AI in Future C2 – Who's in Command When AI Takes Control?

Bjørn Tallak Bakken
Inland Norway University of Applied Sciences (INN)
bjorn.bakken@inn.no

Inger Lund-Kordahl
Inland Norway University of Applied Sciences (INN)
inger.lundkordahl@inn.no

Erik Bjurström
Mälardalen University (MDU)
erik.bjurstrom@mdu.se

Abstract

Artificial Intelligence (AI) will become an increasingly dominant element of future Command and Control (C2) systems and organizations. In this paper we discuss how to make effective use of this technology in time-critical decision-making situations, as we classify the relationships between human decision maker; the AI system; and the task or situation at hand, along two dimensions: Automation and Autonomy. The former concerns the employment of AI to standardize and speed up the decision-making process (efficiency), without necessarily sacrificing accuracy or precision. The latter concerns the influence an AI system will have on the determination of the utility of the outcome of a decision-making process (effectiveness), relative to a human decision maker within the same process. Our research question is whether there is a trade-off between automation and autonomy when using AI to support decision-making in time-critical situations. In other words, can an AI system be trusted to make decisions that are both efficient and effective in time-critical situations, such as encountered by C2 systems?

1 INTRODUCTION

Command and Control (C2) systems and organizations are undergoing a significant transformation with the advent of Artificial Intelligence (AI), as AI will become an increasingly dominant element of future C2 systems. This paper aims to provide an overview of the expanding role of AI in C2 systems and organizations, emphasizing the criticality of effective decision-making in time-critical situations.

Effective decision-making lies at the core of C2 systems, where rapid and accurate responses to complex and dynamic situations are essential [25]. Traditionally, humans have reigned the field of strategic, operational, and tactical decision-making, where the speed and accuracy of decisions relied fully on these humans' expertise and experience. With the rapid advancement of AI technologies, it has been possible to augment human decision-making with intelligent systems that can process vast amounts of data, identify patterns, and provide recommendations or even autonomously execute decisions [8] [9] [18].

The challenges inherent in crisis management are established and well-known. A decision-maker, at any level, is faced with great uncertainty, high stakes, and severe time-pressure [4] [5]. A crisis may be framed as acute, meaning that high efforts concentrated in time and space are needed to handle it [6]. Examples are firefighting, police operations, medical emergencies, search-and-rescue operations. Other crises are long-lasting and may vary in intensity during its timespan. Examples are pandemics, climate crisis, food crisis, energy crisis, financial crisis, and regional military conflicts. During a long-lasting crisis, acute "episodes" of intensified crisis management may be required [24]. For example, a long-lasting military conflict may involve shorter periods of attacks and counterattacks. In between these acute episodes, it may appear that the crisis is no longer urgent, and even non-existent [11] This switching between low and high intensity in a long-lasting crisis, places an even heavier burden on decision makers [6] [10] [43] [49] [51].

We have earlier argued that the main cognitive challenge in dynamic, complex situations, such as crisis management, is in the perception and understanding of delayed, non-linear feedback [10] [46] [54]. This view is grounded in a systems perspective on crises and crisis management [2] [3] [12] [13], and a failure to recognize these aspects of a crisis means that we will inevitably be unable to succeed when attempting to manage complex crises. This leaves us with the question: will AI be able to alleviate some of these cognitive challenges of decision makers in a crisis? Since military C2 constitutes some of the most complex contexts within crisis management, it is natural that this question will be of great interest to planners and executors of military operations. Recently, it has been highlighted that future C2 needs to incorporate a capacity to handle both civilian and military crises, exemplified by the concept "total defence" [41]. This requires radical and swift adaptive organizing of both civilian and military capacities and resources, in response to unexpected events [31], which is an area where AI could provide a substantial contribution to crisis

management.

The intensified usage of AI with C2 systems brings numerous opportunities and challenges. On one hand, AI has the potential to enhance decision-making capabilities by augmenting human cognition, enabling faster and more accurate analysis, and automating routine tasks. On the other hand, the increasing reliance on AI raises questions about the appropriate level of automation and the extent to which AI systems can be trusted to make critical decisions in time-critical scenarios [26] [27]. While the former, mere automation, is a question of relieving the human decision-maker of the burden of tedious, repetitive, mundane, and sometimes computing-intensive processing, the latter concerns how far we can go in trusting AI to make autonomous decisions on behalf of humans. How far can we go in transferring the authority within decision-making, from humans to machines? What are the benefits and risks involved? [30]

Our research question guides the study: Is there a trade-off between automation and autonomy when using AI to support decision-making in time-critical situations, such as those encountered by C2 systems? In other words, can AI systems be trusted to make decisions that are both efficient and effective, while considering the unique demands of time-critical operations?

By examining the increasing role of AI in C2 systems and emphasizing the significance of effective decision-making in time-critical situations [34] [42], we lay the foundation for exploring the trade-offs and challenges associated with automation and autonomy. Understanding the implications of these dimensions is crucial for designing AI systems that can enhance decision-making processes without compromising mission-critical objectives. In the words of Dr. Seun Kolade: "Yes, AI is the "frenemy" that wields a double-edged sword. With one edge it offers so much value; with the other it portends known risks that should be controlled and kept at bay[1]."

In the following sections, we will delve into the concepts of automation and autonomy in decision-making, analyze the potential trade-offs between these dimensions, and follow on to discuss the importance of trust in AI systems when they – in time – assume decision-making authority.

2 A CONCEPTUAL FRAMEWORK

Let's start out with a thought experiment[2]: The setting is a family of parents and a small child. The parents are trying to teach the child table manners, and after a few failed attempts they have come to the point where they want to incentivize the child's behavior when it comes to minimize (presumed accidental) spilling on the kitchen table during meals. The parents explain to the child that they will deduct one piece of candy (or for an older child, it could be a monetary amount) from the child's weekly allowance per stain the child has spilt on the table during dinner. This scheme works well for some time, until one day the child accidentally spills several drops of soup on the tablecloth. The number of stains exceed the remainder of the weekly allowance, and the parents are in doubt what to do. The parents take a moment to discuss away from the table, and when they return, they see just one large stain covering most of the tablecloth, effectively concealing the smaller stains. The child shouts out triumphantly: "Look, there is just one stain on the table, I still have candy left!"

What this small anecdote illustrates, is the difficulty of controlling or regulating human behavior in real life situations with one or just a few simple algorithmic rule(s). We can apply this problem analogously to the field of AI, by considering the challenges of developing autonomous (self-driving) vehicles [26] [28]. Despite billions of miles driven to train the AI system behind the wheel, spectacular accidents still happen. When the AI encounters an obstacle in the road, it must decide within milliseconds how to respond. Should it apply the brake and/or defect, possibly injuring the driver and wrecking the car, in the attempt to save a fellow pedestrian from injury that could be fatal? No wonder that manufacturers of autonomous vehicles have put enormous efforts into training the AI to discriminate effectively between "brake-worthy" obstacles, such as other vehicles, pedestrians and (large) animals, and objects that are harmless to the car, for example when the object in the road is just a plastic bag blowing in the wind [50].

Now let's look at another example, where the same family, still obsessed by cleanliness, have purchased a domestic cleaning robot. The robot is programmable and can be instructed to attend to certain household chores. The family decides that they want the robot to remove stains and spots from the indoor floors (living room, bedrooms etc.), and the robot manufacturer has conveniently devised a reward system to ensure efficiency at the task: the robot receives "points" for each stain successfully removed. After being programmed, the robot gets going with the task. This goes on for several weeks;

[1] https://charteredabs.org/has-chatgpt-signalled-the-end-of-assessment-as-we-know-it/ (27.3.2023)

[2] Thanks to philosophy professor Øyind Kvalnes for sharing this anecdotal example of "loophole" behavior in ethics [36].

the floors are spotless when the family returns home from jobs and school every day, and the robot accumulates "points" for success. One day the family discover a large, sudden increase in the points that the robot accumulates daily. The point balance grows through the roof, but the family remains puzzled. They get the idea to install a video camera to monitor the robot while they are away for the day. They become startled when the video reveals that the robot leaves trails of tiny drops of undiluted detergent on the floor, only to immediately remove each spot and collect one point per instance.

Now, the family proceeds to reprogram the robot, and luckily the manufacturer has upgraded the software accordingly. Instead of rewarding the robot for each spot removed, they now instruct the robot to always keep the indoor floors perfectly clean. Being slightly disillusioned over the AI system being "intelligent", the family is less surprised when they one day come home and discover that their access codes to the digital door locks have been invalidated. They are effectively shut out of their own home! When calling the manufacturer of the central door lock system (which by the way is the same as the robot manufacturer), they are explained that the robot has been upgraded to being able to open the main door and fill up detergent from the outside detergent filling station. Accordingly, the robot has been given the capability to lock and unlock the doors using the centralized control system. An unintended feature of the software upgrade has apparently led the robot to deduce that the most effective way to always keep the house spotless, is to deny the family access to the house altogether[3].

Both the child and the robot examples point to the classical "alignment problem" associated with the pursuance of general AI. This is the phenomenon where the goals of the AI system diverge from the goals of the human decision maker that the AI is supposed to serve and stems from the difficulty of stating algorithmically what a human (or society) wants to achieve, resulting in undesirable behavior from the AI system [40]. Until recently, the use of AI in various areas of society has largely been limited to automating tedious, repetitive, and computing-intensive tasks, where the steps of a task are static, clear and sequential (i.e., linear interactions in the Perrow terminology [45]) and the goals or outcomes to be attained are well-defined and undisputed. However, what we see now – and as a trend for the future – is a transition from automation to autonomy of AI systems. While automation is about transferring cognitive *effort* from human to AI, autonomy is more about transferring cognitive *control* from human to AI [39].

Another well-known phenomenon in AI usage is termed "automation bias". This is when a human decision maker is inclined to over- (or under-)use an AI system, caused not only by a belief that the AI is more (or less) trustworthy than what can be objectively determined, but also because (in case of over-reliance) it appears to be more convenient for a human decision maker to "delegate" cognitive tasks to the AI rather than making an effort themselves [16] [17] [21] [39] [50].

In this paper, we also address whether these two phenomena – the Alignment Problem and the Automation Bias – are more than just minor obstacles to be circumvented on the road to more effective use of AI in for example command and control (C2) systems, or are they in fact fundamental barriers that could stall further development and use of AI in the society altogether? If this latter, rather pessimistic, view becomes the prevalent one, we will need to question the very premises that AI systems are built on: that the human mind – its perception, reasoning and storage abilities – can be implemented in an artifact that basically consist of electric circuits switching on and off according to some mathematical-logical programming. Although it seems that such programs can become immensely complex and interconnected, it is still just the programmatic manipulation of zeroes and ones – bits and bytes [50]. As such, AI systems can never become truly innovative, since every action and every decision an AI system will ever make, will have to be pre-programmed in some way.

The aim of *efficiency* using AI – that is, why we automate for task performance – is to achieve consistency, precision, and accuracy, at speeds and volumes that humans cannot match, and at a lesser cost. The aim of *effectiveness* through autonomy is the belief (or rather hope) that AI may produce even better decisions without the supervision or intervention of a human, to the end that AI will by itself define and pursue "greater goals" and higher-valued end-states when decisions are left to the machines – in part or altogether.

We will argue that automation and autonomy represent two independent, but related, concepts that characterize usage of AI to support decision-making (Table 1). Conceptually the two dimensions may be thought of as continuous, but in the framework model they are both dichotomized as high and low. Since the dimensions are

[3] A variant of this rather classic example, probably inspired by the movie "2001: A Space Odyssey" is given in [50].

independent, they may take on any of the 2 x 2 combinations: low-low, high-high, low-high, and high-low.

We propose the following labels for these combinations:

- Low automation – low autonomy: Consultancy
- High automation – low autonomy: Adaptation
- Low automation – high autonomy: Integration
- High automation – high autonomy: Supremacy

Table 1: A Conceptual Framework of Automation and Autonomy of AI Usage in Decision Making

		Autonomy / authority (Effectiveness)	
		Low	High
Automation (Efficiency)	Low	CONSULTANCY	INTEGRATION
	High	ADAPTATION	SUPREMACY

2.1 The Framework Applied to Military C2

2.1.1 Automation

Automation implies that cognitive effort is transferred from human to machine, and is the key to achieve system efficiency, i.e., maximizing a systems output (utility) relative to effort (cost), subject to a given goal or desired end state. Within C2, automation is used to support human decision makers in relatively linear, stable contexts, to achieve speed, precision, and accuracy. These are contexts where a machine significantly outperforms a human decision maker. Examples are guiding precision missiles; fusing data from various sources to build a digital situation picture; and encryption and decryption of secure digital communication [52].

2.1.2 Autonomy

Autonomy implies that cognitive control (authority) is transferred from human to machine. The underlying assumption is that removing the human from the decision process in part or altogether, could open for the AI system not only to achieve maximum efficiency subject to a stated goal or end-state, but also for the AI to reformulate a goal or end-state if that provides higher effectiveness. Of course, for this to be meaningful it requires that the AI is capable of reasoning along the lines of a "greater good" formulation [37]. No matter how generic or specific a goal or desired end state is described at a lower level of operation (for example tactical), it will always be possible to define a goal or end state at a higher (for example operational or strategic), level [50]. This reformulation can continue until on reaches the highest conceivable "greatest good", which could be thought of as a universal, political, or cultural norm. A possible application of AI in this regard is in planning and executing operations, while continuously assessing risks and trade-offs with regard to selection and prioritizing between end-states – which is a thing of the future.

The level of autonomy can range from limited autonomy, where AI systems provide recommendations that human decision-makers can accept or override, to full autonomy, where AI systems have the authority to make decisions and take actions without human supervision and intervention.

2.1.3 The Configurations of Automation and Autonomy

Taken together, the motivation for automating a process with AI is usually to save human effort (quicker and more accurate decisions). With autonomy, there is also the elusive idea that an AI may make qualitatively better decisions altogether, attaining greater effectiveness, pursuing "better goals" for all of humanity.

2.1.3.1 CONSULTANCY

Within the low automation – low autonomy configuration, we conceive of «AI-as-a-tool». This facilitates and relieves the human of repetitive, tedious everyday tasks. E.g., navigation system in cars; medical diagnosis; internet searches; booking systems; in C2: risk and capabilities assessment; Common Operational Picture (COP) generation; logistics planning for operations.

2.1.3.2 ADAPTATION

Within the high automation – low autonomy configuration, we look at adapting AI to fit human arenas for (time-)critical decision-making in high-risk systems. E.g., safety & security systems, both civilian and military. A civilian example is the "driving assistance" functions in cars that help maintain speed limits, lane tracking, and collision avoidance. Within C2: target detection, identification, and prioritization; encryption and decryption of sensitive, time-critical communication. AI is used as support, to prepare the basis for a decision, while a human makes the (final) decision. Commonly, the amount and complexity of data collected and processed vastly exceeds the human's capability for processing.

2.1.3.3 INTEGRATION

Within the low automation – high autonomy configuration, we look at systems where human and AI interact seamlessly & synergistically for time-critical operational decision-making. In these applications, e.g., advanced weapon systems; unmanned remote-control

vehicles; where we find that the distribution of effort and authority between human and AI is optimized in a task, so that the human makes initial higher-level judgments, while AI takes on the parts where precision, accuracy and speed is vital (automation), including selection and prioritization of targets to engage (autonomy). A civilian example would be autonomous (i.e., self-driving) cars, a concept that is still undergoing experimental testing.

Note that "low" automation in this configuration is not the same literal level of automation as "low" in the Consulting configuration. The level is relative within the dimension.

2.1.3.4 SUPREMACY

Moving on from Integration, where it could appear that a human in the loop could be an unnecessary obstacle towards optimal and time-critical decision making, we conceptualize that AI replaces the human actor entirely, taking full strategic and operational control. Within Supremacy, the AI engages in "ethical" judgments, and determines goals and values to be achieved, without human supervision or intervention. As such, AI mimics human behavior in this respect. Ideally, the AI should strive towards the universal "greater good". This application of AI lies in the future. As we will comment later, we could also call this dimension "Totalitarian".

3 DISCUSSION

3.1 AUTOMATION IN DECISION-MAKING

The concept of automation in decision-making entails the application of AI technologies to streamline and expedite the decision-making process. By automating certain aspects of decision-making, AI systems can enhance efficiency, enabling faster and more consistent responses to time-critical situations.

A key aspect of automation is the standardization of decision-making procedures. AI systems can codify expert knowledge and best practices into algorithms, enabling consistent and standardized decision-making across different contexts. This standardization reduces variability and improves the reliability of decision outcomes. Additionally, automation can eliminate human errors arising from fatigue, cognitive biases, or information overload, also leading to more accurate and reliable decisions [16] [17].

Various approaches and techniques can be employed to automate decision-making processes. These range from rule-based systems that follow predefined decision rules, to machine learning algorithms that learn patterns from historical data and make predictions or recommendations based on those patterns. Rule-based systems are particularly useful in well-structured decision-making tasks where explicit rules and criteria can be defined. Machine learning approaches, on the other hand, excel in complex and data-intensive decision domains, where they can uncover hidden patterns and make data-driven predictions [32] [50].

The benefits of automating decision-making can be summed up as follows: Firstly, automation can significantly reduce the time required to make decisions, which is crucial in time-critical situations where rapid responses are necessary. By processing and analyzing vast amounts of data in real-time, AI systems can provide decision support that accelerates the decision-making process. Moreover, automation can free up human decision-makers to focus on higher-level strategic thinking and complex problem-solving, while routine and repetitive tasks are delegated to AI systems [29].

However, the introduction of automation in decision-making also comes with some challenges. One key challenge is the potential lack of transparency and interpretability of AI systems. As AI algorithms become more complex, understanding their decision-making rationale becomes increasingly difficult. This lack of transparency can hinder human decision-makers' ability to trust and validate the outputs of AI systems, particularly in critical scenarios where explainability is crucial [53].

Another challenge is the potential overreliance on automation (Automation Bias), which may lead to complacency or deskilling among human decision-makers [16]. The reliance on AI systems should be accompanied by appropriate training and skill development to ensure that human operators can effectively interact with and supervise the automated decision processes. Moreover, the inclusion of AI systems into existing decision-making workflows and organizational structures requires careful consideration to ensure compatibility and minimize disruption [47] [48].

While automation holds great potential for enhancing decision-making processes within C2 systems, by standardizing decision-making procedures and leveraging AI technologies, leading to improved efficiency, faster and more consistent responses in time-critical situations, careful attention must be paid to the challenges associated with transparency, interpretability, and the potential impact on human decision-makers. Striking the right balance between automation and human oversight is essential to harness the benefits of automation while maintaining trust, accountability, and effectiveness in

decision-making [16] [17] [20]. This tradeoff will be discussed in a later section.

3.2 Autonomy in Decision-Making

The concept of autonomy in decision-making refers to the degree to which AI systems have the authority to make decisions independently, without direct human supervision or intervention [16] [39].

When AI systems are granted autonomy, they are empowered to analyze information, generate options, and execute decisions without direct human intervention. The level of autonomy can range from limited autonomy, where AI systems provide recommendations that human decision-makers can accept or override, to full autonomy, where AI systems have the authority to make decisions and take actions without human intervention [28].

As with automation, a key advantage of autonomous AI decision-making is the potential for faster and more efficient responses. AI systems can process vast amounts of data in real-time, continuously monitor the situation, and rapidly adapt their decisions based on changing circumstances. In time-critical scenarios, this capability can lead to quicker response times, enabling better coordination and synchronization of operations compared to a "human" decision making system, e.g., a C2 [30] [47] [48] [52].

So why should one not "go all the way" and remove the human entirely from the decision process? Granting autonomy to AI systems raises several ethical and practical questions. One of the primary concerns is the question of responsibility and accountability. When an AI system operates autonomously and makes decisions that have real-world consequences, it becomes crucial to attribute responsibility for those decisions. Determining liability and addressing accountability issues become complex when the decision-making authority is delegated to AI systems. Ethical frameworks and legal frameworks need to be developed to address these challenges and ensure that AI systems can be held accountable for their actions [17] [30] [50].

Transparency and interpretability are also important considerations when discussing autonomy in decision-making. AI systems that operate autonomously should be able to provide explanations for their decisions, allowing human decision-makers and stakeholders to understand the underlying reasoning. This transparency is vital for building trust, verifying the reliability and fairness of the AI systems, and ensuring that decisions align with legal and ethical standards [33].

Moreover, the level of autonomy granted to AI systems should be aligned with the complexity of the decision-making task and the domain expertise required. In certain time-critical scenarios, human decision-makers may possess critical domain knowledge, intuition, or contextual understanding that cannot be easily replicated by AI systems. In such cases, a collaborative decision-making approach, where AI systems provide recommendations and human decision-makers retain the final authority, may be more appropriate [8] [47].

Lastly, the potential risks associated with autonomous AI decision-making should be carefully assessed. AI systems are not immune to errors or biases, and the consequences of erroneous decisions in time-critical scenarios can be severe. Robust validation, testing, and risk assessment mechanisms should be implemented to minimize the likelihood of unintended consequences and ensure the safety and reliability of autonomous AI decision-making [35] [38] [45].

Summing up: autonomy in decision-making within time-critical scenarios presents both opportunities and challenges. Autonomous AI systems can offer faster and more efficient responses, but ethical and practical considerations such as responsibility, transparency, domain expertise, and risk management must be addressed. Striking a balance between human oversight and AI autonomy is crucial to leverage the benefits of AI systems while upholding ethical principles and ensuring the effectiveness of decision-making processes in time-critical environments.

3.3 Is There A Trade-off between Automation and Autonomy?

A trade-off between automation and autonomy in decision-making refers to the potential conflict or tension that arises when seeking to achieve both efficiency and effectiveness in time-critical decision-making processes using AI systems. In this section, we delve into this potential trade-off and explore how the interplay between automation and autonomy impacts decision-making outcomes.

Automation aims to streamline and expedite decision-making processes by leveraging AI technologies. It enhances efficiency by reducing the time required for decision-making, standardizing procedures, and automating routine tasks. However, as the level of automation increases, the potential for reduced human oversight and control also rises. This can limit the ability of human decision-makers to intervene or exercise judgment in critical situations. Thus, a high degree of automation may compromise the flexibility and adaptability required in time-critical scenarios.

Autonomy, on the other hand, relates to the authority and decision-making power delegated to AI systems. Granting autonomy to AI systems allows for faster responses and adaptive decision-making in time-critical situations. However, as the level of autonomy increases, concerns regarding accountability, transparency, and the potential for unforeseen consequences also arise. Balancing autonomy with human oversight becomes crucial to ensure responsible and ethical decision-making [14] [15].

Advantages of achieving a balance between automation and autonomy include improved response times, reduced human errors, and enhanced coordination in time-critical situations. AI systems can process and analyze large volumes of data quickly, identify patterns, and provide timely recommendations. Human decision-makers can then leverage these recommendations to make informed decisions, while retaining the final authority to ensure domain expertise, ethical considerations, and contextual understanding are accounted for.

However, limitations and challenges also emerge when striking this balance. The complexity of decision-making tasks, the context in which decisions are made, and the need for explainability and transparency can pose challenges to achieving an optimal level of automation and autonomy. Decision-making scenarios that require nuanced judgment, intuition, or ethical considerations may require a higher level of human involvement and oversight. Additionally, unforeseen or novel situations may push the limits of AI systems, highlighting the need for human adaptability and decision-making capacity [5] [6] [26] [34] [52].

To address these challenges, it is crucial to carefully design and configure AI systems within C2 environments. This involves considering the specific requirements of time-critical decision-making, developing appropriate decision models, integrating AI systems seamlessly into existing workflows, and providing necessary training and support for human decision-makers to effectively interact with automated and autonomous systems.

In conclusion, achieving a balance between automation and autonomy in time-critical decision-making is a delicate task. It requires careful consideration of the advantages and limitations of each dimension, along with the specific demands of the decision-making context. Striking this balance allows for harnessing the benefits of AI in terms of efficiency and effectiveness while ensuring responsible, accountable, and contextually appropriate decision-making. By analyzing case studies and empirical evidence, we can gain valuable insights to inform the design and implementation of AI systems within C2 environments and enable effective decision-making in time-critical scenarios.

3.4 Trust in AI Decision-Making

Whether AI can and will be used to achieve desirable goals now and in the future – as seen from the perspective of a human decision maker – depends crucially on the element of trust. At one end, the Alignment Problem tells us that AI can pursue undesirable (and unexpected) goals, which may jeopardize trust altogether. At the other end, the Automation Bias phenomenon implies that an AI – given an "objective" degree of alignment that translates into trustworthiness – may still be under- or overused, depending on the human decision makers' belief or perception of this trustworthiness. In the following, we will discuss some aspects of trust and trustworthiness pertaining to the use of AI in C2 systems.

In time-critical scenarios encountered by C2 systems, trust becomes even more critical than with "civilian" systems [44]. The stakes are high, and decisions must be made rapidly. Human decision-makers need to have confidence that AI systems will perform reliably and responsibly in these time-sensitive situations [17] [19]. Trust enables effective collaboration between humans and AI systems, allowing for seamless integration and coordination.

The European Commission has established a set of ethical guidelines for trustworthy AI, with seven key requirements [27]:

- human agency and oversight
- technical robustness and safety
- privacy and data governance
- transparency
- diversity, non-discrimination and fairness
- environmental and societal well-being
- accountability

We will now investigate some of the key factors for establishing and maintain trust in AI: explainability, transparency and accountability.

3.4.1 Explainability

Explainability is a critical factor in fostering trust in AI systems [1] [22] [23] [53]. Decision-making processes should be transparent and understandable to human decision-makers and stakeholders. When AI systems operate autonomously or provide recommendations, they should be able to explain the rationale behind their decisions in a clear and interpretable manner. The ability to provide explanations helps human decision-makers gain insight into the decision-making process and evaluate

the reliability, fairness, and appropriateness of the AI system's outputs. Explainability also facilitates trust-building by allowing humans to validate the system's decision-making against their own domain knowledge and expertise.

3.4.2 Transparency

Transparent AI systems enhance trust by providing insights into how decisions are made, allowing human decision-makers to verify that the system operates in accordance with ethical guidelines and legal requirements [27].

3.4.3 Accountability

Accountability mechanisms ensure that AI systems can be held answerable for any unintended consequences or errors in their decision-making. Establishing accountability not only enhances trust but also encourages responsible behavior and adherence to ethical standards by AI systems [27].

3.4.4 Human Factors

Human decision-makers need to have confidence in the capabilities and limitations of AI systems. Trust can be fostered by providing training and familiarization with AI systems, ensuring that users understand the system's capabilities, limitations, and the decision-making context in which they operate. Furthermore, involving human decision-makers in the development and validation processes of AI systems can help build trust by incorporating their expertise, addressing their concerns, and aligning the system's behavior with their expectations [17].

4 Conclusion and Further Work

We, as a society, have enjoyed tremendous increases in productivity and profitability from automating tedious, repetitive and foremost standardized procedures in manufacturing, transportation and service industries, as well as in areas such as agriculture, medicine and energy production (for example oil platforms and nuclear power plants). Along with the civilian industries being automated, so follows defense organizations as well as the various emergency services. Now, when the full potential for automation appears to have been saturated [50], this means that the race for efficiency and effectiveness now undergoes a fundamental transition, from automation of relatively standardized linear procedures and processes by AI, to allowing AI to make autonomous decisions in complex, dynamic situations – even define and pursue goals by itself – without human supervision or intervention.

We have argued that future C2 systems will need to employ autonomous AI for increased effectiveness, in areas where AI so far has been used only for automation (to increase efficiency). We have discussed the potential impact of the Alignment Problem and Automation Bias, as well as ethics problems with giving autonomy and authority to AI. Our conclusion is that humans should "stay in the loop" to avoid the risks mentioned. Unsupervised learning in open systems should be restricted, and explainable AI – with emphasis on transparency and accountability – should be enforced in critical systems. The more complex and dynamic a system, the more wary and prudent one should be of delegating authority to AI.

To explore empirically the framework presented and discussed in this paper, we propose to conduct experiments with human decision makers, in a context of military C2 with a backdrop of relevant crisis and conflict scenarios. We suggest manipulating both automation and autonomy – the independent variables – of an AI tool to support decision making, in a 2 x 2 experimental design. As dependent variables we measure trust and reliance on AI, as well as perceptions of transparency and accountability. We will also operationalize and measure perceived alignment. Performance and goal attainment will also be measured as dependent (outcome) variables. As potential moderators we suggest using cognitive style, personality factors, expertise and experience of the human decision-makers [4] [5] [21] [25].

5 Closing remarks

We started out by asking whether there is a trade-off between automation and autonomy in AI usage for C2. While efficiency from automation has proven to be cost-effective from the start, it is only meaningful in stable predictable environments with a static knowledge base. Autonomy, on the other hand, presents us instead with a *paradox*. Where it could appear that AI will be most useful to us, in dynamic complex environments – difficult to predict – this is also when one should exercise most caution in using AI. Therefore, the rush to implement autonomy in uncharted territory should be slowed down, pending rigorous research into its effects and side-effects [50] [54].

An example is the climate crisis. If autonomy was a viable path with using AI, we would already be on it. But we are clearly not. We know what kind of processes that should be automated to solve the crisis (reduction of fossil fuel usage, reduced consumption of non-renewable resources, and even carbon caption and storage technology), but we lack both the knowledge and ability

to *implement* such processes effectively. Perhaps the final frontier for using AI effectively in the pursuance of this "greater good" requires that we use AI to solve the "collective action problem" first [50]? After all, putting AI to effective use within C2 will be a (much) simpler problem than solving the climate crisis with AI.

REFERENCES[i]

[1] Adadi, A. & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052.

[2] Bakken, B. T. & Gilljam, J. M. (2003). Dynamic Intuition in Military Command and Control: Why it is important, and how it should be developed. Cognition, Technology and Work, 5, 197-205.

[3] Bakken, B. T. & Gilljam, J. M. (2003). Training to Improve Decision Making: System Dynamics Applied to Higher-level Military Operations. Journal of Battlefield Technology, 6(1), 33-42.

[4] Bakken, B. T. & Haerem, T. (2011). Intuition in Crisis Management: The Secret Weapon of Successful Decision Makers. In Sinclair, M. (Ed.), Handbook of Intuition Research (p. 122-132). Cheltenham: Edward Elgar.

[5] Bakken, B. T. & Hærem, T. (2020). Adaptive Decision Making in Crisis Management. In: Sinclair, M. (ed). Handbook of Intuition Research as Practice, pp. 14-26. Cheltenham, UK: Edward Elgar.

[6] Bakken, B. T. (2020). Between chaos and control: the practical relevance of military strategy in fighting and recovering from a pandemic outbreak. In: Rawat, S., Boe, O. & Piotrowski, A. (eds.): Military Psychology Response to Post Pandemic Reconstruction (pp. 585-599). Rawat Publications.

[7] Bakken, B. T., Johannessen, S., Søberg, D., & Ruud, M. (2007). The Intuitive vs. Analytic Approach to Real World Problem Solving: Misperception of Dynamics in Military Operations. In: Cook, M., Noyes, J., & Masakowski, Y. (eds): Decision Making in Complex Environments (p. 201-211). (ISBN 978-0-7546-4950-2).

[8] Bakken, B. T., Lund-Kordahl, I., & Sandberg, I. (2022). Augmented and Virtual Reality (AR/VR) and Artificial Intelligence (AI) Technology in Systematic Inter-Professional Crisis Management Training. Proceedings of the 2022 International Command and Control Research and Technology Symposium (ICCRTS).

[9] Balasubramanian, N., Ye, Y., & Xu, M. (2022). Substituting Human Decision-Making with Machine Learning: Implications for Organizational Learning. AMR, 47, 448–465. https://doi.org/10.5465/amr.2019.0470.

[10] Bjurström, E. & Bakken, B. T. (2022). Dynamic Decision Making under Uncertainty: A Brehmerian Approach. Journal of Behavioral Economics and Social Systems, 4(2), 55-68.

[11] Boin, A., Ekengren, M. and Rhinard, M. (2021). Understanding and acting upon a creeping crisis, in Boin, A., Ekengren, M. and Rhinard, M. (Eds.), Understanding the Creeping Crisis, Palgrave Macmillan, open access, pp. 1-17.

[12] Brehmer, B. (2000). Dynamic decision making in command and control. In C. McCannn and R. Pigeau (eds) The human in command: Exploring the modern military experience, Kluwer, New York.

[13] Brehmer, B. (2002). Learning to control a dynamic system. Unpublished manuscript, Swedish National Defence College, Stockholm.

[14] Carlsmith. J. (2022). Is Power-Seeking AI an Existential Risk? arXiv:2206.13353v1. https://doi.org/10.48550/arXiv.2206.13353.

[15] Constantinescu, M., Vică, C., Uszkai, R. & Voinea, C. (2022). Blame It on the AI? On the Moral Responsibility of Artificial Moral Advisors. Philosophy and Technology, 35 (2), 1-26.

[16] Cummings, M. L. (2004). Automation Bias in Intelligent Time Critical Decision Support Systems. AIAA 2004-6313. AIAA 1st Intelligent Systems Technical Conference, September 2004.

[17] Cummings, M. L. (in press). Rethinking the maturity of artificial intelligence in safety-critical settings. AI Magazine.

[18] Dahl, F. A. & Bakken, B. T. (2002). How game theory fails to explain man – An experimental study of human decision making and learning in an air campaign simulation model. Military Operations Research (MOR), 7(2), 5-14.

[19] Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. International Journal of Human-Computer Studies, 58(6), 697-718. https://doi.org/10.1016/S1071-5819(03)00038-7.

[20] Eich, A., Klichowicz, A., & Bocklisch, F. (2023). How automation level influences moral decisions of humans collaborating with industrial robots in different scenarios. Front. Psychol., 14, 1107306. https://doi.org/10.3389/fpsyg.2023.1107306.

[21] Goddard, K., Roudsari, A., Wyatt, J. C. (2011). Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc, 19(1), 121-7. https://doi.org/10.1136/amiajnl-2011-000089.

[22] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51(5), 93. https://doi.org/10.1145/3236009.

[23] Gunning, D., Vorm, E., Wang, J.Y. & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. Applied AI Letters, 2: e61. https://doi.org/10.1002/ail2.61.

[24] Guppy, L. & Twigg, J. (2013). Managing chronic crises and chronic hazard conditions. Environmental Hazards, 12(1), 5-18. https://doi.org/10.1080/17477891.2012.689251.
[25] Hærem, T., Valaker, S., Lofquist, E. A., & Bakken, B. T. (2022). Multiteam Systems Handling Time-Sensitive Targets: Developing Situation Awareness in Distributed and Co-located Settings. Frontiers in Psychology, 13,864749-864749. https://doi.org/10.3389/fpsyg.2022.864749.
[26] Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2022). Unsolved Problems in ML Safety. arXiv:2109.13916v5. https://doi.org/10.48550/arXiv.2109.13916.
[27] High-Level Expert Group on AI (AI HLEG). (2019). Ethics Guidelines for Trustworthy AI. The European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
[28] Himmelreich, J. Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations. Ethic Theory Moral Prac 21, 669–684 (2018). https://doi.org/10.1007/s10677-018-9896-4
[29] Jia, N., Luo, X., Fang, Z., Liao, C (in press, published online). When and How Artificial Intelligence Augments Employee Creativity. Academy of Management Journal. https://doi.org/10.5465/amj.2022.0426.
[30] Johnson, J. (2020). Delegating strategic decision-making to machines: Dr. Strangelove Redux?, Journal of Strategic Studies. https://doi.org/10.1080/01402390.2020.1759038
[31] Kalkman, J. P. (in press). Radical and Swift Adaptive Organizing in Response to Unexpected Events: Military Relief Operations after Hurricane Dorian. Academy of Management Discoveries. https://doi.org/10.5465/amd.2020.0213.
[32] Kaplan, J. (2016). Artificial Intelligence – What everybody needs to know. Oxford University Press.
[33] Kern, C., Gerdon, F., Bach, R. L., Keusch, F., & kreuter, F. (2022). Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. Paterns, 3, 100591. https://doi.org/10.1016/j.patter.2022.100591.
[34] Klein, G. (2008). Naturalistic Decision Making. Human Factors, 50(3), 456-460.
[35] Knox, W. B., Allievi, A., Banzhaf, H., Schmitt, F., Stone, P. (2023). Reward (Mis)design for autonomous driving. Artificial Intelligence, 316, 103829. https://doi.org/10.1016/j.artint.2022.103829.
[36] Kvalnes, Ø. (2011). Blurred Promises: Ethical Consequences of Fine Print Policies in Insurance. Journal of Business Ethics, 103(1), 77-86. https://www.jstor.org/stable/41475992.
[37] Langeland, P. A. (2022). Description of a Transsystemic Motivation Theory. Presentation at the NEON 2022 Conference, Drammen, Norway.
[38] Manheim, D. (2019). Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence. Big Data and Cognitive Computing, 3, 21. https://doi.org/10.3390/bdcc3020021.
[39] Neads, A., Farrell, T. & Galbreath, D. J. (2023) Evolving towards military innovation: AI and the Australian Army, Journal of Strategic Studies, https://doi.org/10.1080/01402390.2023.2200588.
[40] Ngo, R., Chan, L., & Mindermann, S. (2023). The Alignment Problem for a Deep Learning Perspective. arXiv:2209.00626v4. https://doi.org/10.48550/arXiv.2209.00626.
[41] Norwegian Ministry of Defence& Norwegian Ministry of Justice and Public Security (2020). Support and Cooperation: A description of the total defence in Norway. Norwegian Ministry of Defence & Norwegian Ministry of Justice and Public Security. www.regjeringen.no.
[42] Okoli, J. (2021). Improving decision-making effectiveness in crisis situations: developing intuitive expertise at the workplace. Development and Learning in Organizations, 35(4), 18-20. https://doi.org/10.1108/DLO-08-2020-0169
[43] Paton, D. and Flin, R. (1999). Disaster stress: an emergency management perspective. Disaster Prevention and Management, 8(4), 261-267. https://doi.org/10.1108/09653569910283897.
[44] Paulsen, J. E. (2021). AI, Trustworthiness, and the Digital Dirty Harry Problem. Nordic Journal of Studies in Policing, 8(2), 1–19. https://doi.org/10.18261/issn.2703-7045-2021-02-02.
[45] Perrow, C. (1999). Normal Accidents: Living with High Risk Technologies. Princeton University Press.
[46] Ragni, M., Steffenhagen, F., & Klein, A. (2011). Generalized dynamic stock and flow systems: An AI approach. Cognitive Systems Research, 12(3–4), 309-320, https://doi.org/10.1016/j.cogsys.2010.12.008.
[47] Stensrud, R., Mikkelsen, B., & Valaker, S. (2023). Orchestrating Humans and Non-human Teammates to counter security threats: Human-autonomy teaming in high and low environmental complexity and dynamism. Intelligent Human Systems Integration (IHSI): Vol. 69. https://doi.org/10.54941/ahfeXXXX.
[48] Stensrud, R., Valaker, S. (2023). Methods to meet changes in the security environment a proposal of qualitative and quantitative assessment attributes for coordination performance. Proceedings of the 20th ISCRAM Conference (pp. 676-691), Omaha, Nebraska, USA.
[49] Sterman, J. (2000). Business Dynamics: Systems Thinking and Modeling for a Complex World. McGraw Hill Higher Education.
[50] Strümke, I. (2023). Maskiner som tenker - algoritmenes hemmeligheter og veien til kunstig intelligens. [Machines that think. The secrets of algorithms and the path to artificial intelligence]. Kagge forlag.
[51] Topper, B. & Lagadec, P. (2013). Fractal Crises. J Contingencies & Crisis Man, 21, 4-16. https://doi.org/10.1111/1468-5973.12008.

[52] Valaker, S., Hærem, T. & Bakken, B. T. (2018). Connecting the dots in counterterrorism: The consequences of communication setting for shared situation awareness and team performance. Journal of Contingencies and Crisis Management, 26(4), 425-439.
[53] Vilone, G., & Longo, L. (2020). Explainable Artificial Intelligence: a Systematic Review. arXiv:2006.00093. https://doi.org/10.48550/arXiv.2006.00093.
[54] Staffas, K., Bjurström, E., & Roxtröm, G. (2021). Challenges in understanding AI. Proceedings of the 26th International Command and Control Research and Technology Symposium (ICCRTS). https://internationalc2institute.org/26th-iccrts-information-central

[i] ChatGPT (https://chat.openai.com/, version dated 14.6.2023) has been used as a tool in the writing of this article.

We thank Ronald Slaatmo and Kristoffer Lie Eide for their contributions to the conceptualization and methodological design of proposed empirical studies following this article.