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.