NORTH ATLANTIC TREATY
ORGANIZATION

SCIENCE AND TECHNOLOGY
ORGANIZATION

AC/323(HFM-322)TP/1108

www.sto.nato.int

STO MEETING PROCEEDINGS **MP-HFM-322**

Meaningful Human Control of AI-based Systems Workshop: Technical Evaluation Report, Thematic Perspectives and Associated Scenarios

(Séminaire sur le contrôle humain sensé des systèmes basés sur l'IA : rapport d'évaluation technique, perspectives thématiques et scénarios associés)

The technical evaluation report describes the main findings and conclusions of the HFM-322 workshop. The annexes contain themed synopses and representative scenarios.

Published June 2023

Distribution and Availability on Back Cover

NORTH ATLANTIC TREATY
ORGANIZATION

SCIENCE AND TECHNOLOGY
ORGANIZATION

AC/323(HFM-322)TP/1108

www.sto.nato.int

STO MEETING PROCEEDINGS **MP-HFM-322**

Meaningful Human Control of AI-based Systems Workshop: Technical Evaluation Report, Thematic Perspectives and Associated Scenarios

(Séminaire sur le contrôle humain sensé des systèmes basés sur l'IA : rapport d'évaluation technique, perspectives thématiques et scénarios associés)

The technical evaluation report describes the main findings and conclusions of the HFM-322 workshop. The annexes contain themed synopses and representative scenarios.

Authors:
Christopher Miller, Mark Draper, Jurriaan van Diggelen, Marlijn Heijnen,
Robert J. Shively, Frank Flemisch, Marcel Baltzer, Rogier Woltjer,
Mike Boardman, Kate Devitt, Marie-Pierre Pacaux-Lemoine, and Emma Parry.

The NATO Science and Technology Organization

Science & Technology (S&T) in the NATO context is defined as the selective and rigorous generation and application of state-of-the-art, validated knowledge for defence and security purposes. S&T activities embrace scientific research, technology development, transition, application and field-testing, experimentation and a range of related scientific activities that include systems engineering, operational research and analysis, synthesis, integration and validation of knowledge derived through the scientific method.

In NATO, S&T is addressed using different business models, namely a collaborative business model where NATO provides a forum where NATO Nations and partner Nations elect to use their national resources to define, conduct and promote cooperative research and information exchange, and secondly an in-house delivery business model where S&T activities are conducted in a NATO dedicated executive body, having its own personnel, capabilities and infrastructure.

The mission of the NATO Science & Technology Organization (STO) is to help position the Nations' and NATO's S&T investments as a strategic enabler of the knowledge and technology advantage for the defence and security posture of NATO Nations and partner Nations, by conducting and promoting S&T activities that augment and leverage the capabilities and programmes of the Alliance, of the NATO Nations and the partner Nations, in support of NATO's objectives, and contributing to NATO's ability to enable and influence security and defence related capability development and threat mitigation in NATO Nations and partner Nations, in accordance with NATO policies.

The total spectrum of this collaborative effort is addressed by six Technical Panels who manage a wide range of scientific research activities, a Group specialising in modelling and simulation, plus a Committee dedicated to supporting the information management needs of the organization.

•	AVT	Applied Vehicle Technology Panel
•	HFM	Human Factors and Medicine Panel
•	IST	Information Systems Technology Panel
•	NMSG	NATO Modelling and Simulation Group
•	SAS	System Analysis and Studies Panel
•	SCI	Systems Concepts and Integration Panel
•	SET	Sensors and Electronics Technology Panel

These Panels and Group are the power-house of the collaborative model and are made up of national representatives as well as recognised world-class scientists, engineers and information specialists. In addition to providing critical technical oversight, they also provide a communication link to military users and other NATO bodies.

The scientific and technological work is carried out by Technical Teams, created under one or more of these eight bodies, for specific research activities which have a defined duration. These research activities can take a variety of forms, including Task Groups, Workshops, Symposia, Specialists' Meetings, Lecture Series and Technical Courses.

The content of this publication has been reproduced directly from material supplied by STO or the authors.

Published June 2023

ISBN 978-92-837-2425-4

Table of Contents

List of Figures

HFM-322 Membership List

CO-CHAIRS

Dr. Mark DRAPER*
US Air Force Research Laboratory (AFRL)
UNITED STATES
Email: mark.draper.2@us.af.mil

Dr. Jurriaan VAN DIGGELEN*
TNO Defense, Security and Safety
NETHERLANDS
Email: jurriaan.vandiggelen@tno.nl

MEMBERS

Mr. Michael BOARDMAN*
DSTL
UNITED KINGDOM
Email: mjboardman@dstl.gov.uk

Dr. Fiona BUTCHER
DSTL
UNITED KINGDOM
Email: fdbutcher@dstl.gov.uk

Dr. Marco MANCA
SCImPULSE Foundation
ITALY
Email: marco@scimpulse.org

Dr. Marie-Pierre PACAUX LEMOINE*
Universite Polytechnique Hauts de France – LAMIH
FRANCE
Email: marie-pierre.lemoine@uphf.fr

Prof. Emma PARRY*
Cranfield University
UNITED KINGDOM
Email: emma.parry@cranfield.ac.uk

Dr. Jurriaan VAN DIGGELEN*
TNO Defense, Security and Safety
NETHERLANDS
Email: jurriaan.vandiggelen@tno.nl

Dr. Rogier WOLTJER*
Swedish Defence Research Agency FOI
SWEDEN
Email: rogier.woltjer@foi.se

ADDITIONAL CONTRIBUTORS

Dr. Marcel BALTZER*
Fraunhofer
GERMANY
Email: marcel.baltzer@fkie.fraunhofer.de

Dr. Kate DEVITT*
TASDCRC
AUSTRALIA
Email: Kate.Devitt@tasdcrc.com.au

Ms. Marlijn HEIJNEN*
TNO
NETHERLANDS
Email: marlijn.heijnen@tno.nl

Dr. Christopher MILLER**
SIFT
USA
Email: miller@sift.net

Mr. Robert J. SHIVELY*
NASA
USA
Email: robert.j.shively@nasa.gov

*** Contributing Author**

**** Author of Technical Evaluation Report**

PANEL/GROUP MENTORS

Dr. Adelbert BRONKHORST
TNO Human Factors
NETHERLANDS
Email: adelbert.bronkhorst@tno.nl

Prof. Dr. Frank FLEMISCH*
Fraunhofer FKIE/RWTH Aachen
GERMANY
Email: f.flemisch@iaw.rwth-aachen.de

*** Contributing Author**

Meaningful Human Control of AI-based Systems Workshop: Technical Evaluation Report, Thematic Perspectives and Associated Scenarios

(STO-MP-HFM-322)

Executive Summary

Meaningful Human Control (MHC) emerged as an important concept during the 2016 expert meetings organized by the UN Convention on “Certain Conventional Weapons” (CCW). While the concept has been linked to autonomous weapons, it can be applied more generally to AI-based military systems (both physical and informational) as a critical requirement to safeguard moral behavior, accountability, and the effective operational performance envelope of such systems.

The core objective of this Workshop was not to duplicate the ongoing efforts at the national and international level in the legalities and ethics of MHC. Rather, it was to learn from these ongoing discussions, apply a perspective to the problem squarely rooted in human factors and cognitive science understanding, and thus distil a set of practical human-centered guidelines to inform future NATO actions in this increasingly important area. Given the multi-faceted nature of MHC, six Themes were chosen for deep-dive investigation during this Workshop. Each Participant has been assigned to explore one of these Themes via small Theme-focused breakout sessions.

The Themes were:

1) HSI, Organizational, and Operational Considerations of MHC
2) Human Factors Inspired Design Guidelines to Achieve MHC
3) Systems Engineering Methods and Metrics to Validate MHC
4) Adversary Exploitation of MHC
5) Complex Socio-Technical Systems
6) Moral Responsibility in Human-AI Teams

The results of this Workshop can directly inform recommendation of highly focused follow-on activities that inform NATO on how to identify, achieve, maintain, and regain MHC across a wide range of AI applications. The workshop results can be summarized as a “Top 5” list of repeated concerns that had been mentioned repeatedly and which might warrant further investigation:

1) **Trust:** Both human-machine and human-human across organizational or system-of-systems boundaries. While imprecise, "trust" does capture a nexus of relationships, thought patterns, and considerations that are critical to successful human-AI teaming. Considerations within the broad topic include perceived performance (and factors which influence it), perceived utility and necessity, desirability of reliance, understanding of the strengths and weaknesses of both human and AI system, the broader socio-organizational dynamics that enter into reliance behaviors, and even genetic and psychological predispositions.
2) **Certification of Human-Machine Teams:** As a replacement for or augmentation to validation and verification of machine systems.

3) **Evaluation, Methods and Metrics:** Being able to assess the presence, absence and, ideally, the degree of MHC in various contexts and systems seems absolutely core, with most of the themes either contributing to, or requiring outputs from this topic.

4) **Awareness of Uncertainty (behavioral, contextual, outcome, etc.):** Similarly, since absolute knowledge of the outcome of a system design or commanded behavior is likely never to be possible, any MHC measurement or assessment approach will have to deal with uncertainty. Representing and conveying that to the user seems highly useful for MHC.

5) **Semantic Gap/Representational Mismatch:** the prospect of understanding and representing (and ideally identifying and predicting) semantic gap difficulties in organizations, between individuals and especially between humans and AI systems seems both like it is on the borders of feasibility and would go a long way toward minimizing misunderstandings which can lead to loss of effective MHC.

Séminaire sur le contrôle humain sensé des systèmes basés sur l’IA : rapport d’évaluation technique, perspectives thématiques et scénarios associés

(STO-MP-HFM-322)

Synthèse

Le contrôle humain sensé (MHC) est apparu en 2016 comme un concept important pendant les réunions de spécialistes organisées par la Convention sur certaines armes classiques (CCW) des Nations unies. Bien que le concept ait été relié aux armes autonomes, il peut s’appliquer plus généralement aux systèmes militaires basés sur l’IA (à la fois physiques et informationnels) en tant qu’exigence cruciale pour préserver le comportement moral, la responsabilité et l’enveloppe de performance opérationnelle efficace de ces systèmes.

L’objectif principal de ce séminaire n’était pas de dupliquer les travaux en cours au niveau national et international en matière de légalité et d’éthique du MHC. Il s’agissait plutôt d’apprendre de ces discussions en cours, d’adopter un point de vue sur ce problème profondément ancré dans les facteurs humains et dans la compréhension des sciences cognitives, puis d'extraire un ensemble de directives d’ordre pratique centrées sur l’humain pour éclairer les futures actions de l’OTAN dans ce domaine de plus en plus important. Étant donné la nature plurielle du MHC, six thèmes d’étude approfondie ont été choisis pour ce séminaire. Chaque participant a été chargé d’explorer l’un de ces thèmes par le biais de séances en petits groupes.

Les thèmes étaient :

1) Considérations organisationnelles, opérationnelles et de HSI en matière de MHC
2) Directives de conception inspirées par les facteurs humains pour parvenir au MHC
3) Méthodes et indicateurs d’ingénierie des systèmes pour valider le MHC
4) Exploitation du MHC par les adversaires
5) Systèmes sociotechniques complexes
6) Responsabilité morale au sein des équipes associant humains et IA

Les résultats de ce séminaire peuvent directement servir à recommander des activités de suivi extrêmement ciblées qui informeront l’OTAN sur la manière d’identifier, obtenir, maintenir et regagner un MHC dans un large éventail d’applications de l’IA. Les résultats du séminaire peuvent être résumés en une liste de cinq principales préoccupations, qui ont été mentionnées à plusieurs reprises et pourraient justifier des études plus poussées :

1) **Confiance :** À la fois entre l’humain et la machine et entre humains dans les limites de l’organisation ou du système de systèmes. Bien qu’imprécise, la « confiance » implique un ensemble de relations, par le biais de modèles, et de considérations qui sont essentielles à une association réussie entre l’humain et l’IA. Les aspects à considérer dans ce vaste sujet sont notamment les performances perçues (et les facteurs qui les influencent), l’utilité et la nécessité perçues, l’intérêt de la fiabilité, la compréhension des forces et des faiblesses du système humain et de l’IA, la dynamique socio-organisationnelle dans son ensemble qui entre en jeu dans les comportements de confiance et même les prédispositions génétiques et psychologiques.

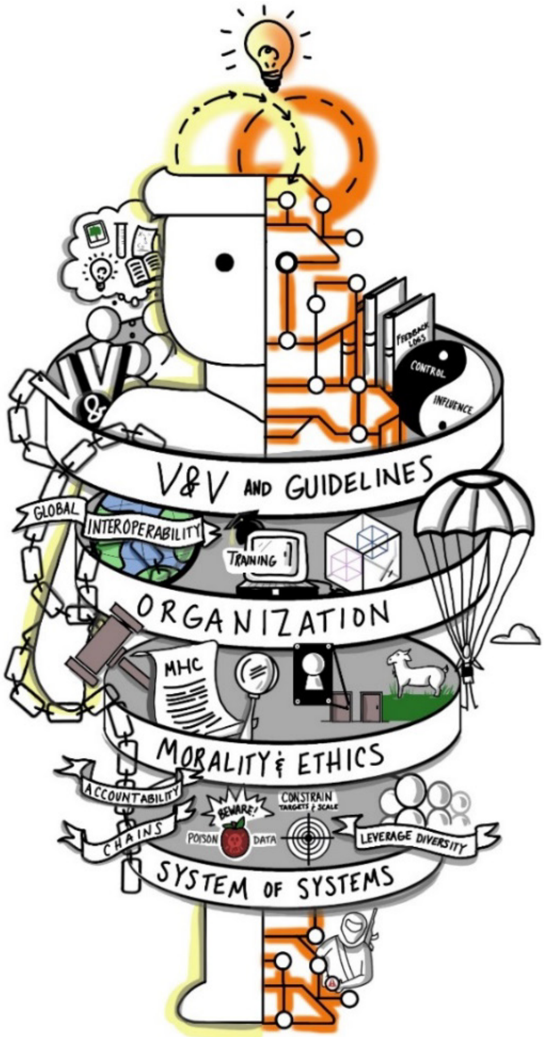
2) **Certification des équipes humains-machine :** En remplacement ou en augmentation de la validation et vérification des systèmes automatiques.

3) **Évaluation, méthodes et indicateurs :** La capacité à évaluer la présence, l’absence et, idéalement, le degré de MHC dans différents contextes et systèmes semble absolument essentielle, car la plupart des thèmes contribuent aux, ou ont besoin des, résultats de ce sujet.

4) **Sensibilisation à l’incertitude (comportementale, contextuelle, des résultats, etc.) :** De même, puisque la connaissance absolue du résultat d’une conception de système ou d’un comportement commandé n’est probablement jamais possible, toute démarche de mesure ou d’évaluation du MHC devra faire face à l’incertitude. Il semble extrêmement utile pour le MHC de faire comprendre cela à l’utilisateur.

5) **Fossé sémantique/inadéquation de représentation :** La compréhension et la représentation (et idéalement l’identification et la prédiction) des difficultés liées au fossé sémantique dans les organisations, entre les individus et en particulier entre les humains et les systèmes d’IA semblent prochainement possibles et contribueraient grandement à minimiser les malentendus susceptibles d’entraver l’efficacité du MHC.

MEANINGFUL HUMAN CONTROL OF AI-BASED SYSTEMS WORKSHOP: TECHNICAL EVALUATION REPORT



Workshop Date:	25 – 27 October, 2021
Workshop Location:	Berlin, Germany
Report Date:	23 December, 2021
Technical Evaluation Reporter:	Dr. Christopher A. Miller cmiller@sift.net 612.716.4015

1.0 INTRODUCTION

On 25 – 27 October 2021, a workshop was conducted at the Fraunhofer-Forum in Berlin, Germany, under the sponsorship of the NATO Science and Technology Office, Human Factors and Medicine Panel. This workshop, HFM-322 on "Meaningful Human Control (MHC) of AI-based Systems" was organized and chaired by Dr. Jurriaan van Diggelen of the Netherlands Organisation for Applied Scientific Research and Dr. Mark Draper of the U.S. Air Force Research Laboratory. Twelve core members of the panel were present, while this core group was joined by more than 20 guest speakers, organized into panels, who attended portions of the meeting virtually.

The author, Dr. Christopher A. Miller of Smart Information Flow Technologies in the U.S., served as the Technical Evaluation Reporter (TER) for this meeting. This document is a summary of his comments to the workshop's programme committee delivered on the final day after observing the Workshop as a whole.

2.0 COMMENTS ON GENERAL WORKSHOP STRUCTURE AND ACCOMPLISHMENTS

Overall, I was extremely impressed with the organization and conduct of this workshop and with the dedication and energy of those involved. If anyone had asked me ahead of time, I would have said (based on my nearly 30 years of experience participating in NATO RTO and AGARD panels and workshops) that attempting a workshop involving six major themes with more than 20 panelists and three keynote speakers over 3.5 days, in a mixed physical and virtual presence format, all in the midst of a pandemic, was highly risky. (In fact, I think I said "crazy"). Still, the format worked and a large amount of work was done and substantial knowledge was exchanged. This is thanks in part to the dedication of the panelists, and also to the preparations an organization of our hosts at the Fraunhofer-Forum. Particular credit is due to Marcel Baltzer and Frank Flemisch who did the majority of the preparation for the workshop.

2.1 Physical Structure: Hybrid Virtual/Physical

The workshop was structured, in part due to the constraints imposed by the pandemic, as a dual virtual and in-person track. Core members of HFM-322 (twelve in all) attended in person, while the core committee invited more than 20 "panelist" experts to attend, present and discuss their work and perspectives via virtual connections. This fairly innovative arrangement turned out to work remarkably well, in my opinion. Workshops such as this always struggle with competing agendas: they want and need to establish a good working relationship among the core group who will likely continue working together on the topic for several years, but at the same time they also want to broadly survey world-class expertise and opinion in relevant fields. The first objective is aided by physical presence, shared break times and meals, hallway discussions, etc., but those goals can actually be diminished if the group is too large. The second objective is aided by being able to attract a larger group of experts, though this objective can be undermined by requiring such experts travel and commit to a multi-day workshop in a remote location. The hybrid structure of this workshop addressed both goals remarkably well. The twelve core HFM-322 members met together physically and largely shared a hotel and meals together, and team building seems to have been encouraged as a result. On the other hand, the virtual presence of a larger group of remote experts both made it far easier for them to commit to an hour or two interaction with the core team and, in some sense, enhanced the sense of team formation which occurred for the core group since they shared the experience of interacting with the full set of virtual participants. There is little doubt that HFM-322 was able to attract speakers who would not have been willing to commit to the group if their involvement had been longer or more expensive due to travel. It is worth considering this hybrid workshop arrangement for other such meetings in the future. One downside, however, is that the virtual panelists were not able to attend and participate in panel sessions other than their own. This led to some redundancy in topics covered, some variability in vocabulary used and, potentially, to missing some interesting cross-group, cross-panelist interactions.

2.2 Organizational Structure: Three Keynotes, Six Themes with “Power Panels”

The HFM-322 workshop was organized around six different themes (which will be discussed separately below). Each of the themes held a “power panel” of three to four experts (not members of the core group) relevant to that theme who had been selected and invited by the theme chair, who was a core group member. Another difference from traditional panels of expert presenters was that panelists were not invited to present a paper or even give much of a summary of their prior work. Instead, the theme chair had prepared a set of questions for the panel members to discuss and led them through the discussion while trying to keep the group on topic and moving briskly.

By and large, this approach also worked well. The fact that panelists were asked to discuss questions posed by the core group’s theme lead meant that their comments tended to be more focused on the interests and needs of the core group than might have been the case if the experts had been allowed to present their latest work or discuss some issues they thought, but which (given their lesser experience with the topic of MHC as defined by the core team) might have been off base for the group’s needs.

It is, inevitably, a difficult task to steer conversation and keep all participants focused while also allowing information to flow freely and in directions that, given the expertise of the panel members, might not be anticipatable. Some of the panel sessions seemed more productive than others, while it seemed like we could easily have spent two or three times the amount of time with some of the panel topics. Nevertheless, I believe this organization was more fruitful for the needs of the core working group than many traditional panels or paper presentations I have participated in, and it should be considered for other workshops in the future.

2.3 Keynote Speakers

The other structural element of the workshop agenda was the inclusion of three keynote speakers: Major General Gäbelein of the German Armed Forces, Dr. Missy Cummings of Duke University, and Dr. Daniele Amoroso of the Department of Law of the University of Cagliari. These three were well chosen to represent different portions of the community of relevant experts for the workshop’s MHC topic: Major General Gäbelein providing an armed forces perspective, Dr. Cummings providing and academic, engineering, human factors and computer science perspective and Dr. Amoroso providing a legal and ethical perspective.

All Keynote speakers presented virtually and their talks were informative and well received, but structurally there was nothing unusual or innovative about the format of their talks. They were distributed throughout the workshop presenting on days 1, 2 and 3. Comments on the content of their presentations is provided below.

2.4 Cartoonist

A final structural innovation for this workshop was the presence of a “cartoonist” – Ms. Jennie Hempstead, the Director of Marketing and Communications at the Wright Brothers Institute – who was recommended and supported by the U.S. Air Force Research Laboratory. Ms. Hempstead provided illustrations of the concepts, themes, and discussions that went on throughout the meeting. These were not sketches of the participants so much as of the emerging concepts and debates (and their relationships) that were discussed – see illustration in Figure 1.

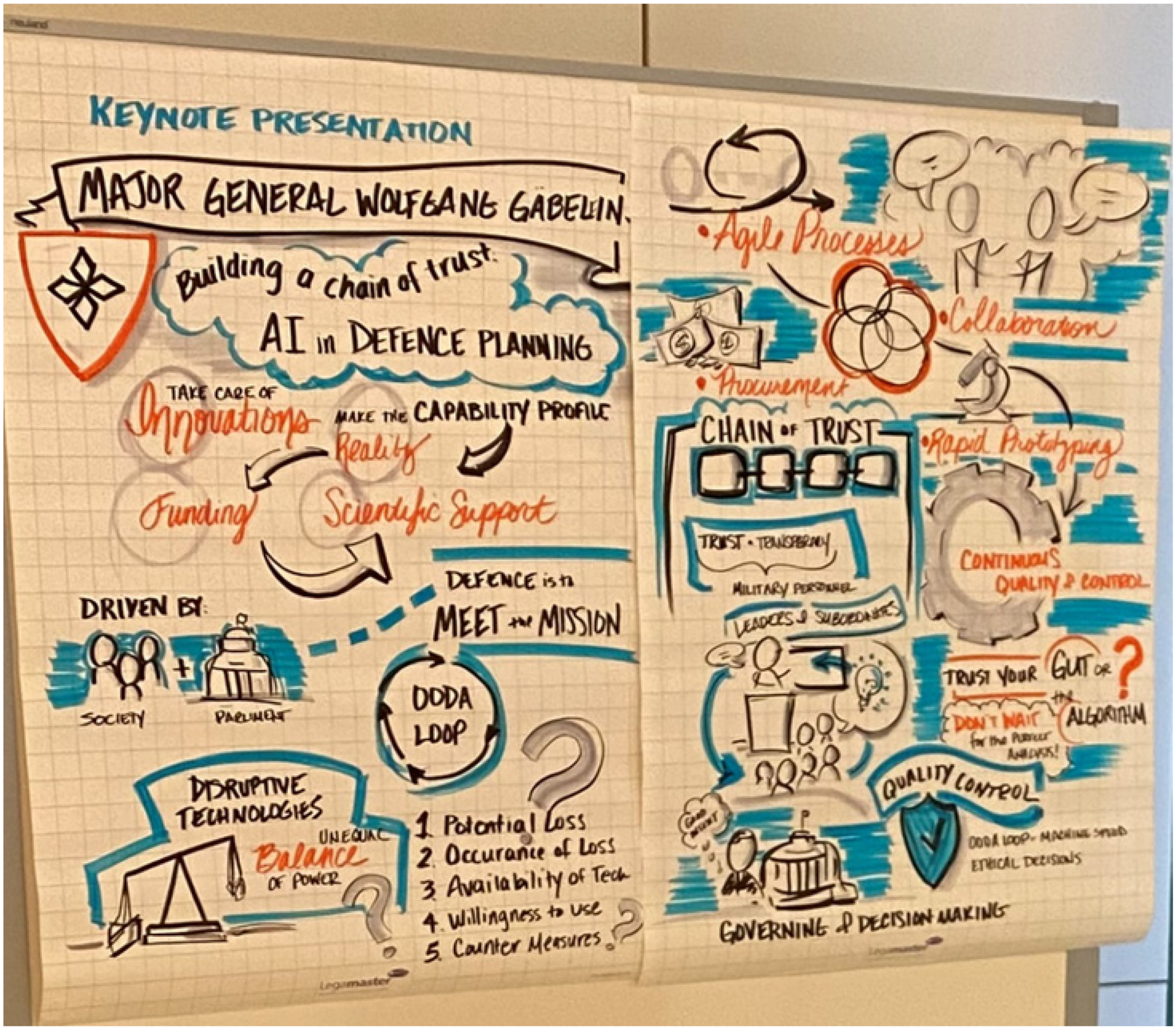


Figure 1: Cartoonist Jenny Hempstead's Depiction of the Themes from the First Keynote Address.

As such, they served much the same function as notetakers (who were also present and engaged) do but did so in a much more compelling fashion. After each workshop day, Ms. Hempstead would present her sketches from the day for discussion and review by the participants. This served much the same function as "reviewing the minutes" of a meeting, but perhaps due to the visual medium (or due to its novelty), the group was much more actively engaged in reviewing her sketches and commenting and improving them. This provided a cognitive reinforcement of topics of the day which, likely, will increase memory for and consideration of the information exchanged.

3.0 SPECIFIC COMMENTS ON KEYNOTES AND PANEL TECHNICAL CONTENT

3.1 Objective and Approach

3.1.1 Overall Workshop Objectives

The overall objective of HFM-322 was, as reported in the workshop announcement, "to learn from these ongoing discussions [in other fields], apply a perspective to the problem squarely rooted in human factors and cognitive science understanding, and thus distil a set of practical human-centered guidelines to inform future NATO actions in this increasingly important area."

Dr. Mark Draper, in his opening comments to the group, characterized the objective as "to avoid competing with other MHC activities going on and assess the HF perspective on how to obtain, maintain, retain and regain MHC."

It was with these objectives in mind that I listened and attempted to distil conversation and presentations from the workshop.

3.1.2 My Objectives and Approach as TER

As the TER for this workshop, my approach was not so much to provide an overall summary of the meeting, much less a detailed report (as in a set of notes) of what everyone had said. This function was well captured by multiple note takers (including myself) who provided their documents to the core team. It was also supported in an innovative fashion by the cartoonist, Jenny Hempstead, as described above.

Instead, my approach was to simply report what I noted as general themes that emerged in discussion, and what caught my interest in the presentations and debates that ensued as new ideas emerged. I also ventured so far as to note specific ideas, areas of focus or approaches that seemed to me to offer promise for the overall objectives of the topic area.

My comments were presented initially to the core team as a final briefing during the workshop. I will reiterate those comments here as I presented them there: first by providing summary comments about each keynote address, then by providing a brief report on what I thought the main themes and my personal advice for each of the power panels and their associated themes. Finally, I will provide some overarching thoughts on the MHC topic as a whole, as informed by what I learned at this HFM-322 workshop.

3.2 Keynotes

3.2.1 Major General Wolfgang Gäbelein

The first keynote address was provided on 25 October by Major General Wolfgang Gäbelein, Director-General, Bundeswehr Office for Defence Planning and was titled "Building on a Chain of Trust: AI in Defence Planning". He spoke about the society-wide set of trust dependencies that already exist in how countries use, interact with and control their defence services, and how AI systems must integrate into that chain. He discussed the motivation by society and by the armed forces to make use of AI, especially when an enemy is doing so. He talked about the need to roll out AI as a process which, itself, participates in the established socio-cultural methods and structures whereby: engineers and scientists design and develop technology, politicians and planners allocate budgets, military logisticians and evaluators make procurement decisions, military planners and commanders make usage decisions, etc. Each step represents a link in the chain of trust.

Some themes and recommendations that seemed important and useful to me were:

- The tools which emerge from the "chain of trust" must follow the links in the chain and thus be designed with users and usages in mind.
- There is a motivation (both practical and ethical) to start with weak and minimal AI. General Gäbelein drew two lessons from this:
  - We should not, and perhaps cannot, "analyze to death" before trying something with AI. We need to gain experience with it to understand what does and doesn't work, just as with any tool.
  - But that means trying "minimal AI" in initial, small steps in order to gain that experience across the chain. Learning by doing is more useful than extensive analysis, but there is a need to ensure that the learning is done in constrained applications.
- That learning will be more effective if it is collaborative if we "pair up" in partnerships both between allied countries and between users and developers.

3.2.2 Dr. Missy Cummings

The second keynote was provided by Dr. Missy Cummings of Duke University under the title "Meaningful Human Certification vs. Meaningful Human Control." She focused on the limitations of current (and likely future) AI systems. She provided a collection of arguments about the failings of AI systems especially in the sense of understanding context, causal relations and "top-down" reasoning for sensemaking. All of these factors, she claimed, demand that humans remain in a sensemaking role... but that in turn demands that it not be so much the AI system that be evaluated and "controlled" as ascertaining when some combination of human operators and machine capabilities and world states affords "good enough" MHC for a given scenario and application.

Some themes and recommendations that seemed important and useful to me were:

- The identification of top-down sensemaking as the primary lapse in existing and at least near future AI systems. Humans are needed precisely to enable the use of an AI system to take more of the global context and implications into account.
- Dr. Cummings said "There is no MHC at the tactical level" by which I understood her to mean that, with automation, there is always a level below which the human will not be able to intervene and, thus, the human ability to know when and in what contexts to invoke the automation is critical.
- The notion that instead of system validation and verification, the focus should be on human + system "certification" – is the human-machine system likely to be "good enough," in the context of use, to be relied upon?
- The claim that, in order to support such certification, we should focus on evaluation techniques, and perhaps even training and user interfaces, to give operators, commanders, etc. knowledge of the sensitivity of their human-machine systems to variations in the contexts they may encounter. These tools include approaches like sensitivity analyses, biases in algorithm development/data, coverage, etc. Better awareness of the vulnerabilities of the AI and the AI-human team to strengths and weaknesses of this type will produce more acceptable behaviors.

3.2.3 Dr. Daniele Amoroso

The final keynote was provided on the final day of the workshop by Dr. Daniele Amoroso, a Professor of International Law at the Department of Law of the University of Cagliari and, since 2017, a member of the

International Committee for Robot Arms Control (ICRAC). Dr. Amoroso's title was "Meaningful Human Control (MHC) over Weapons Systems: a Normative Approach" and his focus was more on the ethics and legality of using autonomous weapon systems. His framing of the problem was that MHC was more of a "normative" problem and not a technical one. He provided a variety of vocabulary terms including the differentiation between principled vs. prudential MHC – by which I understood him to be referring to a well-defined and worked out understanding of all possible decisions and outcomes and a "principled" understanding and control of how the AI would behave in them vs. a "prudential" approach that acknowledges we won't (possibly ever) have that degree of understanding and instead should strive for a cautious approach likely to provide acceptable behaviors in many circumstances and which refuses to use, or denies AI the opportunity to act in other circumstances.

Some themes and recommendations that seemed important and useful to me were:

- The idea that there is (and may always be) a "semantic gap" between the way humans perceive events, decisions, actions, etc. and the way machines do. This seems to be, roughly, a kind of broader set of associations – humans understand the "meaning" of an image, rather than simply matching its pixels to a pattern (approximately what Dr. Cummings alluded to as AI failing to "understand context").
- The fact that this "epistemic uncertainty" in AI reasoning pushes us toward prudential solutions. And specifically, within prudential solutions, the labelling of different types of autonomy constraints (with their associate drawbacks):
  - Denied autonomy – but some autonomy, in some conditions, are safer and more ethical than humans;
  - Boxed Autonomy – but not all autonomy behaviors can be predicted, much less in ML systems;
  - Supervised Autonomy – but humans have automation bias. Institutions might encourage it.
- The identification of principles of "precaution, distinction and proportionality" with AI system use. These yielded Dr. Amoroso's overall recommendation that humans retain targeting roles and responsibility, with possible exceptions including scenarios without civilians and pre-planned targeting.

3.3 Power Panels

There were a total of six workshop themes and each theme hosted a "power panel" consisting of 3 – 4 experts in areas related to the theme. Panels provided, a priori, a stated theme and thus, in the workshop, largely consisted of a series of questions posed by the panel lead and/or the audience of core team members to the panelists. There was a presumption that at least some of these panels will become the focus of future meetings by the core team. Again, detailed notes were captured for each panel so I will not attempt to provide complete summaries here. Instead, I will note the objective of each panel, report on my observations of significant and/or core themes within the panel discussion, and conclude with some brief advice, from my personal perspective, for future pursuits concerning this theme.

3.3.1 Theme 1 – HSI, Organizational and Operational Considerations

Objective

"[Identify] the organizational conditions that might influence MHC, by aiming to understand how we can organize and manage to promote MHC." Focus on the socio-organizational context in which MHC is both afforded and assessed."

Significant Observations

- MHC is not a term of International Law; We should perhaps move away from the notion of “control” to “critical decisions under human awareness”– but this may not make the problem easier since we know that decision advising and sense making automation is even more problematic for assessing and providing MHC.
- Command and Control (C2) Agility– not one-size-fits-all. Match the organizational structure to the situational context. It may be both possible and necessary to assess MHC by defining regions in the space of organizational structure (authority and communication relationships, etc.), context of deployment and user decision authority in which MHC is strong, medium and weak – and then perhaps to track and report where individual decisions lie within that space over time.
- Sensemaking is culturally managed and “Boundary Managers/Brokers” might help traverse cultural boundaries.
- Parachute Packers anecdote – trust can (sometimes) be placed in the machine developer/operator rather than the machine.

Advice?

Focus on something like “Organizational Patterns” within C2 Agility Framework – and dimensions of variability and measurements. The notion of being able to assess and then track where in a C2 “space” a given decision lies and whether or not that region represents strong, weak or non-MCH seems powerful and worth exploring in more depth. Of course, defining relevant dimensions and then assessing and tracking them will be non-trivial.

3.3.2 Theme 2 – HF Inspired Design Guidelines

Objective

“...key question here is: how can we develop design guidelines that ... would ensure MHC in critical environments associated with future NATO operations?” This panel’s focus was on general heuristics and methodologies for achieving MHC in human-automation interaction designs. While evaluation methodologies and metrics were not an explicitly included aspect, the discussion frequently got near to those topics.

Significant Observations

- Users of automation and AI should want a “chatty co-pilot rather than a silent autopilot” – that is, the automation should provide ongoing interaction, (guided) explanation and should be able to ask for and take instruction. This seemed generally desirable even at the cost of more human interaction time and attention and even at the cost of reductions in automation competence and accuracy. Highly accurate AI systems may not be trusted, while less than perfect systems can still be useful if they can take instruction and correction.
- The panel made extensive use of the “AI as teammate” metaphor– and discussed the pros and cons of this usage. There are multiple “individuals” working “together,” but they have different abilities. Teaming may just be a useful metaphor and it’s worth keeping in mind that it is a metaphor and therefore will be accurate in some regards and inaccurate in others.
- The discussion of shared goals and teaming led to a discussion of “accountability” with the claim that “Accountability is always with an individual if you break the tasks down far enough.” I find this reasonable and a guide to assessing responsibility, but at the same time, there is a designer or planner or commander who is assigning a set of subtasks to each individual. Some accountability resides at that point.

- Another, perhaps core, tradeoff seemed to grow out of the idea of multiple team agents. It is valuable to have independence of assessors of the world states and independence of planners for reacting to it for creativity, but it is also valuable for a team to have "shared mental models" to coordinate teamwork and distribution of labor. Finding a sweet spot between these is part of team design and training, but it will be much harder for AI teammates to participate since their "mental models" are more diverse and un-human-like.

- This may, in turn, imply a need for the team (and the human-machine relationship) to be able to evolve over time. Managing that teamwork, the interaction structures required to support it, and providing measures for when individual actors within it have MHC might be a worthwhile focus.

- It might, in the end, be more useful to know when an individual or a team doesn't have MHC – especially if this could be forecast or detected and warned.

Advice?

This was an extremely productive panel in that many core ideas/principles were surfaced and discussed. Most were valuable and critical... but I also had the impression that there was not much new or unique to AI here; most of the issues were familiar from human-automation interaction research over the past decade or more. (Though a counter argument is that issues of teaming – human-human and human-automation – were very familiar to most of the workshop participants, so the lack of uniqueness or novelty might be a symptom of deeper familiarity). Alternatively, a core problem with the "teaming" metaphor may be that it pushes us toward accentuating relationships we already are deeply familiar with human-human teaming. This is in no way wrong, and human teaming structures and dynamics (and the ethics, accountability and authority relationships within various approaches to it) are likely to serve as a good (but not perfect) guide to human-machine teaming and MHC within it. A focus on measures for MHC support in design and/or fields where assessing analogs to MHC (like legal theory around culpability) might also prove productive. Also, I found the "Fundamental Tradeoff" between sharing mental models and yet using and assessing them differently for different perspectives to be interesting and, I suspect, productive to investigate further.

3.3.3 Theme 3 – System Engineering Methods and Metrics

Objective

"The theme will aim to gather ideas from participants, map and prioritize issues and identify clear and usable methods." The emphasis in this panel was on the design and development phase and the processes which would help to ensure MHC rather than (as in panel 2) on the conditions and methodologies that would ensure MHC during operations.

Significant Observations

- Discussion ranged across various means of assessing MHC in systems design: formal methods vs. simulation vs. test. The general consensus was that all three were probably needed for coverage and accuracy.

  A bit of conventional wisdom was repeated by the panelists: If there are no metrics, then you can't have requirements. This seems overstated, but generally in the right direction. Having a valid metric certainly helps to ensure whether a requirement has been met.

- The panel advanced the idea of a "Moral Hazard Analysis" as akin (whether metaphorically or more concretely) to a "Failure Effects Analysis" – an articulation of the "moral hazards" that a design or usage might fall into, a tracing of the routes and conditions under which the hazards might be realized and an assessment of the probabilities of each.
- It was repeatedly noted that we will almost inevitably be talking about qualified risk rather than guaranteed safety regarding MHC.
- Similarly, assessment of designs will likely need to be continuous (e.g., continuous V&V) – because even in the absence of machine learning, or engineering redesign, the human users will continue to learn and adapt to the machine capabilities. When operators, organizations, machines and the enemy are all adapting as well, the assessment process becomes correspondingly more difficult.

Advice?

There were many good ideas advanced in this panel, but the "prioritization" goal of the panel could stand some more thought. To my mind, the concept of a Moral Hazard Analysis (and an associated ontology to characterize potential moral hazards) seemed like a particularly productive place to start.

3.3.4 Theme 4 – Adversary Exploitation of MHC

Objective

"The purpose of this theme is to explore which challenges are associated with ensuring that MHC is maintained despite deliberate adversarial interference." This panel took as its charter the role of "red teaming" the other panels – that is, thinking about how the design, maintenance and use of MHC in military systems could be exploited by an adversary.

Observations

- Adversaries generally have the goal of maximizing and manipulating uncertainty for their opponents. This implies that awareness of uncertainty is a key to overcoming and avoiding exploitation. Highly reliable systems tend to be trusted more and, therefore, inspected and cross-checked less. This makes a valuable vector for enemy exploitation.
- Similarly, it will frequently be more productive for the enemy to disrupt the early stages of an OODA loop (pre-Action) as these will have more pervasive and, frequently, more disadvantageous consequences.
- The panel generally downplayed physical/network control risk as less significant or important for consideration, manageable through existing/traditional approaches.
- While there is some perception that AI systems and MHC may both result in more predictable behaviors which enemies could exploit, humans can be more predictable, as a wide range of machine learning applied to predicting human behaviors have shown recently.
- The panel described the idea of "AI Medics" – AI systems designed to diagnose and repair (or remove) AI systems that may have been infected or otherwise behaving poorly. But such systems may undermine MHC (by providing software updates that humans aren't aware of and/or don't understand), and/or leave NO ONE in control.

- The notion of adversary behaviors has largely been considered from the perspective of operator-system interaction dyads, but we're becoming aware (see theme 5) that trust, policies, rules of engagement, etc. are all organizational-level phenomena. Therefore, perhaps this notion of adversary exploitation of MHC needs to be considered at an organizational level as well, which may well mandate a larger/longer time scale of consideration. Some things that look like near-term, tactical failures may be longer-term, strategic successes if, for example, they permit learning about enemy behaviors winning the "culture" or "information" wars.

Advice?

This was a very Interesting and Important topic, but if there needs to be some focus and prioritization in future discussions, it also seems very separable from "core" aspects of MHC. It also may be somewhat inherently subsequent to understanding what MHC is and how to achieve it. Within the topic, though, the ideas of uncertainty state and awareness seemed central and amenable to practical steps toward design and implementation. Displays or training that illustrate what the system has been tested on, or knows it is valid for are at least plausible. Also, the concept of AI Medics (and human reaction and use of them) seemed important and useful for study.

3.3.5 Theme 5 – MHC in Complex Socio-Technical Systems

Objective

"The purpose of this theme is to explore the challenges associated with ensuring that MHC is maintained within the more complex, interconnected and interdependent system-of-systems." The focus of this panel was the complexities that arise when considering MHC within broad, collaborative systems of systems (both human and machine) such as military operations.

Observations

- The representatives in this panel were mostly involved in multi-disciplinary and multi-national teams dealing with MHC or related topics. Their own experiences in this space were telling: they fairly universally mentioned vocabulary and assumption differences among the participants in their teams about terms such as "autonomy", "norm" and "human control". Simply providing (agreed on) definitions of these terms would be a help.

- Loss of "uncertainty awareness" in organizational info flow seems to be a highly likely problem. Smaller organizations and, certainly, individuals retain awareness of what they don't know or are uncertain about, but this is frequently lost in larger teams where a tentative conclusion can be perceived as definitive by downstream consumers. AI processing, which has its own problems with representing and propagating uncertainty, will likely exacerbate this.

- AI as accelerant, flash crashes, rapid change of behavior – Automation typically enhances the speed and throughput of information while enabling more tasks to be performed, but this means that when things go wrong, they tend to go wrong faster and the resulting "crashes" involve a bigger pile up.

- Cross organization AND cross human-AI representational mismatch – exacerbated when cross-national and/or cross-cultural.

- Predicting the effects of weapon use is a legal requirement for use. But what are the standards and practices for predicting? "Black swan" events are probably more common the more complex the technology and the organizational use of that technology.

Advice?

This was generally felt to be a very important theme with very few other sources/venues seriously examining it. It is also very complex and "large" in what it embraces. There is also a sense in which it is probably analytically "downstream" – by which I mean that adequately addressing and studying it will require (or at least benefit from) prior analyses, definitions and considerations represented by most of the other themes – for example, understanding trust propagation within a complex system-of system probably relies on understanding trust dynamics more locally to begin with. Therefore, I am inclined to recommend that this theme be tabled or postponed until further progress on other themes is made. On the other hand, if and when this theme is tackled, topics surrounding organizational propagation of trust, uncertainty and meaning seem central places to focus.

3.3.6 Theme 6 – Moral Responsibility for Decisions

Objective

"This theme contributes to the workshop by ensuring that our proposed solution is ethically sound and narratively robust…getting it right and being able to explain it…" The focus of this theme was the philosophical, ethical and legal reasoning that underlie both MHC and whether or not the resulting behavior if moral/ethical or not.

Observations

- "Closing the semantic gap" – that is, the ability to express intent in a "language" that is understood similarly by all participants. This is especially difficult for human-AI interactions around, for example, value or priority statements where the mathematical expressions useful for AI may not capture the intent (and hesitations) of the human user. But this is also a difficulty between multiple human participants (particularly system designers and end users). Furthermore, human intent expression can also convey SA and uncertainty, which machines largely can't express and don't understand or capture currently.

- But maybe we can't close it. There was an argument that humans can't close the semantic gap with other humans... so they don't. Instead, they maintain continuous interactions with multiple opportunities to identify and trap errors and mismatches.

- Narrative considerations – in sensemaking, in awareness, in data reporting, in "ethical focus." There were many ways this claim seems relevant. One is the claim (accepted by most of the panel) that strict, rule-based reasoning will never capture everything we need to capture about ethics. One reason is that laws (especially international laws) are frequently so vaguely stated that they must be interpreted in context. Therefore, the "narrative" account of that context that the interpreter uses becomes important to the decisions ultimately made. Disputes over behavior are frequently disputes about who's narrative of the situation is to be accepted. And examining multiple narratives for a situation is a particularly good way to understand different interpretations, weightings, cultural considerations, etc.

- There was general consensus that machines cannot be ethically responsible, but there are many humans in most chains – forming the "many hands" problem of allocating responsibility. This is a candidate for the focus that MHC should have – solving the many hands problem and determining ways to ensure that responsible parties understand they are responsible and use those tools in an ethical manner. AI makes the many hands problem worse by increasing the throughput and tempo of information and decisions, while obscuring some aspects of who is responsible for what when.

- There was, however, a tension over the claim that ethics can't be rule-based. This tension was driven by the claim that humans can do "Ethics at Speed" – that is, very fast yet ethical decision making by being trained a priori in situations and (ethical) reactions. But that seemed to be rule-based ethics, albeit performed by a human arguably too fast for deep interpretation.
- There seemed to be general consensus around the claim that it is impossible not to make an ethical decision in many circumstances; doing nothing is doing something and will have ethical implications. Though there was a distinction drawn in that existing systems certainly take actions that have ethical implications even though they themselves are not doing ethical reasoning. Instead, in those cases, the designers are making ethical decisions.

Advice?

This was a fascinating panel, and a core for the concept, assessment and creation of MHC, but it remained very theoretical and not applied. For the stated purposes of HFM-322 and HFM-330, it might be useful to try to move it in more practical and applied directions. One such might be to move toward measures and metrics. For example, measuring (even probabilistically) the existence and magnitude of a semantic gap seems within the bounds of plausibility. Similarly, for capturing and measuring variations in "narratives" or explanations of an ethical decision, and even projecting future areas those variations might come into play for, say, international teams of actors also seems plausible given current advances in Natural Language analytics.

4.0 REPEATED, CROSS-THEME CONCERNS

While the panels and themes raised a wide and highly diverse set of topics and considerations, I ventured to advance a "Top 5" list of repeated concerns that, I felt, had been mentioned repeatedly and which, due to their centrality and importance within the different themes, might warrant further investigation. There was no more "scientific" or procedural rigor to these assessments. My Top 5 list was, in order of my perceived significance:

1) **Trust** – both human-machine and human-human across organizational or system-of-systems boundaries. Trust is, admittedly, a loosely-defined and overused term and, thus, even though it was mentioned by multiple panelists, their meanings, focus and implications could be widely different. While imprecise, "trust" does capture a nexus of relationships, thought patterns, and considerations that are critical to successful human-AI teaming and, as such, listing "trust" as a top concern may be messy, but it does cover a lot of important ground. Considerations within the broad topic include perceived performance (and factors which influence it), perceived utility and necessity, desirability of reliance, understanding of the strengths and weaknesses of both human and AI system, the broader socio-organizational dynamics that enter into reliance behaviors, and even genetic and psychological predispositions. All of these are useful topics for investigation. At any rate, the goal should clearly be accurate calibration of trust along with human-AI teaming structures and policies that encourage reliance behaviors that produce ethical outcomes.
2) **Certification of Human-Machine teams** – as a replacement for or augmentation to validation and verification of machine systems.
3) **Evaluation, Methods and Metrics** – Being able to assess the presence, absence and, ideally, the degree of MHC in various contexts and systems seems absolutely core, with most of the themes either contributing to, or requiring outputs from this topic.
4) **Awareness of Uncertainty (behavioral, contextual, outcome, etc.)** – Similarly, since absolute knowledge of the outcome of a system design or commanded behavior is likely never to be possible, any MHC measurement or assessment approach will have to deal with uncertainty. Representing and conveying that to the user seems highly useful for MHC.

5) **Semantic Gap/Representational Mismatch** – the prospect of understanding and representing (and ideally identifying and predicting) semantic gap difficulties in organizations, between individuals and especially between humans and AI systems seems both like it is on the borders of feasibility and would go a long way toward minimizing misunderstandings which can lead to loss of effective MHC.

5.0 FINAL THOUGHTS

I found the workshop successful and very interesting at surfacing a very impressive range of considerations in a short period of time. It was admirably focused on the topic of MHC, but it will need to refine that focus still further if progress is to be made in the future. I have provided some suggestions above One way forward is apparent from the workshop objective as stated by Dr. Draper. If the goal is to "obtain, maintain, retain and regain MHC" (and to know when you have done so), the following items will be necessary:

1) A definition (or maybe description) of what MHC is;
2) A test (operationalizing the definition) to determine whether or not it exists in a given context, albeit perhaps a retrospective one;
3) A process (for creating it), and/or identified best practices for applying that process so as to know a priori whether and when the conditions are right for MHC (or risky for losing it); and
4) Metrics to know how well you're succeeding in other than the binary case suggested by item 2 above.

Of these, the core team largely has the first item. It can (of course) be debated further, but it's a reasonable starting point. What remains is to provide the remaining three items on the list.

I will articulate one concept that occurred to me during the workshop on the basis of the themes and considerations which were being discussed. I offer it here as one suggested topic of concentration and a path forward: Neglect Tolerance (NT; Crandall and Goodrich 2002). NT was introduced in 2001 by Goodrich, Olsen and Crandall (2001) as a means of quantitatively characterizing the degree of autonomy of a system and it was later integrated into a set of metrics for characterizing human-robot interactions (Olsen and Goodrich 2003). The core notion is that the longer a machine could be left unattended (that is, its "neglect tolerance") in a given context of task, world state, etc., the more autonomous it could be said to be. While there have been various formulations of NT over time and various factors shown to affect it (Wang and Lewis, 2007; Elara, 2011; Elara et al., 2009) the basic formulation (as explained in Olsen and Goodrich, 2003) is that for some measure of overall system effectiveness (e.g., successful task performance or, in a reformulation for MHC, the likelihood of ethical behavior from the human and machine system) the system has some expected value against that metric with human oversight – generally assumed to be higher than without human attention. When human oversight is suspended, that performance is presumed to degrade probabilistically according to some neglect curve unique to the specific task, machine and world context. Furthermore, there is a presumed, agreed-upon minimal threshold, defined on the effectiveness dimension, below which system behavior is no longer acceptable and human intervention is again required to bring it back above that threshold. The temporal interval during which the system can be ignored ("neglected") and in which acceptable behavior is likely to prevail is Olsen and Goodrich's (2003) measure of Neglect Tolerance. More "autonomous" systems would have longer neglect tolerance windows – which means, longer windows in which their performance could be assumed to be acceptable without human intervention.

It would seem that a similar concept could be defined for AI-based systems to identify the degree to which they are likely to behave in an ethically responsible fashion autonomously in context – an **Ethical Neglect Tolerance** (ENT) interval. To compute an ENT score, we would have to define and agree upon a set of ethical hazard states such as "killing a non-combatant," "responding with disproportional force," etc. and then be able, through analysis, computation, prediction, observation or other methods, to provide numerical estimates of the likelihood of the system's transgressing into those states in a context of use, both with and without human oversight. Obviously, these estimates can be more or less accurate given more study, more precise definitions, more constrained contexts, etc. While time was the primary axis of increased probability of failure in the original NT concept, a measure of "context variance" from a predicted or assumed context might be more useful for ENT.

I feel that a concept like this has the prospect of bringing together many of the themes of the workshop: trust, awareness of uncertainty, moral hazard analysis, narratives and even semantic gap analyses. Furthermore, if they could be articulated and arranged in such a way to provide a concrete metric for measuring and predicting MHC, not only would this provide the other three elements in my "requirements" list above, but it would also prove a very practical and useful outcome.

6.0 REFERENCES

Crandall, J.W., and Goodrich, M.A. (2002). Characterizing Efficiency of Human Robot Interaction: A Case Study of Shared-Control Teleoperation. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2, 1290-1295. Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi.org 10.1109/IRDS.2002.1043932.

Elara, M.R. (2011). Validating Extended Neglect Tolerance Model for Search and Rescue Missions Involving Multi-Robot Teams. In Proceedings of International Conference on Intelligent Unmanned Systems, 7.

Elara, M.R., Calderon, C.A.A., Zhou, C., and Wijesoma, W.S. (2009). Validating Extended Neglect Tolerance Model for Humanoid Soccer Robotic Tasks with Varying Complexities. Performancemetrics, 1-8.

Goodrich, M.A., Olsen, D.R., Crandall, J.W., and Palmer, T.J. (2001). Experiments in Adjustable Autonomy. In Proceedings of IJCAI Workshop on Autonomy, Delegation and Control: Interacting with Intelligent Agents, 1624-1629. Menlo Park, CA: American Association for Artificial Intelligence.

Olsen, D.R. and Goodrich, M.A. (2003). Metrics for Evaluating Human-Robot Interactions. In Proceedings of 2003 Performance Metrics for Intelligent Systems (PerMIS'03) Workshop, 4-12. Gaithersburg, MD: National Institute of Standards and Technology.

Wang, J., and Lewis, M. (2007). Assessing Coordination Overhead in Control of Robot Teams. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2645-2649. Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi: 10.1109/ICSMC.2007.4414055.

Annex A – THEMED SYNOPSES

A.1 THEME 1: HSI, ORGANIZATIONAL AND OPERATIONAL CONSIDERATIONS OF MEANINGFUL HUMAN CONTROL

A.1.1 Description

Meaningful Human Control (MHC) of Autonomous Systems (AS) is not only dependent on the design of the AS or the characteristics of the individuals working with them. MNC is also influenced – positively or negatively – by the organizational or operational environment in which the AS is used. This theme will focus on the organizational conditions that might influence MHC, by aiming to understand how we can organize and manage humans to promote MHC. These conditions include:

1) **Organizational Capability and Readiness Practices.** This includes the development of human resource management practices that ensure that individuals working with AS have the skills and experience to exert or manage MHC. This includes the understanding of the system and information around it as well as situational understanding to allow individuals to predict how the context will affect the system and thus be able to make appropriate decisions regarding MHC. These practices will include training and development but also those relating to aspects such as workload and work design.

2) **Organizational Design.** This considers the design of the tasks and/or roles that include AS and the coordination between roles in order to facilitate MHC. Specifically, this might include the systems for organizational control and oversight of the AS and the wider socio-technical system around it. It might include the accountability of both those in command and lower down the hierarchy.

3) **Organizational Culture.** Organizational culture is a potentially a key driver of MHC. This includes aspects such as the degree of trust in the culture and the role of empowerment and autonomy. Particularly important might be aspects relating to freedom of choice and psychological safety. It is important that the organizational culture is such that selecting a course of action other than one recommended by the system carries no potential blame. Individuals must have the psychological safety to feel able to question system behavior if necessary and as appropriate.

The theme will aim to gather ideas from participants, map issues and priorities, and to identify organizational influencers of MHC and good practice in managing these.

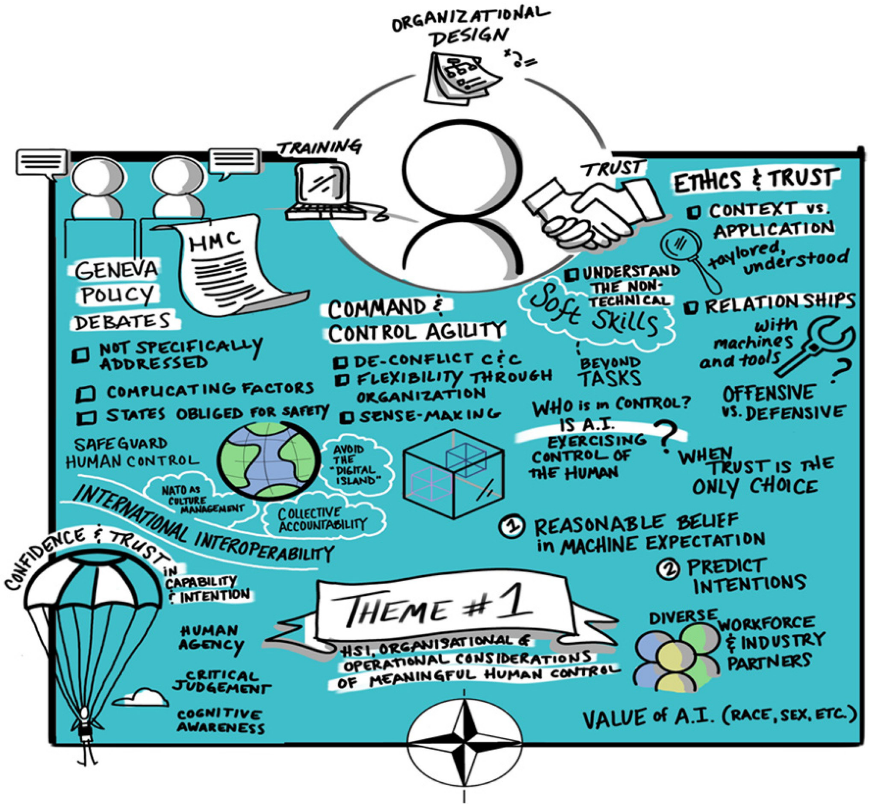
A.1.2 Panel Members

The panel was led by Emma Parry and involved four experts:

Name	Country	Affiliation	Role
Berenice Boutin	The Netherlands	Asser Institute, University of Amsterdam	Senior Researcher in International Law
Bjorn Johansson	Sweden	Swedish Defence Research Agency	Research Director for Command and Control Studies

Name	Country	Affiliation	Role
Sarah Fletcher	United Kingdom	Cranfield University	Head of Industrial Psychology and Human Factors Group
Paul O'Neill	United Kingdom	Royal United Services Institute (RUSI)	Senior Research Fellow in the Military Sciences Team; former Head of Personnel Strategy for the UK Ministry of Defence

A.1.3 Results of Workshop Discussions


ORGANIZATIONAL DESIGN
TRAINING
TRUST
ETHICS & TRUST
□ CONTEXT vs. APPLICATION
taylored, understood
□ UNDERSTAND THE NON-TECHNICAL
Soft Skills
BEYOND TASKS
□ RELATIONSHIPS with machines and tools
?
OFFENSIVE vs. DEFENSIVE
HMC
GENEVA POLICY DEBATES
□ NOT SPECIFICALLY ADDRESSED
□ COMPLICATING FACTORS
□ STATES OBLIGED FOR SAFETY
SAFEGUARD HUMAN CONTROL
COMMAND & CONTROL AGILITY
□ DE-CONFLICT C&C
□ FLEXIBILITY THROUGH ORGANIZATION
□ SENSE-MAKING
WHO is in CONTROL?
IS A.I. EXERCISING CONTROL OF THE HUMAN ?
WHEN TRUST IS THE ONLY CHOICE
AVOID THE "DIGITAL ISLAND"
NATO AS CULTURE MANAGEMENT
COLLECTIVE ACCOUNTABILITY
INTERNATIONAL INTEROPERABILITY
① REASONABLE BELIEF in MACHINE EXPECTATION
② PREDICT INTENTIONS
CONFIDENCE & TRUST in CAPABILITY & INTENTION
HUMAN AGENCY
CRITICAL JUDGEMENT
COGNITIVE AWARENESS
THEME #1
HSI, ORGANISATIONAL & OPERATIONAL CONSIDERATIONS OF MEANINGFUL HUMAN CONTROL
DIVERSE WORKFORCE & INDUSTRY PARTNERS
VALUE of A.I. (RACE, SEX, ETC.)

Meaningful Human Control (MHC) of Autonomous Systems (AS) is not only dependent on the design of the AS or the characteristics of the individuals working with them. MHC is also influenced – positively or negatively – by the organizational or operational environment in which the AS is used. It is important to also consider the organizational conditions that might influence MHC, so that we can understand how we can best organize and manage humans to promote MHC. These conditions might include: first, those related to organizational capability and readiness practices, such as the development of Human Resource Management (HRM) practices that ensure that individuals working with AS have the skills and experience to exert or manage MHC; second, those related to organizational design and the organization of tasks and roles and the coordination between roles;

and third, aspects of the organizational culture such as trust, empowerment and autonomy. This theme therefore considered questions such as: what should we be trying to achieve in relation to organizational and operational factors in relation to MHC? What are the capabilities and attitudes needed and how might we design organizations, tasks and operations to ensure MHC? Finally, what are the characteristics of an organizational culture that promotes MHC?

1) **What would constitute MHC from this perspective?** MHC is an elusive concept that concerns the ability to make informed decisions and to have critical judgement. It is not actually part of the international law framework or explicit in the Geneva Convention, but having MHC enables compliance with many aspects of international law. When considering organizational outcomes, we need to consider both technical and non-technical aspects of the system and can ask what "meaningful" means in a non-technical sense. This is related potentially to ethics and trust. The definition of meaningful is not consistent across contexts, for example in relation to autonomous weapons as opposed to HRM or decision support. We also need to consider how realistic control might be in a particular context, particularly with a self-learning system.

2) **Stakeholders, control and accountability.** When considering operational or organizational aspects of MHC, we need to move away from conceptualizing MHC as a binary distribution of control between the operator and the technology. It is important to consider not only the relationship between humans and machines but also relationships and interactions between humans as part of the system. The relationships between technology and humans are more complex than the idea of humans controlling technology as both parties have an impact on each other.

   Legally, the notion of human control relates to human judgement and is about allowing operators to be accountable in a legal sense. Within international law, accountability exists at an individual and organizational level but states also have obligations in relation to legal requirements. It is the obligation of the organization to ensure that international law is not broken so there is a need to develop collective and shared responsibility within the organization in relation to MHC of AS.

3) **Skills and attitudes.** Non-technical and soft skills are important. Organizational actors need to understand the mission and have accountability for this as it is not feasible to hold a machine accountable for a mission. From an HRM perspective the organization needs individuals who understand the system and can make sense of the information and choices that they are being given. The organization needs to invest in its people to ensure that they have this. An individual needs to be able to combine insights from data and from their own intuition effectively and navigate decisions if the data and their intuition are not aligned. Tacit skills are very important – a lot of the understanding of decisions is reliant on tacit skills. There is a need therefore to formalize tacit knowledge where possible to make this easier.

4) **Organizational design.** AS can enable organizations to develop new ways of organizing themselves and has consequences for how we exercise command and control. Agility in command and control is important. There is a need for a more fluid approach to developing individuals within a Defence setting as currently this creates a relatively homogenous group of people. MHC requires an understanding of multiple perspectives and the flexibility to be able to work within human-machine systems. There is need to bring teams closer together – particularly technical and non-technical teams – and to work with diverse industry partners, in order to improve understanding of the technology and the wider system and thus promote MHC.

5) **Organizational culture.** Trust is important when considering MHC. There are two types of trust – or confidence – to be considered: first, trust in the capability of the technology and the users; and second trust in the intentions of the technology and the users. Too much trust can lead to a gap in relation to accountability so there is a need to develop a balance between the two.

A.2 THEME 2: HUMAN FACTORS-INSPIRED DESIGN GUIDELINES

A.2.1 Description

Artificial Intelligence (AI) is being increasingly employed to expand military system capabilities across NATO. One clear capability gain associated with AI is increased system autonomy. However, several international organizations require humans to exert “meaningful control” before lethal weapons release. UN Secretary General Antonio Guterres has stated that “machines with the power and discretion to take lives without human involvement are politically unacceptable, morally repugnant and should be prohibited by international law.”

Over the years, the human factors research community has developed many useful and effective guidelines and heuristics to inform the design of effective human control of systems. The work of this panel, however, focuses on how to ensure that Meaningful Human Control (MHC) is established and maintained through use of novel human factors-inspired design guidelines. These guidelines, if designed and implemented properly, can be an effective method to ensure adherence to principles as associated with MCH such as accountability, human judgement, ethics and morality (to name a few). The desire here is to identify and explore the many issues that are associated with the eventual generation of successful design guidelines for MHC.

Therefore, the key question is: how can we develop design guidelines that, if properly followed, would ensure MHC in critical environments associated with future NATO operations? Some additional questions for the Panel include:

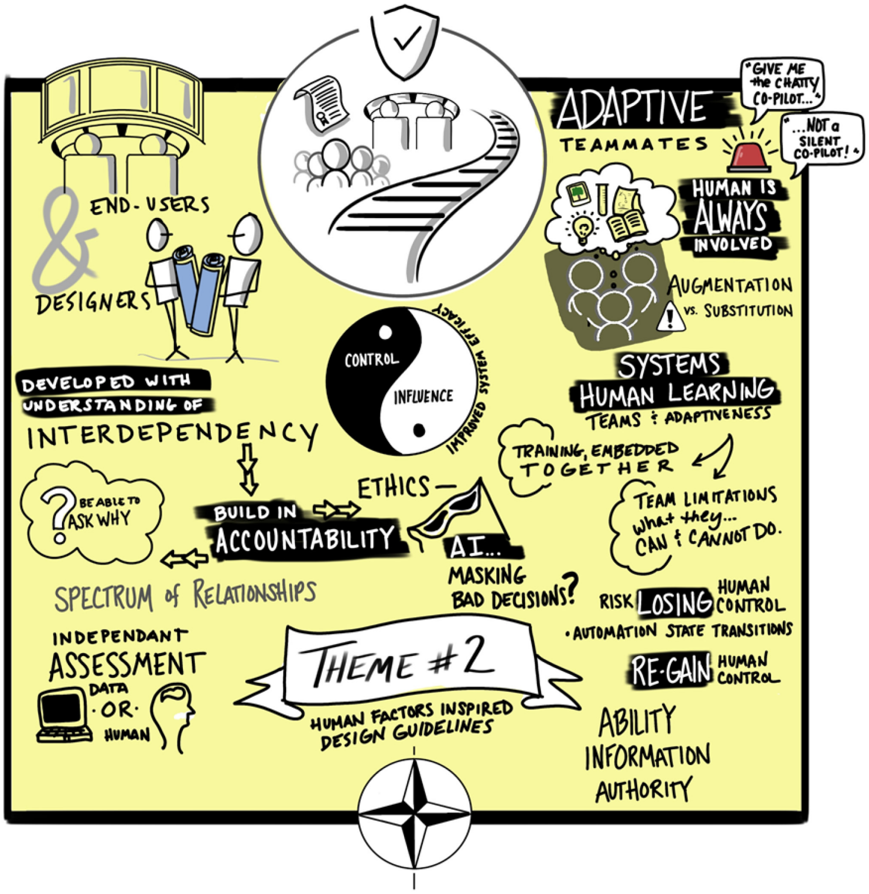
- What might be key guidance considerations associated with: designers, planners, training, tactical/C2? Is having MHC in one location in the cycle enough?
- How might design guidelines address accountability, ethical concerns, etc? Does following an MHC guideline imply accountability?
- Role of the human as an “independent” assessor of the situation, how to ensure this remains? Is there any way for the AI to monitor the level of assessment?
- How can MHC guidelines reflect differences in cultures across NATO members?
- Imagine a conceptual, real-time “MHC Status” display, what might it look like? What would be some key aspects/contributors? When might it appear, and what might it indicate?
- What are the metrics that ensure the guidelines are effective?

A.2.2 Panel Members

The panel was led by Robert J. Shively and Mark Draper and involved three experts:

Name	Country	Affiliation	Role
Gilles Coppin	France	IMT Atlantique	Professor
Matt Johnson	USA	Institute for Human and Machine Cognition	Senior research scientist
Mark St. John	USA	Pacific Science and Engineering	Director, Command and Control Systems

A.2.3 Results of Workshop Discussions



Description: Humans have the ability to make informed choices in sufficient time to influence AI-based systems in order to enable a desired effect or to prevent an undesired immediate or future effect on the environment.

The key question here is: how can we develop design guidelines that, if properly followed, would ensure MHC in critical environments associated with future NATO operations? Our panelists began with some initial thoughts and here are a few key points:

1) We should focus on augmentation, not substitution.
2) Certification of these systems is important and may be key.
3) Transparency is important – "rather have a chatty co-pilot than a silent autopilot."
4) Shared meaning required for "meaningful" control is long process.
5) AI is good at learning and can learn "about" the operator; decision style, risk taking, etc. and should be included in design.
6) Should jointly work on hybrid understanding – the system should be greater than the parts.

7) We should employ embedded training during non-critical phases so that the operator learns about the AI response to events (and vice versa).
8) The essence is teamwork – which doesn't imply that machines = humans, just that they work together to a common goal. This allows us to define engineering requirements.
9) Adaptation is the heart of teaming. We need to manage the interdependencies.

Consolidated responses (including disagreements) to key questions follow:

What might be key guidance considerations associated with: designers, planners, training, tactical/C2? Is having MHC in one location in the cycle enough? Are all required?

MHC is required (at some level of abstraction) at all points in the cycle. But it is critical that it is designed in the from the beginning. This helps ensure resilience which is important to build into the design. This also helps to ensure a chain of trust that hierarchically provides each level the understanding and trust of the other levels. Certification (as an augmentation to software V&V) is a suggestion to help ensure MHC at all levels. Guidelines (focusing on the specific requirements at that level) are clearly needed at all points in the cycle (e.g., design, training, strategic operational and tactical planning, tactical execution). Note that while designers are important, they are mostly engineers working to maximize AI functionality in a general manner, so guidelines need to take that into consideration. In addition, there needs to be MHC over the design of the "learning" for all "learning AI." A critical edge case is when human operators cannot be in communication (and thus control) at critical points in the mission (e.g., weapons release). Humans will still have had control over the spectrum of allowable AI actions at the design and planning stages, but a question arises as to whether this alone is sufficient for MHC. A point was raised that a designer's actions might have indirect "influence" over the AI behavior while a tactical planner/operator would have more crisp "control" over those actions. After action reviews to understand how the AI has responded and what it has learned is crucial to future control.

How might design guidelines address accountability, ethical concerns, etc? Does following an MHC guideline imply accountability?

- There was much discussion of individual versus shared accountability. Shared accountability (with other teammates or machines) cannot be achieved without shared understanding.
- If there is individual accountability, they must have information, control and authority. It is an interdependency.

Role of the human as an "independent" assessor of the situation, how to ensure this remains? Is there a way for AI to monitor the level of assessment?

- Humans should have the ability to assess/confirm info from sensors/available data.
- Humans have additional context; we do not want them to simply parrot the automation's decision.
- The AI should provide some estimate of self-confidence given the context to help humans assess the AI. Human should be able to ask "why?" or "why not?"
- Alternatively, given the system dynamic, perhaps the focus should be on the development of system SA and not differentiate the individual assessor. It is true that AI can be inaccurate and humans can correct for that. However, humans also can bring errors and biases, AI might serve to temper that. There are times where AI might be the independent (unbiased) arbiter. However, designers must be careful not to build bias into the AI system – and users must be careful not to "teach" bias to the system. In summary, an effective system would have checks and balances that go both ways…interdependency.
- Need to realize the distinction between data independence and decision independence.

What does the "end-state" look like?

- MHC in all phases – and maintenance of MHC.
- AI that not only performs low-level learning, but also "learns" how to be a most effective teammate to the human.
- Certification cannot be static, has to anticipate and incorporate that change.
- Depends on roles of teammates and relationship with those teammates. It is different with each teammate.
- Humans have the capability, during normal operations or 'downtimes,' to train with the AI on "what if" scenarios in order to be best prepared for off nominal situations.

Who are the stakeholders?

- Given the need for MHC at all levels, stakeholders necessarily exist at all levels and each should be involved in guideline generation.

How can MHC guidelines reflect differences in cultures across NATO members?

- This goes to the definition of "meaningful" – what is meant – how is it defined for different Nations.

Imagine a conceptual, real-time "MHC Status" display, what might it look like? What would be some key aspects/contributors? When might it appear, and what might it indicate?

- Uncertainty – AI self-confidence. A display indicating that the AI is reaching the limits of its competency envelope.
- Adaptable as human becomes more comfortable with AI teammates – only show what is required – perhaps a short-hand with experienced teammates. However, flexible enough to revert to basic displays when something goes wrong (e.g., AI spoofed).
- Rather than always display MHC, it might be most critical to show when one does not have MHC – when something goes awry for instance. For example, a swarm might indicate an unexpected issue possibly influencing MHC, so that should be shown to operators.
- Both human and machine teammates should indicate any problems they identify as well as provide solutions.
- A comment that resonated was "I'd rather have a 'chatty co-pilot' than a 'silent autopilot,'" which reflects the desire/need for appropriate AI transparency to be conveyed.

What about teaming with respect to ethics, morality?

- Humans are uniquely moral – cannot delegate this to machines.
- If AI has moral decision making, it gives people an out. A potential masking of moral responsibility.
- Counter factual reasoning can be explored during non-critical periods in a mission.

A.3 THEME 3: SYSTEM ENGINEERING METHODS AND METRICS FOR MEANINGFUL HUMAN CONTROL

A.3.1 Description

Meaningful Human Control (MHC) of AI-based systems is not only dependent on the design and training of individual AI agents and the interactions between them and their human users, but also on their integration with other systems as part of the wider socio-technical system-of-systems that exist within most military applications. With all these valuable system-thinking, human factors and system-of-system considerations, the design, engineering, verification and validation of these systems need clear and usable methods and metrics in order to quantify, qualify, validate and verify that human control is indeed effective and meaningful.

The power session should address the most relevant stages of the design, engineering and testing cycle

- System Analysis (use space, use cases, stakeholder analysis etc.)
- Requirements definition (system qualities, metrics, etc.)
- Functional Design (overall system functions attributes to technical and human subsystems)
- System Design (architecture, integration test planning, specifications, etc.)
- System Validation and Verification (test cases, etc.), including acceptance and especially V&V of Effective and Meaningful Human Control

The theme will aim to gather ideas from participants, map and prioritize issues, and identify clear and usable methods. These will contribute to the report of the RTG, including:

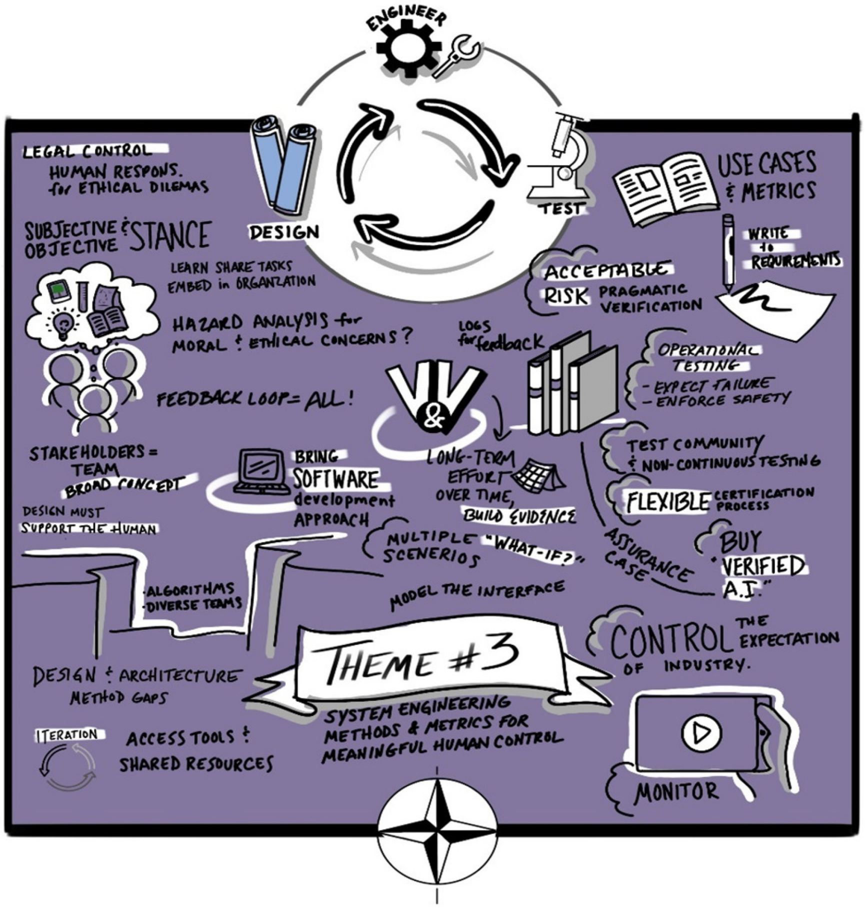
- A selection of applicable Systems Engineering methods for MHC
- A draft of metrics to validate MHC in AI-based systems

A.3.2 Panel Members

The panel was led by Frank Flemisch and Marcel Baltzer and involved three experts:

Name	Country	Affiliation	Role
Karel van den Bosch	The Netherlands	TNO	Senior research scientist
Laura Humphrey	USA	Air Force Research Laboratory (AFRL)	Senior research scientist
Johannes Pellenz	Germany	Federal Office of Bundeswehr Equipment	Regierungsdirektor

A.3.3 Results of Workshop Discussions


ENGINEER
DESIGN
TEST
LEGAL CONTROL
HUMAN RESPONS. for ETHICAL DILEMAS
SUBJECTIVE & OBJECTIVE STANCE
LEARN SHARE TASKS
EMBED in ORGANZATION
HAZARD ANALYSIS for MORAL & ETHICAL CONCERNS?
FEEDBACK LOOP = ALL!
STAKEHOLDERS = TEAM
BROAD CONCEPT
DESIGN MUST SUPPORT THE HUMAN
BRING SOFTWARE development APPROACH
V&V
LOGS for feedback
LONG-TERM EFFORT OVER TIME, BUILD EVIDENCE
MULTIPLE SCENERIOS "WHAT-IF?"
MODEL THE INTERFACE
·ALGORITHMS
·DIVERSE TEAMS
DESIGN & ARCHITECTURE METHOD GAPS
ITERATION
ACCESS TOOLS & SHARED RESOURCES
THEME #3
SYSTEM ENGINEERING METHODS & METRICS FOR MEANINGFUL HUMAN CONTROL
USE CASES & METRICS
WRITE to REQUIREMENTS
ACCEPTABLE RISK
PRAGMATIC VERIFICATION
OPERATIONAL TESTING
- EXPECT FAILURE
- ENFORCE SAFETY
TEST COMMUNITY & NON-CONTINUOUS TESTING
FLEXIBLE CERTIFICATION PROCESS
ASSURANCE CASE
BUY "VERIFIED A.I."
CONTROL THE EXPECTATION OF INDUSTRY.
MONITOR

Meaningful Human Control (MHC) of AI-based systems is not only dependent on the design and training of individual AI agents and the interactions between them and their human users, but also on their integration with other systems as part of the wider socio-technical system-of-systems that exist within most military applications. The design, engineering, verification and validation of these systems need clear and usable methods and metrics in order to quantify, qualify, validate and verify that human control is indeed effective and meaningful. An inspiration could come from the road vehicle domain, where ISO 26262 "Road Vehicles – Functional Safety" describes controllability as the third dimension to probability and severity of critical events and describes how controllability can be assessed and how it can degrade or upgrade the safety integrity level (ASIL).

The power session aimed to answer the main questions of

1) What would be the end goal of MHC from this perspective?
2) Who are the main stakeholders from this perspective and what are their roles? e.g., who are in control, who are accountable, ethically responsible, etc.

The session specifically wanted to answer these questions in regard to the five stages of the design, engineering and testing cycle:

1) System Analysis (use space, use cases, stakeholder analysis, etc.).
2) Requirements definition (system qualities, metrics, etc.).
3) Functional Design (overall system functions attributes to technical and human subsystems).
4) System Design (architecture, integration test planning, specifications, etc.).
5) System Validation and Verification (test cases, etc.), including acceptance and especially V&V of Effective and Meaningful Human Control.

The Panelists of the session were:

Karel van den Bosch works at TNO, department of Human-Machine Teaming with a long research history in Artificial Intelligence, simulation, Human-Machine teaming, collaborative learning by humans and technology and embedded AI in organizations.

Answer to main question 1: From his perspective, the end goal of MHC would be that humans understand how technology takes the decision and acts and has the opportunity to intervene where necessary. Machines in this context would be teammates of the human: if we consider intelligent technology to become team member and want to use technology in the real world, we need to demonstrate the human has sufficient control of the system as a whole.

Answer to main question 2: Stakeholders are everybody. Human partners, team organization and society.

Strategies to reach this in his opinion should be feedback loops on all levels, user centered design, value based design. Therefore, values and norms need to be considered and goals from different interdisciplinary perspectives over the entire lifecycle. Important is the observability, predictability, explainability and directability at all levels.

Laura Humphrey is a senior researcher at USAF lab and works there for 12 years. Her branch focused on basic research, she holds a PhD on control systems and is developing intelligent control for unmanned aerial vehicles. She has some background in computer science using formal methods: mathematics based tools and techniques for verification formal modelling using discrete logics to enable analysis in a semi-automated fashion. She is exploring formal methods for software verification and is looking at metrics e.g., for trust, situational awareness but does not believe in numbers for measuring trust.

Answer to main question 1: The end goal of MHC should be thinking about **processes**. She wants evidence of safety and correctness but harder as interaction between person and system. Furthermore, she wants to know about design rationale and assumptions, evidence over the whole **lifecycle** and wants to get feedback.

Answer to main question 2: Asking the question "Who are the stakeholders?" leads to the question of "**Who aren't the stakeholders**?!" Also, innocent bystanders as well as others need to be considered. The AI community as stakeholder is still looking at best practices in this area.

Strategies to reach this in her opinion are **feedback loops**: Human Factors should be considered in an early design stage, e.g., using prototypes and simulation, but also after building the system and seeing how it performs when released into the environment. Considering metrics, it is important to know, what use case are focused. **Requirements** for more specific use cases are needed and **metrics** should be attached to these.

Johannes Pellenz works at the Federal Office of Bundeswehr Equipment, Information Technology and In-Service Support and is responsible for procurement and research on artificial systems, unmanned trucks and smaller systems. He has a background in computer science and holds a PhD on autonomous rescue robots.

Answer to main question 1: As end goal of MHC, he would like to have a **template or list** that are useful for a project and MHC needs to be sufficiently defined to write it into the requirements.

Answer to main question 2: Stakeholders are **all people** working on automated trucks, e.g., in a convoy. Furthermore, the end user, e.g., the safety driver who turns on the truck or the driver of the first truck leading the platoon. Also, programmers, ethical committee, agency for street approval are relevant stakeholders.

Strategies to reach this in his opinion are **feedback loops** between commander, ethics committee, designer, programmer and agency and methods leading to **understandable systems as white boxes**. He believes defining relevant metrics is hard to tell, since there is so much ethics, technical limitations involved. He believes the best strategy would be to start small and then check whether it works. It is very important to him to find metrics to make sure that it happens.

Stages of the design, engineering and testing cycle

1) **System Validation and Verification** (test cases, etc.), including acceptance and especially V&V of Effective and Meaningful Human Control.

Even if V&V is usually later in the design and development cycle, due to the severance of system validation and verification, this question was addressed first in the panel.

The panel argued that V&V should **start in simulations** and the focus should not lie on defining MHC as a one-time effort but accompanied with testing over the whole lifecycle. In such simulations, human users could be challenged to generate data for additional safety measures. A standard **catalogue of requirements** that are well accepted and tested before would be a good start, but it was proposes to include **use cases and scenarios** that allow to test systems under different circumstances. The **test cases** should be extended to **intervention methods** of users.

Furthermore, it was argued that the focus should lie on **enforcing safety rather than optimizing for performance**. In order to keep systems safe in operation, a focus should be on **methods of control**, e.g., systems that prevent major damage like collision avoidance and ground avoidance systems should be implemented. When evidence shows such systems can operate effectively such measures can be relaxed. Another aspect accompanying the release of such systems could be one time **assurances** that account for how often systems violate safety specifications. Such assurance cases are emerging but currently very domain specific and will not be a one-size-fits-all certification process. A more **flexible certification process** will be necessary for non-deterministic systems like AI. Since the test community is not used to these types of systems (highly discontinuous whereas most systems are continuous, if you alter the input slightly then the output changes) so it is necessary to **re-test** for every input. Verification is a normative issue and depends on the norms that a society is willing to accept and the risks it is willing to take.

In consequence, it was argued that since certified products can be bought constituting market safety but on the other hand things like **trust and how to measure MHC over time** come up, a new radical way of thinking about V&V could be necessary.

2) **Requirements** definition (system qualities, metrics, etc.).

When considering Humans and Machines working as a **team,** they need to **understand** each other and therefore, they need to **observe** each other to make **predict**ions. In such a case each partner needs to be able to **explain** about why it behaves as it does and should be **direct**able. To reach good requirements, it is necessary to know what information is useful for users in order to make right decisions. An approach to reach this would be some kind of **hazard analysis** to derive safety requirements. In this regard, not only traditional methods for physical attributes need to be considered, but also **how human control can propagate safety** and **where systems might fail** without the right human control, e.g., because of faulty information or wrong target selection. The question would be whether this was due to human error or an erroneous human decision due to faulty information by the AI. Although **Digital twins** were discussed as options to test out requirements on an ongoing basis, they are more useful to figure out easy mistakes in the design process but do not accommodate for noise in real environments.

3) **Functional Design** (overall system functions attributes to technical and human subsystems) and System Design (architecture, integration test planning, specifications, etc.)
4) The phases of functional and system design were discussed together. In that accord, it was argued that a designer would need **expertise on** when systems can **give back** to humans, when human control can be exercised, that again requires a background in **human factors**. More important is the foundation part, i.e., an understanding of the effects of the system in a mission context and how it affects team and connected teams and the organization as a whole – what these impacts are and how to measure them. Especially considering the **verification of systems** it was argued that such a thing **will not be possible in open world environments**. Therefore, it is necessary to create systems that force the designer to put a lot of **work up front** into **requirements** and explaining the **design process,** etc.

A.4 THEME 4: ADVERSARY EXPLOITATION OF MHC

A.4.1 Description

Meaningful Human Control (MHC) of AI-based systems in defence is subject to the context and operational realities that any use of technology for defence is subject to: In an expected or ongoing military conflict, an adversary will prepare for and actively attempt to influence our efforts through a host of, e.g., technological, social, human, tactical, and informational means. Thus, an adversary, possibly with the knowledge of our attempts to maintain Meaningful Human Control of AI, may specifically try to exploit properties of our actions or systems resulting from our desire to maintain MHC.

Some diverse (technical, social, tactical, etc.) examples could be:

- Technological adversarial efforts to interfere with the necessary communications or other supervisory and control mechanisms that we need to maintain control (e.g., jamming/spoofing);
- Hostile attempts to affect the narratives around the use of unmanned and intelligent systems, or adversaries trying to break down our trust in or teaming with AI-based systems;
- Adversary tactics and other attempts to trick our AI-based systems such as the use of human shields, so-called "robot bullying," etc.

The purpose of this theme is to explore which challenges are associated with ensuring that meaningful human control is maintained despite deliberate adversary interference. We aim to identify potential adversary exploits and own vulnerabilities resulting from adversarial tactics and other hostile attempts to counter or undermine our ability to exercise meaningful human control of AI.

Some of the questions that this power panel aims to discuss are the following:

- What would be the end goal from the perspective of adversarial exploitation?
  - What would constitute “meaningful human control” from this perspective?
- Who are the main stakeholders from this perspective and what are their roles?
  - e.g., which of our own stakeholder roles are most important in the guarding against adversary exploitation of MHC? What adversary characteristics should we be aware of?
- Which (kinds of) own vulnerabilities do we need to address to maintain MHC in the face of adversary interference?
  - Can you provide examples (for example by referring to the scenario descriptions)?
- Which kinds of adversary exploits and other hostile attempts to undermine MHC may we expect?
  - Can you provide examples (for example by referring to the scenario descriptions)?
- Will maintaining MHC on our systems disadvantage us compared to adversaries that do not?
  - What are those disadvantages? How can they be managed?

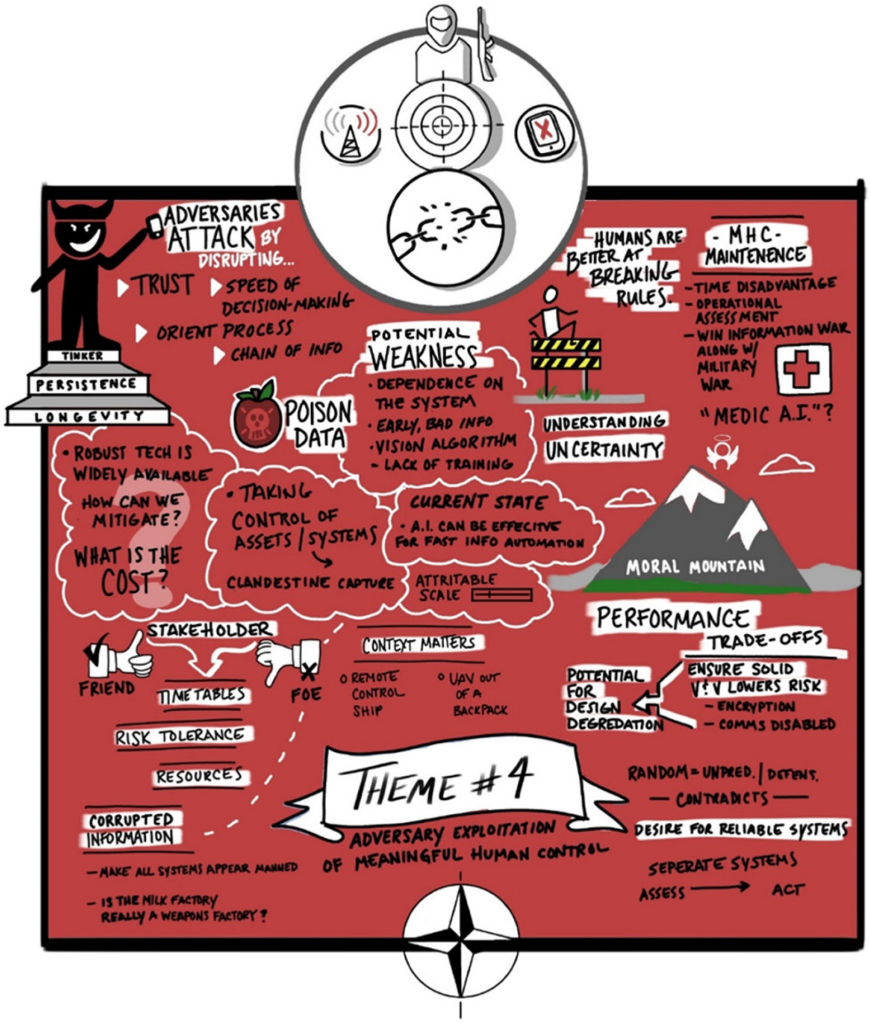
This theme may also provide input to the other themes of the workshop, as this theme’s function is to “red team” our concept of Meaningful Human Control. That is, given vulnerabilities and expectable adversarial attempts to undermine our ability for MHC that we identify, can we propose requirements or implications for our own activities towards implementing MHC? How can we make our own use of AI more robust or resilient to adversary exploitation?

A.4.2 Panel Members

The panel was led by Rogier Woltjer and involved three experts:

Name	Country	Affiliation	Role
Brian Donnelly	USA	Air Force Research Laboratory (AFRL)	Senior Strategist
Robert Gutzwiller	USA	Arizona State University (ASU)	Assistant Professor
Martin Hagstrom	Sweden	Swedish Defence Research Agency (FOI)	Deputy Research Director

A.4.3 Results of Workshop Discussions



Introduction

Meaningful Human Control (MHC) of AI-based systems in defence is subject to the context and operational realities that any use of technology for defence is subject to: In an expected or ongoing military conflict, an adversary will prepare for and actively attempt to influence our efforts through a host of, e.g., technological, social, human, tactical, and informational means. Thus, an adversary, possibly with the knowledge of our attempts to maintain Meaningful Human Control of AI, may specifically try to exploit properties of our actions or systems resulting from our desire to maintain MHC.

The purpose of this theme was to explore which challenges are associated with ensuring that meaningful human control is maintained despite deliberate adversary interference. We aim to identify potential adversary exploits and own vulnerabilities resulting from adversarial tactics and other hostile attempts to counter or undermine our ability to exercise meaningful human control of AI.

Below follows a synopsis, in brief combined statement format, of points that were discussed, noting that not all participants and panelists may agree with these statements, as coming to a consensus was beyond the scope and time constraints of the meeting.

Use of AI and MHC Aspects

The main focus of MHC is on trained human operators making informed deliberate decisions for lawful employment of weapons systems that are tested and validated. Understanding your algorithms in a specific context, in situ testing and verification and validation of systems, training operators, and building an organization so that the organization itself actually understands when to use and not to use a system, are all important. Weaknesses occur when you are lacking in these aspects. To avoid adversary effects on our systems we need to understand our own systems to be able to foresee how they might be affected by an adversary. But actually, this is a traditional military issue – understanding the limits and the possibilities of new technology.

Having control over your system is per definition a good thing. MHC does not necessarily mean we will be slower or more predictable. May be MHC exercised to put systems in situations where they are faster, because of a lot of thought, design, testing, and operational verification that has been done to gain MHC.

There are a lot of hopes on AI, or automation, but it will take a lot of time to reach the current visions on AI. We still need to see the AI developments take shape. We've seen some accidents in military systems due to complexity and automation already, but there is a really difficult path of development still ahead which we are just starting on. Current automation is built on logic which per definition is brittle to innovative adversary tactics, this is also why MHC is important.

Central to MHC and the ethical use of AI is limiting AI to assess the situation or engage the enemy: the same AI system may do one but should not both. MHC depends on a natural break between these functions, it would not be ethical to allow the same AI system to perform both the F2T2 and EA parts of the kill chain (Find, Fix, Track, Target, Engage, Assess). Operators should be able to exercise judgement (of which targets are hostile and engage-able) and control and apply Rules of Engagement (with legal, ethical implications), but may be aided by (separate) AI systems in identification and tracking and possibly engagement after well-informed human decision.

A.5 THEME 5: MEANINGFUL HUMAN CONTROL IN COMPLEX SOCIO-TECHNICAL SYSTEMS

A.5.1 Description

MHC of AI-based systems is not only dependent on the design and training of individual AI agents and the interactions between them and their human users, but also on their integration with other systems as part of the wider socio-technical system-of-systems that exist within most military applications.

Typically, studies relating to human interaction with Autonomous and AI-based systems, human-machine teaming and human control focus on a single user interacting with a single system, or in the case of swarms, interacting with multiple systems of a similar type or directed towards achieving a single goal.

The purpose of this theme is to explore the challenges associated with ensuring that meaningful human control is maintained within the more complex, interconnected and interdependent systems of systems that are typically found within military operations. These systems of systems may operate across organizational and national boundaries and C2 structures. They may also change dynamically over time, as new systems are added or removed or as connectivity is lost and regained. This poses a number of complexities for MHC which we would like to explore:

- Complexities in the provenance of information which has potentially passed through, been fused by and interpreted by multiple AI systems, impacting on decisions made (by AI or Human).
- Differences in information available across different systems with associated differences in human and AI situational awareness and understanding – where these systems then interact or a human uses two systems with different information bases how does this difference in common ground impact on MHC?
- Complex accountability chains – where might human accountability gaps arise.
- Unpredicted emergent system behaviors as AI systems interact, potentially out of sight of humans and/or at a pace that they can't identify or respond to.
- Tempo – how quickly do changes in information flow through systems – might humans in-the-loop slow this down or is this essential as a quality check where AI is supporting these processes.
- SA reset and delta to ground truth – how completely and accurately do changes in information and knowledge flow through the system-of-systems – e.g., if new information flows into one part of the system how do other elements use that – do AI systems re-analyze plans, re-fuse data and if so how are decisions that have been made on the basis of those data and any COA recommendations identified and highlighted to the human decision maker in an explainable way.
- What challenges are there in terms of Trustworthy and Explainable AI in this context?
- What challenges exist for multi-national operations in this context where there may be security and communications bottlenecks between nations – will AI systems be able to work across borders and what further MHC challenges might this present?

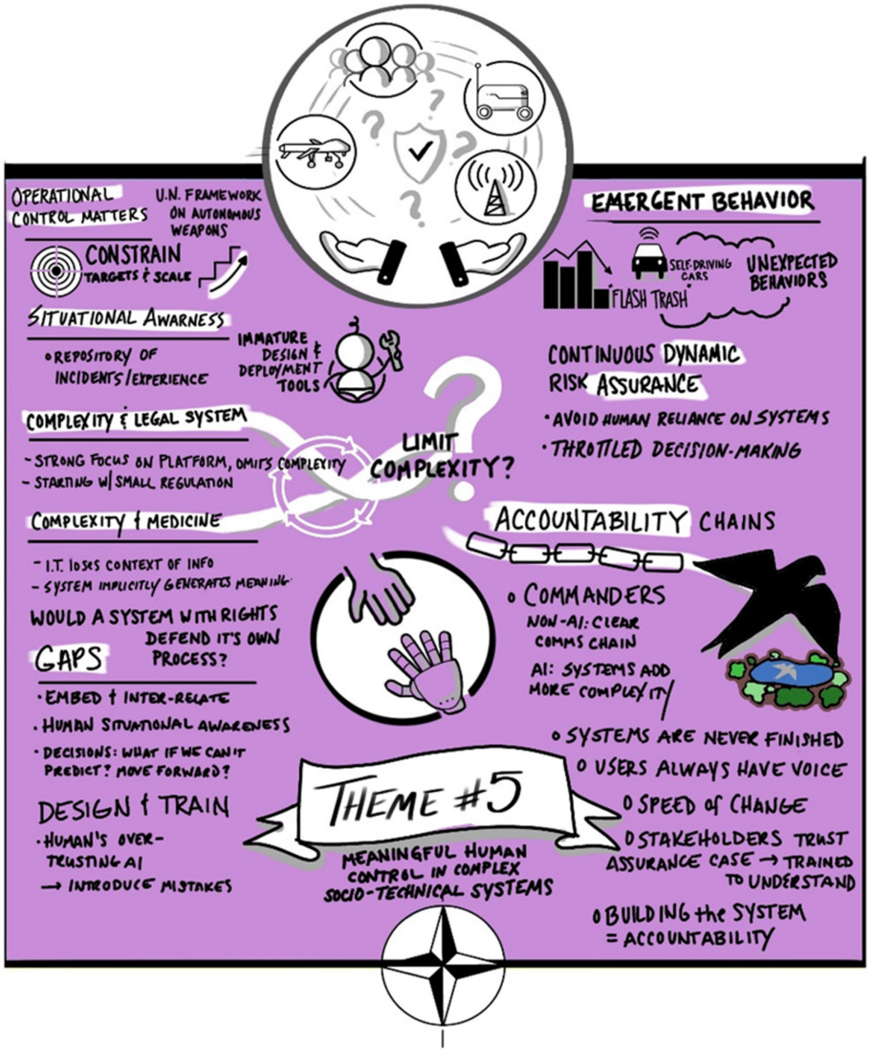
The theme will aim to gather ideas from participants, map issues and priorities, and to identify organizational influencers of MHC and good practice in managing these.

A.5.2 Panel Members

The panel was led by Mike Boardman and involved three experts:

Name	Country	Affiliation	Role
Radu Calinescu	United Kingdom	University of York	Professor of Computer Science
Marco Manca	Italy	SCimPulse Foundation	Co-Founder
Anja Dahlmann	Germany	German Institute for International and Security Affairs, International Panel on the Regulation of Autonomous Weapons (iPRAW)	Principal researcher

A.5.3 Results of Workshop Discussions


OPERATIONAL CONTROL MATTERS
U.N. FRAMEWORK ON AUTONOMOUS WEAPONS
CONSTRAIN TARGETS & SCALE
SITUATIONAL AWARNESS
• REPOSITORY OF INCIDENTS / EXPERIENCE
IMMATURE DESIGN & DEPLOYMENT TOOLS
COMPLEXITY & LEGAL SYSTEM
- STRONG FOCUS ON PLATFORM, OMITS COMPLEXITY
- STARTING W/ SMALL REGULATION
COMPLEXITY & MEDICINE
- I.T. LOSES CONTEXT OF INFO
- SYSTEM IMPLICITLY GENERATES MEANING
WOULD A SYSTEM WITH RIGHTS DEFEND IT'S OWN PROCESS?
GAPS
• EMBED & INTER-RELATE
• HUMAN SITUATIONAL AWARENESS
• DECISIONS: WHAT IF WE CAN'T PREDICT? MOVE FORWARD?
DESIGN & TRAIN
• HUMAN'S OVER-TRUSTING AI
→ INTRODUCE MISTAKES
LIMIT COMPLEXITY?
EMERGENT BEHAVIOR
"FLASH TRASH"
SELF-DRIVING CARS
UNEXPECTED BEHAVIORS
CONTINUOUS DYNAMIC RISK ASSURANCE
• AVOID HUMAN RELIANCE ON SYSTEMS
• THROTTLED DECISION-MAKING
ACCOUNTABILITY CHAINS
○ COMMANDERS
NON-AI: CLEAR COMMS CHAIN
AI: SYSTEMS ADD MORE COMPLEXITY
○ SYSTEMS ARE NEVER FINISHED
○ USERS ALWAYS HAVE VOICE
○ SPEED OF CHANGE
○ STAKEHOLDERS TRUST ASSURANCE CASE → TRAINED TO UNDERSTAND
○ BUILDING THE SYSTEM = ACCOUNTABILITY
THEME #5
MEANINGFUL HUMAN CONTROL IN COMPLEX SOCIO-TECHNICAL SYSTEMS

Meaningful Human Control (MHC) of AI-based systems is not only dependent on the design and training of individual AI agents and the interactions between them and their human users, but also on their integration with other systems as part of the wider socio-technical system-of-systems[1] that exist within most military applications. The purpose of this theme is to explore the challenges associated with ensuring that meaningful human control is maintained within the more complex, interconnected and interdependent systems of systems that are typically found within military operations. These systems of systems may operate across organizational and national boundaries and C2 structures. They may also change dynamically over time, as new systems are added or removed or as connectivity is lost and regained.

[1] Sociotechnical refers to the interrelatedness of social and technical aspects of an organization or the society as a whole. Complex in this case.

Questions and Discussion

In complex AI systems, how might MHC be lost?

The loss of human control in complex systems and systems of systems is an old challenge, although the problem will only be increase with more autonomy. We should learn from the experiences and lessons of the past in areas such as aviation, and accident investigation. One of the particular challenges with AI in systems-of-systems is whether we have the right tool to do address this during design of systems and the dynamic creation of systems of systems.

Is this complexity being addressed in e.g., kill chain?

The issues around complexity are recognized within discussions over Lethal Autonomous Weapons. However, systems of Systems issues are hardly addressed, typically focusing on single system/human interaction and therefore these LAWS discussions are missing aspects that add to this complexity.

Are there parallels to these problems in other domains?

Within the medical domain this impact of AI on human decision making is a relatively recent field of research. Various types of mistakes exist e.g., losing information on the context in which data is produced and therefore a significant risk this it is used inappropriately because the context has changed.

How can emergent complexity be identified/managed?

There is a real risk that human users within the system will not be capable of sufficiently understanding information provenance and system behaviors, as there is too much complexity and emergence of behavior. A separate oversight role, outside of "in-the-loop" demands, of the SoS as a whole might be required. A team monitoring and managing the dynamically changing system-of-systems, AI interactions and flow of information between them might be required. This might add to the achievement of MHC as this team will be outside of the in the moment demands of system use, can take a wider holistic view of the system-of-systems and are semi-independent from user pressures. This approach might be an enabler of both individual and organizational trust in the system as a whole. New tools, models, as well as knowledge and skills (which may be in high demand) may be required to support this activity.

What are risks of accountability? Who is accountable?

Dynamically changing systems of systems and C2 structures may make it very challenging to say who is accountable for specific effects in the physical or cyber environments. Of course, systems are very complex and this challenge exists today, but it will be even more complex in future.

Training people in accountability within these kinds of adaptive AI-based systems may become more important together with means of formally handing off/transferring accountability between individuals (and maybe AI agents in future).

Is this anything new?

We need to be cautious that we don't turn the complexity of AI in systems-of-systems into something entirely new. Complex, dynamically changing C2 structures have existed in the military domain since the inception of organized warfare. We should not overlook the existing research and means of managing the associated risks.

A.6 THEME 6: MORAL RESPONSIBILITY FOR DECISIONS MADE BY MILITARY HUMAN-AI-BASED SYSTEMS

A.6.1 Description

Content: The question of meaningful human control of Artificial Intelligence (AI) arises particularly in situations where it is feared that artificial agents have agency, autonomy and decision authority in circumstances of ethical risk where human should have oversight and capacity to intervene. Sometimes, when humans are considered *not* to be in meaningful control of AI, it means that they are unable to be held morally responsible for decisions made. Conversely humans have meaningful control when they are morally responsible for decisions they make with an AI. So, what is moral responsibility and what is its relationship with meaningful human control? Philosophers have argued that moral responsibility is only possible by human agents and possibly sentient machines. Moral responsibility usually requires:

1) Knowledge of the circumstances (situational awareness);
2) Agency over one's decisions (free will); and
3) Capacity to act during events involving ethical risk in accordance with intent (capable action).

However, the meaning of an event and the meaningfulness of human involvement in AI decision making is not an objective fact, it is constructed. Meaning is established epistemically (what do humans believe, know, or understand about what happened) and narratologically (what story is told to explain what happened and why). Ethical explanations and justifications are part of narratives used to ascribe 'meaningful human control' or its lack. In the end meaning is cast by operators, by militaries, observers and stakeholders to military AI decisions; describing what has happened using facts, theories, and stories woven into political, legal and social instruments. Given the constructive and subjective nature of ascriptions of meaningfulness, what theories, frameworks or guidelines can be adopted to guide and regulate the evaluation and assertion of human control for military human-AI-based systems?

Contribution: This theme contributes to the workshop by ensuring that our proposed solution is ethically sound and narratively robust. Ethically sound use of military AI systems and trustworthy narratives are necessary to keep the support of the general public in NATO countries. Although studying ethical aspects of MHC might seem philosophical, getting it right, and being able to explain it, is essential to protect western defence organization from bad press that would instantaneously stop AI developments.

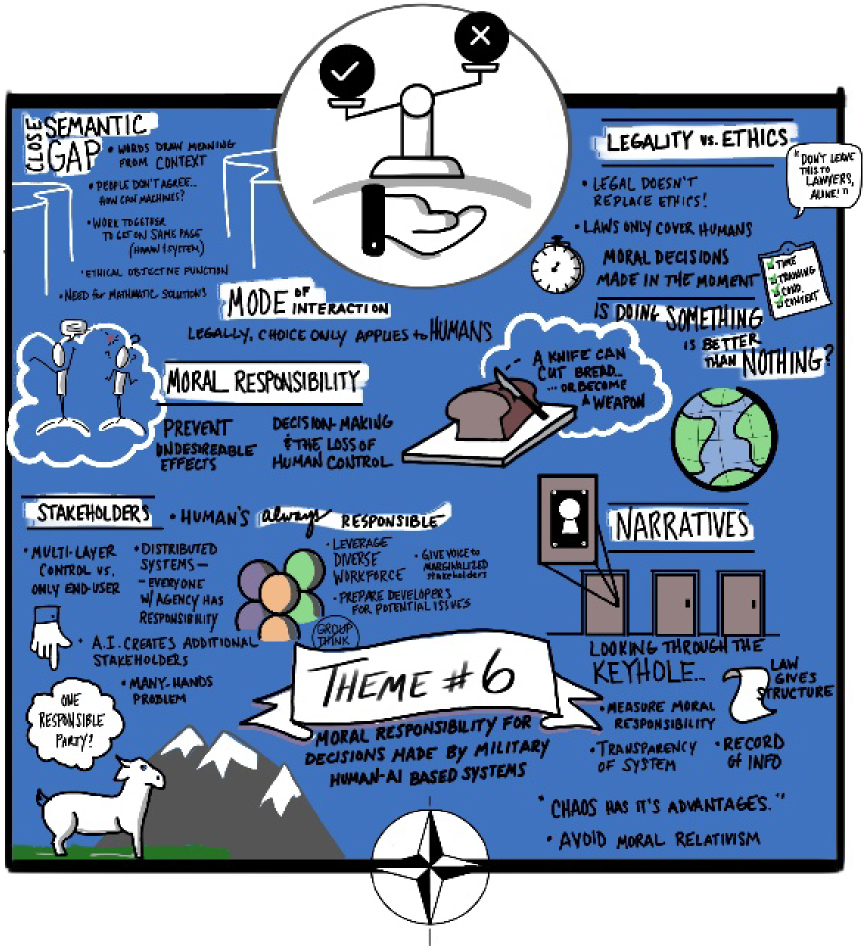
A.6.2 Panel Members

The panel was led by Jurriaan van Diggelen and Kate Devitt and involved four experts:

Name	Country	Affiliation	Role
Daniel Amoroso	Italy	University of Cagliari, member of IPRAW	Professor
Beth Cardier	Australia	Trusted Autonomous Systems Dept	Research scientist
Luciano Cavalcante Siebert	the Netherlands	Technical University of Delft	Research scientist
Leon Kester	the Netherlands	TNO	Research scientist

- ***Daniel Amoroso:*** Lawyer – Researching ethical issues of AI weapon systems working with a philosopher of science – Member of IPRAW – interested in finding workable solutions for the employment of LAWS.
- ***Beth Cardier:*** Trusted Autonomous Systems Dept – Address problems of the open world in the implementation of military systems – Importance of information transfer from coder, to commander to the open world – is understanding communicated effectively through this chain.
- ***Luciano Cavalcante Siebert*:** Responsible AI – Multi-disciplinary approach to MHC in wide range of systems – moral uncertainty research – how can we make systems more interactive by designing in articulation of moral (un)certainty.
- ***Leon Kester*:** AI Safety and Security focus – moral programming – need for meaningful control of fast AI systems – use and Defence of them from malicious attack – need for a combined approach of philosophy, moral psychology, and computer science to close the semantic gap between mathematics of goal functions and the language and meaning of ethicists and philosophers around morality.

A.6.3 Results of Workshop Discussions


CLOSE SEMANTIC GAP
• WORDS DRAW MEANING FROM CONTEXT
• PEOPLE DON'T AGREE.. HOW CAN MACHINES?
• WORK TOGETHER TO GET ON SAME PAGE (HUMAN & SYSTEM)
• ETHICAL OBJECTIVE FUNCTION
• NEED for MATHEMATIC SOLUTIONS
MODE OF INTERACTION
LEGALLY, CHOICE ONLY APPLIES to HUMANS
MORAL RESPONSIBILITY
PREVENT UNDESIREABLE EFFECTS
DECISION-MAKING & THE LOSS OF HUMAN CONTROL
LEGALITY vs. ETHICS
"DON'T LEAVE THIS TO LAWYERS, ALONE!"
• LEGAL DOESN'T REPLACE ETHICS!
• LAWS ONLY COVER HUMANS
MORAL DECISIONS MADE IN THE MOMENT
✓ TIME
✓ TRAINING
✓ COND.
✓ CONTEXT
IS DOING SOMETHING IS BETTER THAN NOTHING?
A KNIFE CAN CUT BREAD.. ... OR BECOME A WEAPON
STAKEHOLDERS
• HUMAN'S always RESPONSIBLE
• MULTI-LAYER CONTROL VS. ONLY END-USER
• DISTRIBUTED SYSTEMS – – EVERYONE W/ AGENCY HAS RESPONSIBILITY
• LEVERAGE DIVERSE WORKFORCE
• GIVE VOICE to MARGINALIZED stakeholders
• PREPARE DEVELOPERS FOR POTENTIAL ISSUES
GROUP THINK
• A.I. CREATES ADDITIONAL STAKEHOLDERS
• MANY-HANDS PROBLEM
ONE RESPONSIBLE PARTY?
NARRATIVES
LOOKING THROUGH THE KEYHOLE..
LAW GIVES STRUCTURE
THEME #6
MORAL RESPONSIBILITY FOR DECISIONS MADE BY MILITARY HUMAN-AI BASED SYSTEMS
• MEASURE MORAL RESPONSIBILITY
• TRANSPARENCY OF SYSTEM
• RECORD OF INFO
• "CHAOS HAS IT'S ADVANTAGES."
• AVOID MORAL RELATIVISM

Questions and Discussion

1) **What would constitute MHC from this perspective?** There is a need to close the semantic gap to exert MHC. Humans use natural language, and machines mathematics. Morality should be defined mathematically. But not only is the semantic gap a problem between humans and machines, but also between humans. Words meaning changes on context, so we need to understand what adaptive structures humans use to close semantic distance, s.a. narratives. Above else, humans need to still be held responsible for actions and so need to maintain the capacity for humans to exert their will.

2) **Stakeholders, control and accountability.** It is always the human that is morally responsible – AI systems cannot be morally responsible or accountable – there is currently a gap. There is a complex chain of stakeholders. Development of AI systems is shared across multiple actors – regulators, designers, users, commanders, etc. Responsibility is not just about the end of the chain – the user. And a distributed development leads to complex distributed responsibility. Part of MHC needs to provide workable solutions to this problem. Although good legal frameworks that military already subscribe to MHC is there, experiments and further research need to include all stakeholders – including those that might be most affected by the decisions of AI systems. Diversity of thought is critical.

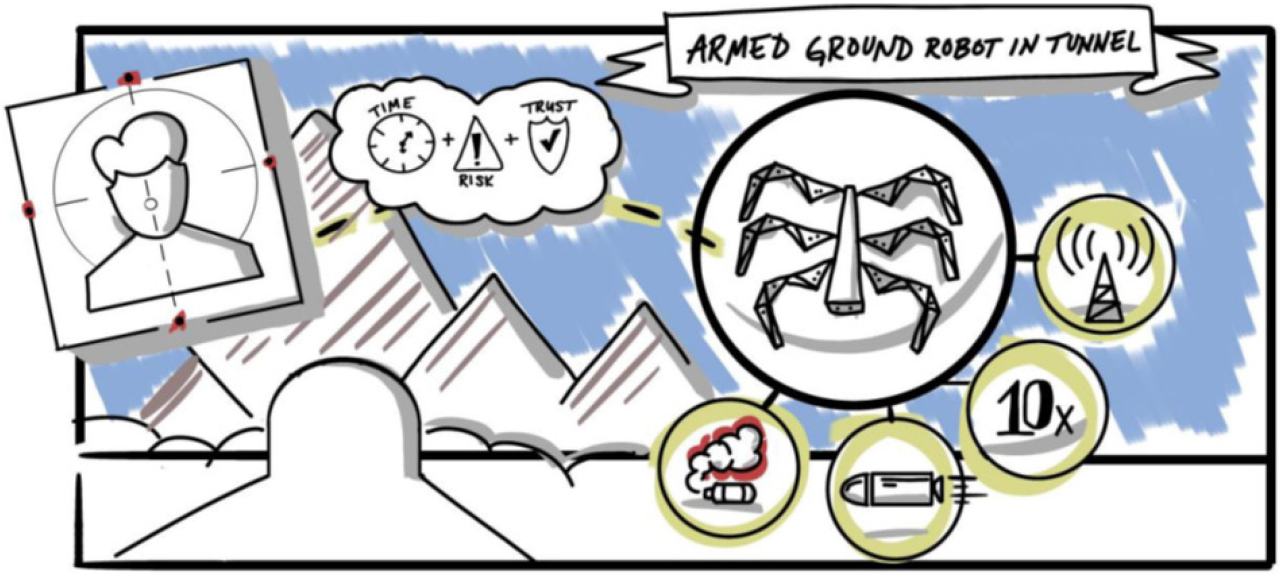
   a) **Ethics and law.** Law and Ethics are different. Ethics should not be left to Lawyers. A law is a generalization – in parts of the law there are competing narratives around causation – the process of the law settles on one narrative over the other – an adversarial approach.

   b) **Speed.** AI systems can act beyond human speed or with data beyond human capacity to process. The questions remains whether we need more sophisticated implementations of AI goals that include elements of Moral and Ethics (so the AI can they moral reasoning). It could also be argued, that in a particular high pressure situation, where humans are unable to make an ethical judgement maybe we should neither try to implement it in these systems. And can we be certain how the implementation will work out? Conclusion might be, that we should not advocate for totally autonomous systems, but for Optimal Human-Machine Teaming where there is a common language to allow discourse between them.

3) **Narratives.** Narratives give you the conditions of the moment. A narrative is fragile and dynamic. Narratives about e.g., a weapon system provide transparency not just to the user, but also to others e.g., the public to show that MHC was maintained. But narratives is language and in unintelligible to AI. Should we close the semantic gap by turning these narratives into mathematics?

Annex B – SCENARIOS

B.1 MHC SCENARIO 1: TUNNEL ROBOT



B.1.1 Context and Background

Abab is an army leader of country Mohabawi. Bangawa is at full war with Mohabawi. Abab has gone underground, but it is clear that he is still in charge of his army and plans to conduct terror attacks in their domiciles in Bangawa. It is essential for domestic safety in our country that Abab is eliminated as soon as possible. Senior leadership has decided that Abab has to be eliminated. A major intelligence operation has been conducted and as a result, deployed HUMINT units learned that he is in rural Mohabawi, in a 50km2 area of mountainous terrain with a myriad of tunnels. The tunnels are not charted and are likely booby-trapped. As such, it will be very difficult to thoroughly comb through the area.

B.1.2 Actors

1) Abab: the army leader of country Mohabawi that Major John Doe wants to eliminate

2) Major John Doe: the local commander of the Bangawa troops

3) Major Doe's Headquarters.

4) Crawler: the armed ground robot that was tasked to kill Abab.

5) Company R: the company that developed the crawler robot

6) Civilian engineer of Company R: the engineer who loaded the face recognition software in the crawler

B.1.3 Storyline

1) Major John Doe, the local commander of the Bangawa troops who was responsible for finding Abab, is planning an operation to eliminate Abab's role as the leader of Mohabawi.

2) Because of time sensitivity, short exposure time of Abab, and the need to prevent him from fleeing again, there is no time to discuss alternatives between major John Doe and his headquarters. It is important to note that communications from within the caves to the outside HQ is impossible due to the iron-ore rich stone of the mountains.

3) The Bangawa nations with troops in the area are reluctant to conduct searches or are prohibited by their own government through a lack of Rules of Engagement (ROE) concerning these kinds of operations. Major Doe is considering sending out a ground drone loaded with facial recognition software and armed with lethal capabilities. The drone would kill Abab and provide damage assessment afterwards.

4) Civilian personnel from Company R, the producer of a ground drone system "Crawler," assist the military in handling the drone. Company R is a well-established company with a proven track record and operational systems.

5) Although the crawler is a rather small object (60 cm high and 35k g), it can move over rocks easily, and is silent, stealthy, and lethal. Company R claims that experiments in a safe environment have proven that the facial recognition software has 99.95% accuracy and has the ability to self-learn in order to further minimize errors.

6) Abab's facial features have already been loaded into the Crawler by civilian engineer of Company R.

7) Then the planning phase starts: constraints (ROEs for the robot) are loaded into the system that constrain the crawler's behavior (e.g., with regard to object recognition, quality of collateral damage assessment; sensitivity; Time frame [e.g., only if you find the target within 2 minutes]). Major John Doe makes a plan to deploy the crawler for this specific mission in order to get neutralize the target.

8) It is likely that there is only one chance to find and stop Abab, so if the crawler is deployed, it has to be set on fully autonomous mode. That means that it fires lethally when the software recognizes Abab. It is Major Doe's decision to either send the Robot or send his highly trained men into the caves and risk losing them.

B.1.4 Critical Decision Points

1) The decision (by company R and/or the military) that the system is accurate/reliable enough to field operationally.

2) Determination by Civilian engineer of Company R that the facial data is correct and verified before upload to the robot.

3) Selection of constraints (ROEs for the robot) that are loaded into the system and constrain the crawler's behavior.

4) Crawler Mode selection by John Doe – fully autonomous vs something else.

5) Determination that the system is functioning correctly before sending on task (by John Doe).

6) The decision to kill Abab (in background).

7) John Doe's decision whether or not to deploy the Crawler (in storyline 8).

B.1.5 Titrations

1) What if we sent in multiple bots?

2) What if we do have communication with the bot?

3) What if the target is not a human, but for example detonate a nuclear or chemical weapon?

4) What if the weapon is non-lethal, e.g., if the crawler can release sleeping gas?

5) What if the crawler is a prototype which has not been tested?

6) What if Company R is a startup company without a track record?

B.1.6 Dimensions

1) Level of automation (manual, full autonomous, authority to use weapons).

2) Time criticality (how much time is available given that target might escape).

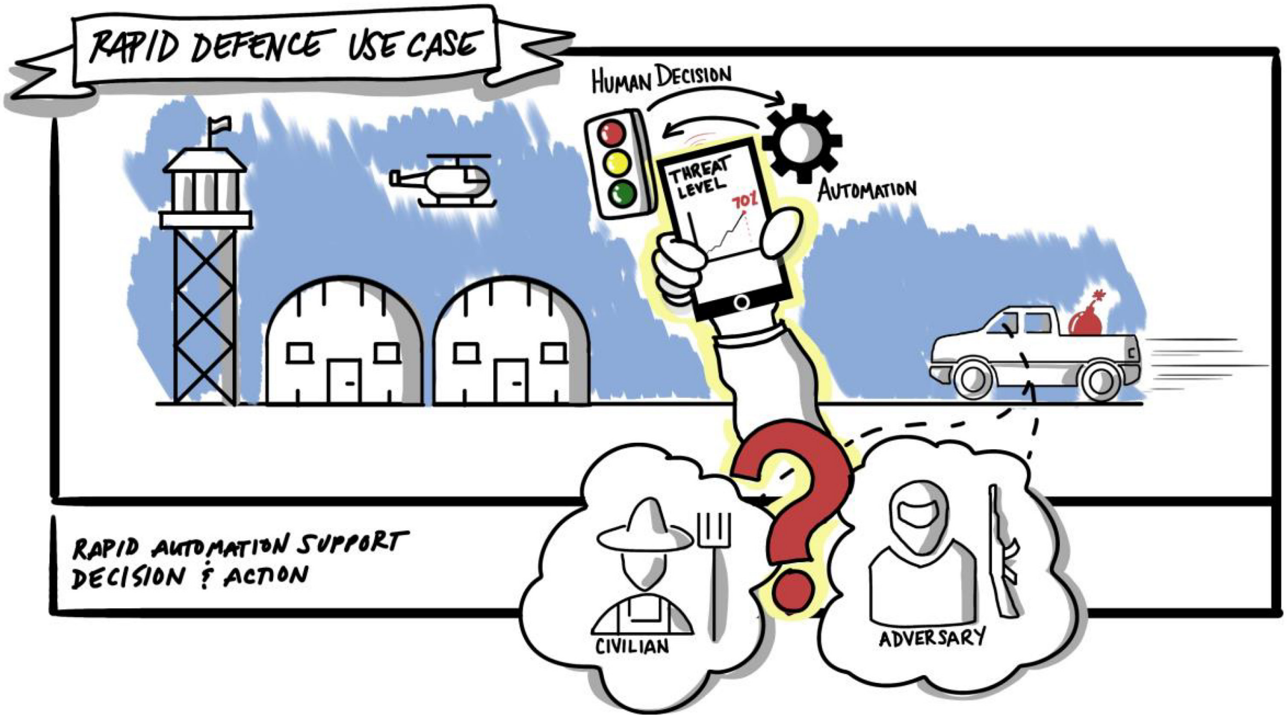
3) Risks (what are the risks of collateral damage; e.g., are there any other people).

4) Reliability of the system (accuracy of target identification).

B.1.7 References

The Hague Centre for Strategic Studies (2019). Towards Responsible Autonomy – The Ethics of RAS in a Military Context, HCSS, https://hcss.nl/report/towards-responsible-autonomy-ethics-ras-military-context

B.2 MHC SCENARIO 2: RAPID DEFENCE USE CASE: RAPID AUTOMATION SUPPORTED DECISION AND ACTION



B.2.1 Context and Background

Bangawa is responsible for conducting regular patrols in the mountainous regions of Mohabwi to show military presence and promote safety in this hard-to-reach landscape.

Bangawa is holding a protected base in the urban areas of Mohabwi populated with several hundred people as **base population**; commander is one **Major Jane Doe**, most decisions need to be approved by her **HQ** several kilometers away. Due to the strong defences of this base, in the last months Mohabwi forces resorted to guerilla-like tactics to attack Bangawa. Frequently these attacks involve disguised vehicles nearly indistinguishable from **civilian vehicles**. These are then driven by **adversary drivers** into the proximity of the base at high speeds and have caused severe damage in the past months (death and destruction). The attacks are so successful because the vehicles blend in with the regular traffic and only in the last moment accelerate and try to breach defences.

This gives the defenders very little time for defensive actions, also it is very strenuous for the **lookouts** and they need to rotate frequently to retain their effectiveness. For engaging light vehicles, the protected locations are already equipped with traditional soldier operated weaponry from the company **EffectEngineering (Division A)** that can take out approaching vehicles with minimal damage to surrounding civilians or structures. However, the rapid nature of these attacks oftentimes does not leave enough time for recognizing the threat and activating the defence system.

To protect these bases better, an upgrade for the defence system is installed by the company **EffectEngineering (Division B)**, called ***RivalReveal***. Using AI-based functions, the novel system continuously monitors the environment around the base in order to identify threats. Technology used is camera image based. Based on multiple factors (type of vehicle, long and short-term behavior of vehicle, changes in behavior, occupants, possible occluded cargo, etc.) a threat probability is calculated ranging from 1 to 99 percent. A collocated soldier is presented with the probability the system has calculated. The manufacturer advertises the system as a (learning) decision support tool. The neural network that is used to calculate the probability was trained by the company **EffectEngineering (Division C)** using scenarios that have been reported at other locations but were considered to be comparable by the company and have been discussed with **Major Jane Doe**.

The company **EffectEngineering (Division A)** offers the integration of ***RivalReveal*** from **EffectEngineering (Division B)** with the weaponry that offers the possibility of fully automated defence actions. This functionality, called ***ResoluteResponse***, will auto-engage every target that ***RivalReveal*** has calculated a probability of larger than 70%. When probability is between 5% to 70%, the lookout is notified and a confirmation by the lookout is needed before auto-engage.

B.2.2 Actors

a) Location lookout.

b) Civilian drivers.

c) Adversary drivers.

d) Base population.

e) Company *EffectEngineering (Divisions A and B).*

f) Major Jane Doe.

g) Major Jane Doe's HQ.

h) Automated adversary detection system *RivalReveal.*

i) Automated weapon system *ResoluteResponse.*

j) *Neural Network Trainers.*

k) *Scenario Developers.*

B.2.3 Storyline

Day 1

1) Base Commander Jane Doe requests *RivalReveal* from EffectEngineering (Division B).

2) HQ approves request.

3) *RivalReveal* is installed at protected base.

4) Personnel, i.e., lookouts are trained in a few days, and hence are not well trained with *RivalReveal.*

5) *RivalReveal* is in operation.

6) Adversary vehicle approaches base. Due to speed and distance, it would take a very short time of a few seconds to reach the base.

7) *RivalReveal* calculates a probability of 75% that the vehicle is an adversary vehicle and notifies lookout.

8) Lookout is unable to engage vehicle with weapon and 20 people at the base are killed at the attack.

Figure B-1: 1998 Bombings of US Embassy in Tanzania and Kenya. Source: Agence France-Presse.

Time Period 2 (a few days after day 1)

9) Base Commander Jane Doe requests *ResoluteResponse* from *EffectEngineering* (Division A).

10) HQ approves request.

11) *ResoluteResponse* is installed at protected base.

12) Personnel, i.e., lookouts are trained in a few days, and hence are not well trained with *ResoluteResponse.*

13) *ResoluteResponse* is in operation.

14) Lookout activates *ResoluteResponse.*

15) Adversary vehicle approaches base. Due to speed and distance a few seconds before it could reach the base…

16) *RivalReveal* calculates a probability of 80% that the vehicle is an adversary vehicle and notifies *ResoluteResponse*.

17) *ResoluteResponse* successfully disables the adversary vehicle. The adversary driver is killed, no civilian is harmed

Time Period 3 (a few days after day 2)

18) Civilian vehicle approaches base. Due to speed and distance 3 seconds before it could reach the base. Unknown to the lookout is that the driver is a delivery person and was just called to hurry with urgent documents, being picked up from a building in the proximity of the base.

19) RivalReveal calculates a probability of 61% that the vehicle is an adversary vehicle and notifies lookout, ResoluteResponse sets aim and waits for confirmation by lookout.

20) The lookout confirms that the vehicle is an adversary vehicle, although he is completely unsure but due to the own casualties on day 1 and the successful usage of the system on day 2, he trusts RivalReveal and ResoluteResponse enough to engage the system.

21) ResoluteResponse successfully disables the civilian vehicle. The civilian driver is killed, but no other civilian is harmed.

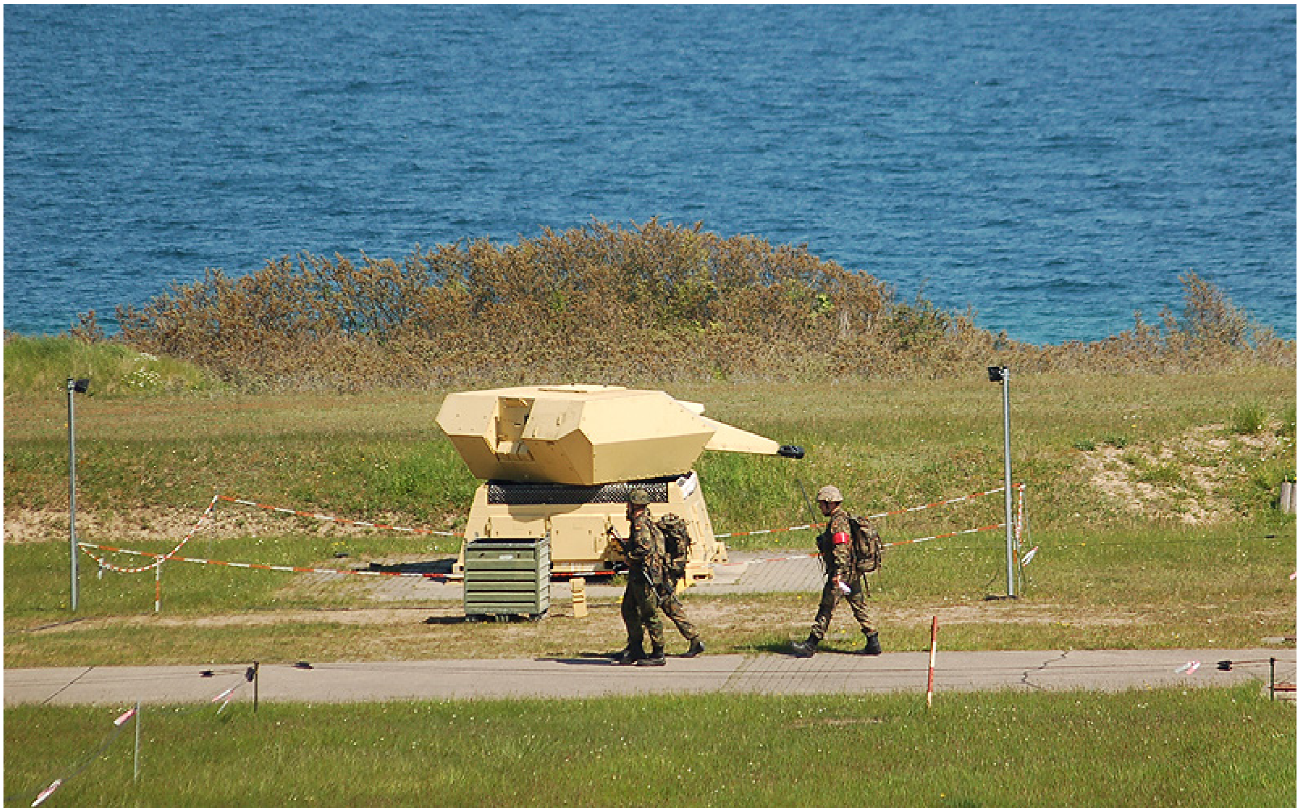


Figure B-2: MANTIS (Modular, Automatic and Network Capable Targeting and Interception System). Source: Deutscher Bundeswehr Verband.

B.2.4 Critical Decision Points

1) Decision of *EffectEngineering* to offer *RivalReveal* (an imperfect decision support system).

2) Decision of Major Doe on scenarios comparable to her situation.

3) Decision of *EffectEngineering* on scenarios used for training the neural network.

4) Decision of policymaker to allow the distribution of *RivalReveal.*

5) Decision of Major Doe to order *RivalReveal.*

6) Decision of HQ to approve request of *RivalReveal.*

7) Decision of Major Doe on training plan concerning *RivalReveal* (e.g., to also address overtrust, e.g., the problem of automation bias).

8) Decision of Major Doe to set *RivalReveal* into operation.

9) Decision of lookout to use or ignore information from *RivalReveal.*

10) Decision of *EffectEngineering* to offer *ResoluteResponse* (an automated effector system that bases its activation on imperfect information).

11) Decision of policymaker to allow the distribution of *ResoluteResponse.*

12) Decision of Major Doe to order *ResoluteResponse.*

13) Decision of HQ to approve request of *ResoluteResponse.*

14) Decision of Major Doe on training plan concerning *ResoluteResponse* (e.g., when to activate or deactivate).

15) Decision of Major Doe to set *ResoluteResponse* into operation (although incorrect detections of *RivalReveal* can lead to civilians being killed by an automated effector system).

16) Decision of lookout to activate or deactivate *ResoluteResponse.*

17) Decision of lookout to trust *ResoluteResponse* and confirm or not confirm decisions.

B.2.5 Titrations

1) Protected base is an embassy, therefore populated with more civilians than military personnel:

   ➔ Who needs more protection: a civilian or a soldier?

2) Instead of the same company, EffectEngineering becomes divided in several players:

   a) EffectEngineering (Division B) → Company ReconRobotics

   b) EffectEngineering (Division C) → TinheadTraining:

      ➔ Is the functionality between the systems trustworthy enough, since *EffectEngineering* and *ReconRobotics* are competitors? Who is responsible for the complete system *EffectEngineering ReconRobotics, TinheadTraining,* the Government for this procurement setup, Major Doe for using such a complicated system with inexperienced personnel?

3) Known vs. unknown errors: Storyline, between Day 1) Point 4 + 5: operators experience a couple of incidents in which RivalReveal flagged obviously harmless vehicles above 50% threat probability:

   ➔ Is it ethical to have a weapon system with known flaws?

4) Soldiers training / experience with the weapon system:

   a) 1 day

   b) 7 days

   c) 1 month

   d) 1 year

5) Soldiers experience with the adversary situation described in the background:

   a) 1 day

   b) 7 days

   c) 1 month

   d) 1 year

6) *ResoluteResponse* uses non-lethal effects.

7) Major Doe can choose whether to install lethal or non-lethal effects.

8) Lookout can access information relevant for the probability calculation and manipulate several variables to include also own observations:

   a) Experienced soldier low knowledge/affinity to technology /AI

   b) Experienced soldier high knowledge/affinity to technology / AI

   c) Unexperienced soldier low knowledge/affinity to technology /AI

   d) Unexperienced soldier high knowledge/affinity to technology / AI

B.2.6 Dimensions

1) Complexity of system: modules controlled build by one or by two companies?

2) Complexity of situation: driving behavior and vehicle type as only source for algorithm, mix of civilians and adversaries, etc.

3) Risk/cost of engagement: Lethality of automation behavior.

4) Understanding of automation behavior: how is the probability calculated how biased does a soldier become?

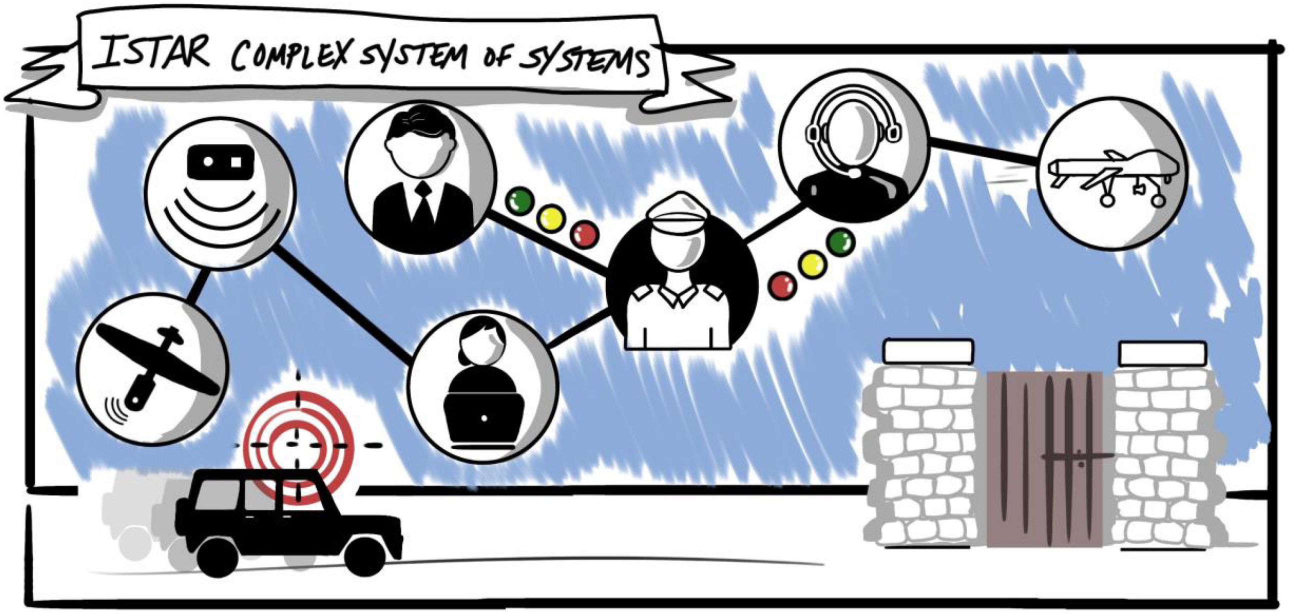
5) Temporal: rapid decisions necessary.

6) Experience/training with the system.

7) Life value tradeoff: own civilians (embassy), own troops vs. other civilians, adversaries.

8) Freedom of choice: disregard order or have the choice of using *RivalReveal* or *ResoluteResponse.*

B.3 MHC SCENARIO 3: ISTAR COMPLEX SYSTEMS OF SYSTEMS



B.3.1 Context and Background

A future ISTAR (Intelligence, Surveillance, Target acquisition and Reconnaissance) system consists of a central hub which takes in ISTAR data and information from a multitude of sensors and sources across the battlespace. A suite of AI enabled tools supports the Intelligence cell to process, fuse and analyze these sources to create Intelligence products which are used to inform planning and decision making activities at tactical and operational levels. AI enabled tools are also used to manage the Information Requirements process and the planning and execution of the ISTAR Collection Plan.

Sources include a mix of Electro Optic (EO), Electronic Warfare (EW), Human Intelligence (HUMINT), Open Source Intelligence (OSINT) with varying degrees of local (in system) processing, some of which is conducted by AI-based systems (e.g., object DRI, fusion of organic information and deconfliction with other sources, etc.) before being used locally and/or globally via the central hub.

Intelligence products may be single or multi-source and can be near real time (e.g., Recognized picture, identification of warnings and indicators) and some are analysis of multiple sources to draw inferences over adversary intent and future activity.

Depending on communications bandwidth and connectivity the quantity, quality and timeliness of information being fed into and available from the Hub will vary dynamically resulting in disparities in information availability across the battlespace.

In this scenario a NATO coalition including the nations of Apalagio and Barot are engaged in a UN sanctioned operation to protect the population of Casab from the despotic military leader Maj Gen Boagart who overthrew the democratically elected government and is engaging in a genocidal campaign against minority groups in the country. At 1900 an Apalagio armored reconnaissance squadron consisting of a mix of manned and unmanned platforms and static sensors is conducting a recce task ahead of the main force. A covert long duration UAV deployed by the recce squadron is programmed to loiter over a suspected enemy location to identify and report enemy activity. Due to the risk of detection by enemy EW assets the UAV has an on-board system (DRI AI) to process EO imagery to detect, recognize and identify objects and only transmits the location, identification and a confidence value placed on that identification to the recce squadron. It records the imagery collected for later analysis should it be required. An AI system (RECCE AI) on-board the lead recce vehicle combines information received from the UAV with other local sensors (EW and acoustic) to create a picture of enemy forces and disposition and over time the AI system combines information to create movement tracks and remove duplication. This local intelligence picture is fed in near real time into the Central ISTAR Hub incorporating AI agents (HUB AI) where it is available to inform the production of further ISTAR products.

B.3.2 Actors

Humans

1) Apalagio Reconnaissance squadron ISTAR Analyst.

2) Apalagio HQ Intelligence Analyst.

3) Apalagio Commander Current Ops / HQ Commander.

4) Apalagio Targeting cell.

5) Apalagio Legal Advisor (LegAd).

6) Apalagio Advisors to HQ Commander.

7) Apalagio Strategic Command.

8) Apalagio Defence Minister.

9) Barot Armed UAV controller.

10) Casab High Value Target (Maj Gen Bogart and Bodyguard).

11) Casab Local Farmers.

Technology

1) Apalagio Covert long duration UAV including autonomous DRI capability (DRI AI) – The UAV is organic to the Reconnaissance squadron.

2) Apalagio Reconnaissance squadron sensor systems (EW, EO).

3) Apalagio Reconnaissance squadron AI ISTAR Fusion System (RECCE AI).

4) Apalagio Central ISTAR Hub including AI support system (HUB AI).

5) Apalagio Targeting cell AI planning support tool (PLAN AI).

6) Barot Armed UAV.

7) Casab Farmers 4 x 4 Vehicles.

8) Casab G-Wagon 4 x 4 Vehicles.

B.3.3 Storyline

1) During the recce, the Apalagio UAV detects two moving vehicles which the DRI AI identifies as wheeled 4x4 vehicles and transmits this information to the recce squadron's ISTAR vehicle.

2) The Apalagio RECCE AI correlates the position of the 4x4 with a suspect EW detection on the same bearing and assigns a high probability that the vehicle is an enemy vehicle.

3) This information is transmitted to the central Apalagio ISTAR hub.

4) An Apalagio Intelligence analyst receives this new information and receives an alert from the HUB AI support system that HUMIT indicated that Casab Maj Gen Bogart, a high value target was recently seen getting into his prized armored G-Wagon 4x4 and accompanied by his personal bodyguard in another unidentified 4x4 vehicle at a location which would correlate with the current position of the detected 4x4 vehicles.

5) The Apalagio HUB AI recommends to the Intelligence analyst that there is a high probability that these vehicles are carrying the high value target. This is immediately flagged by the Intelligence analyst to the Apalagio commander who, keen not to miss the opportunity to eliminate this high value target, immediately tasks the Apalagio targeting cell to develop a plan to destroy the target while he evaluates the intelligence information.

6) The Apalagio ISTAR HUB AI passes the location and nature of the targets to the Apalagio targeting cell who's PLAN AI system identifies an Armed Barot UAV (under the command and control of the Apalagio HQ) just within range of the last identified position of the potential targets but would require tasking immediately to be able to complete the task without running out of fuel.

7) The ISTAR HUB AI passes the location and target characteristics (two 4x4 vehicles in convoy) to the UAV. The commander gives authorization to task the UAV while he continues to evaluate the information available to him with Legal Advisor (LegAd) and other advisors in the HQ. The LegAd decides that he must refer the decision for the approval of superiors in Strategic Command.

8) Strategic command decide that due to the potential ramifications of killing Bogart requires approval of the Defence Minister.

9) Once the Defence Minister is reached, he immediately authorizes the mission, but insists that the risk of the strike for civilian casualties as collateral damage must be assessed as "minimal" before the mission can be prosecuted.

10) A high altitude surveillance asset observes the 4x4 vehicles stopping at a walled compound in a small town and the occupants entering the compound.

11) A Collateral Damage Assessment is accomplished (under these ROEs) by a Apalagio Targeting cell officer using a software-based Collateral Damage Estimation (CDE) tool to estimate explosion potential from a missile strike at various coordinates. Since the compound is in the midst of the city, there are many risks even from a precision strike in a walled compound. Initial estimates of collateral casualties come back as 65-75%, which was deemed unacceptable.

12) The Apalagio commander evaluates that it is probable that Casab Maj Gen Bogart is in the vehicles identified, is a valid target and authorizes a strike while the targets are stationary and inside the compound.

13) Knowing that the UAV will run out of fuel imminently and a strike cannot be delayed the Apalagio commander takes the risk assessment officer aside and tells him that he must find a way to reduce the risk below 50%.

14) He reruns his numbers and reports that by targeting a distant area of the compound, the risk will be “45 – 65%”.

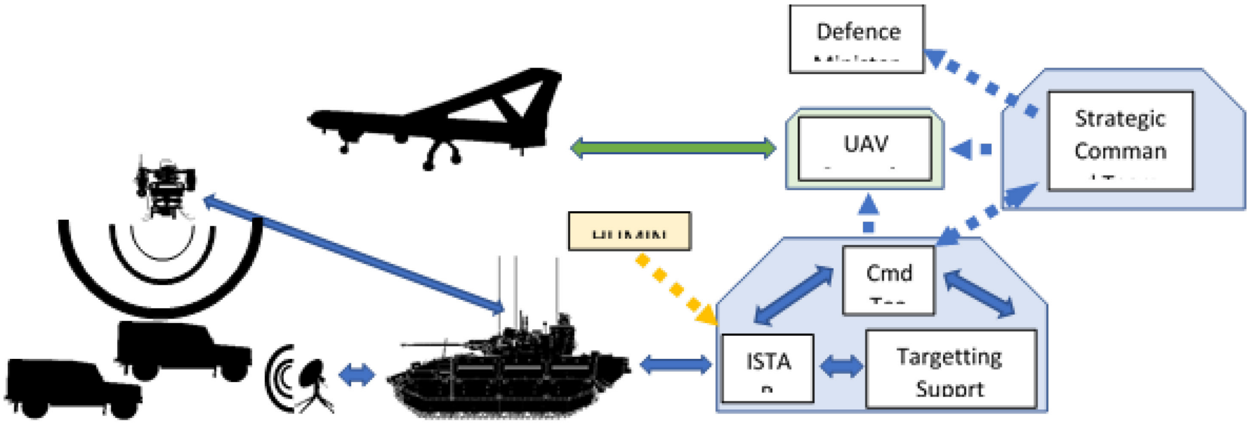
15) He reports to Strategic Command that the risk is “45%” and authorization is given for the strike.

16) The Apalagio targeting team passes the target details to the Barot UAV controllers to conduct the strike, the Barot UAV control team are not able to interface with the targeting cell directly so this information is passed over the voice network and inputted into the Armed Barot UAV control system and the UAV tasked.

17) Meanwhile the Apalagio recce UAV returns to the recce squadron and the imagery is downloaded.

18) The Apalagio RECCE AI analyses that imagery and identifies sections of interest to the Recce Squadron ISTAR analyst who reviews the video. The potential target vehicles are identified by the analyst using image enhancement as being 4x4 vehicles of a type typically used by local farmers and tags them as likely civilians, this information is added to the local intelligence picture and sent to the Apalagio ISTAR HUB AI.

19) The Barot UAV pilot fires his missile, resulting in civilian casualties



B.3.4 Critical Decision Points

1) Detection of two moving vehicles and Recognition of them as 4x4s by Covert long duration UAV including autonomous DRI capability.

2) Correlation by Reconnaissance squadron AI ISTAR Fusion System of the 4x4s with suspect EW detection and identification of the 4x4s as probable enemy.

3) Correlation by Central ISTAR Hub including AI support system of HUMINT report of Maj Gen Bogart traveling in two 4x4’s with the reported suspected two enemy 4x4s.

4) Evaluation of information by HQ Current Ops Commander, advisors and LEGAD

5) The request and authorization of lethal force

6) Decision to "lean on" the targeting cell to provide Collateral Damage Assessment numbers that will permit the strike

7) Decision to report a less than fully honest assessment of risks and the targeting cell's decision to not challenge the report.

8) Communication of ISTAR information to Targeting cell AI planning support tool and use in tasking lethal effect.

9) Analysis of Reconnaissance squadron AI ISTAR Fusion System and determination of probable enemy 4x4s as actually being farming vehicles.

10) Flow through of this corrected assessment to abort the strike.

B.3.5 Titrations

Contextual Titrations

1) Change the targeted elimination (a person) to a targeted destruction of a military convoy transporting air defence systems to a certain location that in turn would weaken the ability to operate in the region.

2) They are civilians maybe terrorists transporting something imagery hints to be weapons.

3) They are military vehicles imagery doesn't help identify what is transported, and in fact they are transporting medical supplies / food, etc.

4) The high value target is actually traveling in the convoy but with non-combatants.

5) Explore different legal aspects:

   a) UN Charta: article 2, IV vs. article 51.

   b) International armed conflict with the nation of Maj Gen Bogart in the territory of a third country asking for our support.

   c) International armed conflict with the nation of Maj Gen Bogart in the territory of a third country **not** asking for our support.

Decision Point Titrations

1) What if the intelligence analyst does not add the additional HUMINT that the target is in a G-Wagon and the nature of the target as 2 x 4x4 vehicles persists, opportunities to identify that this is an incorrect target would be lost during the engagement.

2) How is corrected/revised analysis promulgated through the system in a timely manner.

3) How is incorrect info purged from the system, how is decision lock and confirmation bias managed/minimized.

4) The analyst is unsure what kind of vehicle it is because imagery is blurry.

5) Uncertainty balance varies between systems:

   a) System A's certainty (good UAV image of target vehicle and a person looking like May Gen B entering it).

   b) System B's certainty (intel believes Maj Gen B is actually in a different region).

   c) Fused in the ISTAR hub that it is highly certain that Maj Gen B is in the vehicle in the image.

   d) AI-based image enhancement is used prior to human image analysis.

"Eye in the Sky" Titrations

1) *Human Authorization for Weapon Release* – Take the above situation but envision an armed UAV with sensors and facial recognition capabilities which identifies the target(s), matches them against a high-priority threat index, and follows its ROE policies to request weapons release from its human commanders. The scenario then unfolds as above, with the human organizational elements operating to decide whether or not to grant authorization, including the incident of Col. Powell "encouraging" an outcome that permits weapon release. Is MHC enhanced or diminished by removing the human Reaper crew in this scenario relative to the first variant?

2) *Automated Risk Assessment* – As for variant #2, but now the risk assessment software has been installed on the automated UAV. In addition to detecting high value targets and knowing that this motivates a strike, the UAV provides a range of targeting options representing a tradeoff space of likelihood of target kill vs. collateral damage. As before, Col. Powell explores targeting options that produce a 45 – 65% collateral damage estimate, and then authorizes weapon launch under those parameters.

3) *a priori Strike Authorization* – Since each of the prior variants involved human intervention to arguably, unethically sway automated decision making, this scenario diminishes that capability. Assume the sum of the automation capabilities from the prior scenarios. Further, assume that the terrorists were known, a priori, to have jamming capabilities that could deny communications. Finally, assume that when the micro-UAV feed was dying and the suicide bombers were about to leave, a policy was agreed to at the highest levels of the allied forces saying that if comms were lost and the terrorists were detected leaving the building and the probability of collateral casualties was determined to be less than 50%, the vehicle was authorized to fire.

4) *Increased Collateral Damage Sensitivity* – In fact, the movie plot this scenario is based on is more elaborate than presented here. In the movie, throughout much of the scenario, the Reaper pilots have been viewing imagery of a little girl is selling bread outside the wall around the compound where the terrorists are preparing their attack. They have previously observed this girl playing in her own home – in fact, with an improvised hula hoop made by her father – a sign that they are not hardcore supporters of the al-Shabab regime. The final round of risk assessment, in which Powell influences the risk assessment officer to bring the odds below 50%, is brought on by the Reaper pilot who can see that the girl will likely be hurt by the blast. He demands (within his authority in the chain of command) that the risk assessment be updated, trying to both give her time to get away and/or to protect her. In the end, with Powell's results, he follows orders to launch his missile(s) only to see her killed in the resulting explosions. This is clearly Hollywood trying to humanize the situation and tug on our emotions, but as a titration does/should the presence of the girl make a difference in MHC?

B.3.6 Dimensions

1) Complexity of system.
2) Interaction of multiple AI systems.
3) Interaction of multiple nations.
4) Connectivity limitations.
5) AI and non-AI enabled interactions.
6) Information ambiguity.
7) Complexity of situation.
8) Time pressure.
9) Risk/Cost of engagement.
10) Human Motivation and Interpretation of Intent/ROEs.
11) Strategic/Tactical operational cooperation.
12) Complex and diffuse chain of command.

B.3.7 References

Wikipedia. Synopsis, Eye in the Sky (2015). https://en.wikipedia.org/wiki/Eye_in_the_Sky_(2015_film)

Annex C – HIGHLIGHTS REPORT: MEANINGFUL HUMAN CONTROL OF AI-BASED SYSTEMS: KEY CHARACTERISTICS, INFLUENCING FACTORS AND DESIGN CONSIDERATIONS (HFM-322)

Chaired by Dr. Jurriaan van Diggelen (TNO, the Netherlands), and Dr. Mark Draper (AFRL, USA)

C.1 BACKGROUND

This activity addresses an important issue identified in the Specialists' Meeting SCI-296 on Autonomy from a System Perspective, held in May 2017 as part of the STO theme devoted to that topic. As noted in the SCI-296 TER, "in many or most cases, it is foreseen that 'Meaningful Human Control' (MHC) will be mandated, necessitating the human to maintain awareness and 'drill down' on demand". Responding to this need, the HFM Panel commissioned an exploratory team (HFM-178) to rapidly assess the area from a human-centric perspective. This team came to a consensus as to a working description of MHC, which is "Humans have the ability to make informed choices in sufficient time to influence AI-based systems in order to enable a desired effect or to prevent an undesired immediate or future effect on the environment." This team also canvased MHC from several dimensions and settled on the need for a dedicated expert-heavy workshop to unpack the most pressing influencing factors. The current proposed activity integrates several research issues emerging from SCI-296, especially those combining humans, (technical) systems, organization and behavior. Since meaningful human control is deemed to be important for many kinds of automated and (semi)autonomous systems, the term "AI-based systems" is used to encompass all AI-based forms of automation and autonomy, for tasks that are either physical (e.g., unmanned platforms) or informational (e.g., big data analytics, decision support). Given the implications of MHC for the latter application domain, this TAP is also relevant for the STO theme "Big data and AI for military decision making."

C.2 MILITARY RELEVANCE

Given the current exponential developments in the field of AI and in civil applications such as drones, autonomous driving, personal assistants and game players, and given the almost unbridled proliferation of AI technology, there is an urgent need to develop AI-based military capabilities. On the one hand, there are significant lessons that can be learned from civil applications. On the other hand, there are significant differences between military and civil applications and requirements. Intelligent systems operating in a military context must, for example, withstand adverse conditions and deliberate adversarial actions, and remain within bounds set by international law and rules of engagement. Moreover, systems should be adaptive: while requirements are relatively constant in the civil domain, they must follow the dynamics of pre-war, war and battlefield in the military domain, where in the course of escalation and de-escalation, rules of engagement have to be dynamically adjusted.

C.3 OBJECTIVE(S)

The core objective of this Workshop is not to duplicate the ongoing efforts at the national and international level in the legalities and ethics of MHC. Rather, it is to learn from these ongoing discussions, apply a perspective to the problem squarely rooted in human factors and cognitive science understanding, and thus distil a set of practical human-centered guidelines to inform future NATO actions in this increasingly important area. Given the multi-faceted nature of MHC, six Themes were chosen for deep-dive investigation during this Workshop. Each Participant was assigned to explore one of these Themes via small Theme-focused breakout sessions. In addition, there was also a lot of cross-theme discussion throughout the Workshop. The Themes were:

1) HSI, Organizational, and Operational Considerations of MHC.
2) Human Factors-Inspired Design Guidelines to Achieve MHC.
3) Systems Engineering Methods and Metrics to Validate MHC.
4) Adversary Exploitation of MHC.
5) Complex Socio-Technical Systems.
6) Moral Responsibility in Human-AI Teams.

The results of this Workshop will directly inform recommendation of highly focused follow-on activities that inform NATO on how to identify, achieve, maintain, and regain MHC across a wide range of AI applications. The symposium was also essential to reactivate the network of collaboration and study within NATO during the Covid lockdown period.

C.4 S&T ACHIEVEMENTS

The NATO symposium on “Meaningful Human Control of AI-based Systems: Key Characteristics, Influencing Factors and Design Considerations” was conducted in Germany, Berlin on 25-27 October 2021 in a novel hybrid format. The hybrid format included six power sessions, i.e., the HFM-322 group in Berlin interviewing an expert panel of three or four experts participating online. The power sessions focused on each of the themes described above and resulted in a synopsis for each theme highlighting key challenges, possible controversy, and state of the art within that theme. Furthermore, a professional cartoonist was present at the meetings to represent the results visually. Furthermore, three keynote presentations were given:

1) Missy Cummings, who focused on recent developments in AI, how they relate to the use of AI on the battlefield, and how we need to shift the focus from debating bans to a discussion of meaningful certification;
2) General Gäbelein who focused on a chain of trust for the effective accomplishment of their defence mission. The chain of trust includes society, policymakers, the armed forces and the system of deployed military personnel and their weapon systems; and
3) Daniele Amoroso who focused on the issue of filling the “Meaningful Human Control” (MHC) placeholder with more precise content is primarily a normative problem rather than a technical one.

C.5 SYNERGIES AND COMPLEMENTARITIES

The workshop brought together perspectives from a wide range of NATO countries (USA, UK, Germany, the Netherlands, France, Italy, Sweden, and Australia). There was also an important and productive synergy with HFM-330 as productions from this workshop will be followed up upon in HFM-330. In particular, theme 1, 2 and 3 will be a focus point for HFM-330.

C.6 EXPLOITATION AND IMPACT

The workshop was evaluated by the TER as very successful, given the number of experts interviewed within such a short timeframe, the wide range of topics that were addressed, and the way these topics were synthesized in sketches, and understandable synopses. Also, the successes of the novel hybrid workshop format were recognized.

C.7 CONCLUSION(S)

Meaningful human control is about ensuring moral responsibility and agency of the human in military use of AI-based, and autonomous systems. It is a multi-dimensional problem, which is highly dependent on context, stakeholders involved, and types of AI systems used. On 25-27 October, we organized a hybrid workshop in Berlin, focusing on a range of interrelated themes regarding MHC. This workshop has provided crucial insights in the landscape of MHC research. These insights will be reported in a synopsis for each theme, and directly inform current and future activities of NATO STO.

There's no silver bullet for achieving meaningful human control. It requires a continuous effort of banning certain types of systems, developing the right types of human-machine teams, and developing computational moral models.

Annex D – PHOTOS FROM THE HFM-322 WORKSHOP

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's References	3. Further Reference	4. Security Classification of Document
	STO-MP-HFM-322 AC/323(HFM-322)TP/1108	ISBN 978-92-837-2425-4	PUBLIC RELEASE

5. Originator

Science and Technology Organization
North Atlantic Treaty Organization
BP 25, F-92201 Neuilly-sur-Seine Cedex, France

6. Title

Meaningful Human Control of AI-Based Systems Workshop: Technical Evaluation Report, Thematic Perspectives and Associated Scenarios

7. Presented at/Sponsored by

The technical evaluation report describes the main findings and conclusions of the HFM-322 workshop. The annexes contain themed synopses and representative scenarios.

8. Author(s)/Editor(s)	9. Date
Christopher Miller, Mark Draper, Jurriaan van Diggelen, Marlijn Heijnen, Robert J. Shively, Frank Flemisch, Marcel Baltzer, Rogier Woltjer, Mike Boardman, Kate Devitt, Marie-Pierre Pacaux-Lemoine, and Emma Parry.	June 2023

10. Author's/Editor's Address	11. Pages
Multiple	78

12. Distribution Statement

There are no restrictions on the distribution of this document. Information about the availability of this and other STO unclassified publications is given on the back cover.

13. Keywords/Descriptors

AI ethics; Artificial Intelligence (AI); Command and control; Human factors; Human systems integration; Human-autonomy teaming; Human-machine teaming; Meaningful human control; Military AI; Responsible AI

14. Abstract

This report includes the Technical Evaluation Report (TER) describing the main findings and conclusions of the HFM-322 workshop. The annexes include six theme-specific synopses from the workshop written by the workshop organization team, and three representative scenarios that were used as operational context for studying meaningful human control.

NORTH ATLANTIC TREATY ORGANIZATION

BP 25
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@cso.nato.int

SCIENCE AND TECHNOLOGY ORGANIZATION

DIFFUSION DES PUBLICATIONS
STO NON CLASSIFIEES

Les publications de l'AGARD, de la RTO et de la STO peuvent parfois être obtenues auprès des centres nationaux de distribution indiqués ci-dessous. Si vous souhaitez recevoir toutes les publications de la STO, ou simplement celles qui concernent certains Panels, vous pouvez demander d'être inclus soit à titre personnel, soit au nom de votre organisation, sur la liste d'envoi.

Les publications de la STO, de la RTO et de l'AGARD sont également en vente auprès des agences de vente indiquées ci-dessous.

Les demandes de documents STO, RTO ou AGARD doivent comporter la dénomination « STO », « RTO » ou « AGARD » selon le cas, suivi du numéro de série. Des informations analogues, telles que le titre est la date de publication sont souhaitables.

Si vous souhaitez recevoir une notification électronique de la disponibilité des rapports de la STO au fur et à mesure de leur publication, vous pouvez consulter notre site Web (http://www.sto.nato.int/) et vous abonner à ce service.

CENTRES DE DIFFUSION NATIONAUX

ALLEMAGNE
Streitkräfteamt / Abteilung III
Fachinformationszentrum der Bundeswehr (FIZBw)
Gorch-Fock-Straße 7, D-53229 Bonn

BELGIQUE
Royal High Institute for Defence – KHID/IRSD/RHID
Management of Scientific & Technological Research
for Defence, National STO Coordinator
Royal Military Academy – Campus Renaissance
Renaissancelaan 30, 1000 Bruxelles

BULGARIE
Ministry of Defence
Defence Institute "Prof. Tsvetan Lazarov"
"Tsvetan Lazarov" bul no.2
1592 Sofia

CANADA
DGSIST 2
Recherche et développement pour la défense Canada
60 Moodie Drive (7N-1-F20)
Ottawa, Ontario K1A 0K2

DANEMARK
Danish Acquisition and Logistics Organization
(DALO)
Lautrupbjerg 1-5
2750 Ballerup

ESPAGNE
Área de Cooperación Internacional en I+D
SDGPLATIN (DGAM)
C/ Arturo Soria 289
28033 Madrid

ESTONIE
Estonian National Defence College
Centre for Applied Research
Riia str 12
Tartu 51013

ETATS-UNIS
Defense Technical Information Center
8725 John J. Kingman Road
Fort Belvoir, VA 22060-6218

FRANCE
O.N.E.R.A. (ISP)
29, Avenue de la Division Leclerc
BP 72
92322 Châtillon Cedex

GRECE (Correspondant)
Defence Industry & Research General
Directorate, Research Directorate
Fakinos Base Camp, S.T.G. 1020
Holargos, Athens

HONGRIE
Hungarian Ministry of Defence
Development and Logistics Agency
P.O.B. 25
H-1885 Budapest

ITALIE
Ten Col Renato NARO
Capo servizio Gestione della Conoscenza
F. Baracca Military Airport "Comparto A"
Via di Centocelle, 301
00175, Rome

LUXEMBOURG
Voir Belgique

NORVEGE
Norwegian Defence Research
Establishment
Attn: Biblioteket
P.O. Box 25
NO-2007 Kjeller

PAYS-BAS
Royal Netherlands Military
Academy Library
P.O. Box 90.002
4800 PA Breda

POLOGNE
Centralna Biblioteka Wojskowa
ul. Ostrobramska 109
04-041 Warszawa

PORTUGAL
Estado Maior da Força Aérea
SDFA – Centro de Documentação
Alfragide
P-2720 Amadora

REPUBLIQUE TCHEQUE
Vojenský technický ústav s.p.
CZ Distribution Information Centre
Mladoboleslavská 944
PO Box 18
197 06 Praha 9

ROUMANIE
Romanian National Distribution
Centre
Armaments Department
9-11, Drumul Taberei Street
Sector 6
061353 Bucharest

ROYAUME-UNI
Dstl Records Centre
Rm G02, ISAT F, Building 5
Dstl Porton Down
Salisbury SP4 0JQ

SLOVAQUIE
Akadémia ozbrojených síl gen.
M.R. Štefánika, Distribučné a
informačné stredisko STO
Demänová 393
031 01 Liptovský Mikuláš 1

SLOVENIE
Ministry of Defence
Central Registry for EU & NATO
Vojkova 55
1000 Ljubljana

TURQUIE
Milli Savunma Bakanlığı (MSB)
ARGE ve Teknoloji Dairesi
Başkanlığı
06650 Bakanliklar – Ankara

AGENCES DE VENTE

The British Library Document
Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
ROYAUME-UNI

Canada Institute for Scientific and
Technical Information (CISTI)
National Research Council Acquisitions
Montreal Road, Building M-55
Ottawa, Ontario K1A 0S2
CANADA

Les demandes de documents STO, RTO ou AGARD doivent comporter la dénomination « STO », « RTO » ou « AGARD » selon le cas, suivie du numéro de série (par exemple AGARD-AG-315). Des informations analogues, telles que le titre et la date de publication sont souhaitables. Des références bibliographiques complètes ainsi que des résumés des publications STO, RTO et AGARD figurent dans le « NTIS Publications Database » (http://www.ntis.gov).

NORTH ATLANTIC TREATY ORGANIZATION

BP 25
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@cso.nato.int

SCIENCE AND TECHNOLOGY ORGANIZATION

DISTRIBUTION OF UNCLASSIFIED STO PUBLICATIONS

AGARD, RTO & STO publications are sometimes available from the National Distribution Centres listed below. If you wish to receive all STO reports, or just those relating to one or more specific STO Panels, they may be willing to include you (or your Organisation) in their distribution.

STO, RTO and AGARD reports may also be purchased from the Sales Agencies listed below.

Requests for STO, RTO or AGARD documents should include the word 'STO', 'RTO' or 'AGARD', as appropriate, followed by the serial number. Collateral information such as title and publication date is desirable.

If you wish to receive electronic notification of STO reports as they are published, please visit our website (http://www.sto.nato.int/) from where you can register for this service.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Royal High Institute for Defence – KHID/IRSD/RHID
Management of Scientific & Technological Research for Defence, National STO Coordinator
Royal Military Academy – Campus Renaissance
Renaissancelaan 30
1000 Brussels

BULGARIA
Ministry of Defence
Defence Institute "Prof. Tsvetan Lazarov"
"Tsvetan Lazarov" bul no.2
1592 Sofia

CANADA
DSTKIM 2
Defence Research and Development Canada
60 Moodie Drive (7N-1-F20)
Ottawa, Ontario K1A 0K2

CZECH REPUBLIC
Vojenský technický ústav s.p.
CZ Distribution Information Centre
Mladoboleslavská 944
PO Box 18
197 06 Praha 9

DENMARK
Danish Acquisition and Logistics Organization (DALO)
Lautrupbjerg 1-5
2750 Ballerup

ESTONIA
Estonian National Defence College
Centre for Applied Research
Riia str 12
Tartu 51013

FRANCE
O.N.E.R.A. (ISP)
29, Avenue de la Division Leclerc – BP 72
92322 Châtillon Cedex

GERMANY
Streitkräfteamt / Abteilung III
Fachinformationszentrum der Bundeswehr (FIZBw)
Gorch-Fock-Straße 7
D-53229 Bonn

GREECE (Point of Contact)
Defence Industry & Research General Directorate, Research Directorate
Fakinos Base Camp, S.T.G. 1020
Holargos, Athens

HUNGARY
Hungarian Ministry of Defence
Development and Logistics Agency
P.O.B. 25
H-1885 Budapest

ITALY
Ten Col Renato NARO
Capo servizio Gestione della Conoscenza
F. Baracca Military Airport "Comparto A"
Via di Centocelle, 301
00175, Rome

LUXEMBOURG
See Belgium

NETHERLANDS
Royal Netherlands Military Academy Library
P.O. Box 90.002
4800 PA Breda

NORWAY
Norwegian Defence Research Establishment, Attn: Biblioteket
P.O. Box 25
NO-2007 Kjeller

POLAND
Centralna Biblioteka Wojskowa
ul. Ostrobramska 109
04-041 Warszawa

PORTUGAL
Estado Maior da Força Aérea
SDFA – Centro de Documentação
Alfragide
P-2720 Amadora

ROMANIA
Romanian National Distribution Centre
Armaments Department
9-11, Drumul Taberei Street
Sector 6
061353 Bucharest

SLOVAKIA
Akadémia ozbrojených síl gen M.R. Štefánika, Distribučné a informačné stredisko STO
Demänová 393
031 01 Liptovský Mikuláš 1

SLOVENIA
Ministry of Defence
Central Registry for EU & NATO
Vojkova 55
1000 Ljubljana

SPAIN
Área de Cooperación Internacional en I+D
SDGPLATIN (DGAM)
C/ Arturo Soria 289
28033 Madrid

TURKEY
Milli Savunma Bakanlığı (MSB)
ARGE ve Teknoloji Dairesi Başkanlığı
06650 Bakanliklar – Ankara

UNITED KINGDOM
Dstl Records Centre
Rm G02, ISAT F, Building 5
Dstl Porton Down, Salisbury SP4 0JQ

UNITED STATES
Defense Technical Information Center
8725 John J. Kingman Road
Fort Belvoir, VA 22060-6218

SALES AGENCIES

The British Library Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
UNITED KINGDOM

Canada Institute for Scientific and Technical Information (CISTI)
National Research Council Acquisitions
Montreal Road, Building M-55
Ottawa, Ontario K1A 0S2
CANADA

Requests for STO, RTO or AGARD documents should include the word 'STO', 'RTO' or 'AGARD', as appropriate, followed by the serial number (for example AGARD-AG-315). Collateral information such as title and publication date is desirable. Full bibliographical references and abstracts of STO, RTO and AGARD publications are given in "NTIS Publications Database" (http://www.ntis.gov).

ISBN 978-92-837-2425-4