Human-AI Collaboration in Cognitive Warfare: Enhancing Resilience Against Information Operations

Tim Robin Kosack, Thomas Krüger
German Aerospace Center
Institute for AI Safety & Security
Rathausallee 12, 53757 St. Augustin
GERMANY

tim.kosack@dlr.de, thomas.krueger@dlr.de

ABSTRACT

The rapid evolution of Artificial Intelligence (AI) has significant impact on the speed and accuracy of a broad range of systems. Enabling automated or autonomous workflows, AI bears the possibility to make existing applications more efficient and to design and fulfill entirely new tasks. Nevertheless, the integration of AI entails risks that give rise to calls for meaningful human control (MHC) of these systems. While the cooperation or collaboration between humans and AI is characterized through distinctive challenges, which involve both the capabilities of the human operator as well as the design of the AI system, this connection can tackle multiple ethical and safety-related concerns.

In this paper, we orchestrate the concept of MHC and human-AI collaboration to fit the specific needs of countering information warfare. We investigate the role of AI in countering information warfare, specifically through the lens of a Disinformation Monitoring and Alert System (DMAS). We explore the challenges of fully autonomous systems in the domain of anticipating and countering disinformation, emphasizing the necessity of MHC to prevent adversarial manipulation. Thus, we argue that MHC is not limited to ensure safety and ethical governance, but can also make crucial contributions to the security of an AI system. By examining potential vulnerabilities and the advantages of human-AI collaboration, this paper aims to contribute to the strategic enhancement of disinformation defense mechanisms through MHC.

While information warfare and disinformation campaigns are not exclusive to digital media or social networks, we focus our study on digital-based disinformation campaigns. These pose the greatest threat to NATO members in terms of both their number and their risk potential. In addition, this data-driven area is particularly suitable for AI-based applications for prevention.

1.0 INTRODUCTION

Artificial Intelligence (AI) has transformed the landscape of technology, driving significant advancements across a wide range of applications. Thus, AI has revolutionized the speed, accuracy, and efficiency with which tasks are performed. These AI-driven capabilities not only enhance existing workflows but also enable the development of new, previously unattainable functions. However, with these opportunities come substantial risks, particularly in areas where AI operates with a high level of automation or even autonomy. The growing complexity and automation of AI systems have sparked an urgent debate around the need for meaningful human control (MHC). While AI excels in processing vast amounts of data and automated decision-making, there are ethical, safety, and security related concerns that arise when human oversight and control is diminished or removed altogether.

One particularly critical domain where these concerns intersect is cognitive warfare, especially the spread of disinformation. Disinformation, or deliberately misleading information intended to deceive, has become an increasingly potent tool in the realm of global politics and international conflict. The rise of social media and

digital communication platforms has provided a fertile ground for disinformation campaigns, which can quickly amplify false narratives, undermine trust in institutions, and destabilize societies. The integration of AI into this field — as a tool for creating, disseminating and countering disinformation — raises complex questions. On the one hand, AI-based tools analyze and monitor massive datasets and can be harnessed to detect and prevent disinformation. On the other hand, the same capabilities can be manipulated to spread disinformation at unprecedented speeds, e.g., through the use of generative AI models.

This paper focuses on the role of MHC and Human-AI Teaming (HAT) to address specific risks in this domain. To show the benefit of MHC in AI systems, especially for AI in countering disinformation, we introduce the concept of a *Disinformation Monitoring and Alert* System (DMAS) as a potential tool. The DMAS leverages AI's capacity for pattern recognition and data analysis to identify and flag emerging disinformation campaigns, giving political and military actors the necessary time to respond. However, fully autonomous systems are not without vulnerabilities. They can be susceptible to adversarial manipulation, such as decoy attacks that trigger false alerts, reducing the system's reliability. Additionally, AI systems rely heavily on training data and pre-adjustments, which can introduce biases that create blind spots in the detection of disinformation. These weaknesses highlight the importance of maintaining sufficient human oversight in the form of MHC and HAT to ensure that AI systems remain secure, ethical, and responsive to unforeseen challenges.

In addressing these issues, we argue that human-AI collaboration offers a balanced approach. By combining AI's technical capabilities with human judgment, it is possible to mitigate the risks associated with autonomous systems. Human operators can provide context and oversight that AI systems may lack, particularly in fast-changing environments like information warfare. This paper will explore the framework of MHC within the context of AI-driven disinformation defense, emphasizing that MHC is not only suitable for ethical and safety-related concerns, but also enhances the security and effectiveness of AI systems. By examining the potential and limitations of the DMAS, we aim to contribute to the broader conversation on how AI can be responsibly integrated into the fight against disinformation, safeguarding democratic institutions and international stability.

2.0 THE VIRTUE OF ARTIFICIAL INTELLIGENCE IN INFORMATION DISSEMINATION

The benefits of AI cover numerous areas. From cancer detection through image recognition [1] to financial analysis [2] and autonomous driving [3], AI enhances the efficiency and reliability of processes which could previously only be performed by humans. The analysis of large amounts of data is one crucial aspect in this regard. One distinct example for the impact of AI on everyday life is the use of AI in information dissemination.

While the direct access to news websites and apps has decreased in recent years, people are increasingly turning to social media for information [4]. The high amount of data and the necessity of fast processing in this domain opens the door for AI-based data-analysis. Numerous algorithms are either used for content processing (face recognition, annotation, audio transcription, etc.) or content propagation (recommendation, ad delivery and targeting, content moderation, etc.) [5]. Content propagation algorithms are crucial for information dissemination: a recommendation algorithm can, based on the profile of a user, tailor the content that is shown to him or her. Here, the preferences of a user and his interaction with content enables an algorithm to address an individual person, which could improve the user-experience since relevant content is emphasized while irrelevant content is avoided. This implementation of AI-based systems enables the automated analysis of large datasets, facilitating the customization of content for all users at the same time. This procedure also bears the reason for using such algorithms for content propagation, since the optimized user-experience serves the economic imperatives of the platform operators by most likely increasing the time spend on a social platform, and therefore its advertising revenue.

Thus, the use of AI-based systems and automation facilitates information dissemination in a significant way. Nevertheless, this arises ethical issues as well as risks for the users' safety. The deployment of autonomous AI systems in the domain of information dissemination has led to the discrimination of users in the past [6]. Additionally, questions of digital surveillance and the freedom of users arise. Thus, the use of AI is often concerned with ethical, safety and security issues, which challenge the balance between benefits and risks of using AI-based systems. This is the case in numerous other sectors, including autonomous driving, air traffic control or military applications. Thus, the call for MHC of these systems originates.

3.0 MEANINGFUL HUMAN CONTROL AND HUMAN-AI COLLABORATION

In this section, we discuss the concept of MHC as well as the parameters that determine the integration of MHC into an AI system. To this end, we establish the concept of MHC and levels of control in the process loop, before going on to demonstrate the combination of MHC and HAT.

3.1 The Parameters of Meaningful Human Control

The concept of MHC roots in the discussion about the potentially unethical use of lethal autonomous weapon systems [7]. While the origin of the concept is easily determined, there is no generally accepted definition of what MHC actually contains, in which systems it needs to be integrated and which additional value is created through its integration. But the notion of 'meaningful' indicates that a simple human control of the AI system is not sufficient, either due to restrictions in competency, information, or psychological capabilities [8]. Thus, the approach aims at strengthening the amount of *meaningful* human control through expanding the operator's capability of controlling the system.

In order to identify the potential and possible integration of MHC, the working cycle of a system must be considered. In the military domain, this cycle can be described in a loop, which combines the process steps of observing, orienting, deciding and acting (OODA-loop) [9]. The extent of a human operator's control depends on the degree to which he or she is included into the loop: Human-in-the-loop (HITL), Human-on-the-loop (HOTL), Human-out-of-the-loop (HOOTL), and Human-in-Command (HIC). With a HITL-approach, the operator has the capacity to control each process step, whereas this ability is limited with a HOTL-approach. HIC refers only to the decision of when and how to use AI systems [10], while HOOTL describes an autonomous system without the possibility of a human intervention. Before leveraging any form of human control to MHC, a system needs a starting point of human control. Without this minimum requirement, one must focus on emphasizing the existence of human control in the OODA-loop in the first place. Thus, MHC is not compatible with the HIC-approach, since the operator lacks a significant amount of control to fit the concept of MHC. Accordingly, the HOOTL-approach contradicts the concept of MHC as well since no level of human control is permitted. Therefore, the concept of MHC is compatible with HITL and HOTL, since these approaches bear the potential of strengthening the already existing control of a human operator. This clarifies the systems' requirements if the necessity of MHC is determined.

Besides the possibility of the integration of MHC, the purpose of the integration also indicates the actual conception of MHC. Davidovic (2023) highlights five main purposes of integrating MHC into a system: (1) to increase safety and precision, (2) to ensure responsibility and accountability, (3) to assure morality and dignity, (4) to support democratic engagement and consent, as well as (5) strengthening institutional stability. The purposes for which MHC should be integrated into an AI system may overlap, but emphasizing the main purpose brings clarity into the actual embodiment of MHC [11]. Thus, the systems' overall design as a HITL or HOTL system as well as the purpose of the integration of MHC build the framework for the actual conception of MHC, they determine the appropriate method of controlling an AI system.

3.2 Human-AI Collaboration as a Complement to Meaningful Human Control

HAT refers to the cooperation or collaboration between an AI-based system and the operator to achieve goals. Here, the arrangement of cooperation or collaboration depends on the sophistication of the AI system [12]. While human-AI cooperation is focused on a directive approach and end-user awareness, human-AI collaboration is a concept for sophisticated AI systems, where both shared situational awareness as well as communication between an operator and the system are given. The AI system and the operator work together on a shared goal, supporting each other in a co-constructive approach [12].

The concept of HAT and the sub-category of human-AI collaboration originates from the aerospace and autonomous driving sector. According to this, situational awareness can be widely defined as an awareness about "what is known about the environment, what is happening in it and what might change" [13]. Sharing this awareness between an AI system and a human operator allocates each actor a specific area of authority. For further collaboration, both actors need to be able to communicate with each other, exchanging information and sharing their insights. This exchange is crucial for the collaboration, otherwise both actors would work on their own goals without supporting each other, diminishing the benefits of complementarity. Thus, the overall shared goal can be achieved through sectioning and sharing work steps between the AI and the operator. Human-AI collaboration makes an indirect statement about the control of the operators (e.g., whether it is a HITL or HOTL concept), it defines the requirements for teaming human operators with AI-based systems, conceptualizing rather the nature of working together than the level of control. Thus, human-AI collaboration can act as a complement to the concept of MHC.

4.0 INFORMATION WARFARE AND DISINFORMATION CAMPAIGNS

Davidovic names the increase of safety and precision as a central purpose to integrate MHC into AI-based systems [11]. In the context of countering cognitive warfare and disinformation campaigns, the integration of MHC not only increases safety and precision, it could also serve as a security component. The functional characteristics of disinformation campaigns in information warfare are crucial for the AI-based detection and prevention of such campaigns.

> *"One cannot underestimate the role of the mass media in executing Russia's foreign policy goals." [14]*

In recent years, information warfare became a widely-used addition for conventional warfare. For some actors it even became the prevalent method [14]. Besides cyberwarfare, troll factories and bots, disinformation campaigns are an integral part of information warfare. Disinformation campaigns refer to coordinated efforts by political actors, aiming at deliberately spreading false or misleading information to a target group, influencing their opinion regarding a topic of interest. These could be "[I]ntentional falsehoods or distortions, often spread as news, to advance political goals such as discrediting opponents, disrupting policy debates, influencing voters, inflaming social conflicts, or creating a general backdrop of confusion and information paralysis." [15]. Regarding the resilience of the NATO against disinformation, state actors pose the highest risk. Nevertheless, non-state actors like Jihadist groups still use this practice [16].

Disinformation campaigns are – compared to misinformation or fake news – strategically deployed and have an intentional harmful impact on public discourse, often aiming at specific events. While they are not a phenomenon reserved for modern times, digital media and especially social media allow a more efficient and effective organization.

Three aspects make social media platforms and their audience a prolific target for the spread of disinformation campaigns in the realm of information warfare: 1. A large proportion of political relevant news is disseminated through social media. Accordingly, a large audience can be addressed. 2. The functionality of social media, especially targeted advertising and recommendation systems, allow a precise

targeting of specific groups (e.g. indecisive voters). 3. Lastly, the creation and modification of news and images on social media involves relatively low cost and allows an almost simultaneous dissemination. Since the timing of disinformation campaigns, especially regarding timed events like elections, is crucial, the reduction of the temporal difference between a happening and the moment a person receives the information is conducive for disinformation campaigns.

5.0 BUILDING RESILIENCE AGAINST DISINFORMATION CAMPAIGNS

The manipulation of the public opinion through disinformation is essential in the realm of cognitive warfare. Enhancing the resilience of the NATO and its member states against this kind of threat is of paramount importance. There are multiple ways to build resilience, from strengthening societal cohesion, embedding higher trust in well-established media or political education as well as media competency [17], [18], [19]. Besides these approaches which are focusing on societal aspects, a technical approach in the form of an early warning system can support this process. The use of AI-based systems in the proliferation of disinformation is directly connected to the analysis and handling of large amounts of data. Countering this procedure is connected to large amounts of data as well, which facilitates the instrumentalization of AI for preventing or mitigating the effects of disinformation campaigns.

In this section we provide a theoretical concept for a DMAS. Such a system is designed to detect and track emerging disinformation campaigns, providing an early warning that allows for timely and coordinated countermeasures. The integration of AI in this context offers a powerful tool for analyzing large amounts of data from various sources, including social media and communication networks, to identify potential disinformation campaigns in real-time. However, the deployment of AI-based systems in this domain is not without challenges. We further analyze the inherent vulnerabilities of such a system, including the risks of decoy attacks and biases introduced by training data.

5.1 Conceptualizing a Disinformation Monitoring and Alert System

Cao et al. formulate the idea of a cognitive warfare monitoring and alert system, which could help to identify cognitive warfare campaigns as they arise and before they reach their full potential [20]. Since we focus on disinformation campaigns as a part of information respectively cognitive warfare, we reinterpret this idea as a *Disinformation Monitoring and Alert System*. Key aspect of the DMAS is the early detection of emerging disinformation campaigns which are relevant for the protection of the interests of NATO and its member states. This means creating or extending the timeframe for political and military actors to prepare for such campaigns and to prevent them or mitigate their effects through coordinated countermeasures. Timing is the crucial factor here, both for the effectiveness as well as the prevention/mitigation of disinformation campaigns, especially when those are aimed at specific events like elections. Single disinformation campaigns are aimed more at destabilizing individual NATO member states, but this can weaken cohesion within NATO in the long term. A system such as DMAS should therefore also be introduced at NATO level as part of alliance defense.

A DMAS, acting as a system of systems, combines both the identification of (emerging) disinformation as well as tracking the dissemination of this disinformation. Therefore, the principle of operation is multi-stage: the identification of disinformation includes fact-checking, fake-news and deep-fake detectors, as well as the collection of background data. This process is supplemented by tracking and tracing the data through network-analysis and the collected background data. For fulfilling these tasks, a DMAS must be able to autonomously analyze a high amount of data, coming from broadcasts, social media, communication networks, messaging, etc. [20]. Fulfilling this task through human operators would not be manageable in the timeframe relevant for preventing the effects of disinformation campaigns.

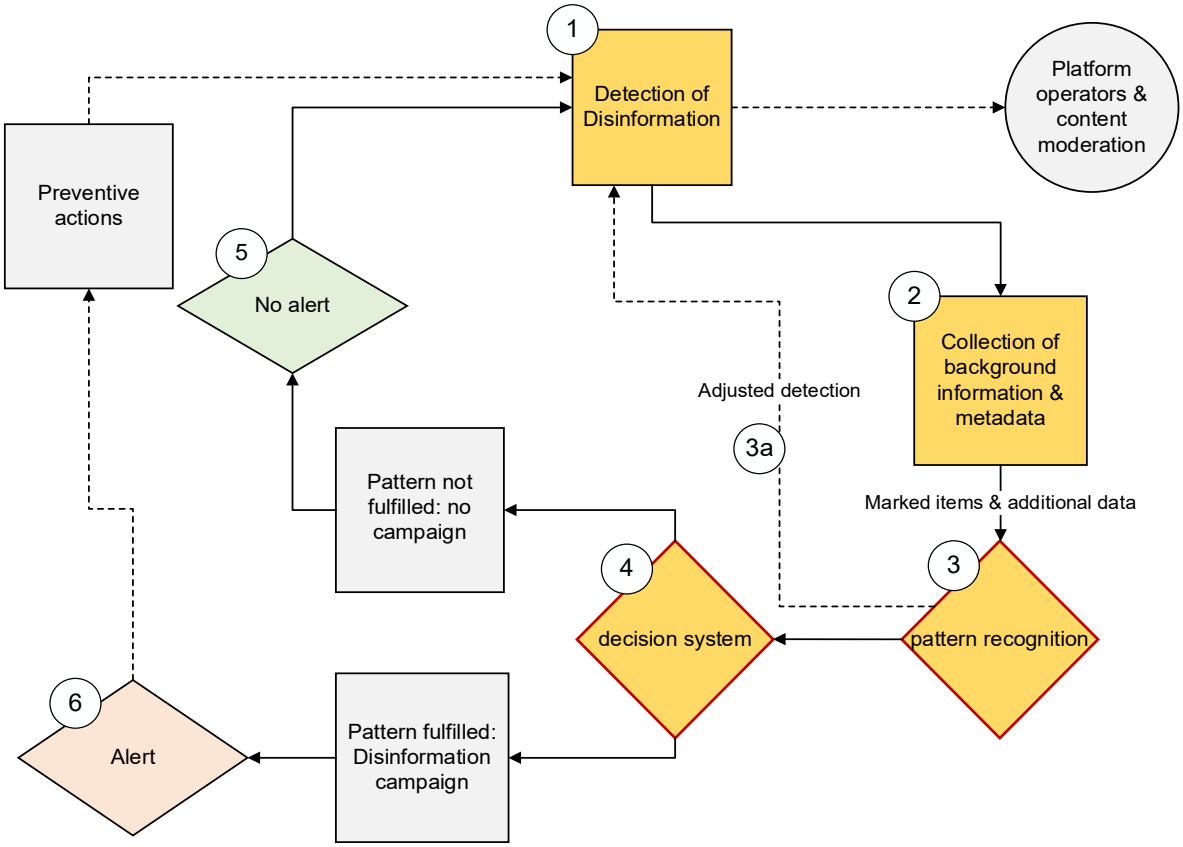

Figure 1: Workflow of the DMAS.

The workflow of a DMAS starts with the identification and detection of suspected disinformation (1). Here, the system should focus on specific content, e.g. international conflicts, election events or other current political issues. In this initial step, the intentionality and level of organization is excluded. A combination of fake-news detectors and web-crawling tools are used for tagging specific articles, videos, images, posts, etc., which creates a pool of items of interest. These tasks can already be addressed through several AI-based systems which use natural language processing or image recognition [21], [22], [23]. Further, a collaboration with the social-media platform operators is conceivable here, even if this is no integral part of the DMAS: Forwarding and reporting flagged items can support platform-internal fake news detectors and content moderation so that content can be deleted at an early stage. Obtaining background information and metadata is essential for a further analysis of the tagged items (2). This includes the geographical and virtual location of actors posting or sharing the tagged items, their usernames, timecodes, and other metadata. Both work steps can be performed by autonomous AI-based systems. Using the collected data, an (3) AI-based pattern recognition system checks the monitored items for a pre-defined pattern. Examples for patterns of disinformation campaigns focus on specific actors [24], topics [25], or the relationships among the spreaders [26]. A pattern contains a specific combination of topics (e.g., elections, conflicts, political leaders), temporal information (e.g., temporal proximity to an event, creation date of the involved accounts, spread of information in a certain period of time) account information (e.g., usernames, virtual and geographical location, interconnection with other accounts), and possible connections to hostile governments and actors. The intentionality of the spread of disinformation and its level of organization is crucial here. Based on the results of the system, statements can be made about the existence of a disinformation campaign (4): if the necessary parameters are not fulfilled, no alert is triggered (5), and the workflow begins again. However, if the parameters are met, an alert is triggered (6). This leads to the initiation of preventive or mitigating actions by actors outside of the DMAS.

5.2 Decoy Attacks and Disadvantages of Autonomous Disinformation Monitoring and Alert Systems

A system like the proposed DMAS could provide the respective authorities and advantage in their response options. This ranges from countering fake news through pre-bunking and de-bunking [27] to legal and regulatory answers [28]. But the AI-based data-analysis contains two weaknesses which are inherent in the functionality of AI systems. First, the analysis of patterns which are typical for a disinformation campaign enables the preparation of decoy attacks. Further, the focus on specific areas may create geographical blind spots. Both weaknesses could mitigate the reliability and efficiency of a DMAS by alerting in situations where there is no real disinformation campaign incoming, or alerting not respectively too late.

5.2.1 Decoy Attacks as a Means of Cognitive Warfare

The DMAS acts as a warning system in the realm of cognitive warfare. Attacking such a warning system with decoy attacks would prevent the identification of real disinformation campaigns. Thus, invalidating the preventive effect of the DMAS and turning it into a potentially harmful tool, consequently fulfilling the goals of cognitive warfare.

Pattern-recognition is the key aspect of the network-analysis work step. A pattern would contain different kinds of information, which, if combined in the framework of required parameters, would cause an alert for a disinformation campaign. This information may be the virtual and geographical position of an account, it's posting and sharing frequency, the source of this content, the involvement of freshly created accounts (bots), individuals and institutions interacting with this content, the topic of the content, etc. If enough parameters are met, an alert will be triggered notifying about an incoming disinformation campaign. It is likely the first goal of the initiators of a disinformation campaign to mask it in such a way as it appears to be a credible and truthful spread of information, this is essential for a functioning campaign. But a reverse approach could aim at the reliability of a DMAS. Other than an actual disinformation campaign, a decoy attack would use the pattern-recognition function of an AI-based system to its advantage, proactively creating alerts: the intentional creation of a disinformation campaign would combine the necessary parameters to mark it as a potential disinformation campaign, but without trying to actually manipulate the publics opinion. The main difference between an intentional disinformation campaign and a decoy attack would be the attention paid to the details: One campaign is actually trying to persuade its audience, the other is just trying to meet the requirements for activating the DMAS. Thus, complementing the quality of actual disinformation campaigns with the quantity of decoy-attacks. While this approach would be a gray- or even black-box attack [29] and could therefore be challenging in the first attempts, the use of generative AI-tools could facilitate this procedure through automatized content creation. Further, while topic and content of disinformation change in a high frequency, other parameters may change seldom (virtual or geographical location), or not at all (the high interaction-frequency of users). Thus, targeting a DMAS through a decoy attack is most likely cheaper, which increases the likelihood of such attacks.

5.2.2 Biases and the Dependency on Input Data

The second weakness of an AI-based DMAS can be found in its strong dependency on training data and pre-adjustments. This may create a bias towards specific regions, which leaves a blind spot for other regions. Further, the timeliness of the input data could influence the classification of politically relevant happenings.

While creating a pattern which is typical for disinformation campaigns, the geographic and virtual position are crucial for tracking and tracing a potential campaign. Looking at the current status of international relations, it is evident that the threat of disinformation campaigns by certain political actors is greater than by others. In addition, certain areas or certain NATO member states are more intensely affected than others, with the geopolitical and historical background as main factors [30]. Under these conditions, it is important that disinformation campaigns from well-known adversarial actors and regions are identified and mitigated

as early as possible. Thus, pre-adjusting a DMAS towards specific geographical and virtual locations, languages, usernames, etc. is a natural choice. Especially linguistic approaches are reasonable [31] while the correct selection of training data sets is crucial here. But while this procedure may increase the efficiency and accuracy of the DMAS for specific regions, it also creates a bias. The over-representation achieved by emphasizing certain regions simultaneously leads to an under-representation of other regions. This unintentional creation of blind-spots could pose a weakness in the reliability of a DMAS: disinformation campaigns which root in these under-represented countries are more likely to be overlooked or misclassified. The timeliness of the input-data also plays a decisive role in the evaluation and classification of information. Even when an AI-based DMAS can be infused with new data, thus, updating political situations and happenings, it will always be based on data of the past. And since continuous real-time updates are challenging, major political single-time events (e.g., a stock market crash, the death of a political leader, election results) may be harder to classify compared to longer-lasting events (e.g., ongoing wars, state oppression of the opposition, environmental events). While this is an insuperable technological condition, it impedes the classification of incoming information, since the contextualization of such events is deficient.

Both weaknesses, the risk of decoy attacks and the strong dependency on training-data and pre-adjustments, are inherent to the way AI-based systems work. It is important to note that an autonomous AI-based DMAS is no exception, and this could undermine the benefits of such a system. Both the dependency on training-data and pre-adjustments and decoy attacks pose a risk to the security of the DMAS. While both problems should also be addressed through technological approaches (e.g., strengthening the systems resistance towards fake campaigns, thoughtful and well-balanced pre-adjustments), the integration of a human operator is promising.

6.0 MEANINGFUL HUMAN CONTROL AS A SECURITY COMPONENT

Many benefits of AI-based systems are based upon the AI's ability to take over tasks that human cannot do at all, or with highly reduced efficiency. The analysis of high amounts of data, as in the workflow of the DMAS, is not different here. But as shown above, an autonomous DMAS bears risks which are inherent in the functionality of AI systems. Mitigating those risks by exchanging the autonomous analysis of large amounts of data through a team of human operators is no option: While skilled human operators could potentially screen the media with some technological auxiliary means for fake-news or disinformation, an AI-based data-analysis is considerably more efficient. In addition, any integration of human operators must be carefully considered so that the speed of the AI-based system is not unnecessarily reduced by human involvement. And since the timing is crucial in the realm of cognitive warfare, the use of AI-based autonomous data-analysis is unavoidable, even with the risks described here. Accordingly, the overarching goal is to minimize the risks of AI-based systems without significantly curtailing their benefits through relying only on human operators. Nevertheless, the collaboration of both AI systems and human operators, using in the concepts of MHC and HAT, can provide a solution to the risks without losing the benefits of AI-based data analysis. The combination of these two authorities connects their specific skills which fosters the resilience and reliability of the DMAS.

6.1 Sub-System Control and Human-AI Collaboration

In this regard, the integration of MHC into the DMAS rather aims at increasing its security than increasing safety and precision or other typical purposes of MHC-integration. The assurance of responsibility and morality is more of a by-product, especially in a system that is intended to protect against hostile measures. The key features that enables a team of human operators to mitigate the risk of decoy attacks as well as the dependency on input data of the DMAS are their capabilities of contextualizing data and accessing current information.

To address the risk of decreasing the reliability of the DMAS through decoy attacks (5.2.1), a human operator must be able to distinguish between actual disinformation campaigns and fake-attacks which just aim at fulfilling the right parameters to trigger an alert. However, the complexity of both disinformation campaigns and fake-attacks to the DMAS occurs: one the one side, a disinformation campaign must reach a minimum level of plausibility. On the other side, a fake-attack must mimic the exact same characteristics of a disinformation campaign, just without the necessary attention to details to actually manipulate the public opinion. The distinction between both can be made with the operators' attention to these details: their task lies in the evaluation of the plausibility of the disinformation campaign. To address the risk of blind spots (5.2.2) through the pre-adjustment of the DMAS, a human operator must equilibrate the areas where the DMAS may underperform. Thus, the operator must focus his or her awareness towards geographical or content-wise areas, which are not in the focus of the DMAS. This creates a shared situational awareness towards incoming disinformation campaigns in its entirety.

To this end, the operator must have a significant level of control over the DMAS. The initial process steps of the DMAS (Figure 1, step 1 & 2) can run fully autonomous without risking the security or reliability of the system. The steps of pattern recognition and the decision system (Figure 1, step 3 & 4) are the sub-systems, in which the integration of MHC is actually applicable. Thus, the human operator must have a significant amount of control over both sub-systems: controlling the pattern-recognition systems enables the operator to shift the systems' focus to under-represented areas (both geographical and content-wise), thus restarting and re-adjusting the process towards a new area, mitigating the risk of blind spots (Figure 1, step 3a). Controlling the decision system (Figure 1, step 4) enables the operator to separate between actual disinformation campaigns and false positives. This human-made decision is based on the results of the previous steps, but solely belongs to the human authority. The subsequent step, triggering the alert, can take place autonomously based on the decision from step 4. Thus, MHC is integrated into the sub-systems of the DMAS in the process steps 3 & 4, following a HITL-approach. Here, HAT in the form of human-AI collaboration splits the authority between the human operator and the system, thus, both authorities work on a shared goal. The communication between both authorities is given, both the system and the operator can share their information and (re-)allocate their tasks dynamically. This includes the necessity for mutual expectability, the capability to anticipate possible reactions from the respective other authority. The shared situational awareness inside the system is directly connected to this bidirectional (data) exchange. Applying MHC and human-AI collaboration to sub-systems of the DMAS allows a sustainment of the AI-based benefits while increasing its security.

6.2 Requirements for Human Operators

The technological integration of MHC is accompanied by specific requirements towards the human operator. Besides having the actual control, the operator must fulfill further requirements to safeguard the controlling process in a non-technical way: starting with a high level of experience in international relations, current politics and knowledge about potential adversaries, the operator must know about common topics and paradigms encompassing disinformation. This enables the operator to contextualize the data collected and analyzed by the DMAS in a holistic way. This is crucial for the evaluation of disinformation. In addition, the operator must receive daily updates on international relations, conflicts, and political events. In contrast to a data-based DMAS, which cannot be enriched with new data on a daily basis, this enables the operator to include the latest events in his evaluation. The process of updating knowledge is significantly shorter for human operators than for AI-based systems. Thus, the operator must be an expert in the politically relevant fields that are tackled by disinformation.

Further, the operator needs to be aware of the risks of human-machine respectively human-AI interaction. Here, the problems of 'rubber-stamping' and the automation bias may interfere a sophisticated integration of a human operator. The problem of 'rubber-stamping' arises from the decision pressure in a time-critical situation. If the operator does not have a sufficient amount of time to actually check the AI-generated results, he may tend to 'rubber-stamp' these results, giving a stamp of approval to save time [10], [32]. This leads to

a superficial, not-MHC conform level of control. While this risk originates from the technical terms of the system, the risk of the automation bias is rather connected to the psychological condition of the operator. The automation bias describes the condition of trusting machine-made results more than results which are based on one's own experience, even when information contradictory to the machine-made results is at hand [33], [34]. This process of over-trusting machine-made results could annul the incorporation of the human operator as a security component, especially regarding decoy-attacks. Even systems which communicate the probability of their evaluation are affected by this risk, since this communication of probability may also be misleading. Accordingly, the operator must have a high-level of experience with the system, knowing its exact capabilities and possible weaknesses. Thus, proactively countering over-trusting the AI system is crucial for the integration of MHC.

7.0 CONCLUSION

In this study, we set out to explore the concepts of MHC and human-AI collaboration in the realm of cognitive warfare, with a specific focus on the integration into a DMAS. Through combining the concept of MHC with human-AI collaboration, we have been able to highlight the security-enhancing effect of the integration of human operators into otherwise autonomous AI systems. Thus, we contribute to a deeper understanding of enhancing the resilience of systems which help to prevent or mitigate adversarial information operations. The key findings of this research include: 1. AI-based preventive systems in information warfare have specific weaknesses which may reduce their reliability or overall resilience. 2. These weaknesses can be addressed through MHC and human-AI collaboration, enhancing the security and of these systems. 3. This integration of human operators is accompanied by demanding conditions for these operators, since they must fulfill high-level requirements.

The implications of this research expand the theoretical and practical debates for the development and deployment of AI-based systems to counter adversarial information operations. Thus, indicating weaknesses and providing possible precautions for developers and responsible authorities. While this research has yielded necessary insights, several limitations must be acknowledged. First, the DMAS is still a theoretical concept. This limits the generalizability of the integration of human operators, since the actual workflow may chance. Secondly, the realm of cognitive warfare is in constant alteration, which may influence the necessity and configuration of preventive systems. Thus, a re-evaluation of the benefits of a DMAS and the integration of human operators is imminent.

While the results of this research emphasize the importance of MHC and human-AI collaboration, further research is needed in the conceptualization and development of AI-system to counter cognitive warfare. Especially interdisciplinary approaches that incorporate information technologies, political science and psychology are promising here, considering resilience and security in a comprehensive way.

8.0 REFERENCES

[1] Gregoor, A. M. S., Sangers, T. E., Bakker, L. J., Hollestein, L., Uyl – de Groot, Carin A., Nijsten, T., & Wakkee, M. (2023). An artificial intelligence based app for skin cancer detection evaluated in a population based setting. *Npj Digital Medicine*, *6*(1), 90. https://doi.org/10.1038/s41746-023-00831-w

[2] Gera, R., Assadi, D., & Starnawska, M. (Eds.). (2024). *Innovations in big data and machine learning. Artificial intelligence, fintech, and financial inclusion* (First edition). CRC Press. https://doi.org/10.1201/9781003125204

[3] Wang, Y., Du, S., Xin, Q., He, Y., & Qian, W. (2024). Autonomous Driving System Driven by Artificial Intelligence Perception Fusion. *Academic Journal of Science and Technology*, *9*(2), 193–198. https://doi.org/10.54097/e0b9ak47

[4] Reuters Institute. (2023). *Reuters Institute Digital News Report 2023.* Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf

[5] Narayanan, A. (2023). *Understanding Social Media Recommendation Algorithms.* Knight First Amendment Institute. https://doi.org/10.7916/khdk-m460 https://doi.org/10.7916/khdk-m460

[6] Grant, N., & Hill, K. (2023, May 22). Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's. *New York Times*. https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html

[7] Robbins, S. (2023). The many meanings of meaningful human control. *AI and Ethics*, 1–12. https://doi.org/10.1007/s43681-023-00320-6

[8] Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, *5*, 15. https://doi.org/10.3389/frobt.2018.00015

[9] Scharre, P. (2019). Army of none: Autonomous weapons and the future of war. W.W. Norton & Company.

[10] Albrecht, L., Hagen, B., Kosack, T. R., & Krüger, T. (2024). *The Limits of Meaningful Human Control of AI in the Maritime Domain [Manuscript submitted for publication].* 4th European Workshop on Maritime Systems Resilience and Security. German Aerospace Center.

[11] Davidovic, J. (2023). On the purpose of meaningful human control of AI. *Frontiers in Big Data*, *5*, 1017677. https://doi.org/10.3389/fdata.2022.1017677

[12] EASA. (2024). *EASA Artificial Intelligence Concept Paper Issue 2: Guidance for Level 1 & 2 machine-learning applications.* European Union Aviation Safety Agency. https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-concept-paper-issue-2

[13] Mutzenich, C., Durant, S., Helman, S., & Dalton, P. (2021). Updating our understanding of situation awareness in relation to remote operators of autonomous vehicles. *Cognitive Research: Principles and Implications*, *6*(1), 9. https://doi.org/10.1186/s41235-021-00271-8

[14] StratCom COE. (2015). Analysis of Russia's Information Campaign against Ukraine: Examining non-military aspects of the crisis in Ukraine from a strategic communications perspectives. NATO.

[15] Bennett, W. L., & Livingston, S. (2020). *The Disinformation Age. Social Sciences*. Cambridge University Press.

[16] Vilmer, J.-B. J., Escorcia, A., Guillaume, M., & Herrera, J. (2018). *Information Manipulation: A challenge for our Democracies.* Policy Planning Staff (CAPS, Ministry for Europe and Foreign Affairs), Institute for Strategic Research (RSEM). https://www.diplomatie.gouv.fr/IMG/pdf/information_manipulation_rvb_cle838736.pdf

[17] Golob, T., Makarovič, M., & Rek, M. (2021). Meta-reflexivity for resilience against disinformation. *Comunicar*, *29*(66), 107–118. https://doi.org/10.3916/C66-2021-09

[18] Teperik, D., Denisa-Liepniece, S., Bankauskaitė, D., & Kullamaa, K. (2022). *Resilience Against Disinformation: A New Baltic Way to Follow?* International Centre for Defence and Security.

[19] Dobrescu, P., Durach, F., & Vladu, L. (2022). *Building Resilience to Disinformation through Media and Information Literacy.* https://doi.org/10.21125/inted.2022.0863

[20] Cao, K., Glaister, S., Pena, A., Rhee, D., Rong, W., Rovalina, A., Bishop, S., Khanna, R., & Singh Saini, J. (2021). *Countering cognitive warfare: awareness and resilience.* NATO. https://www.nato.int/docu/review/articles/2021/05/20/countering-cognitive-warfare-awareness-and-resilience/index.html

[21] Steinebach, M., Gotkowski, K., & Liu, H. (2019). *Fake News Detection by Image Montage Recognition*. https://doi.org/10.1145/3339252.3341487 https://doi.org/10.1145/3339252.3341487

[22] Sufi, F. K., & Alsulami, M. (2021). Automated Multidimensional Analysis of Global Events With Entity Detection, Sentiment Analysis and Anomaly Detection. *IEEE Access*, *9*, 152449–152460. https://doi.org/10.1109/ACCESS.2021.3127571

[23] Cartwright, B., Frank, R., Weir, G., & Padda, K. (2022). Detecting and responding to hostile disinformation activities on social media using machine learning and deep neural networks. *Neural Computing and Applications*, *34*(18), 15141–15163. https://doi.org/10.1007/s00521-022-07296-0

[24] Yarova, A. (2023). Thematic Patterns of Russian Disinformation. *Baltic Journal of Legal and Social Sciences* (4), 158–165. https://doi.org/10.30525/2592-8813-2022-4-19

[25] Teixeira, J., & Martins, A. (2022). Thematic Patterns of Disinformation about COVID-19: The Framing of Checks in the Fato ou Fake and Lupa Agencies. *Journalism and Media*, *3*(1), 27–39. https://doi.org/10.3390/journalmedia3010003

[26] Zhou, X., & Zafarani, R. (2019). Network-based Fake News Detection. *ACM SIGKDD Explorations Newsletter*, *21*(2), 48–60. https://doi.org/10.1145/3373464.3373473

[27] Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(5). https://doi.org/10.1073/pnas.2020043118

[28] United Nations. (2022). *Countering disinformation for the promotion and protection of human rights and fundamental freedoms* (A/77/287). United Nations. https://documents.un.org/doc/undoc/gen/n22/459/24/pdf/n2245924.pdf?OpenElement

[29] Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). *Adversarial machine learning.* National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf https://doi.org/10.6028/NIST.AI.100-2e2023

[30] Shultz, D. (2023). *Who controls the past controls the future: How Russia uses history for cognitive warfare.* NATO Defense College. http://www.jstor.org/stable/resrep58203

[31] Põldvere, N., Kibisova, E., & Alvestad, S. S. (2024). Investigating the Language of Fake News Across Cultures. In S. Maci, M. Demata, M. McGlashan, & P. Seargeant (Eds.), *Routledge handbooks in applied linguistics. The Routledge handbook of discourse and disinformation* (pp. 153–165). Routledge Taylor & Francis Group. https://www.taylorfrancis.com/books/edit/10.4324/9781003224495/routledge-handbook-discourse-disinformation-stefania-maci-massimiliano-demata-philip-seargeant-mark-mcglashan

[32] The Guardian (2024, April 3). 'The machine did it coldly': Israel used AI to identify 37,000 Hamas targets: Israeli intelligence sources reveal use of 'Lavender' system in Gaza war and claim permission given to kill civilians in pursuit of low-ranking militants. *The Guardian*. https://www.theguardian.com/world/2024/apr/03/israel-gaza-ai-database-hamas-airstrikes

[33] Cummings, M. L. (2016). Automation Bias in Intelligent Time Critical Decision Support Systems. In D. Harris & W.-C. Li (Eds.), *Decision Making in Aviation* (First edition, pp. 289–294). Routledge. https://doi.org/10.4324/9781315095080-17

[34] Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association: JAMIA*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089