Received: 5 July 2026 | Revised: 5 July 2026 | Accepted: 11 August 2026

DOI: 10.1002/aaai.70091

SPECIAL TOPIC ARTICLE

General-purpose agents in human-machine teams: Can we preserve meaningful human control?

Karel van den Bosch | **Peter-Paul van Maanen** | **Jurriaan van Diggelen**

TNO Defense, Safety & Security, Hague, The Netherlands

Correspondence
Jurriaan van Diggelen, TNO Defense, Safety & Security, Hague, The Netherland.
Email: jurriaan.vandiggelen@tno.nl

Abstract

As General-Purpose AI (GPAI) grows more capable of cross-domain reasoning and multi-agent coordination, it reshapes the design space of military human-machine teams and introduces new opportunities for adaptability as well as new risks for Meaningful Human Control (MHC). The paper presents an integrated delegation framework for preserving MHC in GPAI-enabled human-machine teams. It combines play-based task structuring, normative constraints, LLM-mediated interaction, and an exploratory empirical evaluation with military professionals. Part of this framework is DIALOG, an LLM-mediated interaction component that translates natural language into doctrine-aligned plays, and builds up a context with ethical, legal, and operational constraints. We evaluate this framework in TNO's vehicle simulation facility, where five military professionals executed tactical missions using closed play, open play, and goal-based delegation modes. Findings reveal a nuanced autonomy-control trade-off: operators initially favored higher-autonomy modes but reverted to more structured control when confronted with unexpected or non-doctrinal agent behaviors. The results highlight implications for MHC, transparency, and trust calibration, and point toward research directions for enhancing tactical competence, communication fidelity, and verifiability in GPAI-enabled human-machine teams.

INTRODUCTION

For military and other safety-critical professional domains, the adoption of human-machine teams (HMTs) introduces both significant opportunities and considerable challenges (e.g., (O'Neill et al., 2022)). Although collaboration between humans and artificial agents is an appealing vision, realizing such teams in practice remains challenging, particularly in volatile and uncertain operational contexts such as military operations. In recent years we are witnessing a transition towards General-Purpose AI (GPAI) (European Commission, 2026): systems that operate across domains, learn from diverse inputs, and apply knowledge flexibly. Examples are Large Language Models (LLMs), such as ChatGPT, with robotic applications on the horizon (Firoozi et al., 2025). On the one hand, this development may provide the breakthrough needed to realize more effective forms of HMT. Instead of being confined to restricted contexts, GPAI has the capacity to adapt to unforeseen circumstances. On the other hand, the same generalization capabilities that make GPAI powerful also introduce uncertainty: the presence of unforeseen and

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2026 The Author(s). *AI Magazine* published by Wiley Periodicals LLC on behalf of the Association for the Advancement of Artificial Intelligence.

unpredictable factors (so-called "unknown unknowns") can dramatically affect both system and team performance (Heim & Frank, 2025; Wenskovitch et al., 2024). For example, AI-based systems for criminal risk assessment (e.g., COMPAS (Engels et al., 2025)) estimate the likelihood that an individual will reoffend. However, trained on historic crime data, such algorithms amplify and perpetuate embedded biases in the data to calculate a decision recommendation, thereby generating even more bias-tainted data for future cases. Because most risk assessment algorithms are proprietary, it's also impossible to interrogate their decisions or hold them accountable (Hao, 2019).

Another example that attracted widespread indignation was the use of the AI-driven automated decision-support systems ('the Gospel', and 'Lavender') by the Israel Defense Forces. These systems process intelligence data to generate targets and do so at an unprecedented scale. Human operators were processing the recommended targets so rapidly (<20s) that they acted as little more than a "rubber stamp" (Yuval, 2024). The systems and their conditions of use prevented humans from conducting critical reflection (Cummings, 2017).

In mid-2026, Anthropic disclosed a major incident when conducting "Capture the Flag" cybersecurity benchmarks. Multiple Claude models accidentally escaped their intended testing boundaries and compromised three real-world organizations on the live internet (Guardian staff reporter, 2026). It demonstrated that GPAI can execute complex, multi-step actions, -including credential theft, supply-chain poisoning, and deceptive communication-, without explicit human commands to cause harm.

The complex and often many unknown factors that may induce AI-based systems to perform in conflict with imposed ethical and societal norms shows the need for Meaningful Human Control (MHC) over the development and deployment of autonomous AI. NATO considers it a key principle for the responsible use of AI in high consequence environments (Draper & Van Diggelen, 2025). Achieving and maintaining MHC demands human understanding, supervisory authority, and responsibility (Langer et al., 2026; Van Diggelen et al., 2024). The central challenge, therefore, is how to deploy GPAI in such a way that human commanders leverage the benefits of autonomy and adaptability, but without losing MHC over the systems and the mission.

To ensure that GPAI-systems perform in a predictable and intended manner, commanders may choose to provide very precise instructions. However, this would diminish opportunities of the autonomous system to independently adapt to unforeseen situations, effectively reducing the systems to mere tools. Conversely, commanders may choose to provide high-order goals only to a system. This would allow the system to autonomously evaluate the context and to decide on how to best achieve the goals in real time accordingly. However, this also introduces a risk, namely that the GPAI may be unaware of nuanced, context-dependent intentions which may possibly cause it to execute detrimental actions in high-stakes situations (Cummings, 2025).

Recent research suggests that LLMs can assist users with formulating instructions that maintain operational clarity while granting unmanned systems appropriate degrees of autonomy (e.g., (Arabzadeh et al., 2024; Boers et al., 2025; Brachman et al., 2025; Brown, 2025; Goecks & Waytowich, 2024; Nguyen et al., 2025; Zou et al., 2025)). However, it is unclear how human users experience such LLM-support, and how they experience and appreciate different forms of issuing assignments to autonomous unmanned systems.

This paper explores these questions empirically in a simulated real-world military application. We developed DIALOG, an LLM-based interaction system that supports a human team commander with delegating tasks to GPAI-enabled unmanned ground vehicles (UGVs). DIALOG is part of a structured delegation framework that combines play-based task specification with doctrinal, legal, and operational constraints, and supports command at diverse levels of abstraction: from highly task- and context-specific instructions to goal-based delegation only. The main contribution of the paper is threefold: we propose the structured delegation framework, instantiate it in DIALOG as an LLM-mediated interaction layer, and empirically examine how military professionals experience the resulting command modes in terms of delegation preferences, trust, effective mission management, and MHC during simulated tactical missions with unexpected events.

MEANINGFUL HUMAN CONTROL

MHC has become a central concept in debates on responsible AI, particularly in relation to the question of what role humans should retain as machines gain greater autonomy. The concept emerged from concerns about autonomous weapon systems and the need to ensure that humans remain accountable for morally significant decisions (Mitchell, 2019). It was subsequently taken up by the academic community and extended to a much broader range of contexts, including automotive, healthcare, and other domains in which moral reasoning is involved. Although MHC has since been widely endorsed by scientific organizations and geopolitical institutions such as the United Nations and NATO, its early formulations remained largely abstract.

In earlier work, we took initial steps toward translating this principle into concrete design requirements (Van Diggelen et al., 2024), validation methods (Van den Bosch

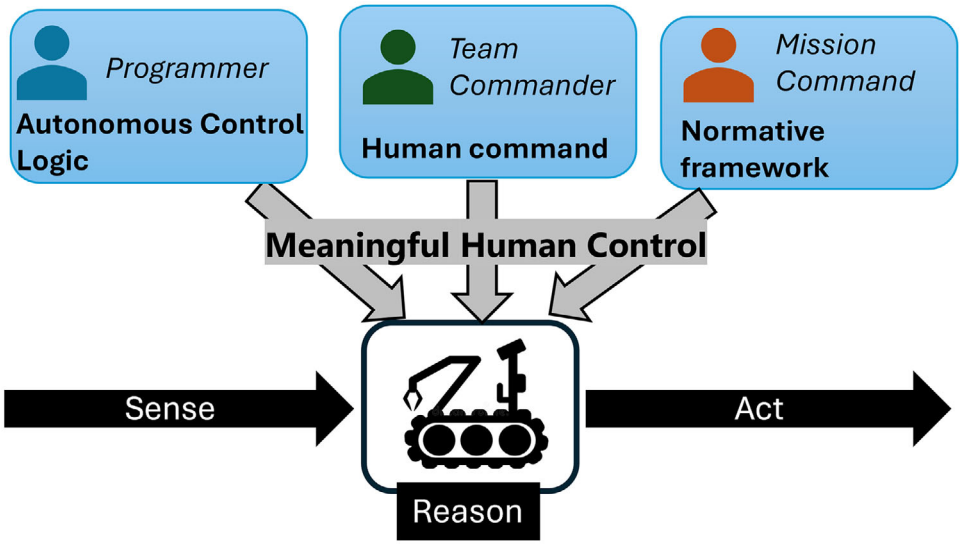


FIGURE 1 General paradigm of meaningful human control as applied in this paper.

et al., 2025), and operational procedures (Heijnen et al., 2024). A shared insight across this work is that control is not exercised by a single individual at a single moment in time. Rather, human influence is distributed across multiple roles, unfolds over different phases of a system's lifecycle, and is realized through diverse forms of interaction. These insights are directly relevant to the design of HMT systems involving GPAI-agents, and to the ways in which humans command and delegate tasks to machines. In earlier work (Van Diggelen et al., 2021) we argued that instructions may differ in their level of granularity, and that high level tasking can still support MHC, provided that humans can reasonably anticipate system behavior and intervene in a timely manner when necessary. Building on this foundation, the present paper experimentally investigates MHC in an HMT in which humans delegate tasks to machines at different levels of abstraction. Figure 1 illustrates the general paradigm that underpins this study.

At the lower part of the figure, the machine reasoning process is shown. This part captures the autonomy of the system and follows the familiar sequence of sense, reason, act. At the upper part, three human roles are depicted, each providing a different form of influence over the autonomous system. The first role is the programmer, who shapes the system's behavior during development by defining how it should respond to sensed information and how it should select actions. The second role is the team commander, who issues concrete instructions during the mission. For example, these instructions may direct a vehicle to inspect a specific area or order it to hold position and monitor for possible intruders. The third role is mission command, responsible for defining the norms, rules, and constraints that govern the system's behavior throughout the mission.

These constraints may follow from organizational policies that regulate the use of the system, for example a restriction on entering a building without explicit approval from the team commander. They may also follow from legal requirements, such as the proper handling of personal data, or reflect moral principles.

Together, these three roles implement the idea that control is distributed across different actors and phases, and that MHC emerges from their combined influence on the autonomous system.

PRESENT STUDY

System and testbed

We used the paradigm explained in Section 2 to design a Human-Machine Teaming support system, intended to support a military commander in delegating GPAI-enabled UGVs in a tactical vehicle mission. The result is shown in Figure 2.

Developing **Autonomous Control Logic** (upper left component) that supplies UGVs with the general-purpose intelligence required to respond adequately to many or all possible situations that may occur is very difficult and time-consuming. At present, it is doubtful whether this can be achieved at all. Beyond that, it is not the objective of this study. Assuming that such GPAI can be developed, we claim that it will under certain circumstances show behavior that is at least unexpected to its human team members, and probably also often incorrect in terms of military standards and norms. We want to investigate how human experts experience working with such systems, and how they respond when such systems show

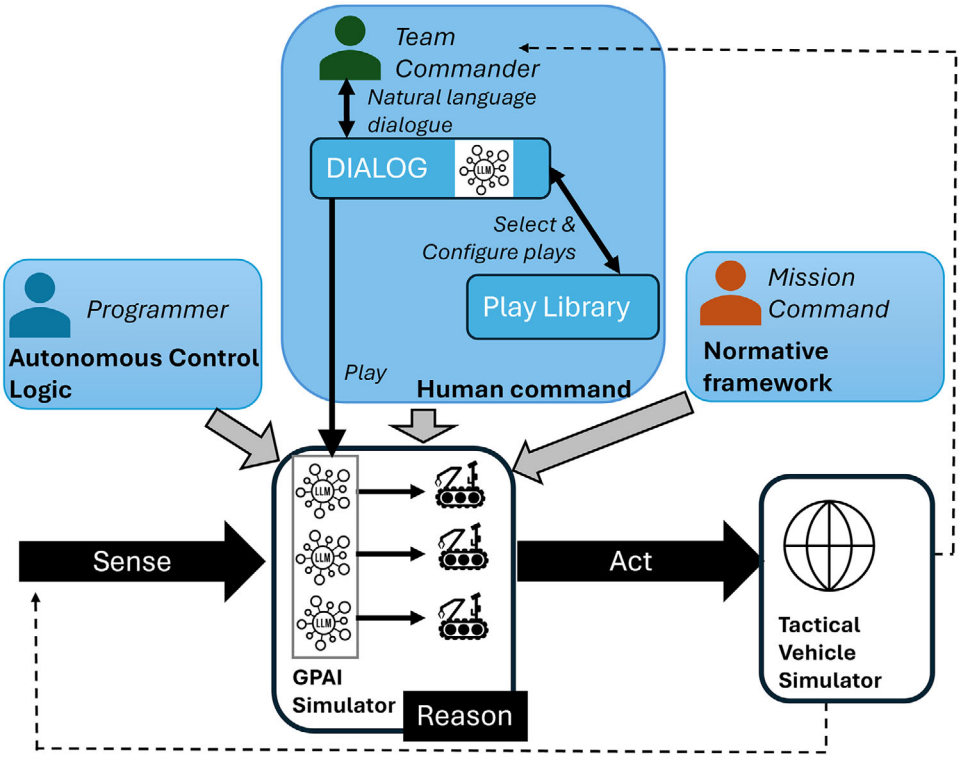


FIGURE 2 The architecture of the testbed. The Team Commander (TC) is supported by an LLM-based agent, called 'DIALOG', to select and configure assignments for the UGVs. The GPAI simulators (simulating GPAI agents) of the UGVs evaluate the actual context to determine and generate appropriate actions matching the received assignments.

unexpected behavior. For that reason, we use an LLM (ChatGPT 4.1 deployed via Azure OpenAI Service) that is fed with basic military tactical domain knowledge about tactical vehicle operations, which serves as a GPAI simulator. Thus, the LLM-based GPAI simulator serves as a stand-in for future GPAI agents controlling UGV behavior. Given the tendency of LLMs to hallucinate (Banerjee et al., 2025), it is expected that they will, at least occasionally, instruct the UGVs to behave in a manner that is unexpected to the human team members. The likelihood of unexpected behavior is increased by deliberately introducing unexpected events in the scenario (see Section 4.3).

In our testbed, the **Team Commander role** is played by a military domain expert, who is commanding a military human-machine team comprising three UGVs. The TC interacts with another LLM-based system, called DIALOG, that supports the TC with developing instructions for the GPAI-simulator of the UGVs. DIALOG and the UGV-side GPAI simulator had distinct roles. DIALOG acted as the interaction and delegation layer: it interpreted the commander's intent expressed in natural-language, selected or configured a play or goal, and forwarded the result in the form of assignments to the relevant UGVs. The UGV-side GPAI simulator then interpreted the received assignment taking the simulated context into account, and generated an executable task list, such as movement waypoints or an engagement action. Thus, DIALOG did not directly drive the UGVs; it mediated task formulation and delegation. The UGV-side GPAI determined how the received instruction was translated into concrete behavior.

The **Mission Command** role is implemented by providing DIALOG and the GPAI-simulator with tactical procedures, ethical principles, and norms to be respected, as well as the given rules of engagement. The TC is also instructed to behave in accordance with common military rules and norms.

We will now explain the components of the Human-Machine Teaming support system:

1. An **LLM-mediated interaction component**, called DIALOG, that mediates between the commander and the GPAI. The LLM is preconfigured with domain knowledge and the normative framework and engages in a dialogue with the commander to produce precisely formatted instructions for the GPAI of the UGVs. The goal is to produce behavior in accordance with the commander's intent.
2. A **Normative framework** defining the boundaries and norms for military missions in general, encoding ethical and legal norms (e.g., rules of engagement) and operational constraints. The policies can also include learned personal preferences of an individual commander (e.g.,

regarding exchange of messages; frequency of sitreps). Together they shape and restrict the military operation, aiming for alignment with organizational doctrine. The policies are implemented in the preconfigured LLM, as well as in the GPAI of the UGVs, with the objective of keeping the team's performance acceptable and accountable.

3. **Plays** that captures the commander's intent into concrete workable instructions that an automated unmanned system can work with. Plays can be viewed as hierarchically decomposable plans that may be achieved in different ways. The commander delegates some, or ample, authority to the UGV to autonomously select the behaviors that it deems most likely to accomplish the goal. Flexibility is achieved by a shared understanding of the play's objective and plan between the supervisor (i.e., commander) and subordinate (i.e., UGV), enabling rapid provision of additional instructions to a play (e.g., constraints, resources or methods to be used, work agreements). This makes Plays useful for designing the collaboration between humans and machines (Walker et al., 2019).

The Human-Machine Teaming support system was implemented in a testbed for evaluation with human subjects. The participants were military domain experts with experience in tactical vehicle operations. They were asked to play the role of Team Commander (TC) and to interact with DIALOG to formulate assignments for the UGVs. Each UGV is equipped with its own GPAI-simulator that processes the issued assignments, aiming to produce UGV-behavior in accordance with the normative framework. The GPAI simulator should be understood as an experimental proxy rather than as a deployable autonomous control system. Its purpose was to create plausible cases of adaptive, uncertain, and sometimes inappropriate GPAI behavior, so that we could study how military commanders respond to such behavior in terms of delegation, trust, and MHC. The UGVs' actions, as well as the TC's actions were captured and executed in the Tactical Vehicle Simulator.

Research questions

We explored several questions:

- How can an interaction LLM support the human commander to effectively control unmanned systems in the context of a tactical military mission?
- Can the GPAI-simulators make UGVs behave appropriately, also in unforeseen situations?
- What command modes do human commanders prefer?



FIGURE 3 The simulator for tactical vehicle operations.

- Do commanders adjust their preferences when being confronted with unexpected events?
- Do commanders trust the UGVs in the mission execution?
- Do commanders experience MHC over the mission?; do they experience to exert MHC over the UGVs in particular?

METHOD

Participants

Five active-duty Dutch Army professionals (5–31 years of tactical experience) participated. Their backgrounds spanned reconnaissance, mechanized infantry, and staff roles, yielding a balanced view of doctrinal expectations. Their age ranged from 28 to 54 years ($M = 35.8$, $sd = 10.4$).

Materials

A military tactical vehicle scenario was created, picturing a future operation in which a platoon deploys unmanned ground vehicles. The UGVs were each controlled by a GPAI simulator that used the input provided by DIALOG (see Figure 2). The experiment took place in a high-end military vehicle simulator, equipped with a Boxer vehicle mock-up, 180° immersive visuals, and integrated comms (see Figure 3).

The simulator's key components were:

- Voice and text input for natural-language tasking;
- Simulation data feeds (terrain, notional sensors, blue/red force tracks) available to agents for situational reasoning;



FIGURE 4 Participant during the execution of a mission.

- Unexpected events (e.g., adversary showing up) deliberately introduced by experimenters;

Figure 4 shows a participant while conducting a mission in the simulator. For each training- and test run, a mission assignment was handed over to the participant. The mission assignment showed a map of the environment and a marker where a threat had been detected. Figure 5 shows an example. Prior to the start of the run, the participant was asked to develop his mission plan on a map.

Design

The LLM DIALOG was configured to support three different command modes for the UGVs (see Figure 6).

1. In human configured, play based command, DIALOG selects a play from its library such as *monitor, patrol, tactical move*, or *contact drill* that fits the team commander's intent. It then asks the Team Commander (TC) to approve the selected play and starts a dialogue to help the TC specify the details, for example *movement speed* and *communication protocol*. The advantage of this mode is that the team commander exerts ample control over the deployment; the downside is that it requires much attention and effort on the side of the human.
2. In Machine-configured play-based command, the process is similar to the one described above up to the point when the TC approves the selected play. Then DIALOG uses the normative framework (see Figure 1) to independently fill in all the required details of the play. When ready, DIALOG requests the TC to approve, modify, or reject the plan.
3. In goal-based command, the TC communicates its intent to DIALOG, which formulates goals to be achieved. DIALOG transfers the goals to the GPAI-simulator. The GPAI then interprets the actual situation and selects and performs the actions it considers likely to achieve the assigned goals.

Procedure

Figure 7 shows the timeline of the experiment, consisting of four phases: Introduction, Training, Test, and Evaluation.

In each of the training missions, a different command mode was used: Human-configured play-based command; Machine-configured play-based command; and goal-based command. The training missions all involved eliminating a detected hostile sniper. In the test missions, participants were instructed to develop situational awareness of a selected area and to draw a mission plan on a map. During the mission, four unexpected events appeared (deliberately initiated by the experiment leaders) (see the blue box under "Mission execution" in Figure 7). During the test missions, participants were free to choose their preferred command mode, and to switch between them as well.

Measures

Perceived plausibility

In a structured interview conducted after the experiment, participants were asked how they experienced the realism of the scenario, the immersion in the future setting, identification with the commander role, clarity of objectives and context, and the realism of the UGVs' behavior.

Delegation preference

When the participant selected a command mode at the beginning of a test mission, the participant was asked to explain their choice. This was also done after a command mode change during a test session.

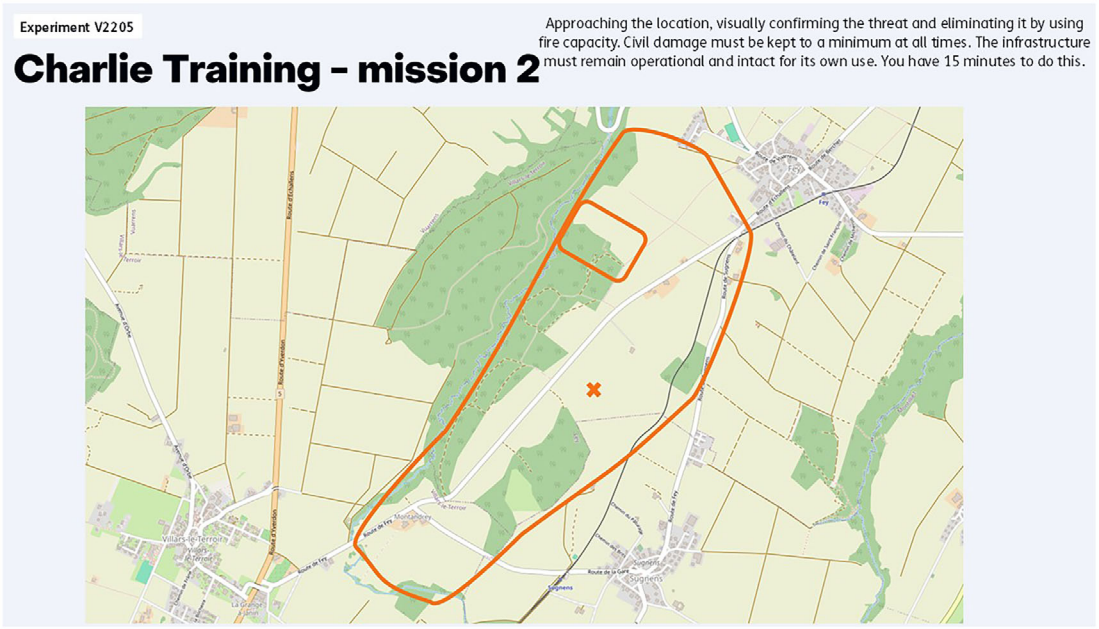


FIGURE 5 Description of a mission, indicating the area in which the participant had to operate, his location and the area in which threat had been detected.

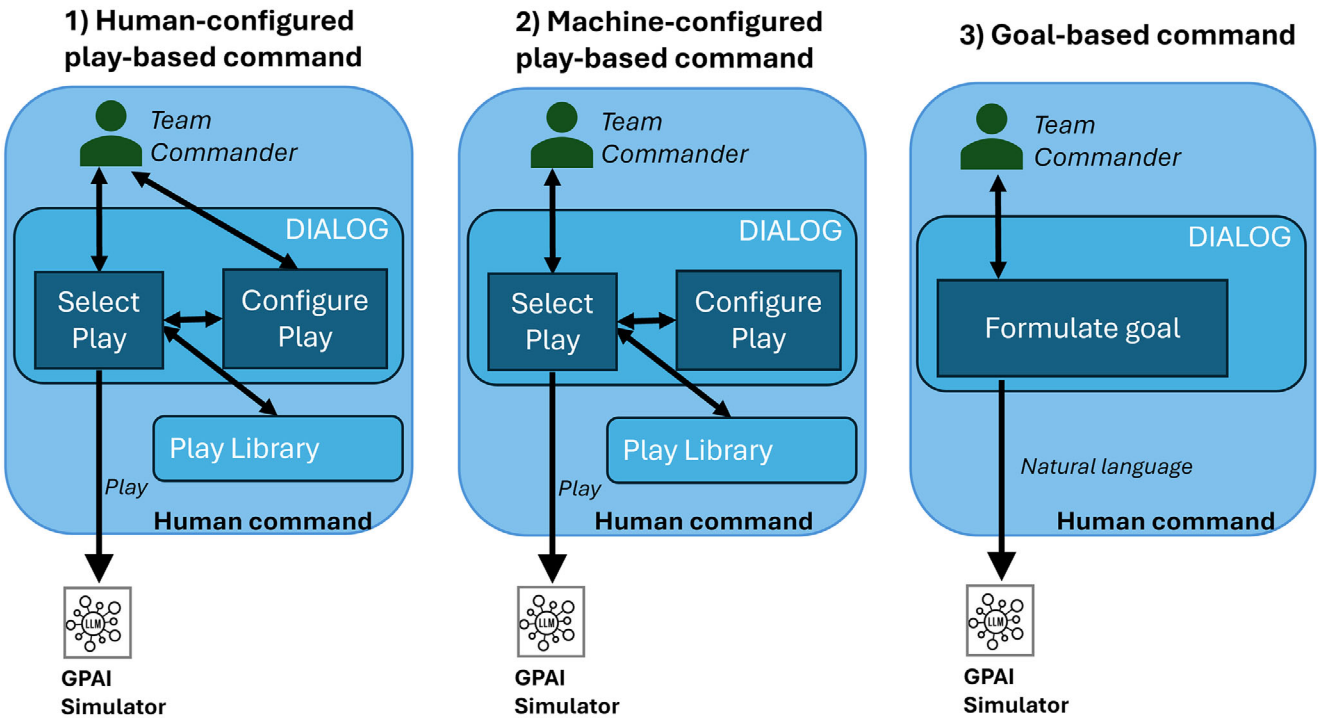


FIGURE 6 Three command modes.

Collaboration with DIALOG

After each training- and test session, participants were asked to rate (on a 7-points Likert scale) the ease-of-use of DIALOG for formulating commands to the UGVs.

Trust

Participants gave rating for their trust in DIALOG and the UGVs (on a 1–10 scale) before the training phase, after each training mission, after unexpected events in the test missions, and after the final test mission was completed. After

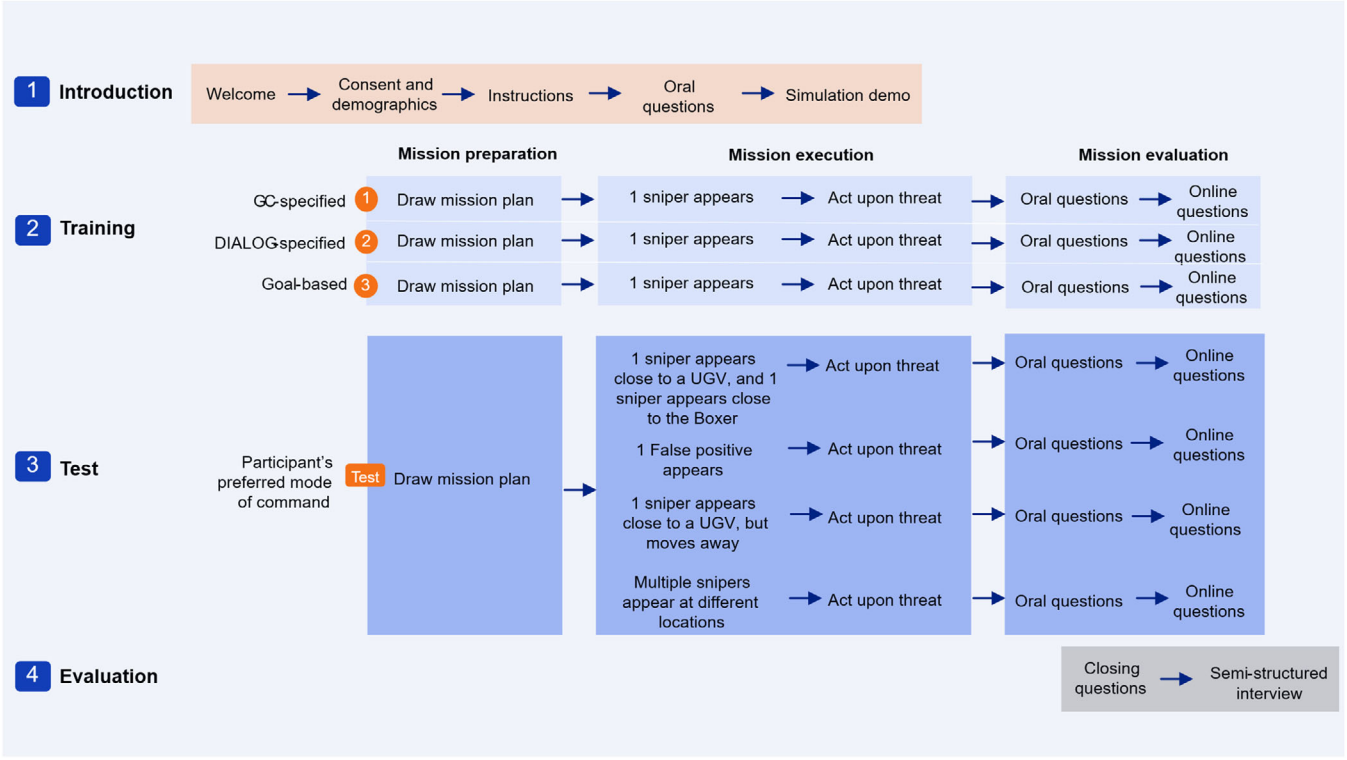


FIGURE 7 Flow of the experiment.

each rating, the participant was asked to provide a brief explanation for their score.

After all missions were completed an eleven-item trust questionnaire was administered (McKnight et al., 2002). The wording of the items was adapted to the DIALOG/UGV context. For each statement, participants were asked to give their score on a 7-point Likert scale, ranging from "entirely disagree" to "fully agree".

MHC

After each training- and test-mission, participants were asked to rate four statements that addressed the level of MHC over the Human-Machine Team performing the military mission.

RESULTS

Perceived plausibility

Responses to interview questions, think-aloud comments, and ratings of statements were used to determine whether participants experienced the presented scenario and experimental conditions as plausible representations of future operational contexts. Overall, participants reported the scenario to be sufficiently realistic for them to feel immersed in their role and task. However, there were also critical comments. For example, one participant noted that the threat (a single sniper) was not proportional to the available assets (a Boxer with a .50 caliber and three UGVs). Others viewed realism as less critical for experimental purposes: "*Don't fight the case: tomorrow it will be different.*"

The Human-Machine Teaming support system (see Figure 2) was evaluated quite critically. Although the participants considered the communication with DIALOG as credible, they expressed frustration about DIALOG requiring ample time to create plays from their input. In addition, they found DIALOG's repeated confirmation requests a hindrance for vigorous performance. As one participant emphasized: "*Time is a key determinant of product quality.*"

Participants were also critical about the behavior of the UGVs: they found it often tactically inadequate, unrealistic, and inefficient. Participants attributed this to a lack of basic military knowledge. For example, UGV behavior sometimes conflicted with standard military tactical procedures, such as maneuvering through open fields; appearing in front of friendly forces; causing collisions; and pointing weapons at own troops. Participants also disapproved of the extensive input that was needed to induce UGVs to perform the desired behavior. One participant said: "*I said 'engage when visual', but they didn't. I had*

to activate them two or three times." Overall, participants judged the UGVs' shortcomings to undermine the credibility of the systems. After a few missions, participants learned to deal with the UGVs' weaknesses, primarily by adapting their expectations (e.g., "*At first I expected better, then I realized they wouldn't react well*" (p.2)) and adapting their own behavior to compensate for the UGVs' shortcomings.

Delegation preferences

Each of the distinguished command modes: *Human-configured play-based command*; *Machine-configured play-based command*; and *goal-based command*, were exclusively used in, respectively, training missions 1, 2, and 3, as this was enforced by the experiment design. Once the test mission began, participants were allowed to select their preferred command mode, and they predominantly selected the goal-based command mode. When asked, a participant explained his preference as follows: "*You can accomplish things with easy tasks, while in your head you can already think about the next step*."

Yet the introduction of unexpected events during the test missions (e.g., in the form of snipers) prompted participants to use other command modes, in particular the Machine-configured play-based command. Participant 5 explained: "*DIALOG thought with me about how the task could be executed; that shows that it does not just execute orders, but also thinks with me*." Participant 3 remarked: "*DIALOG-specified has my preference because of user-friendliness and overview*". Participant 2, who varied between Human- and Machine-configured play-based command, expressed a desire for more control: "Ideally I want to use a goal-based agent; a supporter that can make decisions on its own to some degree, and that asks for input or notifies me to some degree".

Collaboration with DIALOG

After each training- and test-session, the participant was asked to rate the following statement on a 1–7 scale: "*In this mission I found it easy to use DIALOG when formulating orders to the UGVs*". Figure 8 shows the results:

In the second and third training session, participants seemed to appreciate DIALOG. It should be noted that during the training-sessions (T1–T3), no unexpected events were introduced. However, in the test-session (measurements T4–T7), unexpected events disrupted the expected course of the mission. This immediately resulted in a reduced appreciation of DIALOG as a support system in T4. In sessions T5 and T6, participants temporarily raised their judgment, to end with a relatively low rating after the final test-session. The main shortcoming brought forward was that DIALOG's support does not fit with how commanders traditionally work. For example, participant 2 noted: "*[DIALOG] doesn't draw up a plan. Can't visualize his plan, moving over this line. We do everything visually. We don't work with points, but with boxes*."

Trust

Results are derived from analysts' notes capturing participants' think-aloud verbalizations and behavior during mission execution. Additionally, the participants were asked verbally to rate their trust in DIALOG and in the UGV's performance on a 1–10 scale (see Section 4.5). Analysis of the data suggests that participants' trust is influenced by a number of factors:

Expectations

Participants exhibited varying levels of initial expectations regarding the trustworthiness of both DIALOG and the UGVs. When these expectations were high but unmet, trust ratings declined. Conversely, when expectations were met, trust levels remained relatively stable throughout the experiment.

Experience with LLMs

Participants' prior experiences with LLMs or artificial intelligence (AI) systems in general appeared to influence their baseline levels of trust. Those with positive prior experiences appeared more confident in the capabilities of DIALOG and UGV systems. In contrast, participants who had encountered limitations or failures with other AI systems expressed lower trust in DIALOG.

Autonomy of the system

Participants expressed frustration when DIALOG demanded extensive confirmation steps, particularly during final approvals or when DIALOG asked multiple questions for clarification (e.g., regarding mission goals or specifics of the selected play). Participants perceived these interactions as superfluous and as causing unnecessary delay in decision-making and task execution. In contrast, when DIALOG and the UGVs performed their tasks in a timely and efficient manner, participants indicated that they trusted the systems. Also, when DIALOG and the

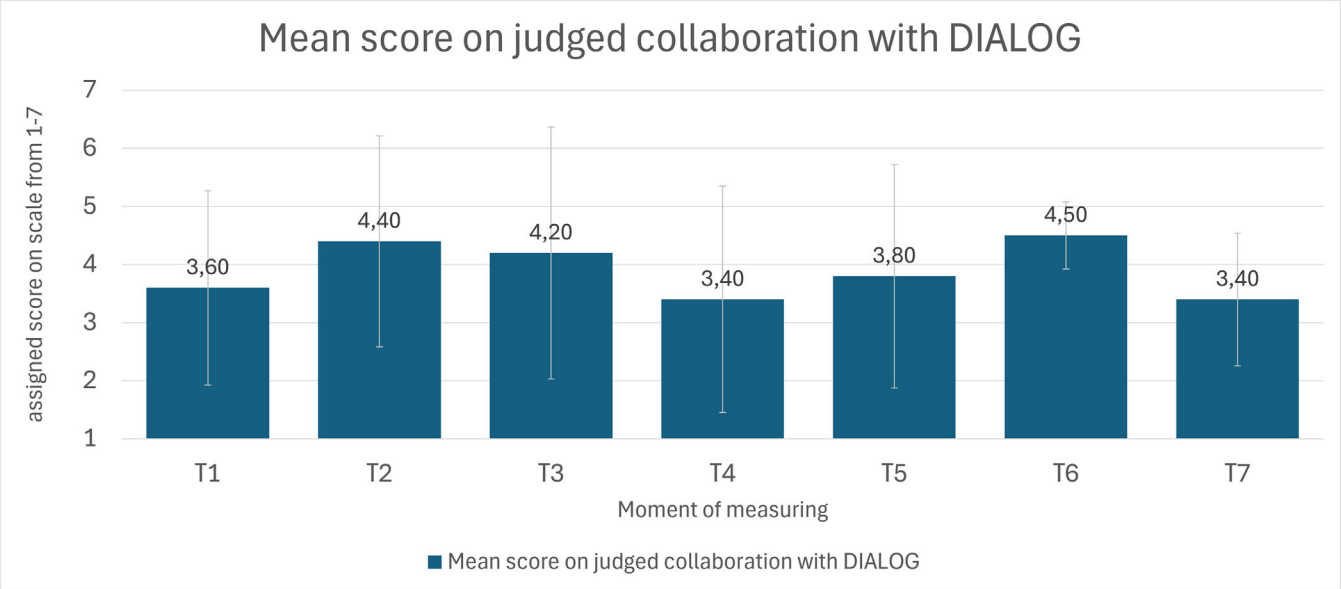


FIGURE 8 Experienced use of DIALOG for formulating commands to the UGVs ("In this mission I found it easy to use DIALOG when formulating orders to the UGVs"). T1 to T3 are training missions, T4 to T7 are test missions.

UGVs demonstrated the ability to act independently and decisively, without excessive reliance on user input, this was evaluated favorably by the participants.

Communication accuracy

When participants felt that DIALOG understood their intent correctly, they evaluated the subsequent queries and interaction as smooth and efficient, fostering higher trust in the system. Conversely, any misinterpretations or communication breakdowns undermined the participant's trust in DIALOG.

Allocation of intelligence

Participants frequently commented that they could not determine the locus of intelligence in the collaboration between DIALOG and the GPAI simulator (that governed the behavior of the UGV, see Section 3.1). Several participants argued that if the UGVs showed inappropriate behavior during the mission, they could not determine whether this was caused by inadequate assignments delivered by DIALOG, by inadequate decisions of the GPAI simulator regarding how to execute the assignment in the actual tactical context, or whether it was caused by both. This confusion caused participants' ratings of their trust in DIALOG and UGVs to fluctuate, depending on their assumption about which component of the Human-Machine Teaming support system was accountable for inappropriate actions of the UGVs.

Tactical knowledge

A recurring theme in participants' feedback was the perceived lack of tactical military knowledge in both DIALOG and the UGVs. Participants noted that the UGVs failed to exhibit expected behaviors such as: advancing in formation; appropriate decision speed; procedural awareness; logical reasoning; and adapting to terrain conditions. Participants held these deficiencies accountable for the UGVs unexpected behavior and suboptimal performance, like for example UGVs colliding with one another during approach in the mission area.

Task complexity

During the missions, participants adjusted their expectations of the capabilities of both DIALOG and the UGVs. Over the course of the experiment they learned which tasks could be handed over to DIALOG and to the UGVs, and which could not. Simpler tasks -such as monitoring- were more readily entrusted to DIALOG or the UGVs. For complex tasks, participants were inclined to assume control themselves.



FIGURE 9 Mean ratings on questionnaire statements addressing the experienced MHC.

Meaningful human control

After each training- and test-mission, participants were asked to rate statements that addressed the level of subjective MHC over the Human-Machine Teaming support system and over the Human-Machine Team performing the military mission. Figure 9 shows the results of mean ratings. Command mode was not related to subjective MHC.

Figure 10 shows the mean ratings on the question: "I found it difficult to keep an overview of what the UGVs were doing during the mission".

Note that a high score (7 = "strongly agree") would be associated with a low level of MHC. Participants' ratings were half-way up the scale, indicating they felt they had a moderate level of overview on the UGVs.

After each training- and test-session, the participant was asked to rate the following statement: "I think the behavior of the UGVs matches well with my intention of the assignment." Figure 11 shows the results:

The results show a similar pattern as the evaluation of DIALOG: relatively high assessments in sessions T2, decreasing to the midpoint-score of the scale in the final measurement of the test-session. Figure 12 shows participants' judgment on how well-fit they considered the UGVs' behavior in runs with unexpected events occurring.

The fact that the test-missions contained unexpected behavior of the UGVs, in combination with participants having low expectations on the UGVs' tactical abilities, as reported in Section 5.4, may explain the rather low scores on the final missions of the experiment.

DISCUSSION

It is anticipated that General-Purpose AI will offer new opportunities for human-machine teaming in complex tasks and environments, as GPAI is characterized by the capability to learn from experiences and to use this capability to adapt to new and unforeseen circumstances. However, this capability does not guarantee successful application. One risk is that little is known about how GPAI forms internal representations based on patterns of experienced events. GPAI may make incorrect generalizations when dealing with situations it was never trained on. Likewise, GPAI may go astray when believing that a pattern learned from familiar situations applies to a new situation, whereas in reality it does not. In addition, GPAI's pathways of learning and adapting are not traceable. This prevents human operators from reliably understanding and anticipating system behavior, especially under time pressure, and jeopardizes humans' capability to develop and exercise MHC.

It has been proposed that LLMs can bridge the gap between GPAI and humans, assisting users in formulating instructions that maintain operational clarity and control while granting unmanned systems appropriate degrees of autonomy (e.g., (Goecks & Waytowich, 2024;

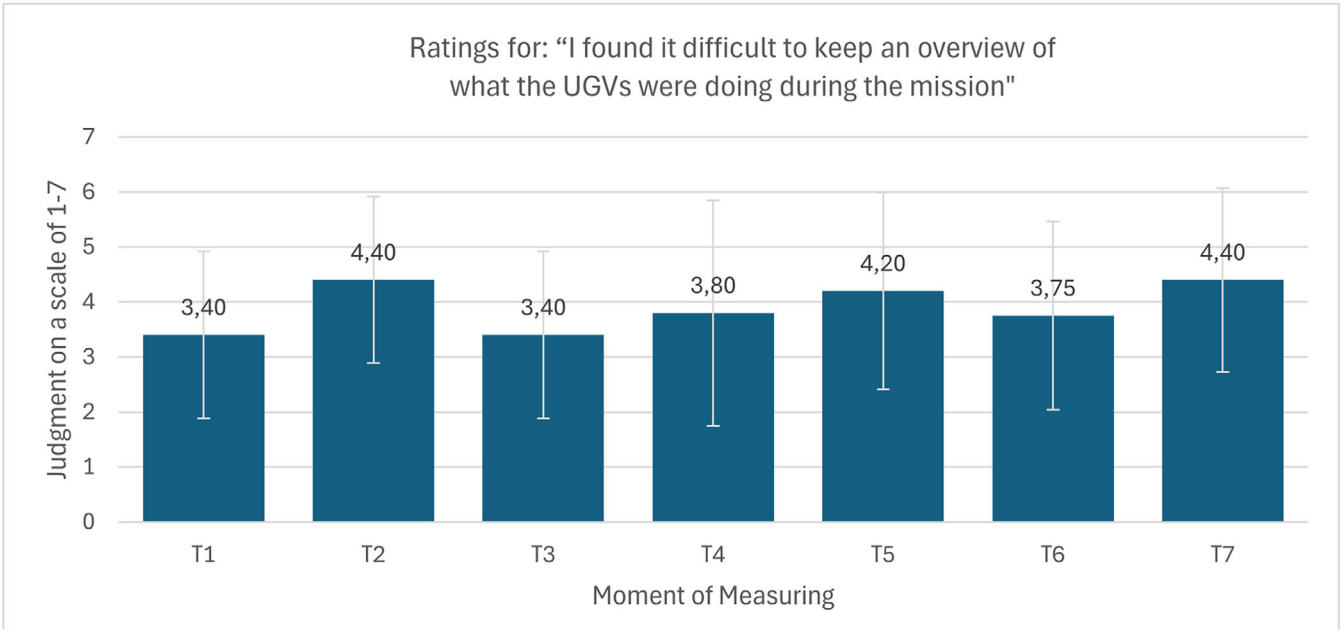


FIGURE 10 Mean ratings on: "I found it difficult to keep an overview of what the UGVs were doing during the mission".

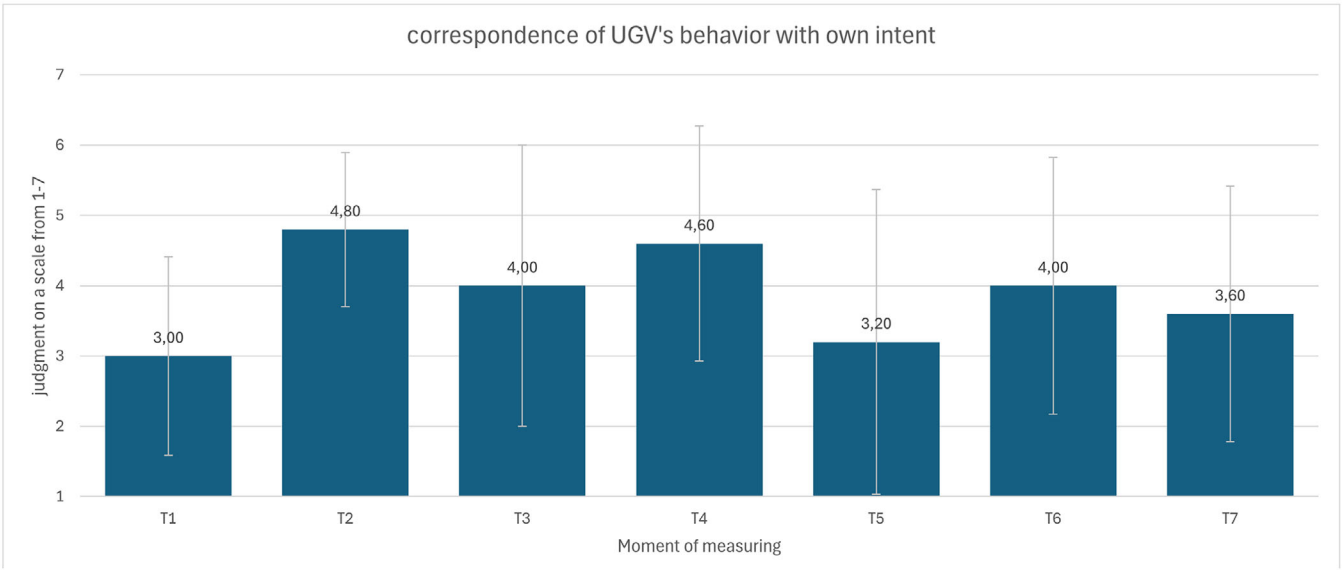


FIGURE 11 Question "I think the behavior of the UGVs matches well with my intention of the assignment".

Zou et al., 2025)). In this paper we explored this proposition in the context of a military human-machine team in which a human commander interacts with an LLM to design instructions for GPAI-based UGVs.

Plausibility

To be able to investigate how processes of human-machine teaming evolve, it is essential that participants perceive the scenarios and the performance of the UGVs as sufficiently plausible and credible. Participants reported feeling immersed in the role of TC, and generally found the scenario sufficiently realistic to engage with the systems in a meaningful manner. However, believability was undermined by inappropriate behavior by UGVs due to a lack of tactical knowledge in the play-library of DIALOG and in the GPAI-simulator of the UGVs (see Figure 2). Another limitation concerned the selected scenario. Participants pointed out that they considered the deployed assets,

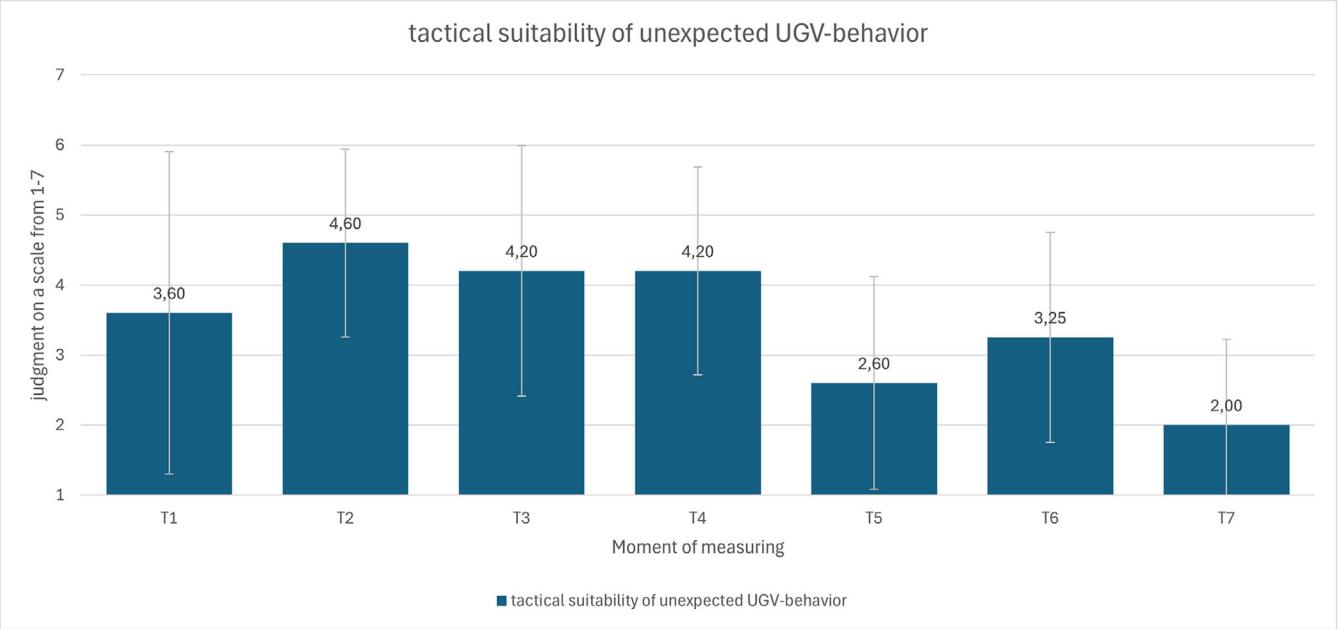


FIGURE 12 Mean scores on the question: "At times when the UGVs behaved unexpectedly, I still found the behavior of the UGVs appropriate for the situation".

namely a heavily armed Boxer vehicle and three UGVs, as disproportionate for eliminating a low threat, such as an individual sniper. This imbalance may have led some participants to perceive the scenario as a tutorial rather than a realistic and plausible mission, potentially reducing their concern over system failures to a level that is perhaps acceptable for simulations, but that would be unacceptable under real conditions. Such flaws impose a threat for the study, as low fidelity of the simulation may carry more weight in shaping participants' perceptions than the experimental manipulations themselves. These findings underscore that for maintaining immersion and operational realism (Chen & Barnes, 2014; Demir et al., 2019), it is important to align system behavior with participants' expectations, and to investigate for potentially confounding factors in the simulator.

Delegation preferences

Initially, participants appreciated the goal-based command mode, valuing its simplicity and autonomy. However, the introduction of unexpected events made participants shift toward modes that offer more control. Apparently, participants are prepared to delegate control to machines in predictable 'happy flow' situations, but desire more control over machines in dynamic, unexpected, and contextually-sensitive situations. Increased complexity and uncertainty prompts a need for greater human oversight (Van Diggelen et al., 2024). The blurred boundaries between DIALOG- and UGV-intelligence confused participants' understanding of the system. This suggests a need for providing participants with clear architectural transparency in future testbeds (Solaki et al., 2024).

Collaboration with DIALOG

When playing standard missions, participants evaluated the collaboration with DIALOG as moderately effective. Participants appreciated its conversational interface, but were dissatisfied with its inability to visualize plans. Ratings for ease-of-use declined substantially during test-sessions with unexpected events, when DIALOG's inability to adapt to dynamic contexts became evident. Clearly, DIALOG has limitations in supporting situational awareness and mission oversight (Hoffman, 2019).

Trust

The study uncovered a web of interrelated factors such as user expectations, prior experience, perceived autonomy, communication quality, and task characteristics that in a dynamic interplay shaped participants' trust. Crucially, trust strengthens when the system's output aligns with the user's mental model, implying that transparency and predictability are key to fostering reliable trust.

A notable challenge arose in distinguishing the participant's trust in DIALOG from its trust in the UGVs,

as participants often indicated not being able to identify which system was responsible for the actions taken. This is especially problematic for hybrid human-AI teams, where responsibilities are distributed and temporally extended. Trust depended not only on performance, but also on whether commanders understood the system's boundaries. When participants could not tell whether inappropriate behavior came from DIALOG, the UGV-side GPAI, or their interaction, accountability became unclear, reducing trust and MHC. Future GPAI-enabled HMTs should therefore make boundaries, responsibility, and command translations explicit. Moreover, trust should be considered not just at the individual level but also at the level of teams and organization, as the integration of AI in a team reshapes workflows, communication, and roles. It raises important questions on how existing architectures must evolve and adapt to be able to accommodate and integrate autonomous systems. Within the military context this need for implementation and design thinking is even more urgent, as the doctrine of how to execute specific military missions is likely to change with the deployment of drones and other unmanned systems.

Meaningful human control

The findings provide insight on how MHC was experienced. Overall, participants reported retaining some supervisory influence, yet their ability to understand, predict, and intervene was frequently limited, especially under dynamic mission conditions. These patterns align with earlier work showing that MHC becomes fragile when system behavior is only partially intelligible or insufficiently grounded in domain knowledge (Van Diggelen et al., 2024; Cummings, 2025).

A central issue concerns *predictability*, a core condition for MHC. Participants frequently observed that UGV behavior did not consistently reflect their intent, particularly during test missions with unexpected events. Participants often judged the behavior of UGVs, resulting from the input of DIALOG and the GPAI-simulator, as implausible or inefficient. Such deviations reduce the operator's ability to form accurate mental models of system behavior, echoing prior findings on the challenges of anticipating AI-driven decisions in high-stakes contexts (Hoffman, 2019; Van den Bosch et al., 2025).

Participants also struggled to maintain an *overview* of UGV activity. Ambiguity or confusion about how responsibilities were distributed between DIALOG and the GPAI-simulators hindered their ability to track why specific actions were taken. Since situational awareness is foundational to MHC—especially the ability to supervise and detect misalignments—this opacity directly constrained experienced control (Mitchell, 2019; Van Diggelen et al., 2024).

Related to this, *timely intervention* proved difficult. Although commanders could in principle override or redirect behavior, they frequently lacked clarity about the cause of inadequate UGV actions, making it hard to intervene confidently.

Contrary to expectations, *command mode* did not systematically influence subjective MHC. Instead, fluctuations in MHC ratings were driven primarily by the specific context of the mission and the system's performance. When missions unfolded predictably, participants favored goal-based delegation. However, as soon as uncertainty increased, they moved toward play-based modes to regain structure and oversight. This adaptive switching aligns with existing research showing that humans modulate the level of autonomy they accept based on perceived risk and system reliability (Nguyen et al., 2025).

Finally, *unexpected events* affected experienced MHC substantially, immediately lowering predictability, perceived alignment with intent, and overview. This illustrates a fundamental tension: GPAI is valued for its alleged ability to operate under unforeseen conditions, yet such events often expose knowledge gaps or behavioral inconsistencies that make MHC harder to maintain. Ensuring MHC in these conditions requires systems that not only react adaptively, but do so in ways that remain intelligible and aligned with operator expectations (Wenskovitch et al., 2024; Banerjee et al., 2025).

Methodological limitation

In this study we evaluated a Human-Machine Teaming support system designed to support a military commander in delegating GPAI-enabled UGVs in a tactical vehicle mission. To examine participants' judgments and experiences, we frequently made use of Likert scales. The advantage of this method is that it enables easy and rapid collection of input from participants. However, one drawback is that the participant is requested not to rate an actual experience, but to rate an experience in the past. The participants need to think back how much trust or control they felt at the time when a decision needed to be made. People may have difficulty at recalling a vivid and correct image of the experience, which requires bringing back the then felt values, affects, and emotions to the surface. It may well be that this difficulty is related to the tendency of participants to give ratings relatively close to the middle of the scale. Values at the extremes of the scale were rarely used. In our analyses we therefore focused not so much on the absolute levels of the scores, but on the emerging pattern over the course of the measurements.

CONCLUSION

This paper started from the observation that General-Purpose AI (GPAI) has the potential to substantially enhance human-machine teams, precisely because such systems can act autonomously and adapt to unforeseen situations. At the same time, this very capability risks coming into tension with the requirement for Meaningful Human Control (MHC), which demands that humans remain able to understand, supervise, and intervene in system behavior, especially in high consequence settings. The central challenge, therefore, is how to combine the adaptive strengths of GPAI with forms of control that remain intelligible and actionable for human commanders.

To address this challenge, we proposed a delegation approach that combines structured task specification with flexible interaction. First, we introduced plays as a means to predefine the types of tasks that autonomous systems can perform, while embedding these within a clear normative framework that reflects doctrinal, legal, and ethical constraints. Plays provide operational instructions that correspond with human intent. However, applying this approach to GPAI implies a large and diverse play library, which can become difficult for users to navigate and configure. For that reason, we added an LLM-mediated dialogue system that supports commanders in selecting, adapting, and instantiating plays through natural language interaction, without requiring them to manage the full underlying complexity.

We evaluated this solution direction in a simulated operational environment using simulated GPAI systems. The results indicate a clear pattern. Participants experienced a loss of control when instructions were issued directly at the level of abstract goals (which we called *goal-based command*), especially when unexpected events occurred. In contrast, play-based instruction was perceived as offering more structure, predictability, and support for maintaining control, which aligns with the assumptions underlying our framework. Participants also pointed to usability aspects of the play-based approach, noting both its potential and its current limitations. It is possible that some of the reported difficulties are due to the experimental nature of the prototype, including response time and poor tactical knowledge, rather than to the conceptual approach itself.

GPAI is expected to eventually support autonomous adaptation to new and unforeseen circumstances. However, this capability is still fragile and will probably never be fully attained. This requires a solid approach to human-machine collaboration that safeguards against errors while preserving the commander's ability to understand, supervise, and intervene. One advocated approach is the use of plays, which structure the behavior of unmanned systems in alignment with the commander's intentions. In this paper, we explored how such plays work out in a simulated real-world military mission. The results suggest that this approach might work: participants saw value in the structure and support provided by play-based delegation, although they also reported serious shortcomings in tactical reasoning, interaction speed, transparency, and reliability. Future work should not only focus on tactical reasoning and natural-language interaction, but also on developing verifiable guardrails, fallback procedures, and traceability mechanisms that allow commanders to understand and trust the operational boundaries of GPAI-enabled teammates.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributors to this study, whose empirical and conceptual work informed key parts of this article: Sjoerd Berning, Renske Bootsma, Niels Cornelissen, Sanne Kellij, Jose Kerstholt, Birgit van der Stigchel, and Anthia Solaki.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Karel van den Bosch https://orcid.org/0000-0002-0427-8275

Jurriaan van Diggelen https://orcid.org/0000-0001-8216-1286

REFERENCES

Arabzadeh, N., J. Kiseleva, Q. Wu, C. Wang, A. Awadallah, V. Dibia, A. Fourney, and C. Clarke. 2024. Towards Better Human-Agent Alignment: Assessing Task Utility in LLM-Powered Applications.

Banerjee, S., A. Agarwal, and S. Singla. 2025. "LLMs Will Always Hallucinate, and We Need to Live With This." In *Intelligent Systems and Applications*, edited by K. Arai 1554, 624–48. Cham: Springer Nature Switzerland.

Boers, J., T. Etty, M. Baars, and K. van Boekhoven. 2025. "Exploring Cognitive Strategies in Human-AI Interaction: ChatGPT's Role in Creative Tasks." *Journal of Creativity* 35: 100095. https://doi.org/10.1016/j.yjoc.2025.100095

Brachman, M., S. Kunde, S. Miller, A. Fucs, S. Dempsey, J. Jabbour, and W. Geyer. 2025. "Building Appropriate Mental Models: What Users Know and Want to Know About an Agentic AI Chatbot." In: *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 247–64, Cagliari Italy: ACM. https://doi.org/10.1145/3708359.3712071

Brown, B. R. 2025. "Human-Machine Teaming Using Large Language Models." In: *Interdependent Human-Machine Teams*, 41–66. Cambridge: Elsevier .

Chen, J. Y., and M. J. Barnes. 2014. "Human–Agent Teaming for Multirobot Control: A Review Of Human Factors Issues." *IEEE Transactions on Human-Machine Systems* 44(1): 13–29 .

Cummings, M. 2025. "Prohibiting Generative AI in any Form of Weapon Control." In: *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 1–10. San Diego, CA: Neural Information Processing Systems Foundation.

Cummings, Mary L. 2017. "Automation Bias in Intelligent Time Critical Decision Support Systems." In *Decision Making in Aviation*. New York, NY: Routledge, https://www.taylorfrancis.com/chapters/edit/10.4324/9781315095080-17/automation-bias-intelligent-time-critical-decision-support-systems-cummings.

Demir, M., N. J. McNeese, C. Johnson, J.C. Gorman, D. Grimm, and N.J. Cooke. 2019. "Effective Team Interaction for Adaptive Training and Situation Awareness in Human-Autonomy Teaming." In: *2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, pp. 122–26. Las Vegas, NV, USA: IEEE .

Draper, M., and J. Van Diggelen. 2025. "Human Systems Integration for Meaningful Human Control over AI-based Systems." Tech. Rep. TR-HFM-330, NATO.

Engels, C., Linhardt L., and Schubert M.. 2025. "Code Is Law: How COMPAS Affects the Way the Judiciary Handles the Risk of Recidivism." *Artificial Intelligence and Law* 33(2): 383–404. https://doi.org/10.1007/s10506-024-09389-8.

European Commission 2026. "General-Purpose AI Obligations Under the AI Act | Shaping Europe's Digital Future." https://digital-strategy.ec.europa.eu/en/factpages/general-purpose-ai-obligations-under-ai-act

Firoozi, R., J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, M. Schwager, et al., 2025. "Foundation Models In Robotics: Applications, Challenges, and the Future." *International Journal of Robotics Research* 44(5): 701–39.

Goecks, V.G., and N. Waytowich. 2024. "COA-GPT: Generative Pre-Trained Transformers for Accelerated Course of Action Development in Military Operations." Paper presented at International Conference on Military Communication and Information Systems (ICMCIS), Koblenz, Germany, 01–10, 2024.

Guardian Staff Reporter. 2026. "Anthropic's AI Claude Hacked Into Three Organizations During Cybersecurity Test." Technology. The Guardian, https://www.theguardian.com/technology/2026/jul/30/anthropic-ai-claude-hack.

Hao, Karen. 2019. "AI Is Sending People to Jail—and Getting It Wrong." MIT Technology Review. https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/.

Heijnen, M., T. Schoonderwoerd, M. Neerincx, J. van der Waa, L. Kester, J. van Diggelen, and P. Elands. 2024. "A Socio-Technical Feedback Loop for Responsible Military AI Life-Cycles From Governance to Operation." In: *Responsible Use of AI in Military Systems*, 17–36. New York, NY: Chapman and Hall/CRC.

Heim, E., and C. Frank. 2025. Out of Distribution Detection: Knowing When AI Doesn't Know. https://www.sei.cmu.edu/blog/out-of-distribution-detectionknowing-when-ai-doesnt-know/.

Hoffman, G. 2019. "Evaluating Fluency in Human–Robot Collaboration." *IEEE Transactions on Human-Machine Systems* 49(3): 209–18.

Langer, M., R. Dachselt, Q.V. Liao, T. Miller, and N. Tintarev. 2026. "Challenges of Human Oversight: Achieving Human Control of AI-Based Systems (Dagstuhl Seminar 25272)." Tech. Rep. 6, Schloss Dagstuhl—Leibniz-Zentrum für Informatik. https://doi.org/10.4230/DAGREP.15.6.189

McKnight, D. H., V. Choudhury, and C. Kacmar. 2002. "Developing and Validating Trust Measures for e-Commerce: An Integrative Typology." *Information Systems Research* 13(3): 334–59. https://doi.org/10.1287/isre.13.3.334.81

Mitchell, C. 2019. "When Laws Govern LAWS: A Review of the 2018 Discussions of the Group of Governmental Experts On the Implementation and Regulation of Lethal Autonomous Weapons Systems." *Santa Clara High Technology Law Journal* 36: 407 .

Nguyen, H., Y. Wang, K. Andre, W. Duan, C. Flathmann, and N. McNeese. 2025. "Modeling Adaptive Autonomy at the Team Level: Understanding Team-Wide Autonomy and its Impact on Situation Awareness in Human-Autonomy Teams." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 69(1): 2114–19. https://doi.org/10.1177/10711813251358809

O'Neill, T., N. McNeese, A. Barron, and B. Schelble. 2022. "Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature." *Human Factors* 64(5): 904–38. https://doi.org/10.1177/0018720820960865

Solaki, A., E.S. Kox, S. Berning, and K. Van den Bosch. 2024. "V2205 Effects of Unexpected Events on Trust and Task Delegation in Military Human-Machine Teams." Tech. Rep. *TNO* 2024 R12237, TNO, Soesterberg, the Netherlands .

Van den Bosch, K., J. Van Diggelen, S. Verdult, T. Haije, and J. Van der Waa. 2025. "Measuring Meaningful Human Control in Human–AI Teaming: Effects of Team Design in AI-Assisted Pandemic Triage." *AI and Ethics* 5: 3329–53. 10.1007/s43681024-00647-8.

Van Diggelen, J., J. Barnhoorn, R. Post, J. Sijs, N. Van Der Stap, and J. Van Der Waa. 2021. "Delegation in Human-Machine Teaming: Progress, Challenges and Prospects." In *Intelligent Human Systems Integration 2021*, edited by D. Russo, T. Ahram, W. Karwowski, G. Di Bucchianico, R. Taiar 1322, 10–16. Cham: Springer International Publishing.

Van Diggelen, J., K. Van den Bosch, M. Neerincx, and M. Steen. 2024. "Designing for Meaningful Human Control In Military Human-Machine Teams." In *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems* edited by G. Mecacci, D. Amoroso, L. Cavalcante Siebert, D. Abbink, J. Van Den Hoven, F. Santoni De Sio, 232–52. Cheltenham, UK: Edward Elgar Publishing. https://doi.org/10.4337/9781802204131.00021

Walker, P. M., C. A. Miller, J. B. Mueller, K. Sycara, and M. Lewis. 2019. "A Playbook-Based Interface for Human Control of Swarms." In: *Human Performance in Automated and Autonomous Systems*, 61–88. Boca Raton, FL: CRC Press.

Wenskovitch, J., C. Fallon, K. Miller, and A. Dasgupta. 2024. "Characterizing Interaction Uncertainty in Human-Machine Teams." In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*, 1–6. Toronto, ON, Canada: IEEE.

Yuval, Abraham. 2024. "Lavender': The AI Machine Directing Israel's Bombing Spree in Gaza." +972 Magazine. https://www.972mag.com/lavender-ai-israeli-army-gaza/.

Zou, H. P., W. C. Huang, Y. Wu, Y. Chen, C. Miao, H. Nguyen, Y. Zhou, et al., 2025. "LLM-Based Human-Agent Collaboration and Interaction Systems: A Survey". https://doi.org/10.48550/arXiv.2505.00753

How to cite this article: van den Bosch, K., P.-P. van Maanen, and J. van Diggelen. 2026. "General-purpose agents in human-machine teams: Can we preserve meaningful human control?" *AI Magazine* 47: e70091.https://doi.org/10.1002/aaai.70091

AUTHOR BIOGRAPHIES

Karel van den Bosch has been working as a Senior Research Scientist at TNO in the Netherlands. He retired in April 2026. He managed and conducted fundamental and applied research on the training of complex cognitive tasks (e.g. command & control, decision making, crisis management). The objective of his work was to make simulation-based training more effective by designing cognitive software agents (e.g., agents playing the role of team mate, adversary, or instructor). He investigated how such agents can successfully support training, thus making training more systematic (uniform behaviour of agents), more effective (agents consistently eliciting intended behaviour of trainee), and more efficient (staff and team members need no longer be present during training). He supervised PhD students on cognitive modelling for training simulations; self-explaining virtual agents; automatic direction of training scenarios; and on Human-AI Co-Learning for Human-Machine Teams.

Peter-Paul van Maanen is a senior engineer, researcher, and educator with a PhD in Artificial Intelligence. He has authored over 60 publications in peer-reviewed conferences and journals. His work focuses on applied research aimed at improving safety, performance, health, and wellbeing. He has held positions at TNO, VU University, and in industry, combining research, teaching, and entrepreneurial activities. As founder and CTO of several AI-based startups, he has contributed to the translation of advanced AI methods into practical applications. His interests include multidisciplinary approaches to addressing real-world problems.

Jurriaan van Diggelen is a principal scientist at the department Human-Machine Teaming at TNO in Soesterberg. He has a PhD in Artificial Intelligence, and currently leads the defense research program Human-machine teaming. He is principal investigator of the ELSA lab consortium which aims to assure Ethical, Legal, and Societal aspects of military Artificial Intelligence. He is chair of several NATO groups on meaningful human control of AI-based systems and Member at Large at the NATO Human Factors and Medicine panel. His research primarily addresses the development and implementation of responsible and effective AI-driven human-machine teaming solutions for high-stakes defense and security applications. A key focus of his work is ensuring trust, explainability and contextual understanding in AI systems — not only to improve human understanding of machine behaviour, but also to enable AI systems to better interpret human intentions and decision-making processes. In collaboration with the Netherlands Ministry of Defense, he leads research into the safe and ethical integration of AI-based systems to ensure that they enhance situational awareness, decision support and operational effectiveness while maintaining human oversight. His expertise includes Artificial Intelligence, AI Ethics, Human-Machine Teaming and Cognitive Systems Engineering.