Trust Between Humans and Intelligent Machines and Induced Cognitive Biases

Gilles Desclaux

NORTH ATLANTIC TREATY ORGANIZATION

SCIENCE AND TECHNOLOGY ORGANIZATION

www.sto.nato.int

Cognitive Warfare: The Future of Cognitive Dominance

First NATO scientific meeting on Cognitive Warfare (France) – 21 June 2021.
Symposium organized by the Innovation Hub of NATO-ACT and ENSC,
with the support of the French Armed Forces Deputy Chief of Defence,
the NATO Science and Technology Organization / Collaboration
Support Office, and the Region Nouvelle Aquitaine.

Scientific Editors
B. Claverie, B. Prébot, N. Buchler and F. Du Cluzel.

Chapter 5 – TRUST BETWEEN HUMANS AND INTELLIGENT MACHINES AND INDUCED COGNITIVE BIASES

Lieutenant General Gilles Desclaux[1]

"Humanity has learned a lot from the machines built by itself,
except perhaps how to live better with them."

The strategic field of crisis management is based both on knowledge of the most complete information possible, confidence in the best technologies that deliver them, and the decision-making ability of the commander who relies on a strong organization and effective.

In the context of massive information, these three dimensions require the development of so-called "intelligent" software agents capable of selecting, merging, and representing relevant information and of delivering decision-making solutions at high speed. These agents are developed by large industrialists; they are progressing steadily towards greater autonomy. Despite this progress and faced with an increasing complexity of the criticality of the situations, the project of purely autonomous systems is moving away from realistic prospects in the short and medium term. Experts in crisis management and these artificial systems must increasingly work in a collaborative manner, each bringing the best of their skills to the human-system duo. The notion of trust is therefore central for the I2HM (Human-System Interaction/Integration), and the collaboration between humans and machines. The strength or weakness of this collaborative relationship is a key security issue, and therefore one of the targets of cognitive warfare (Cyber Warfare).

5.1 HUMAN-MACHINE COLLABORATION FOR CRISIS MANAGEMENT

The management of defence systems or military operations is a field as complex as it is codified. One of the strategic areas is rapid crisis management. Doctrine, the law of war, the responsibility for minimal human attrition for adequate tactical material effectiveness limit the action of the decision maker who must nevertheless act quickly and well. Managing a crisis means mobilizing in the most effective way possible the means made available to imagine, evaluate and implement the most relevant measured and measurable solutions leading to a favorable solution as quickly as possible. Crises can be ad hoc, in place or in time, or more global and lasting, requiring adjustments or solutions whose complexity evolves with multiple evolutionary dimensions to be taken into account.

For this, knowledge is the real "fuel" for measuring, anticipating, and driving action. It is a major criterion of differentiation to control the criticality of situations. It is developed from masses of data which today exceed human capacities for global representation or comprehension and requires recourse to techniques using "Big Data," "Artificial Intelligence" and "Visualization" of potential and changing solutions upon which the decision is based.

In recent years, the development of "intelligent" software agents has progressed towards greater autonomy. Many obstacles remain to be overcome in order to achieve the prospect of real systems capable of effectively replacing human experts. In the near future, these experts and artificial systems will have to continue to "work as a team," in an even more collaborative way. The concept of "Human-Autonomy Teaming" (HAT) was proposed for this by NASA teams in 2018 (O'Neill et al., 2020) to account for this "strange collaboration," which mixes Artificial Intelligence (AI) and Natural Intelligence (IN). It contributes to the emergence of

[1] Gen. Gilles Desclaux is Air Force Lieutenant General (ret), president of RACAM (Civil Aviation – Military Aviation Interface). He is researcher at the Human Engineering for Aerospace Laboratory (HEAL – ENSC Bordeaux-INP / THALES, FR). There, he coordinates the "Anticipe" program: AI-human decision support processes for "Air C2".

hybrid, anthropotechnical systems, a form of dual and shared intelligence, which is not without posing concrete problems of fragility and reliability in the cognitive domain.

5.2 COOPERATION BASED ON DIFFERENT COGNITIVE PROCESSES

The decision-making process implemented by humans is radically different from that of intelligent machines. Identical cognitive architectures could facilitate communication, but unlike humans, machines are restricted to well-defined objectives and priorities, without the capacity for improvisation or interpretive adaptation, and without real inventiveness beyond the algorithmic proposition of unexpected solutions. Humans, on the other hand, can develop these qualities but remain mediocre in accurately describing their intentions, goals, and priorities as intelligent machines demand. Likewise, their capacities for attention, memory or reliability of reasoning are fragile and frequently compromised, whereas artificial systems are particularly reliable in this area.

Within a HAT-type "decision-making network," humans and machines continually modify their own roles, tasks, and relationships with other actors, natural and artificial, partners and external alike. This activity is called "centered networks." When the usual processes do not seem to correspond to their expectations, new strategies are implemented: machines open procedures for consulting external databases, while humans form or restructure informal or ad hoc working groups and are looking for new experts.

Intelligent machines remain and will remain, at least for the foreseeable future, partially incomprehensible to humans. It is obviously the same with humans for machines. Establishing trust between the two types of entities is therefore difficult. Intelligent machines are susceptible to cyber intrusions that can compromise their "perceptions," the relevance of their "decision making," and their data management and communication capabilities. Humans have other weaknesses, such as fatigue, limited memory, and fragile and easily influenced cognitive abilities. In such a context, one solution is to foster the establishment of constructive performance monitoring relationships between human experts, between machines and, in both directions, between experts and machines.

5.3 THE PROBLEM OF INTERPRETABILITY

Interpretability has two dimensions. The first aspect corresponds, for the user of an automated or autonomous system, to the user's degree of understanding of what the system does, how it does it and why it does it. The interpretability of the system can lead to the development of a cognitive model that is as complete as possible in order to provide an understanding of how it works, and the ability to predict what it would do under certain circumstances. Two approaches make it possible to facilitate interpretability:

- System feedback improves the experience of interacting with users and facilitates their sense of control. Users usually want the system itself to provide understandable information about its own level of trustworthiness, in order to know whether to trust it or not.
- The post hoc explanation, known in the English-speaking world as eXplainable AI (Adadi and Berrada, 2018) or XAI, provides the user with an explanation that justifies the decision making, thus making the system more interpretable and facilitating feedback (Retex).

The second aspect, interpretability, concerns the limitation, for the user or the human partner, to behaviors or decisions that are understandable for the machine, or consistent with its own knowledge registers. This limit is necessary to maintain the effective collaboration link. This dimension is not without problems of acceptability for naive human users, who must learn to collaborate with machines to facilitate the competence and maintenance of the efficiency of the HAT system. Here again, the learning systems are frequented by experts and must be able to identify them in order to adapt to their peculiarities and the specifics of their cognitive characteristics: personality, age, greater or lesser mnemonic performance, visual

or formal, sensitivity to sounds or images, field dependence or independence, attentional saturation, resistance to fatigue, stress control, etc. To address this issue, the use of portable technologies (wearable tech.), sensors and auto-quizzes on tablets is now being studied by the laboratories of the US Army (Buchler et al., 2016) and within the framework of collaborations between certain industrialists and university or defence engineering schools in NATO countries.

Although this avenue is still exploratory, we can expect to see technologies capable of facilitating the collaboration and efficiency of the human-system pair and the performance of the mission in terms of making the human partner recognized and identified by the machine, and continuously informing the machine of the evolution of the human partner's cognitive state and his knowledge.

5.4 THE ASSESSMENT OF UNCERTAINTY

To date, most decision-making automations work well for specific situations, and for which they are designed, but require the use of human expertise when it comes to managing situations outside certain defined or limited environments. In particular, when computer algorithms are confronted with uncertainty and ambiguity in data, they are often overwhelmed by decision making.

Humans surpass machines in understanding context. Machines remain incapable of exercising nuanced judgment in complex or ambiguous and evolving environments. Additionally, as machines are programmed or trained using sets of information relevant to a specific task or problem, encountering a new problem tends to lead to ambiguities or even to failure. The human capacity to adapt to new situations is much greater and even incomplete or imperfect responses are likely to perform well. Humans use mental surrogate abilities and estimations from familiar skills or tasks, and can thus provide approximate answers, which AI technologies are not yet able to do.

Humans also surpass machines in their ability to assess the quality of their cognition. Metacognition is a hallmark of the human mind. It escapes the machine for now. Work is being undertaken in order to understand the cognitive expertise of this human phenomenon, to give it a structure that can be understood by the machine, and to endow the machine with "metaprogrammatic" capacities to evaluate itself, to be able to evolve, and especially to evaluate human cognition in order to adapt to its evolution or its performance in a dynamic HAT relationship.

5.5 LACK OF TRANSPARENCY

When stand-alone systems lack understandability and predictability, there is a problem of lack of "transparency." This notion refers to the inability of humans to understand why the system takes such action or, on the contrary, does not take the decision of an expected action. Lack of transparency produces a lack of awareness, in particular it does not allow operators to know what information is used to perform a task.

This lack of transparency is sometimes the origin of a lack of trust which leads both to underuse of the system through mistrust or on the contrary to overuse due to blind trust (Clark et al., 2014). This confidence problem must be able to be assessed on an objective basis, with clear indicators.

These areas of difficulty are not independent problems and can combine in often dangerous ways (Endsley, 2016). Intelligent systems are fragile, and can quickly go from good operation to rapid, global degradation. It is therefore the responsibility of the human operator to monitor the occurrence of such failures, and to anticipate their consequences. But monitoring a system that appears to be working properly is a job that humans are ill-prepared for. We are talking here about phenomena of "taking out of the loop," or "OOTL" (Out-Of-The-Loop, in English – cf. Suhir, 2021), which induce a restricted awareness, even very reduced of the situation (Endsley, 2015).

5.6 TRUST AT THE HEART OF THE HUMAN/INTELLIGENT MACHINE RELATIONSHIP

In the HAT context, trust must be examined at two levels.

For the machine, the quality of the relationship is based on statistical algorithms for psychophysiological monitoring or on the quality and quantity of information exchanged. Monitoring human partners can allow the implementation of automated processes or operator reminders. This type of process is particularly studied in driving assistance and the detection of sleepiness or loss of driver attention, but also the non-detection of imminent dangers (pedestrian, obstacles, ice, etc. .). The required computational formalism requires a cognitive model of the driver (Bellet et al., 2011). The cyber defence of these programs remains one of the major concerns in view of the need for continuous evolution and updating of software.

For the human partner, trust is generally defined as "the degree to which a user believes that a system will behave as expected." Without this appropriate level of trust, operators may refuse the use of stand-alone systems or, on the contrary, completely offload onto them. These phenomena of overdependence that can lead to failure, followed by underdependence on automation, are well documented. The main factors that promote the development of trust are acceptability, tolerance, transparency, and the bidirectional nature of Human-System communication.

Confidence depends on the specific context of a human/intelligent system interaction and is influenced by the environment and the mental state of the operator. The perceived usefulness of an autonomous system in terms of the ability to perform a difficult or demanding task influences an individual's decision to trust it. But operators with a high workload also tend to rely more on the machine, regardless of their actual level of confidence in the system. The automaton, apart from simple tasks, generally does not completely replace humans. On the contrary, he changes the nature of his work by relieving it of certain tasks for which he is more efficient. This clearly poses the problem of reciprocal acceptability. The understanding, usability, and expectation of users of an intelligent system are correlated with the likelihood of trusting.

Confidence is built over time, and as a result, for the human partner, education and training foster the familiarity necessary to use the system. As for the artificial system, it must now be programmed due to the lack of scalable algorithms, or even adaptive machines.

5.7 COGNITIVE BIASES IN THE HUMAN-AUTONOMY DUO

Transparency is what allows the operator to determine if the autonomous machine is likely to provide the right response in a given complex situation. Transparency allows the machine to know if the information given by the human is trustworthy, or contain incongruities that need to be clarified.

But this transparency goes beyond the simple provision of information to the human operator or to the autonomous artificial partner. To be transparent, the automaton must present the information in a way adapted to the mental model of the operator, taking into account the operator's preferences and cognitive constraints, while, conversely, the human partner must adapt to the mental model of the program designer. Therein lies a first cognitive bias: the machine is not a partner like any other, it has been programmed by someone. It can also be deprogrammed, reprogrammed, be influenced by patches or additional programs, and therefore viruses, Trojans, and other malware. This cognitive dissonance bias is all the more, thanks to the fact that it imposes itself without any real solution, in the face of computer scientists or industrialists convinced that their way of thinking is the best for others.

Cognitive biases are spontaneous distortions of the rational thinking that humans adopt and which are the source of many errors (Kahneman and Tversky, 1974). They are studied by economists and psychologists,

especially with regard to decision making, but they are the subject of new attention from these experts and those of information processing, with the study of machine bias (Bertail et al., 2019) and the algorithmic creation of inequity, or even discrimination, posing unavoidable ethical problems.

In the context of big information, and for the users of the systems, humans most often focus on sources and methods of selection that they know well and trust, thereby introducing a different dreadful type of bias. This is an area where machines are nevertheless very efficient, providing a high speed of acquisition and processing of large volumes of information, as well as consistent, rigorous, and impartial data management. But without a level of transparency that makes it possible to recognize the sources of information and analyze their quality, the effectiveness of such systems will remain insufficient, and doubt remains underlying the relationship between humans and machines.

An example illustrates this notion. A semi-autonomous system presents several options that it has generated, along with evaluations of potential effectiveness as to the adequacy of each. Such a transparency facilitation device must be accompanied by a capacity for the operator to add information that the autonomous device does not know. The operator must be able to suggest solutions and have them evaluated by the controller. Collaborative problem solving is therefore a back-and-forth, "Wargaming"-type process. This type of two-way communication promotes partnership and helps assess favorable solutions to potential problem solving.

A third type of bias concerns the spontaneous feeling of human superiority over the machine. A low level of cognitive engagement makes it inherently difficult for an operator to understand what is going on when he is only performing passive surveillance of an autonomous system. Passivity in performing a task is then an obstacle to the effectiveness of intelligent human-machine interaction. This challenge depends on what some authors (Endsley, 2016) refer to as the "automation conundrum." Thus, the more automation you add to a system, and the more reliable and robust this automation is, the less likely it is that human operators will oversee it. They will then be unable to understand the situation and will tend to regain control of the system. The system then becomes degraded, restricted to the simple limited capabilities of the operators, which is obviously a significant advantage for the potential enemy. The automation conundrum creates a major obstacle to autonomy in areas where security is critical.

5.8 CONCLUSION

Today, the complexity of crisis management requires processing a large amount of data and making critical decisions in ever shorter times and increasingly constrained contexts. Decision makers at the head of crisis management organizations must therefore increasingly rely on hybrid systems. The help of intelligent systems has become indispensable. Despite the indisputable performance of such systems, they are still uncertain in several areas, and humans, who will continue to play an important role in this collaboration with machines, have a tendency not to master a set of biases generated by the exchange HAT. Ways forward lie on the one hand in the capacity of these machines to better explain, to establish a supported confidence, to communicate more easily, even to understand the hidden intentions and the emotions of the human actors, and on the other hand in a new culture of acceptance of machines by humans.

In a seminal article (2017), Kott and Alberts wrote, "Welcome aboard smart things. Whatever our respective shortcomings, we will be stronger and more agile by working together in decision-making organizations."

5.9 REFERENCES

Adadi, A., Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160.

Alberts, D.S., Haye, R.E (2006). Understanding Command and Control. Washington (DC-USA): DoD CCRP Publication Series.

Bellet, T. Mayenobe, P. Gruyer, D., Bornard, J.C., Claverie, B. (2011). The Living Cognition Paradigm: An Application to Computational Modeling of Drivers' Mental Activities. US-China Education Review, 1, 4, 568-578.

Bertail, P., Bounie, D., Clémençon, S., Waelbroeck. P. (2019). Algorithmes: Biais, Discrimination et Equité. Rapport de la Fondation Abeona et de Télécom ParisTech. Paris (France): Telecom ParisTech.

Buchler, N., Fitzhugh, S.M., Marusich, L.R., Ungvarsky, D.M., Lebiere, C., Gonzalez, C. (2016). Mission Command in the Age of Network-Enabled Operations: Social Network Analysis of Information Sharing and Situation Awareness. Frontiers in Psychology, 7, 937, 1-15.

Clark, B.B., Robert, C., Hampton, S.A. (2014). The Technology Effect: How Perceptions of Technology Drive Excessive Optimism. Journal of Business and Psychology, 15, 4-18.

Claverie, B. (2005). Cognitique, Science et Pratique des Rapports à la Machine à Penser. Paris (France): L'Harmattan.

Claverie, B., Desclaux, G. (2015). La Cybernétique: Commande, Contrôle et Comportement dans la Gestion des Systèmes D'information et de Communication. Hermès, 71, 72-79.

Endsley, M. (2015). Situation Awareness Misconceptions and Misunderstanding. Journal of Cognitive Engineering and Decision Making, 9, 1, 4-32

Endsley, M. (2017). From Here to Autonomy: Lessons Learned from Human-Automation Research. Human Factors, 59, 1, 5-27.

Gutzwiller, S.R., Espinosa, S.H., Kenny, C., Lange, D. (2018). A Design Pattern for Working Agreements in Human-Autonomy Teaming. In D.N. Cassenti (Ed.) Advances in Human Factors in Simulation and Modeling: Proceedings of the AHFE 2017 International Conference on Human Factors in Simulation and Modeling. New-York (NY, USA): Springer, 12-24.

Kahneman D., Tversky, A. (1974). Judgment Under Uncertainty: Heuristics and Biases. Science, New Series, 185, 4157, 1124-1131.

Kott, A., David S.A. (2017). How Do You Command an Army of Intelligent Things? Computer, 12, 96 100.

Le Guyader, H., Eshelman-Hayne, C., Irandoust, H., Lange, D., Genchev, A., Cakir, M., Verstraete, E., Brill, J.C., Desclaux, G. (2022 in press), Human Considerations for Artificial Intelligence in Command and Control. H. Le Guyader (Ed.). Technical Report of the NATO Science and Technology Organization Research Group IST-157, NATO. Paris (France): NATO-STO Collaboration Support Office.

O'Neill, T., McNeese, N., Barron, A., Schelble, B. (2020). Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature. Human Factors. 2020 Oct 22, 18720820960865.

Shively, R., Lachter, J., Brandt, S.L., Matessa, M., Battiste, V., Johnson, W. (2018). Why Human-Autonomy Teaming? Proceedings of the AHFE 2017 International Conference on Neuroergonomics and Cognitive Engineering, July 17 – 21, 2017, Los Angeles (CA, USA). Advances in Neuroergonomics and Cognitive Engineering, 586, 3-11.

Suhir, E. (2021). Human-In-The-Loop: Probabilistic Modeling Approach in Aerospace Engineering. Boca Raton (FL, USA): CRC Press.

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's References	3. Further Reference	4. Security Classification of Document
		ISBN 978-92-837-2392-9	PUBLIC RELEASE

5. Originator

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

6. Title

Cognitive Warfare: The Future of Cognitive Dominance

7. Presented at/Sponsored by

First NATO scientific meeting on Cognitive Warfare (France) – 21 June 2021. Symposium organized by the Innovation Hub of NATO-ACT and ENSC, with the support of the French Armed Forces Deputy Chief of Defence, the NATO Science and Technology Organization / Collaboration Support Office, and the Region Nouvelle Aquitaine.

8. Author(s)/Editor(s)	9. Date
B. Claverie, B. Prébot, N. Buchler and F. Du Cluzel	March 2022
10. Author's/Editor's Address	**11. Pages**
Multiple	118

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

Cognition; Cognitive bias; Cognitive domain; Cognitive war; Cognitive warfare; Cyber-psychology; Human

14. Abstract

This document, published by the NATO-CSO, brings together articles related to the presentations given during the first Symposium on Cognitive Warfare, held in Bordeaux, France, in June 2021, on the initiative of the NATO-ACT Innovation Hub and the Bordeaux-based ENSC, with the support of the French Armed Forces Joint Staff, the NATO-STO-CSO, and the Region Nouvelle Aquitaine. This first Symposium reflected on human cognition, its strengths and weaknesses, its collaborative organization for military decision-making, its relation with and dependence on digital technology, and its social and political dimensions within the context of fierce international competition. The Supreme Allied Commander for Transformation (SACT) and the French Armed Forces Vice-Chief of Defence expressed their views on the topic. This first Symposium was the starting point of a series of meetings and workshops further exploring the subject, on the initiative of NATO CSO and ACT.