Trustworthiness for AI in Defence (TAID)

-

Human Factors Case Study Annex

Submission date: 09/05/2025

Document version: 1.0

Authors: TAID Working Group (TAID WG)

Table of Contents

The Case-Study - The Impact of AI on each level of the STS

The purpose of this case Study is three-fold:

1) To provide a narrative and context of use of an AI system to enable the readership to understand how the humans and other stakeholders may benefit from the addition of AI and how any challenges may be incurred by the wider system.
2) To present the context with detail for a readership that is perhaps not familiar with military defence operations (i.e. the public readership).
3) To demonstrate the complexity of a socio-technical system and to aid the readership is seeing the scenario from multiple perspectives.

It is hoped that readers understand why it is paramount that multiple stakeholders are included in the entire design lifecycle and that AI systems are design from both value-based and systems-based approaches. This case-study is cross-referenced to Human Factors (Chapter 6 of TAID White Paper) and Ethics (Chapter 7 of TAID White Paper). It is recommended that further scenario-based work is done for future research to continue this work.

1.1. System of Interest: Drone with Autonomous Weapon Capability

The Stakeholders (as defined in Chapter 2):

AI Customers

- Drone Operator (full list of stakeholders in impact section)
- Team members (human) supporting drone operator in the field
- Team member (non-human) providing decision support for drone operator
- Commanding officer(s)

AI subjects (neutral)

- Civilians around target area
- The environment
- Future generations

AI partners - Other stakeholders involved in the design, testing and certification of the technologies such as software engineers, technology manufacturers and those involved in safety management such as risk assessors, lawyers etc. A full list of these is available in Chapter 2 Stakeholders in AI for Defence.

The application of AI in Unmanned Aerial Vehicles (UAVs) for military purposes has seen a rapid growth and focus in modern conflict. However, the increasingly electronically contested battlespace hinders successful use of, in particular, FPV (First Person View) drones for tracking and engaging hostile targets by drone operators. Furthermore, jamming of GPS can also prevent accurate navigation towards intended observation locations. Currently, many of these UAVs are considered an instant loss, as most require a human operator to keep them airborne or on track. To counter a loss of capability due to signal loss, armed drones have recently been enabled by AI to not only track but also adjust their final trajectory towards a

target. The application of AI in these systems transforms a remote-controlled system to one that has (full) autonomy and that functions as a kinetic effector.

The Case study narrative is presented below, followed by two sections: 1) detail of the positive impact on each level of the socio-technical system and 2) the negative impact of AI on each level of the socio-technical system.

1.2. Case-Study Narrative

This case study presents a scenario of a drone operator presented with novel challenges due to the abrupt and likely unexpected loss of signal and therefore control over such systems. A drone operator has been tasked to support her battalion in the defence of a strategic position. She is experienced in flying FPV drones mounted with shaped charges to immobilise enemy armoured vehicles. Recently, she has been trained to also fly a drone with AI-based autonomous tracking and engagement capabilities. This is her third time using a drone with this new technology in the field and her task is a critical one as the hostile forces outnumber their own. The mission goal of drone operator(s) in this area is to block the enemy attack and immobilise the Armoured Personnel Carriers (APCs). She is positioned several kilometres from the frontline and supported by her teammate who can watch the same video feed while also keeping a watchful eye in the sky. Drone operators are regularly targeted by enemy FPV drones and wearing a head-mounted display may strongly inhibit the ability to maintain situational awareness of her direct surroundings.

She flies her drone over her own forces towards the frontline, aware that she must be absolutely certain the autonomous engagement mode is currently disabled. A loss of connection close to her own forces is not uncommon during electronic warfare (EW). Unfortunately, the current AI on-board the drone is incapable of distinguishing between friend or foe because similar military equipment is used on both-sides and processing power is limited. Although the interface requires confirming not only the autonomous mode but also the target to be engaged, she does not want to risk a possible human error resulting in her drone targeting her own forces. She continues her flight while being informed by her teammate about the situation in the engagement area. A fixed-wing drone high up in the sky has been monitoring the area for some time. Several hostile APCs have been identified approaching a possible infantry drop-off point. Being able to immobilise the APCs would be an important step in preventing the advance of the adversaries.

Now that her drone has the enemy APCs in sight, her interface is automatically starting to track the moving APCs. She experiences variable video quality due to the distance between her and the drone, but possibly also due to EW. The closer the APCs get, the more time-critical her decision becomes. Does she choose to attempt a manual engagement on the moving APCs, or does she enable its autonomous engagement capabilities? She trusts her own skills of hitting a moving target, but a loss of signal negates all of that. She could wait for the APCs to move closer in the hopes that the signal quality improves, yet this also means having her drone closer to her own troops. An autonomous engagement is an option, but the APCs are kicking up large dust clouds and are moving at high speeds. This could be a challenge for the AI.

Her teammate reports a fellow drone operator decided to manually engage the APC but lost connection in the final approach. She decides to enable the autonomous engagement mode using her controls. Her interface informs her of this mode and requests to confirm engaging

the APC she is tracking. Her teammate confirms that the APCs are still at a distance from friendly forces. She confirms the engagement of the tracked APC and manually approaches it. The video feed cuts out several hundred meters before reaching the APC. She hasn't had a signal loss this far away yet. Will her drone be able to bridge this distance autonomously? Her teammate, is still receiving a feed from the fixed-wing drone and informs her that the APC has been immobilised.

There's little time to reflect on what happened. Her next FPV drone is already ready to fly.

1.3. Positive Impact of AI on the Socio-technical System:

1.3.1. Positive Impact at Task Level

Decreased Task Complexity – The final approach towards a target is a complex and stressful task requiring a skilled operator. Even then, latency, poor video quality, moving targets and other human and environmental factors can negatively impact the chance of a successful hit. AI can positively impact the chance this task is successful because it is locally processing sensor information and adjusting the drone's trajectory at speeds and precession unachievable by humans.

Increased target capability – A potential reduction in collateral damage and preservation of life may also be realised. The system can be instructed to target infrastructure only, and not when any human life is detected.

Increased agility – Able to reach previously considered "unreachable" targets (location, range, terrain etc.).

Fewer resources are required to perform tasks (e.g. time, energy, human resources).

1.3.2. Positive Impact on the individual

Increase situational awareness and reduced cognitive workload – Having multiple forms of real-time system feedback (e.g. video feed as well as trajectory view of where drone is in relation to target as well as potential feedback from other meshes of sensors) supports the operator in being able to anticipate drone behaviour further in advance, thus increasing situational awareness and reducing cognitive workload. Furthermore, the specific capability of having AI to engage a target purely based on video that the drone itself receives, without a connection to the operator.

Fewer boots on the ground - fewer human resources are required in the field.

Enhanced operational picture – this frees up the human operators from numerous communication tasks in trying to establish what is happening e.g. liaising with those above and below in chain of command, remote team members, members in the field etc. These communication tasks are known bottlenecks for decision-making in safety critical systems.

Humans are seen as "overseers" or "managers of the operation" and data from all system levels can be used for proactive risk management. Thus, smaller more incremental adjustments are made in relation to management of the operation/mission as opposed to larger, more dramatic adjustments being required much later in the mission when less time is available.

Proactive Workload Management (incremental Adjustments further in advance) Real-time first-hand feedback to operator feeds may afford greater time for decision-making for the operators.

Ability to train for complex threat and error management scenarios – Tailored training for individuals and teams without damage to persons, infrastructure, or the environment. The system itself will learn as part of this process and provide further decision-support based on the data obtained during training and de-briefing sessions.

1.3.3. Positive Impact at Team Level

Enhanced team situational awareness – as a result of a far richer, more timely and more detailed common operational picture.

Potential for greater collaboration co-ordination from both team and command perspectives (i.e. with improved situational awareness, fewer communication bottlenecks – potential for less pressurised decision-making).

Enhanced Inter and intra-team training – AI facilitates the ability to provide nuanced training for teams (both inter and intra team training) (i.e. based on data from previous sorties, missions, training sessions as well as incidents and accidents). This allows teams to train for complex scenarios, across a variety of skills bases and operational domains.

Technological supremacy – Having a modern military with technological advancements that surpass that of your adversaries generates confidence and trust in one's ability to achieve victory over them.

1.3.4. Positive Impact at Organizational Level

The system itself can be used for the enhanced testing of designs, functions, and capabilities across the design lifecycle. This can be advantageous for future system and technology acquisition as interaction between numerous levels (task, individual, team, organization etc.) may be tested from an integrated systems perspective. This enhanced testing and scenario-based training also benefits the organization from risk assessment, safety management, procedure writing and accident investigation perspectives.

Enhanced system training facilitates the inclusion of those indirect stakeholders who were involved in the design, testing, verification, and certification of the system as well as those responsible for the procurement of the technologies, recruitment, selection and training of the humans in the organization. It is important that they are included as part of a co-design approach and aware of the ethical guidelines around all stages within the design lifecycle. Being able to demonstrate real operational scenarios and highlight how different practices, features, capabilities etc. impacted the outcome and consequences of a mission provides a powerful tool for indirect stakeholders to grasp the criticality and relevance of their input to the wider system.

Organizations will be able to specify how they will V&V their AI-based systems, considering that this can and should include the human operator and other human stakeholders. This necessitates deciding on key human factor requirements that influence trustworthiness in AI.

Value-alignment – The AI-system could be pre-trained with a valued-aligned model that optimizes autonomous decisions between the mission goals and ethical goals. Also see **Positive Impact on the Individual**.

Technological supremacy – Having a modern military with technological advancements that surpass that of your adversaries generates confidence and trust in one's ability to achieve victory over them.

1.3.5. Impact at Regional/National Level

Standards, best practices, and guidelines can be applied to the full life cycle of AI systems (Concept, design, implementation, evaluation, Certification etc. AI can enable the system to learn potential mission recovery options based on data from previous missions, training exercises etc. This system learning could be aligned to IHL, doctrine and rules of engagement at regional / national and industry levels.

Enhanced training, testing, and learning facilitates a more efficient process for defining and maintaining Standards.

Standards are critical to the entire STS as they allow commonality of both national and international standards and facilitate common terms, definitions, taxonomies. All of these are necessary and indeed critical to ensure same standards of quality and safety around design, use and testing of technologies and systems.

1.3.6. Impact at Industry and Regulatory Levels

Standards, guidelines, and best-practices can and should be developed for the full life cycle of AI systems in regard to trustworthy AI-systems. Agreed standards, best practices and guidelines are also essential for ensuring measurements and metrics are applied in the same way (impact on validation, verification, risk assessment etc.) This is not only relevant to testing of technological systems, but also of the human resources and collective competence – this is critical with regards to human-machine teaming.

A more efficient process for defining and maintaining standards, best practice and guidelines facilitates greater operational clarity between operational domains, national forces, and international communities (e.g. EDA, NATO).

1.4. Negative Impact of AI on the Socio-technical System:

1.4.1. Impact at Task Level

Lack of control – there is a lack of control over the recovery of the scenario which may lead to task/mission failure and potentially catastrophic consequences (i.e. friendly fire).

Edge-cases – AI might have difficulty dealing with edge-cases where humans may not (e.g., dust clouds, low-lighting conditions, or adversarial AI attacks).

Abort procedure – Without a human controlling the final trajectory it is impossible to abort the final engagement.

1.4.2 Impact on Individual

Lack of intent information to and from human-machine team members may compound the difficulty in being able to anticipate likely behaviour of the system resulting in increased workload (e.g. cognitive workload, increased communication tasks required in order to maintain situational awareness etc.)

Increased Cognitive workload-When the connection with the drone is lost, the operator has lost the feedback mechanism from the drone and thus has no means of knowing if the drone

is proceeding to target (or not). They may also not know the reason that connection is lost (e.g. technical malfunction, out of range, Cyber jamming etc.) or whether the AI capability itself is malfunctioning. The operator is likely to have significantly reduced situational awareness and increased cognitive workload in trying to ascertain a full operational picture, rehearse scenario recovery possibilities (e.g. abort mission, continue to target, hover until connection has been reestablished, self-destruct).

Increased Stress – During time-critical, safety critical missions where potential lives are at risk, lack of situational awareness may impact operator stress which could negatively impact their ability to make decisions.

Meaningful Human Control – Meaningful Human Control (MHC) generally refers to the need for humans to maintain control, directly or indirectly, over decisions made by (autonomous) systems. Because there is no direct way to control the drone once its connection to the operator has been lost a human-in-the-loop solution seems impossible. As such, clarity around rules of engagement, doctrine and prioritisation of recovery actions need to be implemented beforehand, to facilitate a human-before-the-loop approach.

Lack of Transparency – It may not be clear to the human which level of autonomy the system is currently in, nor what the expected human input is during the transition between those levels. It is important that human operators know when they are able to take back control from the automation. The authority that the human has over this, has an impact on their trust in the system. This may also have a detrimental impact on the time required for the human to make decisions and for humans being out of the loop.

Calibrated trust – Calibrated trust refers to a degree of trust by the user in the system that corresponds to the actual trustworthiness of the system. Inappropriate trust calibration can result in insufficient monitoring or misunderstanding of a system's behaviour. The AI algorithm operates autonomously, making decisions based on real-time data. However, the inner workings of these algorithms are often opaque to its operator. This lack of transparency creates a barrier for trust calibration leading to suboptimal decision making. Operators need to trust that the AI's decisions align with its actual trustworthiness. An operator therefore needs to be able to correctly trust the system to act not only execute its task effectively, but also according to the moral and ethical standards set by their organization.

Explainability – For the human to understand the decisions and behaviour the autonomous component of the drone makes, the AI needs to be able to explain itself. For our use-case, little information was available to the operator to base her decisions on. Instead, it would be beneficial for a human operator to have an estimation of the drone's ability to, for example, successfully track and hit a target given the context. This information would need to be explained in a way that fits the operators needs, to avoid cognitive overload or distraction.

Training – Operators will need to feel both responsible, capable, and supported in their role. They need to be trained to know the capabilities and limitations of the (AI)-systems they work with to understand when, in this case, it is appropriate to enable the autonomous capabilities and when not (e.g., possible civilian or friendlies in the operational area, known environmental conditions that negatively impact the AI's capabilities). Additionally, they need to understand what the possible consequences of enabling this mode are as well. For example, what happens when the AI loses track of its intended target? Does the drone attempt to reacquire the same target? Or will it attempt to find another target? Furthermore, The AI-driven component is likely to evolve over time and behave differently due to newer version of the AI being developed. This adaptive capability adds complexity, as the AI system's behaviour may

change over time. Operators must adapt their expectations via training as the system learns and improves.

1.4.3 Impact at Team Level

Increased unease for other team/unit members – due to friendly forces with an armed drone with full autonomous capabilities is flying over them or in close proximity.

Increased workload for other remote support (other units, command levels) – in order to maintain situational awareness, the drone operators may seek clarity in operational picture from other team members and teams. This may add additional pressure to those team members and distract them from other tasks at hand.

The impact at individual levels (see above) are all relevant and become more pronounced and compounded due to the complex and distributed nature of the team(s). Remote operators and those in the chain of command are attempting to understand what is happening – a common operational picture and this requires increased communications, more time make decisions and take appropriate action. The tasks of leadership and command become more difficult due to the reduced situational awareness and loss of control. This can have a negative impact on all aspects of team co-ordination demands (Burke,2006) such as communication, situational awareness, decision-making, leadership, adaptability and assertiveness within the team itself.

1.4.4 Impact at Organizational Level

Increased uncertainty and risk – The co-ordination of tasks, operations and missions for a single operational domain is a highly complex. To manage this across more than one operational requires a novel approach to risk management – especially where greater uncertainty around human-machine teaming is relevant. The management of risk, safety and operational performance should be approached as “socio-technical systems risk” to reflect risk from an integrated perspective in highly complex, time and safety critical operations such as defence operations.

Individual human operators may be considered the 2nd victim if devastating consequences arise from their actions in the field. Post-traumatic stress and a further sense of helplessness arising from an incident may be experienced due to the perceived lack of control that the human may have had. In this case-study, the APC was immobilised and prevented adversaries from further advance. However, if there was a case of friendly fire due to “unforeseen circumstances” when “full automation” was engaged – the uncertainty and lack of intent and feedback from the system may compound the sense of profound responsibility that the operator may carry for the rest of their lives. The support offered to operators in these circumstances should be focussed on a just culture and a collective responsibility of the defence force and wider STS rather than on the individual operator. It is critical that operators are trained to understand likely intent and behaviour for recovery scenarios (for connection loss) so that they are able to rationalise what may or what did happen. The ethical values and moral obligation of the organisation around the 2nd victim need to be made clear at regional level (i.e. rules of engagement, doctrine etc.) and supported at both industry and regulatory levels.

1.4.5 Impact at Regional/National and Industry Levels

The ethical values and moral obligation of the organisation around the 2nd victim need to be made clear at regional level (i.e. rules of engagement, doctrine etc.) and supported at both industry and regulatory levels.

Humans as a main form redundancy in the system- Consensus is required on how to handle system inconsistencies and fallibilities in relation to human redundancy. This should include agreement on what is fair to place on the shoulders of the operators, the programmers, those that construct the technologies, procure them etc. This consensus requires adequate backing under regulation and law so that the humans in the system are fully supported and done so as part of a just culture.

There are numerous ethical concerns re how defence forces choose to engage and what to do if adversaries engage in unethical behaviour/ rules of engagement using AI systems. It is the responsibility of the EDA, and wider international communities to reflect and obtain consensus on how this will be approached.

1.4.6 Impact at Regulatory Level

Regulation and legal framework for ensuring practice and defence operations are ethical need to be aligned with novel practices, tools, and technologies. A full understand of how these function on an integrated level is still unclear in practice. As highlighted in Chapter 7 on Ethics, the importance of inviting member states and their forces to sign up to these ethical codes of conduct is ever more pressing given the current political climate. It is the responsibility of the international community to ensure that national and international humanitarian law is upheld. This requires member states and forces to lead by example and best practice.

Regulation has yet to make it ultimately clear who has authority and who is responsible across levels of autonomy (to include transitions between levels)– this has yet to be fully understood and modelled adequately to understand the complexity of the human and machine interactions. Where this is yet to be understood, the human should be the overseer and in control – however, they should know that they are supported by their team, their unit, their organisation, and their national defence forces. This will enable them to be as confident as possible in the decisions they make under time and safety critical conditions with far-reaching and potentially catastrophic consequences. It is the responsibility of the international community to make this a priority for all current and future stakeholders.

This case study demonstrates the complex nature of the decision-making process for stakeholders and why it is necessary that a variety of stakeholders are included throughout the design lifecycle. Please refer to the recommendations section (Chapter 9 of TAID White Paper) for final perspectives on both Human Factors (Chapter 6 of TAID White Paper) and Ethics (Chapter 7 of TAID White Paper) relevant to trustworthy AI for future design, practice and military defence operations.

All references for Chapters and abbreviations are exposed in TAID White Paper.