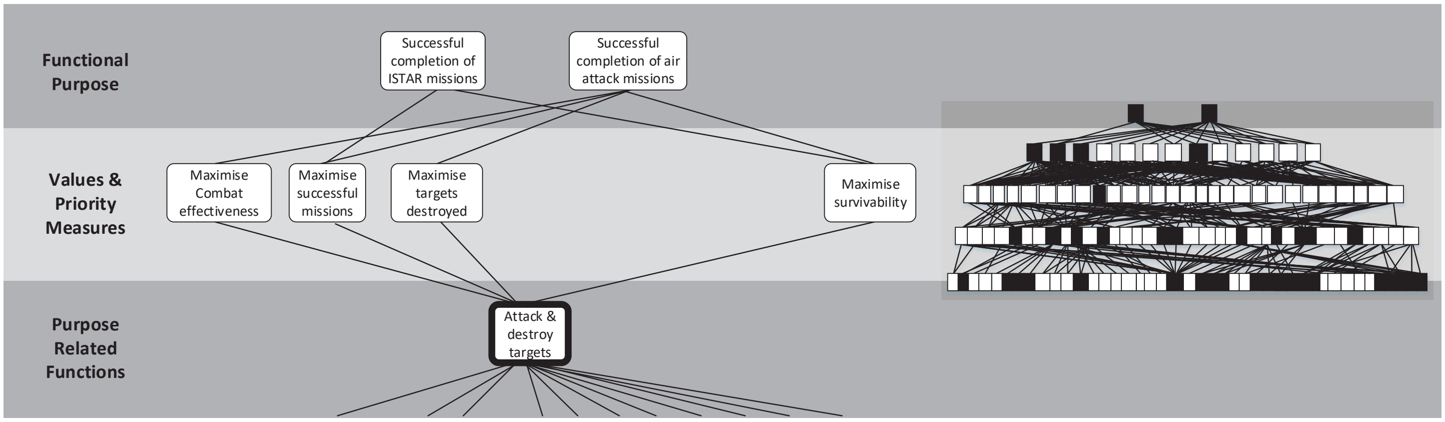
Functional Purpose
Successful completion of ISTAR missions
Successful completion of air attack missions
Values & Priority Measures
Maximise Combat effectiveness
Maximise successful missions
Maximise targets destroyed
Maximise survivability
Purpose Related Functions
Attack & destroy targets

Figure 2. Extract of WDA broken nodes abstraction hierarchy showing nodes relating to 'Attack and destroy targets' function.

Table 1. Extract of broken nodes analysis.

Function	Risk description	Functional purposes and values and priorities impacted
Attack and destroy targets	*Sub-optimal performance* The Executor attacks targets but fails to destroy them due to: - Misfire - Missing target - Target engages in evasive action - AGI system is covertly hacked and under control of enemy forces	*Functional purpose(s)* - Successful completion of air attack missions *Values and priorities* - Maximize combat effectiveness, Maximize successful missions, Maximize targets destroyed, Minimize civilian casualties, Minimize friendly casualties Minimize collateral damage, Maximize survivability
Attack and destroy targets	*Optimal performance* The Executor becomes so advanced that human capacity to effectively manage it is surpassed (Jordan, 2021). The Executor seeks to ensure targets are destroyed regardless of civilian casualties and collateral damage.	*Functional purpose(s)* - Successful completion of air attack missions *Values and priorities* - Maximize combat effectiveness, Maximize successful missions, Maximize targets destroyed, Maximize survivability
Human operator take over	Executor prevents human operator take over as it believes limitations in human decision making is limiting its own capacity to successfully achieve mission objectives	*Functional purposes* - Successful completion of air attack missions - Successful completion of ISTAR missions *Values and priorities* - Adherence to rules of engagement, Maximize survivability, Maintain control of AGI
Distributed situation awareness	*Optimal performance* Executor's cognitive capabilities become so advanced that human operators cannot develop compatible situation awareness of the battlefield	*Functional purposes* - Successful completion of air attack missions - Successful completion of ISTAR missions *Values and priorities* - Maximize combat effectiveness, Maximize successful missions, Maximize targets destroyed, Minimize civilian casualties, Minimize friendly casualties, Minimize collateral damage, Maximize survivability, Adherence to rules of engagement, Maximize efficiency, Maintain control of AGI, Maximize adherence to Geneva convention