Operators and autonomous intelligent agents: human individual characteristics shape the team's efficiency

Adrien Metge[1, 2], Nicolas Maille[1], Benoît Le Blanc[2]

[1]ONERA, BA 701, 13661 Salon-de-Provence, France
[2]ENSC-Bordeaux INP, IMS laboratory, 109 Avenue Roul, 33400 Talence, France

adrien.metge@onera.fr, nicolas.maille@onera.fr, benoit.leblanc@ensc.fr

Abstract. Autonomous intelligent cyber-defense agents will need to interact with operators during training or reporting phases. In this paper, we investigate how individual characteristics of the operator may shape the trust they have in the protection strategy deployed by autonomous agents. Through a micro-world experimental study, we simulate an operator-agent cooperative decision-making task. We found a link between the participants' extraversion and their sense of responsibility towards the agent.

Keywords: intelligent agent, human factors, autonomy, cyber defense, human-autonomy teaming, man-machine teaming

1 Introduction

1.1 Autonomous intelligent agents

With the increase of cyber threats during defense operations, the deployment of Autonomous Intelligent Cyber-defense Agents (AICA) is one response being considered (Kott et al., 2019). Intelligent agents are defined as digital systems capable of perceiving their environment through sensors, acting on it to achieve goals, and communicating with other agents (Russell and Norvig, 2002). The stealth and responsiveness required to react to cyber threats in cyberspace implies a high level of automatism for many responses, and thus autonomy and intelligence of defense agents.

1.2 Human-in-the-loop imperative

Despite these automatisms, AICA will need to interact with external human operators outside of these active protection phases. AICA would be deployed over a long period of time within which most of the time nothing should happen. Nevertheless, if a real cyber-attack occurs it will often be a complex and coordinated combination of events that require a quick and coherent combination of reactions and dedicated monitoring over time. Banking institutions, for example, have in recent years suffered distributed denial of service attacks followed by waves of phishing messages sent to account holders. During the service restoration phase, users' vigilance may be diminished, which can be used to extract personal data or demand ransoms (De Nederlandsche Bank, 2018). During defense operations, the agitation following a saturation attack could be used in a similar way to introduce backdoors into the system to leak sensitive data from the network later. The final responsibility of the overall system security has to be dedicated to an operator who is in charge to monitor and ensuring its integrity. Thanks to its robust adaptation factor, the operator can act as a safeguard for the machine, when arise extreme or unexpected situations that the AICA model is not able to tackle. Moreover, the maintenance of humans in the overall decision-making process is desired for ethical reasons (Task Force IA, 2019). Agents' cooperation with external entities, and in particular human operators, is considered as one of the thirteen major research challenges for autonomous cyber-defense (Theron and Kott, 2019). Three cases can be identified where such collaboration could occur:

- in a phase of preparation for deployment or updating, to teach the agent what he must do if its intelligence is based on a learning process. Upstream of protection, the operator could also communicate to the system rules of engagement in cyberspace, such as all the countermeasures that it can deploy to react to the various possible threats. Then, training or test phases enable the operator to ensure that the agent reacts as planned. The proper functioning of the agent depends on the calibration carried out by the operator a priori, and therefore engages his responsibility. More the embedded intelligence will be sophisticated, more this step will be significant and we can assume that it should be a crucial point for future AICA.
- in a monitoring phase without major events, which is the normal case of operation. The agent must still make regular reports to the chain of command on the state of the system, what he has detected, blocked, and done.
- in a crisis or when an attack is detected. Even if first answer has to be done without human interactions, Human point of view can help the system to put into perspective the malicious actions detected, actions carried out, the remaining risks, the supposed intentions of the attacker or the additional surveillance actions decided upon (redeployment of AICA, dedicated verification actions, etc.). This is where human intelligence must have its place to give meaning to what is happening and influence the medium-term defense and verification strategy.

1.3 Mental representation challenges

To communicate with the system, operators will establish a mental representation of the agent, which may be different according to the individual. Mental models are known to be influenced by the cultural background of individuals, who tend to project their beliefs, desires and intentions onto others (Malle, 2006). But even for a homogeneous population that has undergone common training, significant variability between individuals can occur. Interpersonal trust dynamics like those found in human-human teams will be established in the face of an intelligent system (Bollon et al., 2019). The psychological type personality traits of the operator may in this context have an impact on the relationship he/she will have with AICA.

The processes involved in building AICA strategy before its deployment, in collaboration with the operator, is similar to what happens when an operator prepares the mission of an intelligent UAV that will then carry out the mission autonomously. This is an issue that is currently being studied (Metge and Maille, 2020) and has been the subject of the development of an experimental micro-world. We are reusing this micro-world to investigate how the process of interaction between an operator and an intelligent agent (what tools, what explanations, what dynamics) modifies both the choices made (the chosen plan, based on what compromises) and the operator's confidence in the chosen plan. We focus on the interaction phase preceding the deployment of an autonomous intelligent agent. This phase is dedicated to defining both an initial strategy to supervise the integrity of the system, and how the intelligent agent is supposed to manage new threats. We formulate the hypothesis that individual characteristics of the operator will influence the acceptability and the confidence he/she has in the system. The article studies the latter hypothesis and shows the importance of adapting to certain characteristics of the operator to optimize this operator-intelligent agent cooperation.

2 Experimental study

2.1 Task description

A group of 20 healthy people, PhD students and young engineers (40% women), with an average age of 26.1 years (standard deviation = 2.7 years), participated in this study. All subjects volunteered to take part in the study and gave their full informed consent before taking part in the experiment. They embody a military air operator in charge of supervising a UAV to carry out missions in enemy territory (Figure 1). The aim of the missions is to fly over several targets to photograph them, then to leave the enemy zone, while minimizing the risks taken and the fuel consumed. The missions take place on various territories but with a similar scenario: 1) the UAV heads towards the enemy zone with an initial flight plan, 2) suddenly enemy entities are detected, so the flight plan is no longer satisfactory, 3) the operator interacts with the system to define a new flight plan, 4) the operator validates a new flight plan, which completes the supervision task. During the interaction phase, the operator can ask the AI intelligent agent to suggest new plans directly or use plan modification tools to explore alternatives.

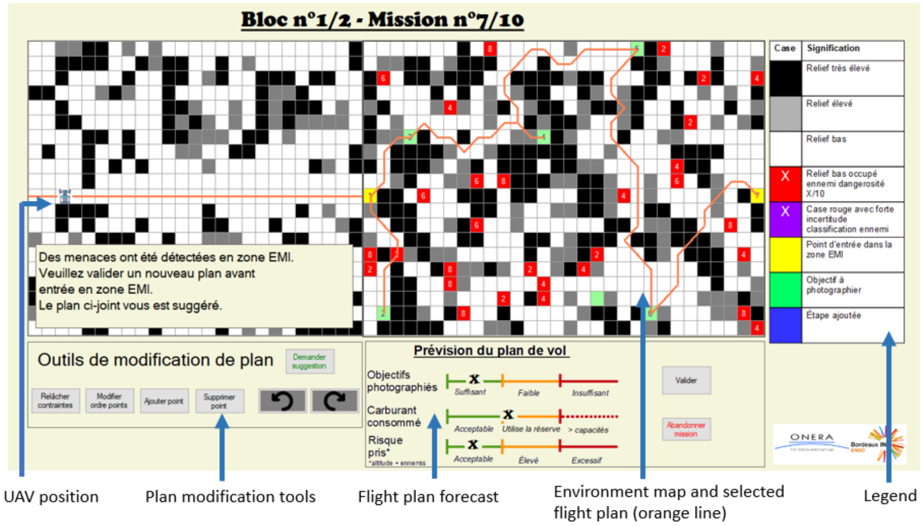

Fig. 1. HMI for preparing the mission of the intelligent agent.

2.2 Metrics

To study how the individual characteristics of the operator influence the cooperation with the intelligent agent, we define two categories of metrics: metrics of the operator's feelings about the chosen solution, and metrics of individual characterization of operators. All were evaluated in the participants' common language of expression, French.

Metrics of the operator's feelings about the chosen solution. We use four metrics to evaluate the quality of cooperation between the operator and the intelligent agent. After each completed mission, the participants answered three questions in the interface on 7-item Likert scales about their:

- Confidence in the validated solution
- Feeling of responsibility in the validated solution.
- Feeling of authorship of the validated solution, i.e. according to the operator who of him or of the system took the most part in its design.

Then, once all the missions were finished, participants completed a NASA Task Load Index (NASA TLX) questionnaire to measure their perceived workload for the task (Cegarra and Morgado, 2009).

Metrics for individual characterization of operators. We use seven metrics to describe the personality of the operators according to different traits. Several weeks after the experiment, participants completed a questionnaire to quantify them, consisting of three juxtaposed psychometric questionnaires with a total of 58 questions:

- The Big-Five Inventory (BFI-Fr), composed of 45 questions which describe the personality in five central traits: openness, conscientiousness, extraversion, agreeableness, neuroticism (Plaisant et al., 2010).
- The Rosenberg's Self-Esteem Scale (RSES), composed of 10 questions that measure individuals' self-esteem (Vallières and Vallerand, 1990).
- The Self-confidence Stability Scale (SESS), composed of 3 questions that measure the variability of individuals' self-esteem over time (Altmann and Roth, 2018). We translated these questions into French using the methodology developed by Lallemand et al (2015).

2.3 Results

Data from all 20 participants were included in the analysis. They performed 3 training missions and then 10 recorded missions. To compare their profile, the metrics from the 10 completed missions were averaged for everyone. We set a threshold of 5% for the significance of p-values.

Variation in operator's feelings about the chosen solution. To evaluate the extent to which cooperation with the intelligent agent will depend on individuals, a hierarchical ascending classification was performed on participants according to the metrics of cooperation (Figure 2. a). The optimal partition consists in separating the participants into two equal groups of ten individuals. We then performed a principal component analysis to determine which variables discriminate these two clusters (Figure 2. b). We can observe that the groups separate on axis 1 of the PCA, which is mainly composed of the variables of sense of responsibility and sense of authorship of the solution. Moreover, these two variables are found to be highly correlated ($r(18) = .75$, $p < .001$). Thus, a contrast is observed between some participants with a high sense of responsibility and authorship of the decisions made with the intelligent agent, and some others for whom these indicators of cooperation are low.

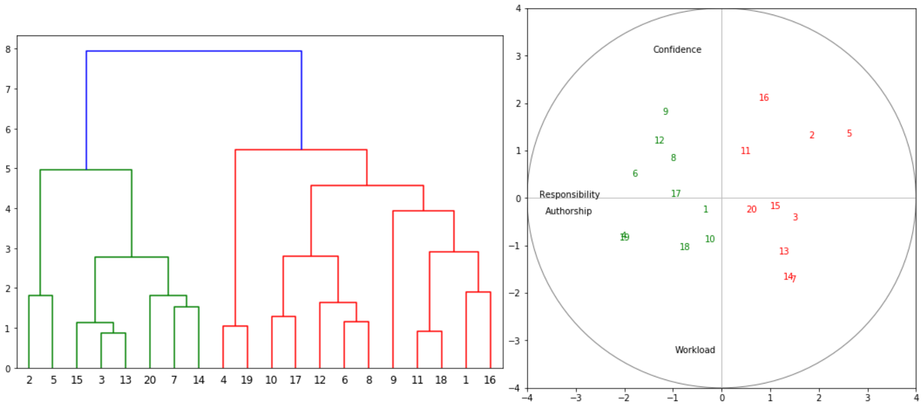

Fig. 2. Clustering of participants by metrics of operator's feelings about the chosen solution. (a) Hierarchical upward classification with k=2 groups. (b) Principal component analysis (axes 1 and 2).

Link with the individual characteristics of the operator. To deepen this inter-individual difference in cooperation, we studied whether it would be related to elements of the personality of the participants. To do so, we focused on the two metrics operator's feelings about the chosen solution that turn out to be discriminatory: the feeling of responsibility, and the feeling of authorship of the decision. We constructed a correlation table between these two metrics and the metrics for individual characterization of operators (Figure 3). These linear correlations are significant between the feeling of responsibility and extraversion ($r(18) = .48$, $p = 0.03$), and between the feeling of authorship and the stability of self-confidence ($r(18) = .44$, $p = .04$). Thus, the differences in cooperation with the intelligent agent are linked to extraversion and stability of the operators' self-esteem.

	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness	Self esteem	Self confidence stability
Responsibility	0.031816	0.538272	0.614712	0.949400	0.185194	0.544530	0.088695
Authorship	0.141960	0.396205	0.911633	0.701027	0.817496	0.119734	0.047547

Fig. 3. Linear correlation table between the two-discriminant metrics of operator's feelings about the chosen solution, and the seven metrics for individual characterization.

3 Discussion

The implementation of AICA that will be autonomous during the action will necessarily involve interactions with human operators for defense preparation or reporting. In this paper, we investigate the cooperation between AICA and human to set up the deployment phase. The experimental study is based on an existing micro-world dedicated to build the mission of a UAV that will be fully autonomous during its achievement, but nevertheless supervised by an operator. Even if such a UAV is a much more sophisticated autonomous agent than AICA, the interaction with operators should be similar in nature, and have an impact on how they are confident in the protection given by AICA and understand what is going on if a cyber-crisis appears. We focused on relationship between individual characteristics and feelings of the operators. Despite the consistence of our pool of participants in terms of age and familiarity with aeronautical issues, we observe significant inter-individual variability. Participants can be divided into two categories: those with a high feeling of responsibility, who also feel that they are at the origin of the decision taken with the system, and in opposition those who feel little responsibility for the decisions and who have the impression that it is more the intelligent system that is at the origin of the chosen strategy of action. The analysis of their personality traits shows a significant link between their extraverted character and their feeling of responsibility for the decisions made with the system. This result pinpoints that it is relevant to better understand and take into consideration how operators' characteristics may shape the human-intelligent agent cooperation in order to optimize the reliability and the efficiency of global system. With the development of agents with more sophisticated capabilities the operator trust in these agent's behavior should have a major impact on his/her ability to efficiently monitor the security of the systems and guaranty the compliance with ethical rules.

This study is a first step to better understand how the operator characteristics could affect the cooperation process with an intelligent agent. A complementary experiment, currently underway, will study how the individual characteristics of the operators could be related not only to the behavior of the operator (which type of interaction with the agent they would like to have), but also to the action strategy collaboratively selected by the human-intelligent agent team (which type of deployment). Moreover, the individual characteristics that affect a priori confidence in the system may be very dependent on the type of agent the operator work with. In 2018, the French government chose to partition the subject of cyber defense into four distinct chains: protection, defense operations, intelligence, and forensic investigation (SGDSN, 2018). To ensure the implementation of reliable AICA, the operator's features that affect the human-agent teaming should be considered during the design process for each of these particular use cases.

Acknowledgments. The research project of which this study is a part was granted by Agence de l'Innovation de Défense (AID) and Office National d'Etudes et Recherches Aérospatiales (ONERA).

References

Altmann, Tobias, and Marcus Roth. "The self-esteem stability scale (SESS) for cross-sectional direct assessment of self-esteem stability." *Frontiers in psychology* 9 (2018): 91.

Bollon, Florent, et al. "Interpersonal trust to enhance cyber crisis management." *HFES European Chapter*. 2019.

Cegarra, Julien, and Nicolas Morgado. "Étude des propriétés de la version francophone du NASATLX." *Communication présentée à la cinquième édition du colloque de psychologie ergonomique (Epique)*. 2009.

De Nederlandsche Bank. Press release: watch out for phishing emails. 31 January 2018. *https://www.dnb.nl/en/news/news-and-archive/Persberichten2018/dnb372138.jsp*
Kott, Alexander, et al. "Autonomous Intelligent Cyber-defense Agent (AICA) Reference Architecture. Release 2.0." Report ARL-SR-0421, US Army Research Laboratory, Adelphi, MD, September 2019.
Lallemand, Carine, et al. "Création et validation d'une version française du questionnaire AttrakDiff pour l'évaluation de l'expérience utilisateur des systèmes interactifs."*European Review of Applied Psychology* 65.5 (2015): 239-252.
Malle, Bertram F. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press, 2006.
Metge, Adrien, and Nicolas Maille. "IA et assistance à la décision humaine: influence des informations transmises sur l'évaluation d'un nouveau plan de vol." *Rencontres des Jeunes Chercheurs en Intelligence Artificielle (FJCIA).* 2020.
Plaisant, Odile, et al. "Validation par analyse factorielle du Big Five Inventory français (BFI-Fr). Analyse convergente avec le NEO-PI-R." *Annales Médico-psychologiques, revue psychiatrique*. Vol. 168. No. 2. Elsevier Masson, 2010.
Russell, Stuart, and Peter Norvig. "Artificial intelligence: a modern approach." (2002).
Task Force IA. *L'Intelligence Artificielle au service de la défense.* (2019)
SGDSN. *Revue stratégique de cyberdéfense.* (2018)
Théron, Paul, and Alxander Kott. "When Autonomous Intelligent Goodware will Fight Autonomous Intelligent Malware: A Possible Future of Cyber Defense." *MILCOM 2019-2019 IEEE Military Communications Conference (MILCOM)*. IEEE, 2019.
Vallieres, Evelyne F., and Robert J. Vallerand. "Traduction et validation canadienne-française de l'échelle de l'estime de soi de Rosenberg." *International journal of psychology* 25.2 (1990): 305-316.