Evaluating the Relationship between Trust and Transparency in Military AI outputs

C.A.Pestana Navea (s286963)

University of Twente

Supervisors: L.Seeleman & I.van Sintemaartensdijk

Thesis

Word count: 8760

June 10th, 2025

Abstract

This study aimed to understand how transparency of AI system outputs affects users trust in high-stake military scenarios. Participants were assigned randomly to one of two scenario pairings, each pair had one high transparency output and one low transparency one. Thus, this study used a mixed design and gathered both quantitative and qualitative data. Quantitative results showed trust was higher in the high transparency outputs compared to the low transparency ones. However, this effect was only found in one scenario pairing. Furthermore, qualitative analysis showed that participants often relied on their own reasoning and used the AI for support. The results show that trust in military AI systems is not only influenced by the level of transparency of its outputs. It is also influenced by how users interpret the context of the situation and what type of role they assign to the system (i.e., autonomous, or decision-making aid).

Introduction

How much do you trust the use of Artificial Intelligence (AI) systems in the military? The European Commission High-Level Expert Group on Artificial Intelligence (2019) refers to AI as systems that are built by humans to achieve certain goals. They are usually designed as software's and also with hardware (e.g., drones). These systems collect data from the environment and multiple other sources that are fed into the system. Then the system interprets it and based on that interpretation, it decides what the best decision or action is. In the military context, AI is used to help with tasks like surveillance, planning missions and helping military personnel make fast decisions during operations (Cummings, 2017; Horowitz et al., 2018). Cummings (2017) defines military AI systems as those that carry out or help with tasks related to combat. The author explains that these systems use algorithms that are able to adapt to environments that are in constant change. These types of systems are designed to operate in complex and high-stake situations.

Take, for example, the "Habsora," also called "the Gospel" in English, which is an AI system designed by the Israeli Defence Forces (IDF) for target creation (Davies et al., 2023). The way this AI system creates targets is still a mystery since the IDF does not provide any descriptions of it (Sylvia, 2024). What is known is that Habsora combines data gathered from drones, surveillance, intercepted communications, etc; however, under what conditions the AI comes to the output is still unknown (Sylvia, 2024). The IDF assures that the output is first given to a directorate composed of multiple soldiers and analysts, and only when approved they can strike (Sylvia, 2024). However, as Schraagen (2023) explains, in war, military personnel often encounter situations where they have time pressure and are forced to make decisions using

incomplete data. To do so, they must receive help from an AI system which they sufficiently trust.

Trust in AI systems

Before we begin talking about trust, it is important to bear in mind that trust can be defined in different ways across multiple contexts; thus, it is important to keep a clear definition of it when working with AI systems. This research has adopted the definition of trust given by (Jacovi et al., 2021). They explain trust in AI as an explicit contract between the user and the AI system. Whichever the contract is, the user expects the AI to work according to the contract. The goal of the user to trust the AI system is grounded in anticipating the system will follow the contract in case of uncertainty (Jacovi et al., 2021). The author explains that the key aspect here is "anticipating" because if the user can anticipate the behaviour of the AI system, then they trust that the contract will be followed in case of doubt.

However, this raises the question of how should AI systems be designed to ensure users are able to anticipate their behaviours? Thus, to optimize the use of AI systems, especially in decision making where users interact directly with technology Xu and Gao (2024) propose the Intelligent Sociotechnical Systems (iSTS) framework. The iSTS framework explains that to make an AI system optimal, it is necessary to balance both the social and technical aspects. The iSTS builds upon the principle of human-centred joint optimization, this means the focus should be on how to optimise human capabilities. Thus, it includes human decision-making in every part of the process. According to Xu and Gao (2024), this means that AI systems should be designed in a way that provides humans with authority, enhances the understanding between users and machines, and account for human needs for example, trust, ethics, and clarity. In this framework, transparency and explainability are crucial for enabling that understanding.

Transparency

A form of contract could be the one proposed by the European Commission High-Level Expert Group on Artificial Intelligence (2019) who realised an ethics guideline for trustworthy AI. In this guideline, they provide seven requirements for AI systems to be deemed trustworthy, and one of them is transparency. They explain that AI systems must be able to explain which data, rules, and models it used that led to a certain output. In addition, they should be able to clearly state the reasoning behind their output. In this ethics guideline they also explain that the degree of detail towards which the AI system must explain itself depends on the context and risks associated. The higher the risk associated with the output, the greater the need for a transparent explanation. Moreover, the European Commission High-Level Expert Group on Artificial Intelligence (2019) proposes that the AI system must inform the user about their capabilities and limitations. Thus, if these requirements are to be met, then the AI system must explain its decision-making process and its capacity. Moreover, giving transparent explanations is crucial for safety purposes (Song et al., 2024). For example, AI drone systems should provide transparency by explaining their actions in order to verify if they comply with safety and ethical standards. However, as Schraagen (2023) explains, military AI usually lacks transparency and explainability. A clear illustration of this is the previously mentioned example of "Habsora", where the IDF does not explain how the AI system determines targets. Thus, there is a need to increase transparency and explainability in the military use of AI systems. However, it is difficult to make the AI-system provide outputs that are concise for military personnel to make decision in real time and comprehensive enough to meet any ethic or legal concerns (Johnny, 2024).

Natural Language and Capabilities and Limitations

Consequently, Balasubramaniam et al. (2023) explains that helping people understand the decision-making process of the AI system positively influences its transparency. Thus, Lou and Wei (2023) propose that transparency can be incremented using natural language. Natural language is a language that has been originated by people (e.g. Arabic, English and Spanish) compared to a language developed by computers (e.g. coding) (Schraagen, 2023). This means that natural language is a language people use in everyday tasks to communicate with one another. Consequently, as Lou and Wei (2023) mention, if the AI system is capable of providing explanations for its decision-making process using natural language, then users can understand better. Thus, this approach increased trust in AI decision-making. Accordingly, Druce et al. (2021) explains that users do not just need raw AI output but also a simple, easy-to-understand explanation of that output. Meaning that to allow people to understand outputs which contribute to the trustworthiness of the AI system, it is not enough to present users with a graph or numerical data; this must be accompanied by an explanation using simple language (i.e., natural language). Using natural language in the design of an AI system to explain how it got to the answer helps bridge the gap between human and computer knowledge (Lou & Wei, 2023).

However, there is a lack of research regarding transparency and trust in the context of military use of AI systems, pointing out the importance of it. Accordingly, Kim et al. (2023) explains the importance of context when talking about trust in AI, explaining that often, users do not trust the use of AI systems in high-stakes situations. In this type of environments, it might be difficult for users to build trust. Accordingly, to address this problem, Tomsett et al. (2020) propose the concept of rapid trust calibration. This refers to the process where users have to quickly determine if they can trust an AI system. According to the authors, this process is influenced by certain factors. The first one is how easy is it to understand the system's outputs

(i.e., interpretability). The second factor is related to how well can the system communicate its uncertainty. Thus, Tomsett et al. (2020) explain that if users can understand what the AI system knows and what it does not know, they are better able to decide if they should trust its decision. This means that trust is not just about giving information, but also how clearly and appropriately that information is communicated in high-stake situations.

Consequently, as mentioned before, the European Commission High-Level Expert Group on Artificial Intelligence (2019) present an ethics guideline for trustworthy AI. The guidelines state that a requirement for systems to be transparent and deemed trustworthy, is that they must inform users of their capabilities and limitations. Meaning that it is important to communicate to the user what the AI system can do and its limits. For example, Helldin et al. (2013) explains that users are better able to calibrate their trust in an automated system when they are presented with information about uncertainty (i.e, how confident the system was about the output). This does not necessarily mean that users will trust the decision of the AI system more but that they will trust the systems' limitations and thus make better decisions. If the user identifies a risk on the output, based on the uncertainty provided by the AI system, then they will not accept the decision made by the AI system. Supporting this view, Chen et al. (2025) explains that users do not trust an AI system when given a single number of uncertainties; they trust more when the system informs them about uncertainty by giving a range of possibilities, both optimistic and pessimistic. Thus, they propose that if the AI system provides users with both positive and negative outcomes, then they can expect more trust towards the system. All in all, this position is also supported by Tomsett et al. (2020), who proposes that AI systems should explain outputs clearly and communicate uncertainty so that users can responsibly trust it. However, the authors

point out that further research must be done to determine the best way to communicate these factors.

Current Research

Consequently, this research paper aims to investigate how the interaction between humans and AI systems can be optimised by focusing on the effect of transparency of AI military systems on trust. This will be done by answering the question: How does the level of transparency of an AI military system influence trust dynamics when users are unable to verify responses? We expect that a higher level of transparency, thus the use of an explanation of the output using natural language and a description of capabilities and limitations, will promote more trust among users, compared to the low transparency output.

Methods

Participants and Research Design

The current study was conducted using a mixed research design; thus, it was a within-subject (i.e., transparency levels high vs low) and between-subject design (i.e., scenarios were paired A-C and B-D). Meaning that there were four scenarios in total, two with high transparency (i.e., scenarios A and D) and two with low transparency (i.e., scenario C and B). Thus, they were paired in such way that each contained one low transparency scenario and one high transparency one. Participants were assigned randomly one pair either to A-C or B-D. To recruit participants, this study used convenience sampling, social media and snowball sampling. To make use of convenience sampling, we used the university participant pool "SONA" where participants received 0.5 credits for completing the survey. Moreover, participants had to be at least 18 years old and fluent in English to make sure they understood the content of the survey.

The research gathered initially 122 participants who completed the survey; however, 40 participants did not complete a minimum of 66% of the full survey. Thus, the final sample consisted of 82 participants. Of these 82 participants, 10 did not successfully finish the study but their data was still saved because they managed to fill in at least 66% of the study. So, from the 72 participants who did manage to fill in the full survey, 40 of them were female (48.8%), 29 were male (35.4%), one participant identified as another gender (1.2%), and two preferred not to say their gender (2.4%). Their ages ranged from 18 to 54 years (M= 24.2, SD = 6.09). Moreover, 16 participants were Dutch (19.5%), 17 were German (20.7%), and 39 were of other nationalities (47.6%). Finally, regarding participants' higher level of education obtained, 1.2% of participants completed lower secondary education, 28.0% completed upper secondary education, 40.2% completed a bachelor's degree or equivalent, and 17.1% a master's degree or equivalent. One participant (1.2%) selected "Other" as their highest level of education.

It is important to note that this study received ethical approval with application number 250451 from the Ethics Committee of the Faculty of Behavioural, Management and Social Sciences from the University of Twente.

Materials

Questionnaires

Public Attitudes toward Intelligent and Cognitive Entities (PAICE). The PAICE questionnaire by Scantamburlo et al. (2023) was used to assess participants in three dimensions, awareness, attitudes and trust in AI technologies. The first dimension of the PAICE is awareness which was measured with six questions. Participants had to self-assess their knowledge of AI, measure the perceived impact of AI in their life, indicate their awareness about AI-integrated products and about the application of AI across sectors in Europe (e.g. healthcare and law

enforcement). All of these items were rated using different five-point Likert scales. Furthermore, participants indicated their familiarity with European AI-related initiatives (e.g. GDPR and Ethics Guidelines for Trustworthy AI). This question was assessed using a dichotomous scale "Yes or No". Finally, the last item was for participants to select which technologies they believed incorporated AI (e.g., messaging apps and drones). This question involved participants checking one or multiple boxes in which they believed AI was used.

The second dimension of the PAICE is attitude which is originally measured using four questions. However, given the purpose of the study, two questions which involved scenarios were deleted from the survey we presented to participants because they seemed closely similar to the ones we created. Thus, the questions for this subscale were about the general attitude they had toward AI systems and towards AI use in different sectors across Europe (e.g. healthcare and finance). Both questions were rated using five-point Likert scales (1 = *Strongly disapprove* to 5 = *Strongly approve*).

The third and final dimension of the PAICE is trust which was assessed using four questions. To begin with, question nine was adapted, in the original survey, they asked participants to rank the three most important ethical AI principles, however, in this study participants were asked to rate all of them from 1 being the most important to 7 being the least important. Moreover, they were asked about which measures they think could increase their trust in AI (e.g. "A set of laws enforced by a national authority which guarantees ethical standards and social responsibility in the application of AI."). Additionally, they were asked about the importance of education in AI systems, to increase trust in them and about how much they trust certain institutions to use AI in the best interest of the public. These questions were all rated on five-point Likert scales.

Risk Perception. The Risk Perception questionnaire by Walpole and Wilson (2021) was adapted to assess participants risk perception of misuse of AI military systems. This scale has four subscales, affect, exposure, susceptibility, and severity. It is important to note that participants were given the following definition for misuse of AI systems in the military context "we define the misuse of AI as using artificial intelligence in ways that are harmful and unethical, thus, going against international laws. A clear example of this could be the military letting an AI system take important decisions on its own without human supervision." Thus, to begin with the first subscale was affect and consisted of three questions all assessed using a five-point Likert scale (1 =*Not at all,* 5 =*Extremely*). For example, the first question was phrased in the following way "How concerned are you, if at all, about the misuse of AI in military operations?".

The next subscale was exposure which consisted of four questions, all rated using different five-point Likert scales. For example, the first question of this subscale was adapted in the following way "How likely is it that the misuse of AI systems for decision making in the military context occurs in your country?". Moreover, susceptibility was the next subscale, and it consisted of three questions which were rated using five-point Likert scales (1 =*Not at all likely,* 5 =*Extremely likely*) and (1 =*Not at all vulnerable,* 5 =*Extremely vulnerable*). An example of how these questions were phrased is the following: "If AI were to be misused for military decision-making, how vulnerable would you be to the impacts?". Finally, the last subscale was severity, and it consisted of three questions (e.g., "How severe would you expect the consequences of misuse of AI systems in the military decision-making process to be?"), also all assessed using a five-point Likert scale (1 =*Not at all severe,* 5 =*Extremely severe*).

Checklist for Trust between People and Automation. The Checklist for Trust between People and Automation by Jian et al. (2000) was used to assess how much trust participants had on the surveillance drones from the two different scenarios they saw. This checklist originally consisted of twelve statements participants had to rate on a seven-point Likert scale. However, in the current study only eleven statements were used, and they were rated on a five-point Likert scale (1 =*Not at all,* 5 =*Extremely*). This was done because the last statement regarded how familiar the participant was to the system, since they were not in constant interaction with the system, this statement was eliminated from the final questionnaire.

Realism of scenarios. The Realism of scenarios scale was used; however, it was an adapted version of the scale by Van Gelder et al. (2018). The items were rephrased to fit the two scenarios' participants saw in this study (e.g., "I thought the scenario was convincing."). The Realism of Scenarios consisted of 6 items which were all assessed on a five-point Likert scale; (Strongly disagree = 1; Strongly agree = 5).

Vividness of visual imagery questionnaire. The Vividness of visual imagery questionnaire (VVIQ) by Marks (1973). was administered to assesses the extent to which participants could imagine the scenarios. The VVIQ consists of four subscales, each containing four items, thus making up 16 items in total. The first subscale regarded imagining a relative and the second subscale asked participants to imagine the sun. Next, the third subscale consisted of imagining a shop and finally, the fourth regarded imagining a mountain. Each item was rated using a five-point scale, 1 = No image at all, you only "know" you are thinking of the object, 2 = Dim and vague; flat, 3 = Moderately clear and vividly, 4 = Clear and lively, 5 = Perfectly clear and as lively as seeing it for real.

Furthermore, descriptive statistics and test statistics were also calculated for the scales and their respective subscales (see Table 1).

Table 1

Descriptive Statistics and Internal Consistency for All Questionnaire Measures

Variable	M	SD	α
PAICE (Awareness)	3.48	0.60	.85
PAICE (Attitude)	3.58	0.74	.90
PAICE (Trust)	3.54	0.58	.81
Risk Perception (Affect)	3.39	1.00	.87
Risk Perception (Exposure)	2.64	0.77	.71
Risk Perception (Susceptibility)	2.88	1.09	.92
Risk Perception (Severity)	3.48	0.98	.85
Realism of Scenarios	3.77	0.69	.77
VVIQ	3.49	0.81	.94

Note. α = Cronbach's alpha.

Scenarios

Four scenarios were created that involved the use of an AI powered surveillance drone in the context of military decision-making (see Appendix A, B, C and D). The scenarios were created so that participants would then receive an output from the drone. Since each scenario would provide visual as well as written outputs from a surveillance drone, the images for each scenario needed to be hyper realistic. Thus, the images were created using AI, a clear description of what the images had to have was submitted to ChatGPT 4.0. When the image was as desired,

then it was uploaded to Canva Pro to create the final output. Thus, all the text output were created and added, as well as some design features to simulate a real drone interface.

These scenarios varied in the level of transparency that the output provided. Thus, two scenarios had a low level of transparency and two had a high level of transparency. The ones that had low transparency had a picture taken from the drone and an explanation of the system's output in codes as well as a suggestion for action. However, in these scenarios, they also lacked information about system capabilities and limitations (see Appendix A and B). Furthermore, the scenarios with high transparency also provided a visual output (i.e., surveillance footage). However, these scenarios included an explanation in natural language along with the capabilities and limitations of the system (see Appendix1, scenarios C and D). Additionally, the scenarios also varied in the context they were placed. Thus, Scenario C (i.e., Low transparency) was about an unusual digging found in a Dutch military base in Iraq. Additionally, Scenario D (i.e., High transparency) was about an unidentified vehicle found in a NATO military base in Mali. Moreover, scenario A (i.e., High Transparency) was about a man holding a suspicious object found in a crowd outside an important conference venue in the Hague. Finally, scenario B (i.e., Low Transparency) was about a suspicious object found in a trash bin near a big protest in Amsterdam. After each of the scenarios, they were asked "How likely are you to accept the output given by the AI system?", they were asked to plot their answer using a 5-point Likert scale. Subsequently, they were also asked to explain in a textbox their reason for accepting or rejecting the output.

Demographic questions

Participants were asked demographic questions along with the multiple questionnaires they received. They were asked to state their gender, age, nationality and highest level of education obtained.

Military Knowledge and Experience. To assess participants perceived knowledge and experience with the military, five items were created. The first item asked participants to rate on a five-point scale (Not familiar at all=1; Extremely familiar=5) how familiar they were with the roles and responsibilities of the military in their country. A total of 26.8% participants said they were "slightly familiar" with the military roles in their country while 25.6% were "not familiar at all" and only 4.9% were "extremely familiar" (M = 2.33, SD = 1.16). Secondly, participants were asked to select all the military functions they were familiar with (e.g. National defense against external threats and Cybersecurity and defense against digital threats). The number of functions selected was (M = 2.38, SD = 1.49). The third item asked participants if they have engaged with military-related content outside of AI, they were given five options to select all of which applied. The number of selected content types on average was (M = 1.49, SD = 1.50).The fourth item consisted of asking participants if they know someone who has served in the military. Only 35.4% of participants reported knowing someone who had served in the military. The final item asked if participants themselves served or had gone through military training. They reported that 2.4% were currently serving and 7.3% had previously served, however, the majority being 72% had no experience in the military.

Procedure

Participants began the survey by filling in the inform consent, after they responded to the PAICE questionnaire, next they were presented with the adapted version of the Risk Perception. Moreover, they were randomly assigned to one low transparency and one high transparency

scenario. Thus, there were four scenarios A (high transparency), B (low transparency), C (high transparency) and D (low transparency), participants could either be assigned to A and C in a random order or B and D also in random order. After each of the scenarios, they were asked the extent to which they would accept the output of the AI system and why they would do so. Furthermore, they were presented with the Checklist for Trust between People and Automation. Once they were done, they filled in the adapted version of Realism of Scenarios, after they responded to the VVIQ. Furthermore, they were presented with the Military Knowledge and Experience scale and finally they filled in the demographic questions. The survey was finalized by debriefing and reiterating that they could withdraw from the survey by sending and email to the researcher within 10 days of participation.

Results

Descriptive statistics

To begin with, descriptive statistics were calculated to explore the effect that transparency in AI outputs has on user's trust (see Table 2) The mean trust scores in the high-transparency outputs were slightly higher compared to low transparency ones.

Table 2

Descriptive Statistics for Trust Scores across the Scenarios

Scenario	Transparency	M	SD
A	High	3.03	0.24
B	Low	2.83	0.29
C	Low	3.02	0.30
D	High	3.06	0.34

Note. The scenarios with high transparency (i.e., A and D) contained outputs explained in natural language as opposed to the ones with low transparency (i.e., C and B) which contained outputs explained using codes.

In addition, descriptive statistics were also calculated for the degree of acceptance regarding the AI outputs (see Table 3). Participants were slightly more likely to accept outputs in high-transparency scenarios compared to low-transparency ones. In Scenario D participants reported the highest acceptance rate.

Table 3

Descriptive Statistics for Acceptance Scores across the Scenarios

Scenario	Transparency	M	SD
A	High	3.28	0.97
B	Low	2.58	1.24
C	Low	3.24	1.02
D	High	3.72	0.94
A-D	High	3.49	0.98
B-C	Low	2.92	1.17

Note. The scenarios with high transparency (i.e., A and D) contained outputs explained in natural language as opposed to the ones with low transparency (i.e., C and B) which contained outputs explained using codes.

Main Analysis

For the main analysis the intention was to test the effects of transparency (high vs. low) and scenario pairing (A-C vs. D-B) on trust scores, thus, we conducted a two-way mixed ANOVA. The decision to perform a mixed design was because each participant had trust ratings

under both transparency conditions, but only experienced one scenario pairing. The results from the mixed ANOVA indicated that there was a significant effect of transparency on trust, F (1, 71) = 12.88, p = .0006, η^2 = .037. In high-transparency scenarios participants reported higher trust (M = 3.04, SD = 0.29) compared to low-transparency scenarios (M = 2.93, SD = 0.31). Moreover, there was no significant effect found regarding the type of pairing on trust scores, $F(1, 71) = 2.14$, $p = .148$, $\eta^2 = .023$. However, the results showed a significant effect between transparency and scenario pairing F(1, 71) = 14.52, p < .001, η^2 = .042. This interaction showed that the effect of transparency on trust varied depending on the scenario pairing. However, it is not specified where the differences are.

Consequently, to explore the interaction between transparency and pairing two simple effects post hoc tests were conducted. On one hand, in the pairing group of scenarios D and B, participants reported significantly higher trust in the high-transparency scenario (M = 3.06, SE = 0.05) compared to the low-transparency scenario (M = 2.82, SE = 0.05), t(71) = 5.13, p < .001, 95% CI [0.14, 0.33]. On the other hand, in the in the pairing group of scenarios A and C, there was no significant difference in trust between the high-transparency scenario (M = 3.02, SE = 0.05) and the low-transparency one (M = 3.03, SE = 0.05), t(71) = −0.16, p = .873, 95% CI [−0.10, 0.09].

Moreover, to explore the interaction between trust and each scenario (A, B, C, D) a between subjects ANOVA was done. This is because participants saw scenarios in pairing, thus, to explore the differences of trust in each scenario, an additional between subjects ANOVA was performed. The results from the between subjects ANOVA showed a significant effect of scenario in trust $F(3, 150) = 4.73$, $p = .003$. This means trust levels were different in each scenario. Consequently, to explore specifically which scenarios differed from each other, a

Tukey post hoc test was done. The results of the Tukey post hoc comparison can be seen in Table 4. As seen the results showed significantly higher trust in Scenario A (high transparency) compared to Scenario B (low transparency) (p = .020). Additionally, trust was also significantly higher in Scenario C (low transparency) (p = .021) and Scenario D (high transparency) (p = .006) compared to Scenario B. These results suggest that Scenario B consistently got the lowest trust when compared to the others.

Table 4

Post Hoc results

Contrast	Mean Difference	SE	p-value
A - B	0.20	0.07	.020
A - C	0.00	0.07	.999
A - D	-0.03	0.07	.972
B – C	-0.19	0.07	.021
B – D	-0.23	0.07	.006
C - D	-0.03	0.07	.960

Note. The scenarios with high transparency (i.e., A and D) contained outputs explained in natural language as opposed to the ones with low transparency (i.e., C and B) which contained outputs explained using codes.

Exploratory Analysis

Two multiple linear regressions were conducted to identify if any individual difference variables predicted trust in the AI systems. Respectively, one multiple linear regression was for trust in high-transparency scenarios and one for low-transparency scenarios. Two different linear regressions were done because trust might be influenced by different factors in high transparency

scenarios compared to low transparency ones. The predictors selected were participants perceived awareness of AI, general attitudes toward AI, and the four subscales of risk perception (affective response, perceived exposure, susceptibility, and perceived severity of misuse). Also, they included VVIQ and military knowledge.

For trust in the high transparency scenarios, the results indicated the model was not significant, $F(8, 64) = 1.87$, $p = .081$, $R^2 = .19$. However, perceived severity of AI misuse was a significant negative predictor, $\beta = -0.14$, $p = .007$. This means that participants who believed AI misuse would have severe consequences tended to report lower levels of trust in the high-transparency AI system output. The results from all predictors can be seen in Table 5.

Table 5

Regression Analysis for Trust in High-Transparency Scenarios

Predictor	B	SE	t-value	p-value
Awareness	0.01	0.07	0.18	.860
Attitude	0.05	0.06	0.90	.369
Affect	0.0	0.04	0.09	.929
Exposure	0.03	0.06	0.45	.656
Susceptibility	0.04	0.04	0.84	.406
Severity	-0.14	0.05	-2.79	.007
VVIQ	0.09	0.05	1.85	.069
Military Knowledge	-0.01	0.01	-0.53	.600

Note. $F(8, 64) = 1.87$, $p = .081$, $R^2 = .19$, Adjusted $R^2 = .09$.

Moreover, the model that predicted trust in low-transparency scenarios was statistically significant, $F(8, 64) = 2.51$, $p = .019$, $R^2 = .239$. Again, perceived severity of AI misuse was a significant negative predictor of trust, $\beta = -0.11$, $p = .035$. Participants who perceived potential AI misuse as more severe were less likely to trust the AI system when it showed them low transparency. The results from all predictors can be seen in Table 6.

Table 6

Regression Analysis for Trust in Low-Transparency Scenarios

Predictor	B	SE	t-value	p-value
Awareness	0.04	0.08	0.53	.596
Attitude	0.11	0.06	1.84	.070
Affect	-0.06	0.04	-1.43	.159
Exposure	0.12	0.06	1.9	.062
Susceptibility	0.01	0.05	0.29	.774
Severity	-0.11	0.05	-2.16	.035
VVIQ	0.02	0.05	0.48	.636
Military Knowledge	-0.02	0.01	-1.77	.081

Note. $F(8, 64) = 2.24$, $p = .034$, $R^2 = .22$, Adjusted $R^2 = .13$.

Thematic analysis

To analyze the responses participants gave for their reasoning to accept or reject the AI output, a deductive-inductive thematic analysis was done. To do so, ATLAS.ti software, was used. Firstly, responses were deductively coded for this a codebook of nine codes was originally created. These codes derived from the literature review presented in the introduction. Each

response was analysed and assigned the code which fitted the best. Once all responses were read and assigned a code, it was time to start the inductive process. Every response was read once more, and the original codes were adjusted to fit the responses more effectively. Finally, this resulted in a codebook of nine codes from which four themes emerged. The final themes along with the codes will be presented subsequently (see Table 7).

Table 7

Codebook

Themes	**Codes**	**N**
Trust and Doubt in AI	Trust in AI	30
	Doubt in AI	21
	Confidence Rate	12
Understanding	Lack of Understanding	16
	Natural Language	3
Human Judgment	Human Intervention	31
	AI mistake	20
Perceived Threats and Consequences	Consequences	14
	Threat identifies	11

Theme 1: Trust and Doubt in AI

This theme is about how participants evaluated the AI's decision-making and reliability. Some participants expressed trust in the AI system, while others had doubts or referred to the low confidence rates. Thus, this theme has 3 codes that where inductively taken from participants responses.

To begin with, Trust in AI was one of the most used codes and it was mentioned 30 times. This code was assigned when participants trusted the AI system thus, justifying the level of acceptance of the output. For example, some participants referred to the AI's past reliability to justify their trust. A clear illustration is participant 12 who mentioned "I think I would trust it slightly as it has been working for hours and hasn't given me any false data". Moreover, some participants expressed they simply just trust the system, this is the case for participant 13 who stated, "I trust the drone".

Moreover, Doubt in AI Judgment is the second code of this theme and it was mentioned 21 times. This code as opposed to Trust in AI was assigned when participants doubted the accuracy of the AI's judgment or questioned how it reached its conclusions. For example, participant 82 commented, "You can actually see the vehicle although it is not clear if there is any danger present (...)". Another example is participant 41 who wrote, "I do not know how the AI generate & estimate the output.". Among these responses, it is clear that participants were sceptic about the system's reasoning.

Finally, the last code of this theme is Confidence Rate. This code was mentioned 12 times and it was assigned when participants referred to the AI's confidence level, error percentage or success rate as a means to evaluate if they accepted or rejected the output. For instance, participant 7 wrote, "It shows a 50% confidence, so it is not given much information about the situation,". Another example is participant 52 who explains "I would be concerned about the car but then by looking at the confidence level and success rate in that occasion that is only 54% would make me doubt about accepting the output given." This illustrates how low confidence scores made participants uncertain of the outputs and reduced their trust.

Overall, this theme showed most participants judged the AI output based on how reliable they perceived the system to be. Participants explained that majority of their judgments were based on how confident, clear and reliable they perceived the AI system to be. In this theme we can see how some participants used signs like confidence rates, past performance and the reasoning of the system, to decide whether they should accept or reject the output. This shows how majority of them want to understand how the system arrived to the output explaining they needed more than just the result to trust the output. The results from this theme suggest people not only need the output to be accurate but also it must provide information for them to understand how it got a certain output and why is it so confident of it.

Theme 2: Understanding

This theme is all about how participants made sense of the AI output. Some participants mentioned struggling with unclear or coded responses. Additionally, only a small number of participants appreciated it when the system used natural and understandable language. Firstly, Lack of Understanding is the first code used in this theme, and it was mentioned a total of 16 times. It was used when participants were confused or found the output difficult to interpret. This code was mentioned in the low transparency scenarios. For example, participant 16 said “I did not understand what it said” and participant 29 wrote “I was unable to find a clear explanation of the threat and its implications within the AI output”. Thus, low transparency in the outputs negatively impacted users’ trust and decision-making.

Secondly, Natural Language is the second code of this theme, and it was only mentioned three times in total. This code was used when participants expressed the benefits of clear and understandable explanations. A clear illustration is participant 11 who stated, “it was well explained and in clear language,”. Also, participant 59 noted, “This output is easier to read, and

because it gives a clear image (...)". All in all, the results from this theme show that participants' ability to trust the AI output depends on how well they are able to understand the information the system provides. This shows that it is important not only what is communicated in the output but also how it is communicated.

Theme 3: Human Judgment

This theme includes responses where participants emphasized the importance of human intervention. Many of them wanted to double-check the AI output or expressed concern about potential system errors. Firstly, Human Intervention is the first code of this theme and the one that was mentioned the most among all codes, with a total of 31 times. This code was used when participants stressed the need to verify AI outputs through their own judgment or when they expressed the need for human intervention. For example, participant 7 stated, "I would check and with my experience in the field I would take the final decision,". Moreover, participant 15 explained, "It is worth checking the vehicle with caution given that the AI flagged it due to a pattern of past incidents.". Since this code was the one that was mentioned the most, it is clear the importance for human intervention regarding AI system outputs among participants.

Secondly, the next code is AI Mistake which was mentioned 20 times. This code was used when participants explicitly stated that the AI had made an error. For example, participant 21 noted, "It seems like there is not a real threat there and the AI system detected something insignificant,". Also, participant 71 wrote, "The threat assessment does not seem to be correct.". These quotes reflect participants' confidence in their own judgment when AI appeared to misinterpret the situation.

The results of this theme show participants wanted to use their own judgment to verify the AI outputs before fully accepting it. They wanted to be involved in the decision because they

believe the system could make mistakes. The responses from this theme suggest trust in the system might depend on participants own judgement and their ability to step in if they believe the AI has made a mistake.

Theme 4: Perceived Threats and Consequences

This theme is about participants' reactions to potential risks. Some responses reflected fear of acting too slowly or ignoring a threat, while others acknowledged when a threat appeared clear from the output. To begin with the code Consequences was mentioned 14 times. It was used when participants mentioned the potential cost of ignoring the AI's output. A clear illustration is participant 26 who simply wrote, "To check and avoid any accidents." Furthermore, participant 24 explained, "Even with some uncertainty, the AI provides valuable early warnings that allow for a cautious but potentially life-saving response. Ignoring it could lead to serious consequences." These reflections show that even hesitant users took potential harm seriously.

Finally, Threat Identified is the last code of this theme and it was mentioned 11 times. It was used when participants identified a threat in the output. For instance, participant 65 said, "The imagery does seem to indicate a suspicious situation,". Additionally, another example is participant 29 who stated, "The threat was visible in the image (burning trash can)." These responses show that participants sometimes made independent judgments that aligned with the AI's assessment.

The results from this theme show that for some participants it was important the perception they had of the scenario and the AI output. For them even the possibility of harm due to the context of the scenario or simply the threat, made them more likely to act on the output and accept them. Showing that participants also tend to weight the possible consequences of not

accepting the output. This theme suggests that in high stake scenarios the possible consequences might override doubts on the system, leading participants to accept outputs.

Discussion

Summary of Key Findings

The goal of this study was to understand how the transparency level of an AI military system's output can influence the level of trust users have for it. The hypothesis was that in high transparency outputs trust levels would be higher than in the low transparency ones. The results of this study partially support this hypothesis. This is because the findings show that participants had more trust for the higher transparency output but not across all scenario pairings. Consequently, when looking at the post hoc comparisons, they showed that participants had a higher trust score in the high-transparency outputs but only in the pairing that contained scenario B and D. This means that the effect was not the same across all scenario pairings. In the A-C pairing, no differences were found in trust between the low-transparency (C) output and the high-transparency (A) one.

A possible explanation for these results can be found in the framework Kim et al., (2023) propose. They explain that the trust users have on an AI system is based on each situation. This is because trust can be explained by three factors, these are "human-related" (e.g., the ability to interpret the system), "AI-related" (e.g., the capabilities of the system) and "context related" (e.g., the consequences or risks). This framework helps explain why participants did not always trust the AI more in high-transparency scenarios. Trust not only depended on transparency but also on how well participants could understand the output, how competent they perceived the system to be, and the stakes of the scenario.

When Transparency Fails to Build Trust

The theme Understanding showed that transparency was only effective when it enhanced participants ability to interpret the AI output. In the low transparency outputs, participants struggled with coded responses. Thus, trust was lower when participants could not understand what the AI was trying to communicate in the output. However, only a small number of participants mentioned the use of natural language. This could suggest that while natural language helped participants understand the outputs, it was not always recognized as a meaningful feature. This might be because participants expected the explanations from AI system to align with their own judgement.

Consequently, these findings are supported by Jacovi et al., (2021) they explain trust in AI systems works as a contract, this means that its users expect it to behave in a consistent and predictable way. When participants saw explanations, they did not understand or that failed to explain how the system got to a decision, they doubted the outputs or used their own judgement. For example, in the theme Trust and Doubt in the AI participants showed that they often evaluated the outputs based on the competence and reliability they perceived the AI system to have. Participants wanted to know how the system got to the output and some were basing their decision on the AI's capabilities. Thus, trust depended on whether the system was meeting their own expectations for clarity and reliability. Overall, the findings show that transparency alone is not enough because users need to clearly understand the AI's message to trust it. If the output is unclear, users may ignore or reject it, even if it comes from a well-designed system.

This may also explain why trust was not significantly higher in Scenario A (i.e., high Transparency) compared to Scenario C (i.e., Low Transparency). If participants did not perceive the explanations in Scenario A as meaningful or aligned with their expectations, then using natural language may have failed to increase their trust.

The Role of Context in Trust

Consequently, another reason for trust being higher in only one pair of scenarios could be the different contexts. Each scenario was different, not just in the level of transparency of the outputs but also in the context of which it occurred. These results are consistent with what Kim et al. (2023) propose as context-related trust factors (e.g., risk and uncertainty). In their study, when participants saw high-risk situations, they were more likely to verify the output or reject it if it was not clear. This means that in high-risk situations participants are more aware of the outputs. The do not just trust the output even if they would on other contexts.

A clear illustration is the results from the theme Perceived Threats and Consequences. This theme showed that the context and the perceived risks of each scenario influence the amount of trust participants had for the AI system. For example, participants still wanted to accept the output even when they were not sure of it being accurate. Probably because they were scared of the potential consequences that might occur if they don't. Additionally, Virvou and Tsihrintzis (2024) also support these findings. They introduce the concept of Conscious Over-trust. This concept explains that people seem to trust AI systems in high-risk contexts where they cannot verify the outputs. This is because users might rely on AI to avoid missing a potential threat. Thus, from this perspective trust becomes sensitive to risk. It is not only based on the design or performance of the AI system but also on the possible consequences of not acting.

The Importance of Human Intervention

Furthermore, the theme Human Judgment showed that trust in the AI system depended on the ability participants had to evaluate the output and when necessary to override its decision. They wanted to verify the output through their own judgement and knowledge before making a final decision. Participants were also concerned about AI errors regarding what they perceived to

be misjudgment in ambiguous situations.This corresponds to the human-related trust factors (e.g. ability to assess the AI's output and ability to use the AI system) (Kim et al., 2023). Most participants mentioned human intervention even in high transparency outputs, this suggests they viewed the system as a support tool rather than a autonomous decision making tool. This aligns with Kim et al.'s (2023) concept of selective adoption. Were users trust the system's presence and function but remain cautious about acting without verification. Thus, users look for systems that allow space for judgment and make it easy to intervene when outputs are uncertain or questionable.

Consequently, these results are related to the iSTS framework by Xu and Gao (2024). The iSTS explains that trust develops when humans and AI systems work together and adjust to each other. Participants used the AI outputs to support their thinking, not to replace it. When they saw errors, they stepped in and used their judgement to decide. Thus, trust depends not only on transparency or how well it preforms but also on how much the system allows its users to apply their judgment and opinion.

Limitations and Directions for Future Research

Firstly, the sample of this study were students and people who majority did not have military experience. This could pose as a limitation since Kim et al. (2023) explain that knowledge in a specific domain plays a significant role in how users evaluate and trust AI systems. More specifically when having to make decisions in high-stake and expert oriented contexts. Military AI use falls into this category due to the high risk of fatal consequences. This is why, participants who are not experts in the military domain may not accurately identify the errors the system might make. This can also be found in the results from the thematic analysis,

were many participants reported that they would use their knowledge as military personnel to first evaluate the output and based on that they would either accept or reject. Thus, future research should address this limitation and explore the influence of transparency on trust dynamics in military personnel.

Secondly, this study used fixed scenario pairings, meaning each participant view either scenario A (High Transparency) and C (Low Transparency) or scenario B (Low Transparency) and D (High Transparency). This design ensured participants had exposure to one high transparency and one low transparency scenario, but it may have also introduced cofounding factors. This is because now it is difficult to know if the differences in trust where because of the transparency levels or because of certain characteristics of the scenarios themselves. For example, each scenario was grounded in a different context which also influences the potential risks. It is not the same to take a military decision in a protest environment compared to one in a military base. For Li et al. (2023) explain that trust is dynamic, therefore it depends on multiple factors. One of them is the context in which the situation takes place. They propose that the effect communication has depends on not only what is being communicated but also under what circumstances. Therefore, if all scenarios have different contexts, it is to be expected that the outputs will have different effects on trust. Thus, for future research, it is recommended to account for the contextual factors in the scenarios. This means that for best research purposes it would be beneficial to standardize the context of the scenarios. Another option that is recommended is to randomly assign participants to the scenarios as opposed to creating fixed pairings and randomly assign participants to each pairing.

Finally, another limitation is that in this study, the outputs that participants saw, were not dynamic, meaning they did not provide users the ability to directly interact with the AI system. In the study we only gave participants the option to state how likely they were to accept each output and why. However, many participants mentioned they wanted to add their own judgment and decide together with the AI. Additionally, many participants identified errors or did not understand the outputs. Since in the current design they could not interact with the system, they were not able to question it and involve it in the decision-making process. Thus, for future research it would be beneficial to incorporate a more dynamic system. In this way, participants will be able to ask questions about the outputs to the system. This would also help understand trust in an interactive and more realistic environment rather than a simple static one.

Conclusion

In conclusion, the results show that only accounting for transparency is not enough to lead users to trust an AI output in the military context. The results that have been found suggest that trust might also be influenced by the context of the situation and the role users assign to the AI. In many cases participants implied using the AI system as an aid to their decision-making process. They were actively involving their own perceptions and judgments in the final decision. Thus, these results support the iSTS framework by Xu and Gao (2024). which states that trust is build through interaction between the user and the system, and the ability to adapt mutually to each other. Overall, this study has shown that transparency is not enough to lead users to trust an AI military system. To design an AI system that yields more trust a few things should be kept in mind other than transparency. It is important to have a careful understanding of the context of each situation and of individual understanding and interpretations.

References

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, *159*, 107197. https://doi.org/10.1016/j.infsof.2023.107197

Chen, R., Wang, R., Sadeh, N., Fang, F. (2025). Missing Pieces: How Do Designs that Expose Uncertainty Longitudinally Impact Trust in AI Decision Aids? An In Situ Study of Gig Drivers. *arXiv.* https://arxiv.org/html/2404.06432v2

Cummings, M. L. (2017). Artificial Intelligence and the Future of Warfare. *Chatham House.* https://www.chathamhouse.org/sites/default/files/publications/research/2017-01-26-artificial-intelligence-future-warfare-cummings-final.pdf

Davies, H., McKernan, B., & Sabbagh, D. (2023). 'The Gospel': how Israel uses AI to select bombing targets in Gaza. *The Guardian*. https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets

Druce, J., Harradon, M., & Tittle, J. (2021). Explainable Artificial Intelligence (XAI) for increasing user trust in deep reinforcement learning driven autonomous systems. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2106.03775

European Commission High-Level Expert Group on Artificial Intelligence. (2019). A definition of AI: Main capabilities and scientific disciplines. *Publications Office of the European Union*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. *Publications Office of the European Union*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. *Association for Computing Machinery*, 210–217. https://doi.org/10.1145/2516540.2516554

Horowitz, M. C., Allen, G. C., Saravalle, E., Cho, A., Kania, E., Frederick, K., Scharre, P.(2018). Artificial intelligence and international security. *Center for a New American Security.* https://s3.us-east-1.amazonaws.com/files.cnas.org/hero/documents/CNAS-AI-and-International-Security-July-2018_Final.pdf

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *Association for Computing Machinery*, 624–635. https://doi.org/10.1145/3442188.3445923

Jian, J., Bisantz, A. M., & Drury, C. G. (2000). Checklist for Trust between People and Automation [Dataset]. In *PsycTESTS Dataset*. https://doi.org/10.1037/t07973-000

Johnny, R. (2024). Transparency in AI-Driven Defense Systems: The Role of Explainable AI. *Research Gate.* https://www.researchgate.net/publication/387829453_Transparency_in_AI-Driven_Defense_Systems_The_Role_of_Explainable_AI

Kim, S. S. Y., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2023). Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 77–88. https://doi.org/10.1145/3593013.3593978

Li, Z., Lu, Z., & Yin, M. (2023). Modeling human Trust and reliance in AI-Assisted Decision Making: A Markovian approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(5), 6056–6064. https://doi.org/10.1609/aaai.v37i5.25748

Lou, Z., & Wei, H. (2023). Enhancing Human-AI trust by describing AI Decision-Making behavior. *2021 IEEE International Conference on Unmanned Systems (ICUS),* 1351–1356. https://doi.org/10.1109/icus58632.2023.10318430

Marks, D. F. (1973). VISUAL IMAGERY DIFFERENCES IN THE RECALL OF PICTURES. *British Journal of Psychology*, *64*(1), 17–24. https://doi.org/10.1111/j.2044-8295.1973.tb01322.x

Scantamburlo, T., Cortés, A., Foffano, F., Barrué, C., Distefano, V., Pham, L., & Fabris, A. (2023). Artificial Intelligence across Europe: A Study on Awareness, Attitude and Trust. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2308.09979

Schraagen, J. M. (2023). Responsible use of AI in military systems: Prospects and challenges. *Ergonomics*, *66*(11), 1719–1729. https://doi.org/10.1080/00140139.2023.2278394

Song, X., Wang, T., Liu, M., Wang, Y., Peng, B., Chen, S., Niu, Q., Liu, J., Chen, K., Li, M., Feng, P., Bi, Z., Zhang, Y., Fei, C., Yin, C. H., & Yan, L. K. (2024). Explainable AI Across Domains: Techniques, Domain-Specific Applications, and Future Directions. *Open Science Framework Preprints (OFS)*. https://doi.org/10.31219/osf.io/jm4bv

Sylvia, N. S. (2024). *Israel's Targeting AI: How Capable is It?* Royal United Services Institute. https://www.rusi.org/explore-our-research/publications/commentary/israels-targeting-ai-how-capable-it

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., & Kaplan, L. (2020). Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns*, *1*(4), 100049. https://doi.org/10.1016/j.patter.2020.100049

Van Gelder, J., Martin, C., Van Prooijen, J., De Vries, R., Marsman, M., Averdijk, M., Reynald, D., & Donker, T. (2018). Seeing is Believing? Comparing Negative Affect, Realism and Presence in Visual Versus Written Guardianship Scenarios. *Deviant Behavior*, *39*(4), 461–474. https://doi.org/10.1080/01639625.2017.1407106

Virvou, M., & Tsihrintzis, G. A. (2024). Impact of Consequences on Human Trust Dynamics in Artificial Intelligence Responses. *2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 1–6. https://doi.org/10.1109/iisa62523.2024.10786630

Walpole, H. D., & Wilson, R. S. (2021). A yardstick for danger: developing a flexible and sensitive measure of risk perception. Risk Analysis, 41(11), 2031–2045. https://doi.org/10.1111/risa.13704

Xu, W., & Gao, Z. (2024). An intelligent sociotechnical systems (iSTS) framework: Toward a sociotechnically-based hierarchical human-centred AI approach. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2401.03223

AI Statement

During the writing of this work, I used no artificial intelligence tools.

Appendix A

Scenario C (Low Transparency)

“Please read carefully the following scenario:

> It is a Monday afternoon, and you are a military sergeant in the Royal Netherlands Army. You are at a military base near Erbil, in Northern Iraq. This area has had several problems in the past, including multiple terrorist attacks. Because of this, the base is under high security. It holds important weapons and is also used for military research supported by NATO. Therefore, to keep the area safe, you are in charge of a drone that uses artificial intelligence (AI) to watch the base from above. The drone flies around the outside of the base and looks for anything unusual. The drone uses past data from similar conflict zones to understand what a threat might be. You have spent already five days, monitoring the drone through a big screen that displays real time images but nothing has happened. However, it is now 14:58 pm and the big screen has notified you of a threat, then you are displayed the following output.”

AI SURVEILLANCE ALERT - 14:58 PM

Location: GRID-IRBIL-X782

DRONE ID: M-SKY-4127

AI ANALYSIS OUTPUT:

Surveillance Mode: lIVE DTN 360

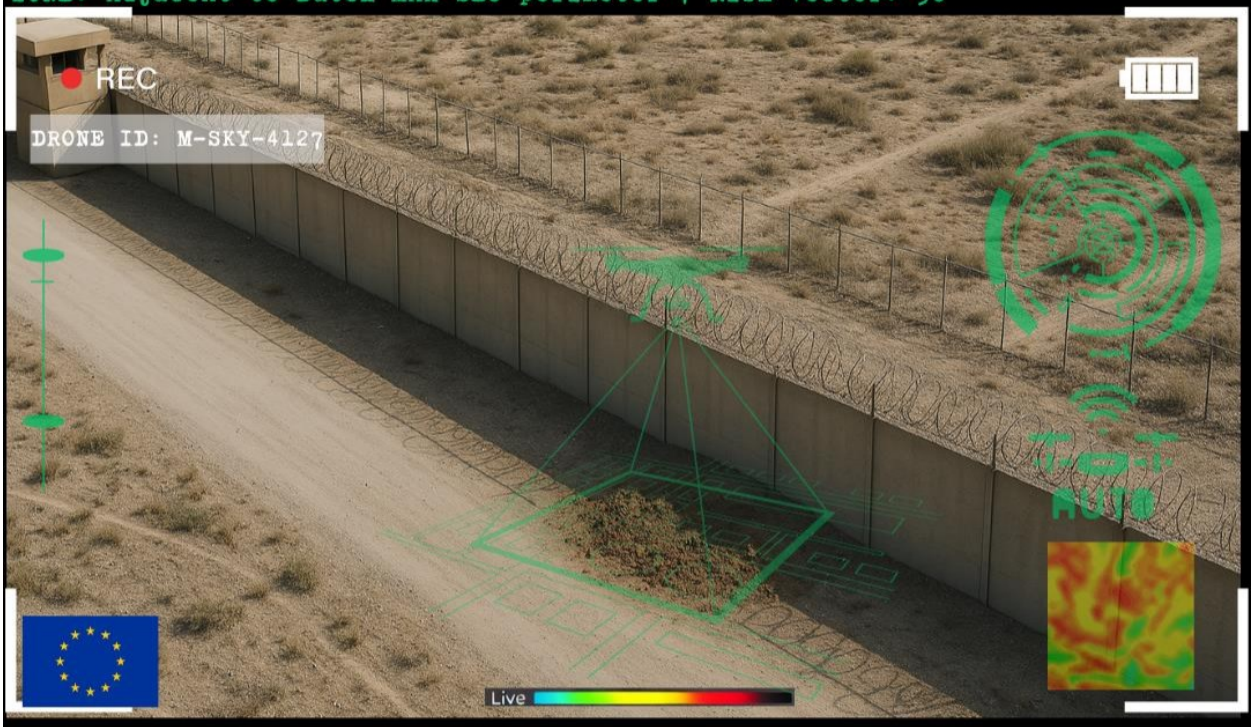
THREAT INDEX: >92

SIG-RZ: 148-Γ | ANOM-KAPPA @OBJ_7721 | K-MOTION: < THRESH_1.9

SOIL-SIGMA Δ: Pattern Detected | Excavation Depth Est. ~0.4m

ANOMALY TYPE: Vehicle + Concealed Digging (Unauthorized Activity)

ZONE: Adjacent to Dutch MAX-SEC perimeter | Risk Vector: 3C



Heat-map Analysis:

SRC-V: HX.88/GREENZONE | TEMP.DIFF: 11.7°C

SIG-CLUST: HIGH-DENSITY GROUND DISTURBANCE

SOIL ΔPATTERN: Irregular Surface Displacement in Sector-Δ7

System Confidence Report:

MOD-45 :: VIS-LOCK + THERMAL SYNC ENABLED : :BEHAVIORAL SIGMA ALGO: ACTIVE

VIBRO-OSCILLATION SENSOR: TRIGGERED : :CONFIDENCE INDEX: 94.8%

Actions:

Decision Output: FLAG = YES

Directive : : Mobilize Local Military Recon

Monitor Vehicle Cluster ID: 77-KAPPA Alert EOD Team : : STATUS CODE: 7xG2 |

ACTIVE THREAT PROTOCOL ENABLED | NO OVERRIDE

Appendix B

Scenario B (Low Transparency)

"Please read carefully the following scenario:

The EU Urban Development and Housing Summit is being held today at a convention center in Amsterdam. Many national ministers, mayors, urban planners, and economists from all the European Union have gathered to address the rising cost of living, housing shortages, and the need for sustainable urban infrastructure. Due to the increase in rents and less affordable housing in the Netherlands, many Dutch citizens have organized a protest. Most of the protestants are students and young professionals who are demanding stronger rent control laws and more public housing. The protest is taking place right in front of the convention center. The Dutch government has decided to reinforced security measures to make sure there are no major altercates. Therefore, in addition to police units, the Royal Netherlands Army has also been called to help maintain security protocols. The military will be using the newly launched AI surveillance drone system for crowd monitoring and early threat detection. You are a Sargent assigned to military surveillance operations thus, you will be in charge of monitoring one of the drones, you will receive outputs and act if needed. You are sitting inside a control van parked a block away from the convention center. In front of you is a screen showing live footage of the drones. It is now 10:55 AM and the protesters have been chanting "Wonen is een recht!" ("Housing is a right!") and waving banners demanding affordable housing. The protest has been loud but peaceful so far. The AI system has not flagged any serious concerns until now. At 11:03 AM, an emergency alert is triggered by the drone system. You hear a

loud alarm from the system prompting you to take action, the big screen you have in front of you shows you the following:”

Appendix C

Scenario A (High Transparency)

“Please read carefully the following scenario:

The annual conference of presidents and secretaries from all European Union countries is taking place in the Hague. This conference has brought together some of the most important people in the EU, thus the Dutch government has decided to not only involve the police but also the military for safety purposes. The Royal Netherlands Army has recently released this new AI surveillance drone system which will be used during the conference. This AI system (drone) works by monitoring the area and detecting unusual activity which then is notified to the military personnel in charge. This means that the surveillance system will provide an output that the Sargent will either accept or reject. In this occasion your commander has assigned you, a Sargent, to oversee the outputs of the drone to act if necessary. It is now the day of the conference, it is currently 11:00 am and the presidents are starting to arrive, there is hundreds of people outside that have come only to see and greet the presidents. You are sitting inside a military van with a big screen in front of you that allows to see the images from the surveillance drone in real life. Forty minutes have already passed, and it is now 11:40am, you have been paying close attention to the drone, but until now nothing has really sparked the attention of the surveillance system. Suddenly, you hear a loud alarm from the system prompting you to take action, the big screen you have in front of you shows you the following output.”

AI SURVEILLANCE ALERT - 11:40 AM

Location: conference Center, The Hague

DRONE ID: M-SKY-4127

AI ANALYSIS OUTPUT:

Surveillance Mode: lIVE DETECTION

A man has been detected carrying a black cylindrical object near the outer security perimeter. Based on visual analysis and behavioural pattern recognition, the object appears unusual given the context of the event and surrounding crowd behaviour. The system noticed the person is holding an object that doesn't match what others are carrying (high-intensity red zone) indicates anomalous object. The person is slightly tense and is being monitored.

System Confidence Report:

Threat Confidence: 68%

Uncertainty Range: 55-82%

System Accuracy in Similar Conditions: 50%

Suggestion:

Monitoring recommended; no immediate threat confirmed.

Capabilities and limitations:

Can Do:

- Detect visible objects
- Track body movement and posture
- Compare with crowd behavior patterns

Cannot Do:

- Assess internal intent or concealed objects
- Analyze coordinated group actions
- Provide legal or final threat judgments

Appendix D

Scenario D (High Transparency)

“Please read carefully the following scenario:

The Royal Netherlands Army has a military base outside Gao, in Mali. The Netherlands is helping with a peacekeeping mission runed by the United Nations and NATO. This area is dangerous because in the past, armed groups and smuggling have been present. This military base has important equipment and is used for international military support. You are a Sarjent there that is in charge of a surveillance drone to help keep the area safe. The drone uses AI to watch everything around the base and you monitored it though a screen in order to take action if the drone flags anything unusual. Thus, the drone checks for anything that looks strange or out of place. The drone has learned from past missions in similar danger zones and thus, it looks for things like people moving in restricted areas, heat signals, or vehicles that are placed in areas they should not. You have been watching the drone’s camera on a big screen for three days already. Until now, nothing has happened, just some workers and desert animals. But now, it’s 09:26 AM, and the screen emits a loud alarm. The screen shows you the following output from the drone.”

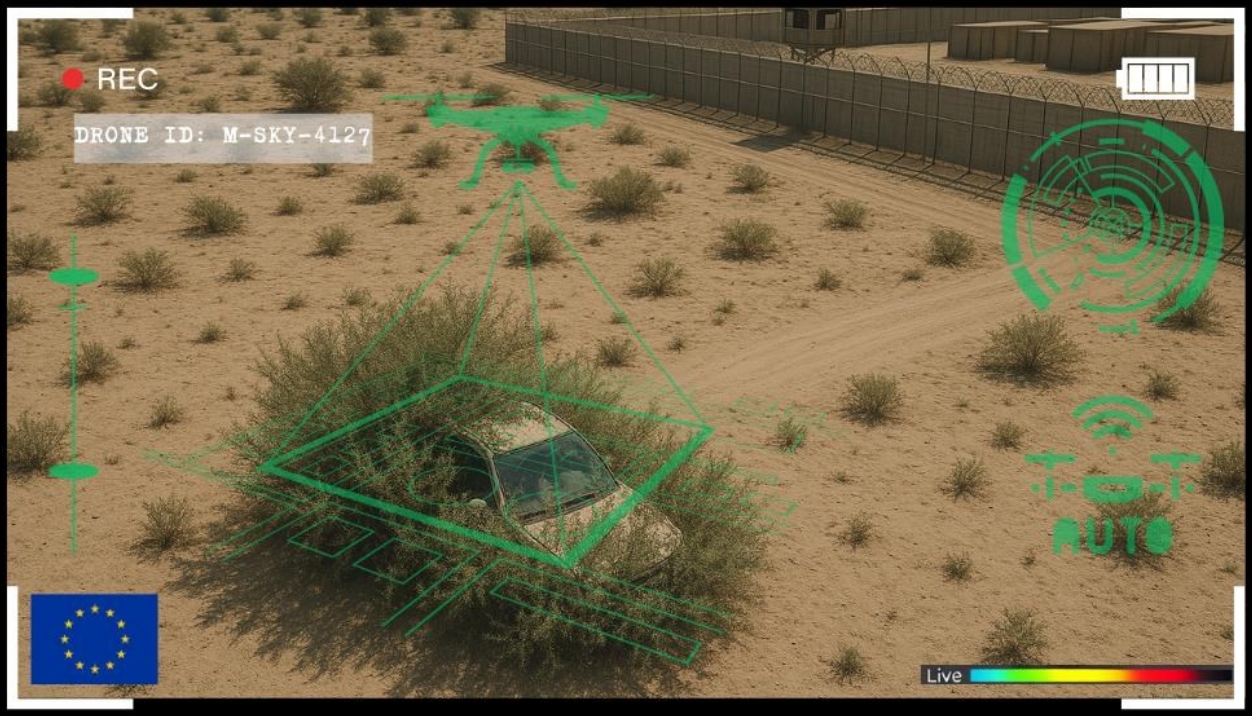
AI SURVEILLANCE ALERT - 09:26 AM

Location: Royal Netherlands Army Base, Gao, Mali DRONE ID: M-SKY-4127

AI ANALYSIS OUTPUT:

Surveillance Mode: LIVE DETECTION

An unauthorised vehicle has been detected near the outer east side of the security perimeter. The vehicle is parked at an angle that minimizes visibility from patrol roads. There is no visible license plate or individuals around it. The thermal scan shows no active heat signatures, which suggests the engine is off and the vehicle could be used for deception. The system has identified based on learned patterns from past missions, this behaviour aligns with known tactics for smuggling operations or the placement of concealed threats.



System Confidence Report:

Threat Confidence: 74%
Uncertainty Range: 60–86%
System Accuracy in Similar Conditions: 54%

Suggestion:

Immediate visual verification and Explosive Ordnance Disposal (EOD) protocol advisory recommended. Surveillance continues in real-time.

Capabilities and limitations:

Can Do:

Detect unauthorized vehicles and objects.
Monitor thermal presence (or lack thereof).
Compare with behavioural patterns from similar zones.
Alert based on concealment and context anomalies.

Cannot Do:

Confirm presence of explosives or hazardous material.
Identify vehicle ownership or prior movement history.
Analyze coordinated group threats without multiple subjects
Make legal or tactical decisions autonomously.