Decision-Making in Threat Intelligence

R_82SfsCGxc0AxI9p

Plagiarism statement: I confirm that the anonymous ID is the one that has been assigned to me and I have not 'confused' it with ID of other students. I also confirm that this assignment is my own work, is not copied from any other person's work (published or unpublished), and was not shared with any other person that could copy its contents

I. Experiment Description

A. Goal

Responding to the discovery of a software vulnerability in one's systems is a critical issue for both private and public organizations [1] which have to decide which updates to prioritize possibly to no avail [2]. To address these shortcomings, many governments have pushed towards the automation of the vulnerability assessment process, starting from Biden's executive order [3].

In a nutshell, an organization receives some information on threat intelligence either AI-based or expert-based which might be vital for one's organization and yet, *must just trust it.* Human decision-making about security risks involving threat intelligence are made in the face of uncertainty [4], leaving space for subjective and possibly biased judgment [5]. For instance, past studies [6], [7] found evidence of algorithm aversion, that is systematic preference to human judgement (even when algorithms outperformed humans). Since the intelligence may come from "non-human" sources (e.g., AI program used for prediction, or a field sensor), algorithm aversion may actually be present in the decision making process. In addition, one might postulate that the very perception of a "biased source" differs depending on the source type (human vs algorithmic), regardless of intelligence correctness.

The objective of this research are to determine, in the field of vulnerability assessment, whether there is an effect on decision-making of algorithmic judgment in comparison to human judgment. We are interested to investigate whether algorithms are held to a higher standard when they are used to provide threat intelligence recommendations. In addition, based on the importance of consistency of information within classical intelligence scenarios [8], consistency with the displayed information is investigated.

B. Procedure

At the beginning of the experiment, participants are asked to give informed consent. Thereafter, the participants are able to read the task description and answer questions about previous experience.

The participant has to assess eight abbreviated report. For each report, the participants were

a) asked to read the report
b) asked to give a likelihood assessment. This assessment is phrased in two different ways. First, the participant is asked whether they agree with the likelihood recommendation in the report. Second, the student has to select a likelihood assessment in terms of *low, medium* or *high*.
c) asked to give an impact assessment. This is done to ensure that the participant takes the likelihood assessment and impact assessment as two separate classifications.
d) asked to explain their answers
e) asked to answer questions about the report in question and the current assessment
f) After all reports are assessed, asked to answer a post-questionnaire about the quality, usability, and framing (in terms of expert power) of the different sources. Subsequently, questions were asked about the perception of the entire experiment.

C. Balanced Design

The purpose of the experiment is to understand whether human assessors will react differently between an alleged human and an alleged AI advice. There are a number of factors that may confuse the result. For example, a recommendation might be perceived to be inconsistent with the given information. Assessors might also react differently to high or medium risk assessment. Generating all possible combination for a pure randomized design would require an exponential number of configurations. Hence we opted for a Taguchi design.

The participants are randomly assigned to one of twelve groups. These groups reflect the different order of manipulations to ensure a balanced design (i.e. the measurements obtained in each condition are roughly equal). See Table II for an overview. The recommendation could come from a human (H) or a machine learning algorithm (A) source. Furthermore, the recommendation could be consistent (C) or inconsistent (I) with the NCSC assessment. Additionally, the NCSC likelihood assessment corresponding to the report could be either medium (M) or high (H).

The scenarios correspond to the different threat intelligence reports that were selected. See Table I for an overview and the main CVE discussed in those reports. For example, a participant assigned to group A will analyze:

- four cases where the threat was assigned by a human corresponding to threat report 1, 2, 7, and 8.
- four cases where the threat was assigned by an algorithm corresponding to threat report 3, 4, 5, and 6.

This participant will also receive four cases in which the report was inconsistent corresponding to the scenario 2, 4, 5, and 7.

Additionally, this participant will receive four cases in which the report was consistent corresponding to the scenario 1, 3, 6, and 8. In addition, the participant receives four cases in which the NCSC likelihood assessment of the original report was medium corresponding to the scenario 2, 3, 6, and 7. The participant also receives four cases in which the NCSC likelihood assessment of the original report was high corresponding to the scenario 1, 4, 5, and 8.

D. Re-aggregation of Data

Once the experiment is completed for all participants the results must be re-aggregated to compare the intervention of interest (for example AI vs Human or Consistent vs Inconsistent).

For example, to obtain all data points relative to the results of the human recommendation one would need to consider the results of the experiments for the pairs: A:1, A:2, A:7, A:8, B:1, B:4, B:5, B:6, C:3, C:4, C:5, C:6, D:2, D:3, D:6, D:7, E:1, E:4, E:5, E:8, F:1, F:3, F:6, F:8, G:2, G:3, G:6, G:7, H:2, H;4, H:5, H:7, I:1, I:4, I:5, I:8, J:1, J:2, J:7, J:8, K:2, K:3, K:6, K:7, L:3, L:4, L:5, and L:6.

At this point there is a possible choice:

- consider each individual assessment as a data point (e.g. A:1 and A:2 are distinct data points for each assessor)
- consider each individual assessor as an aggregate data point (in which case for each assessor we can compute the average of his/her result for A:1, A:2, A:7, and A:8, and so on for the assessor doing B:1, B:4, B:4, B:5, B:6)

The results relative to the scenario corresponding to the analysis of threat intelligence report 1 will correspond to column 1.

E. Experimental artefact

The Dutch National Cyber Security Centre (NCSC) is the governmental single point of contact when it comes to cyber threats and incidents and has the legal obligation to analyze and research cyber threats and incidents [9]. NCSC advisories are governmental reports with the purpose of describing what a specific vulnerability entails and what could potentially happen if this vulnerability is exploited. In addition, practical information for mitigating the vulnerability is also provided if available [10]. The advisories include a risk assessment of the chance of exploitation (”Kans”) and the impact or damages that possible exploitation could entail (”Schade”) [10]. These assessments result in a categorical value of low, medium or high. The likelihood judgments are established by using a classification matrix [11].

Eight reports were selected based on their NCSC impact and likelihood risk classifications. Reports were only included in the study if they had a high impact/decision risk NCSC classification and a previous version of the same report where a medium likelihood risk NCSC classification existed. This choice was made to operationalize the step from medium/high priority to the highest and most critical priority. These two treatments (high impact - medium likelihood vs. high impact - high likelihood) were taken as factors in the study. See Figure 1 for an example of a report.

Abbreviated Vulnerability Report

Vulnerability C	
Description	Fortinet has fixed vulnerabilities in FortiOS, as used in FortiGate, FortiProxy and FortiWeb. An attacker could exploit the vulnerabilities to conduct attacks that could result in the following categories of damage: • Denial of Service (DoS) • Bypassing security measure • (Remote) code execution (Administrator/Root rights) • Access to sensitive data • Access to system data The most serious vulnerability is in the VPN-SSL used by FortiGate and allows an unauthenticated remote attacker to execute arbitrary code on the vulnerable system and potentially take over the system. This vulnerability has been assigned attribute "Vulnerability C".
Request	**Please provide assessment**

Background Threat Intelligence		
12.06.2023	Fortinet	Fortinet attributes abuse of older, similar vulnerabilities to an actor and estimates that this actor can and will soon abuse this specific vulnerability.
13.06.2023	NCSC	The NCSC expects Proof-of-Concept code and/or active abuse in the short term.

Final Internal Recommendation
On the basis of the report, A. Milliband assessed the vulnerability as having a **medium** chance of being exploited.

Fig. 1: Report 3: An example of a high report where the recommendation is inconsistent and the underlying source is human. The advice given in the recommendation is to assess the likelihood risk of the vulnerability at ’medium’. This is inconsistent with the second piece of threat intelligence where it is alluded that the availability of a Proof-of-Concept code and active abuse is expected in the short term.

For the study, the selected advisories were abbreviated to operationalize the problem of dealing with missing and incomplete information but maintaining ecological validity [10]. The abbreviated report consisted of a description of the main vulnerability discussed, pieces of threat intelligence updates (if available), and a recommendation for the likelihood assessment. The date and, where available, the source of the threat intelligence were also presented.

II. Metrics of Success

To measure the success of the intervention using the given data, we need to define specific metrics that can quantify the impact of the intervention. Here, we will replace the terms ’first metric’ and ’second metric’ with ’Confidence Level’ and ’Objectivity Level’, respectively.

A. High Level Explanation

```
.....................................................................
  In this section I have received the following
comments:
 1) first
 2) second
 3) third
 4) fourth
and in response to that I have done the following
changes:
```

- some changes here

TABLE I: Chosen Reports and Main CVE

ID	Report	Main CVE
1	NCSC-2023-0428	CVE-2023-38035
2	NCSC-2023-0277	CVE-2023-20887
3	NCSC-2023-0282	CVE-2023-27997
4	NCSC-2022-0368	CVE-2022-22972
5	NCSC-2022-0334	CVE-2022-1388
6	NCSC-2022-0056	CVE-2022-23131
7	NCSC-2023-0256	CVE-2023-2868
8	NCSC-2023-0346	CVE-2023-29300

TABLE II: Experiment design.

For each randomized order group, combinations of the interventions of a human or AI source recommendation (H or A), a consistent or inconsistent recommendation (C - I), a corresponding NCSC likelihood assessment (Hi - M) were provided. Refer to this Google sheet for the justification of the balancing of the design.

Report	1	2	3	4	5	6	7	8
ID Group								
A	H-C-Hi	H-I-M	A-C-M	A-I-Hi	A-I-Hi	A-C-M	H-I-M	H-C-Hi
B	H-I-M	A-C-M	A-I-Hi	H-C-Hi	H-C-Hi	A-I-Hi	A-C-M	H-I-M
C	A-C-M	A-I-Hi	H-C-Hi	H-I-M	H-I-M	H-C-Hi	A-I-Hi	A-C-M
D	A-I-Hi	H-C-Hi	H-I-M	A-C-M	A-C-M	H-I-M	H-C-Hi	A-I-Hi
E	H-C-Hi	A-C-M	A-I-Hi	H-I-M	H-I-M	A-I-Hi	A-C-M	H-C-Hi
F	H-I-M	A-I-Hi	H-C-Hi	A-C-M	A-C-M	H-C-Hi	A-I-Hi	H-I-M
G	A-C-M	H-C-Hi	H-I-M	A-I-Hi	A-I-Hi	H-I-M	H-C-Hi	A-C-M
H	A-I-Hi	H-I-M	A-C-M	H-C-Hi	H-C-Hi	A-C-M	H-I-M	A-I-Hi
I	H-C-Hi	A-I-Hi	A-C-M	H-I-M	H-I-M	A-C-M	A-I-Hi	H-C-Hi
J	H-I-M	H-C-Hi	A-I-Hi	A-C-M	A-C-M	A-I-Hi	H-C-Hi	H-I-M
K	A-C-M	H-I-M	H-C-Hi	A-I-Hi	A-I-Hi	H-C-Hi	H-I-M	A-C-M
L	A-I-Hi	A-C-M	H-I-M	H-C-Hi	H-C-Hi	H-I-M	A-C-M	A-I-Hi

- some changes there
- and more changes

Define the intuition behind your metrics of success

1) Confidence Level
2) Objectivity Level

The relations are clarified in Figure2

B. Confidence Level

(i) Definition: The Confidence Level metric is defined as the average confidence score of the participants in their vulnerability assessments. In this context:

X represents the confidence scores given by the participants. The formula calculates the sum of all confidence scores and divides it by the number of participants to get the average confidence level.

(ii) Direction: For the intervention to be considered a success, the Confidence Level should increase. Higher confidence levels indicate that the participants feel more assured about their assessments, which is a positive outcome of the intervention.

(iii) Reason: A successful intervention should make the participants feel more confident in their vulnerability assessments. Increased confidence suggests that the participants have gained knowledge and skills from the intervention, making them more capable of evaluating vulnerabilities accurately.

C. Objectivity Level

(i) Definition: The Objectivity Level metric is defined as the average objectivity score of the participants in their vulnerability assessments. It is calculated as follows:

$$ObjectivityLevel = \frac{\sum ObjectivityScores}{Numberof Participants} \quad (1)$$

(ii) Direction: For the intervention to be considered a success, the Objectivity Level should increase. Higher objectivity levels indicate that the participants' assessments are less biased and more based on factual information.

(iii) Reason: A successful intervention should enhance the participants' ability to make objective and unbiased evaluations. Increased objectivity suggests that the participants are relying more on factual information and less on subjective biases, leading to more accurate and reliable vulnerability assessments.

By defining and measuring these metrics, we can evaluate the success of the intervention and determine its impact on the participants' confidence and objectivity in vulnerability assessments.

III. Metrics Computation

The purpose of this section is to explain the process to move from the data that is collected to the metrics of success. Eventually, the process described in this section would lead to a .csv that has the fields as shown in Table III.

A. First Metrics

In this section I have received the following comments:

1) first
2) second
3) third

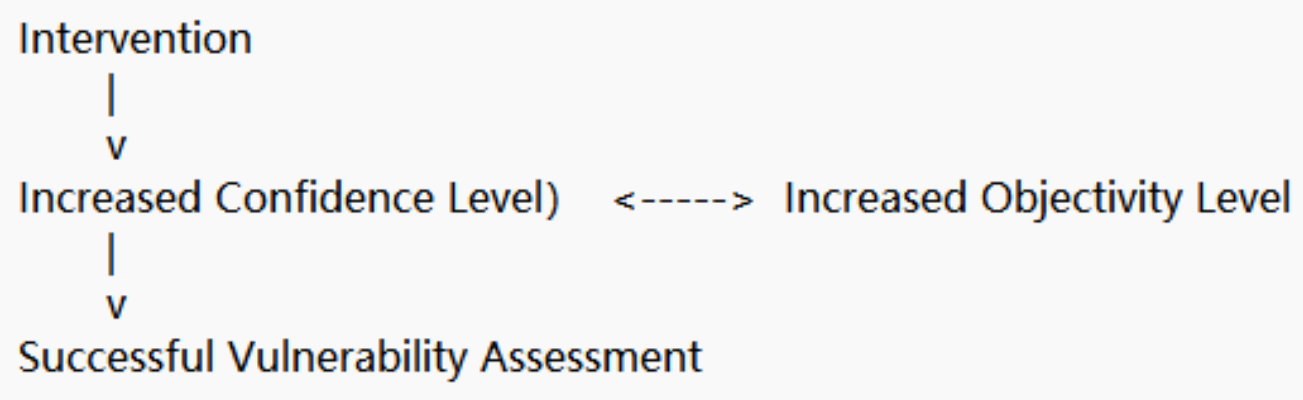


Fig. 2: Causal Relations Behind Confidence Level & Objectivity Level

ID	Groups	Scen_Danger	Demo_Confidence	Demo_Objectivity	Interv_Human	Interv_AI	Perc
1862	I	Medium	3	4	2	2	4.0
1940	E	High	4	3	1	2	3.0
2106	B	Medium	3	4	1	1	3.0
2513	A	Medium	4	5	2	2	3.67

TABLE III: Final results of the Metrics Computation

```
 4) fourth
and in response to that I have done the following
changes:
  • some changes here
  • some changes there
  • and more changes
```

The process to transform the raw data into metrics is described below:

1) **Load the Data:** Read the raw data from `data.csv`, ensuring all columns are initially treated as strings for consistent processing.
2) **Define Relevant Columns:** Identify the key columns used to compute metrics, including `Duration (in seconds)`, group assignment columns, danger scenario columns, and specific metric columns.
3) **Apply Mappings:** Map textual responses to numerical values for columns like `Demo_Confidence`, `Demo_Objectivity`, `Interv_Human`, and `Interv_AI` using predefined mappings.
4) **Calculate Derived Metrics:**
   - **Groups:** Assign each participant to a group based on their responses in columns `FL_20_DO_RandomizationA` to `FL_20_DO_RandomizationL`.
   - **Scen_Danger:** Determine the most frequent danger level (`High`, `Medium`, `Low`) across related columns.
   - **Perc:** Compute the average of mapped values across perception-related columns.
5) **Save Final Output:** Compile the computed metrics into a new table and save as `upload.csv`.

Example 1: Participant 1862 was assigned to Group I, with a medium danger level. Their `Demo_Confidence` score was 3 (somewhat confident), and they achieved a `Perc` score of 4.0. This indicates moderate confidence and good perception alignment, reflecting a positive intervention result.

Example 2: Participant 1940 was assigned to Group E, with a high danger level. Their `Demo_Confidence` score was 4 (fairly confident), but their `Perc` score was only 3.0. This indicates a potential mismatch between confidence and perception, suggesting room for improvement in the intervention's effectiveness.

B. Second Metrics

```
  In this section I have received the following
comments:
 1) first
 2) second
 3) third
 4) fourth
and in response to that I have done the following
changes:
  • some changes here
  • some changes there
  • and more changes
```

The steps for this metric are analogous to the first metric, with additional validations and edge-case handling.

IV. Other Variables

```
  In this section I have received the following
comments:
 1) first
 2) second
 3) third
 4) fourth
and in response to that I have done the following
changes:
  • some changes here
  • some changes there
  • and more changes
```

A. Intervention of Human

It is the confidence level of participants based on their responses to Q1123, where "I'm not confident at all" = 1, "I'm slightly confident" = 2, "I'm somewhat confident" = 3, "I'm fairly confident" = 4, and "I'm completely confident" = 5.

B. Intervention of AI

It is the objectivity level of participants based on their responses to Q1126, where "Strongly Disagree" = 1, "Disagree" = 2, "Neither agree nor disagree" = 3, "Somewhat agree" = 4, and "Strongly agree" = 5.

C. Perception Variable

The perception variable, `Perc`, was computed as the average of several perception-related columns. It is the participant's subjective feeling about the experiment as a whole. It is the value of Q293, Q294, Q295, Q296, Q32, Q34, Q36 and QID13. Each column was mapped using the same scale as `Demo_Objectivity`, ensuring consistency across the dataset.

V. Demographics of the Participants

Use the Colab notebook template to compute the descriptive statistics and run the analysis. Fill this section with the results. Export figure and data from the notebook to the report.

In this section I have received the following comments:

1) first
2) second
3) third
4) fourth

and in response to that I have done the following changes:

- some changes here
- some changes there
- and more changes

To present key characteristics of the population being analyzed, we computed basic descriptive statistics for the demographic variables. These variables include confidence levels, objectivity ratings, risk assessment experience, and familiarity with machine learning algorithms.

Comment on the data of Figure 3 and e.g. Figure 4, in particular if there are clusters and holes into it. For example, the lower part of the box is empty, etc.

Comment on the data. The scatter plot indicates that the dataset is not entirely balanced, with notable clusters and voids that could bias the intervention. The lack of data in the lower-right quadrant might indicate an underserved or underrepresented demographic. The intervention might require additional data collection or tailored approaches to ensure fairness and accuracy in addressing the needs of all groups. Possible criticalities include over-representation of certain categories, which could skew outcomes or lead to unintended biases.

VI. Results of the Intervention

In this section I have received the following comments:

1) first
2) second
3) third
4) fourth

and in response to that I have done the following changes:

- some changes here
- some changes there
- and more changes

Using the same Colab template, the distribution among the intervention groups is summarized in Table IV. Figure 5 provides a visual representation of the same information. If the box plots do not overlap, there might be a significant difference between the type of intervention.

TABLE IV: Distribution among intervention classes

This table summarizes the statistical distribution of Demo_Confidence and Demo_Objectivity across intervention groups.

	Demo_Confidence			Demo_Objectivity		
Group	Mean	Median	Variance	Mean	Median	Variance
A	3.00	3.0	1.00	4.33	4.0	0.33
B	3.17	3.0	1.77	3.50	4.0	2.70
C	3.40	3.0	1.30	4.40	4.0	0.30
D	4.00	4.0	0.00	4.25	4.0	0.25
E	3.33	3.0	0.33	3.33	3.0	0.33
F	3.75	4.0	0.25	4.25	4.0	0.25
G	3.00	3.0	1.00	4.20	4.0	0.20
H	3.33	4.0	1.33	3.00	2.0	3.00
I	3.33	3.0	0.33	4.00	4.0	0.00
J	4.00	4.0	1.00	4.00	4.0	0.00
K	4.00	4.0	0.00	4.60	5.0	0.30
L	4.00	4.0	0.50	3.40	4.0	1.80

A. Comment on the Data

Intervention Effectiveness: Groups D, J, and K exhibited the highest mean scores with the lowest variance in Demo_Confidence and Demo_Objectivity, indicating strong and stable intervention outcomes. Groups B and H, however, displayed higher variance, suggesting less consistent results.

Relation Between Metrics: Groups such as D and K show an alignment between high confidence and high objectivity, which may suggest the intervention's ability to positively influence both metrics under certain conditions.

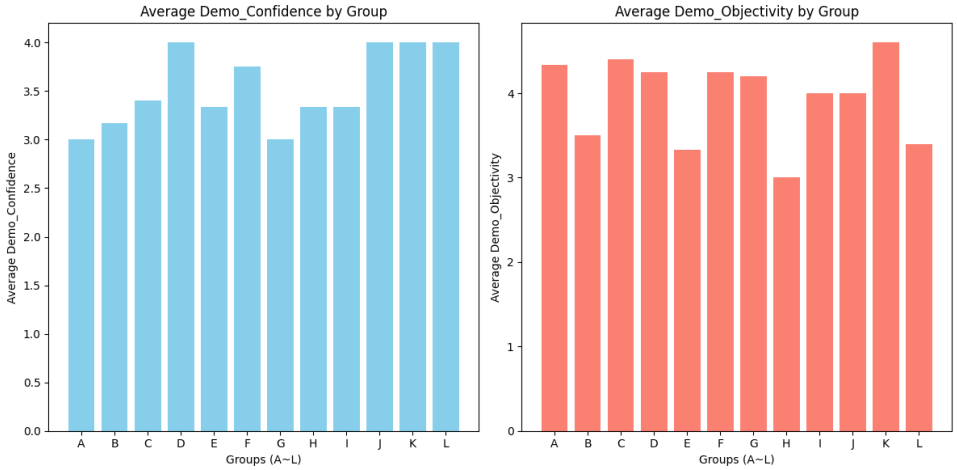
Recommendations:

- Reassess groups B and H to identify the reasons for high variability, such as scenario-specific factors or participant diversity.
- For lower-performing groups, consider implementing tailored feedback mechanisms to enhance intervention outcomes.

B. Control of scenarios (code fragments) as confounding factors

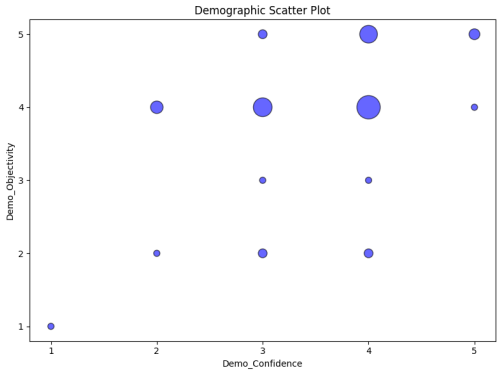
C. Impact of Scenarios on Experimental Results

Based on Figures 5 and 6, there is evidence to suggest that some scenarios (e.g., different code fragments or contextual



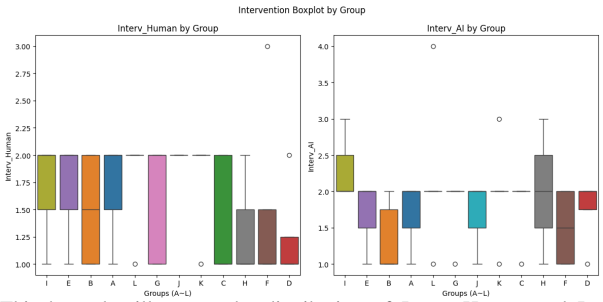
The histogram illustrates the distribution of Demo Confidence and Demo Objectivity across participants. The majority of participants reported medium-to-high confidence levels, while objectivity ratings are more uniformly distributed. This highlights varying levels of self-reported confidence in demographic variables, a key factor influencing other responses in the study.

Fig. 3: Distribution of Demo Confidence and Demo Objectivity



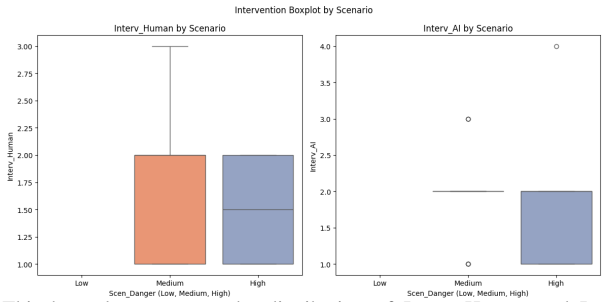
The scatter plot reveals clusters of points that represent demographic categories A and B. Notably, there is a visible gap in the lower-right quadrant, suggesting a lack of representation in this region of the demographic space. Additionally, the top-left region has a dense cluster of points, potentially indicating over-representation of certain attributes. These patterns highlight significant imbalances in the dataset that should be addressed for equitable intervention analysis. The plot emphasizes the need for further exploration of these imbalances.

Fig. 4: Confidence Level vs. Objectivity Level



This box plot illustrates the distribution of Interv_Human and Interv_AI across groups. Key insights include the variation in medians and interquartile ranges, which highlight the effectiveness of the intervention. For example, Groups I and K show relatively high Interv_AI medians, while Groups D and F have lower values, indicating potential group-specific impacts of the intervention.

Fig. 5: REGENERATE THIS BOX PLOT WITH THE REAL DATA



This box plot compares the distribution of Interv_Human and Interv_AI metrics across different scenarios (Low, Medium, High). Key observations include the difference in medians and interquartile ranges. For example, the High scenario shows a narrower range and higher median for Interv_AI, while the Medium scenario exhibits wider variation in Interv_Human. This highlights the potential impact of scenario-specific factors on intervention outcomes.

Fig. 6: REGENERATE THIS BOX PLOT WITH THE REAL DATA

settings) may significantly contribute to the differences observed in the metrics, independent of the intervention. Below are the key observations and interpretations:

1) Interv_Human by Group vs. Scenario: **Figure 5 Insight (by Group):**

- Most groups show a relatively narrow range of variation in the Interv_Human metric, suggesting consistent responses to the intervention across groups.
- Outliers in certain groups (e.g., G and H) indicate context-specific factors influencing the results.

Figure 6 Insight (by Scenario):

- The **Medium-risk scenario** exhibits a broader interquartile range and higher variability in Interv_Human, suggesting more diverse participant responses in this condition.
- The **High-risk scenario** shows concentrated responses, indicating that the perceived risk level might standardize participants' behavior regardless of the intervention.

Conclusion: The broader variation in the Medium-risk scenario suggests that the context plays a significant role in shaping participants' confidence levels, potentially overshadowing the actual effects of the intervention.

2) Interv_AI by Group vs. Scenario: **Figure 5 Insight (by Group):**

- Variability in Interv_AI is more pronounced between groups compared to Interv_Human, with some groups (e.g., F and H) displaying wider interquartile ranges.
- Groups such as I and K exhibit relatively higher median values, suggesting more positive intervention effects in these groups.

Figure 6 Insight (by Scenario):

- The **High-risk scenario** shows a higher median and a more compact distribution in Interv_AI, while the **Medium-risk scenario** displays significant outliers and variability.
- The absence of significant results in the **Low-risk scenario** implies that interventions in this context may not effectively differentiate participants' responses.

Conclusion: The tighter distribution and higher median in the High-risk scenario suggest that scenario-specific factors, such as perceived urgency or complexity, may amplify the effects on Interv_AI. Conversely, the Medium-risk scenario introduces more noise, potentially confounding the observed group-level differences.

3) General Observations and Recommendations: **Scenario as a Confounding Factor:**

- Scenarios, particularly the High- and Medium-risk ones, appear to influence participants' responses beyond the intended effects of the intervention.
- Some observed differences may stem more from the scenario context than from the intervention itself.

Intervention-Scenario Interaction:

- Certain groups (e.g., K) perform well regardless of the scenario, while others (e.g., H) show high variability, possibly due to interactions between the intervention and the scenario's characteristics.

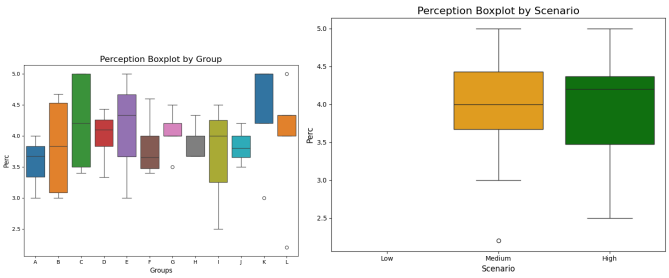
Recommendations:

- **Scenario Balancing:** Ensure equal distribution of scenarios across intervention groups to minimize their confounding impact.
- **Separate Scenario Analysis:** Analyze the effect of scenarios independently from interventions to better isolate the intervention's true impact.
- **Customized Interventions:** Tailor interventions to the specific challenges of each scenario type (e.g., providing more support in Medium-risk scenarios).

Summary: The figures strongly suggest that scenarios significantly influence both metrics, potentially overshadowing the effects of the intervention. Accounting for these contextual factors is crucial for accurate interpretation of experimental results.

D. Perception Metric

Include a discussion on the perceived metrics for the experiment by using a figure like Figure 7.



(a) Perception by Group (b) Perception by Scenario

The box plots illustrate the distribution of the Perception metric (Perc) across groups and scenarios. Subfigure 7a shows variability across groups, with some groups (e.g., B and K) exhibiting higher medians and interquartile ranges. Subfigure 7b highlights differences between scenarios, where the High-risk scenario shows a more concentrated distribution compared to the Medium-risk scenario.

Fig. 7: REGENERATE THESE BOX PLOTS WITH THE REAL DATA

TABLE V: Wilcoxon test results

The first part of the table compares the intervention, and the second part compares the scenarios. If everything goes well, the difference between the intervention should be significant, and the difference between the scenarios should not be significant. Either way, comment on the key results. The U value is the value of the statistics, and p is the p-value from the colab.

		Metric 1 (Interv_Human)		Metric 2 (Interv_AI)	
List 1	List 2	U	p	U	p
Intervention					
Group I	Group E	NaN	NaN	1.091	0.275
Group I	Group B	NaN	NaN	1.807	0.071
Scenarios					
Medium	High	NaN	NaN	1.515	0.130

E. Statistical Test

In this section I have received the following comments:

1) first
2) second
3) third
4) fourth

and in response to that I have done the following changes:

- some changes here
- some changes there
- and more changes

The statistical tests are provided in the colab file. There are many different tests around but the purpose of everybody using the same (simple) tests is to have a shared view of what happens so that the difference between significance and no significance is not in the particular test (this may happen in science), but it can only be caused by the different decisions to transform the raw data into numbers to put into a statistical test.

The colab will give you the value of the Chi-square test for Metric 1 and Metric 2. Each is reported as $\chi^2 = \ldots, p = \ldots$. The results are:

- Metric 1: $\chi^2 = 0.0$, $p = 1.0$.
- Metric 2: $\chi^2 = 0.0$, $p = 1.0$.

1) Conclusion:

TABLE VI: Zero vs. Non-Zero grouping for interventions and scenarios

The first part of the table compares the intervention, and the second part compares the scenarios. The n values are the number of occurrences of each of the cases.

		Metric 1 (Interv_Human)		Metric 2 (Interv_AI)	
List 1	List 2	0	¿0	0	¿0
Intervention					
Group I	Group E	5	3	4	4
Group I	Group B	4	4	3	5
Scenarios					
Medium	High	6	2	5	3

1) **Metric 1 (Interv_Human):** No significant differences were observed between intervention groups or scenarios. The high number of zeros in the data likely reduced statistical power.
2) **Metric 2 (Interv_AI):** While most pairwise comparisons showed no significant differences, the Medium vs. High scenario comparison ($p = 0.13$) suggests potential differences worth further exploration.
3) **Scenarios:** Both metrics showed no significant differences across scenarios. Insufficient data and high zero values likely contributed to the lack of statistical power.

VII. Summary

In this section I have received the following comments:

1) first
2) second
3) third
4) fourth

and in response to that I have done the following changes:

- some changes here
- some changes there
- and more changes

The experiment aimed to evaluate the impact of algorithmic and human-based recommendations on decision-making in vulnerability assessment. Through a balanced design and controlled scenarios, we investigated the effects of consistency, risk levels, and recommendation sources on participants' confidence and objectivity levels.

Key findings include:

1) **Confidence Level**: No significant differences were observed across groups or scenarios. High variability in certain groups (e.g., B and H) suggests room for improvement in intervention design and feedback mechanisms.
2) **Objectivity Level**: While no significant differences were found overall, the Medium vs. High scenario comparison suggested a potential trend (p = 0.13) worth further exploration.
3) **Scenario-Based Impact**: Scenarios, especially those with medium and high risks, were identified as potential confounding factors. The Medium-risk scenario exhibited broader variability, indicating that contextual elements might influence participant responses.

The primary limitations of this study include:

- **Zero-Heavy Data**: A significant portion of the data contained zeros, reducing statistical power and obscuring potential effects.
- **Imbalanced Distribution**: Certain demographic groups and scenarios were underrepresented, leading to potential biases.
- **Metrics Sensitivity**: The chosen metrics (Confidence and Objectivity Levels) might require refinement to better capture the nuances of the intervention's impact.

Future work should focus on improving data balance, refining metrics, and tailoring interventions to specific scenarios to achieve more robust and actionable insights.

Reflect on this experiment and describe what are the key criticalities and limitations of this particular experiment design. The discussed limitations and critical points must be specific to the experiment (e.g. identified co-founding variables), and not general limitations of experimentation.

End the report with your interpretation of the results and provide your main findings. Make sure that the main findings are in fact supported by the analysis described in the previous sections and explain how you derived those conclusions.

Exercise your judgement in determining whether the effect is actually practically significant or just statistically significant or insignificant from all perspectives.

VIII. Artefacts

In this section I have received the following comments:

1) first
2) second
3) third
4) fourth

and in response to that I have done the following changes:

- some changes here
- some changes there
- and more changes

The following artefacts were produced and will accompany this report:

1) **Data Cleaning and Preprocessing**:
   - Microsoft Excel was initially used to clean the raw data. This involved:
     - Removing incomplete rows and irrelevant columns.
     - Merging related columns to ensure data consistency.
   - The cleaned data was saved as `data_after_clean.csv`.
2) **Data Transformation and Feature Engineering**:
   - A Python script was used to transform the data in `data_after_clean.csv` and generate a structured dataset containing the following columns:
     - **ID (Example)**: Derived from the column `Duration (in seconds)`.
     - **Groups**: Grouped participants into A to L based on the columns `FL_20_DO_RandomizationA` to `FL_20_DO_RandomizationL`.

- **Scen_Danger**: Defined as the most frequent value (High/Medium/Low) among columns `Q1035`, `Q1037`, `Q1167`, `Q1173`, `Q1179`, `Q1185`, `Q1191`, and `Q1197`.
- **Demo_Confidence (Metric 1)**: Confidence level mapped from column `Q1123`:
  - "I'm not confident at all" → 1.
  - "I'm slightly confident" → 2.
  - "I'm somewhat confident" → 3.
  - "I'm fairly confident" → 4.
  - "I'm completely confident" → 5.
- **Interv_Objectivity (Metric 2)**: Objectivity level mapped from column `Q1126`:
  - "Strongly Disagree" → 1.
  - "Disagree" → 2.
  - "Neither agree nor disagree" → 3.
  - "Somewhat agree" → 4.
  - "Strongly agree" → 5.
- **Demo_Human**: Risk assessment experience mapped from the `Risk Assessment` column:
  - "No experience" → 1.
  - "As part of University projects" → 2.
  - "Performed one or two risk assessment projects outside of university" → 3.
- **Demo_AI**: Machine learning algorithm experience mapped from the `ML alg` column:
  - "No experience" → 1.
  - "As part of University projects" → 2.
  - "Used one or two ML algorithms outside of university" → 3.
  - "Used several ML algorithms outside of university" → 4.
- **Perc (Perception 1)**: Overall subjective perception mapped from columns `Q293`, `Q294`, `Q295`, `Q296`, `Q32`, `Q34`, `Q36`, and `QID13`:
  - "Poor", "Strongly Disagree" → 1.
  - "Acceptable", "Neutral", "Neither agree nor disagree" → 3.
  - "Good", "Strongly agree" → 5.

- The final dataset was saved as `upload.csv`.

3) **Colab Notebook**:
   - The notebook contains all analyses, including statistical tests (Chi-square and Wilcoxon), data visualization (box plots and scatter plots), and detailed metric computations.
   - Key figures such as:
     - Box plots for intervention and scenario comparisons.
     - Pairwise Wilcoxon test results.
4) **Figures and Tables**:
   - Exported figures and tables for the report, including:
     - Table I: Summary statistics for confidence and objectivity levels.
     - Table II: Chi-square and Wilcoxon test results.
     - Figure 1: Box plots for Metric 1 (Confidence) and Metric 2 (Objectivity) by groups and scenarios.
5) **Documentation**:
   - Detailed documentation explaining the data cleaning process, metric definitions, and experimental procedures.
   - Annotations for potential limitations and recommendations for future studies.

These artefacts ensure reproducibility and transparency of the analysis, providing a robust foundation for further research and practical applications.

REFERENCES

[1] S. de Smale, R. van Dijk, X. Bouwman, J. van der Ham, and M. van Eeten, "No one drinks from the firehose: How organizations filter and prioritize vulnerability information," in *2023 IEEE Symposium on Security and Privacy (SP)*, 2023.

[2] G. Di Tizio, M. Armellini, and F. Massacci, "Software updates strategies: A quantitative evaluation against advanced persistent threats," *IEEE Transactions on Software Engineering*, vol. 49, no. 3, pp. 1359–1373, 2022.

[3] J. R. B. Jr., "Executive order on improving the nation's cybersecurity," *The White House*, May 2021. [Online]. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity/

[4] V. Bier, "The role of decision analysis in risk analysis: A retrospective," *Risk Analysis*, vol. 40, no. S1, pp. 2207–2217, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.13583

[5] J. G. Jaspersen and G. Montibeller, "Probability elicitation under severe time pressure: A rank-based method," *Risk Analysis*, vol. 35, no. 7, p. 1317–1335, 2015.

[6] B. Dietvorst, J. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *Journal of Experimental Psychology: General*, vol. 144, no. 1, p. 114–126, 2015.

[7] N. Castelo, M. Bos, and D. R. Lehmann, "Task-dependent algorithm aversion. journal of marketing research," *Journal of Marketing Research*, vol. 56, no. 5, p. 809–825, 2019.

[8] D. R. Mandel, "Assessment and communication of uncertainty in intelligence to support decision-making," *NATO STO TECHNICAL REPORT, TR-SAS-114*, 2020.

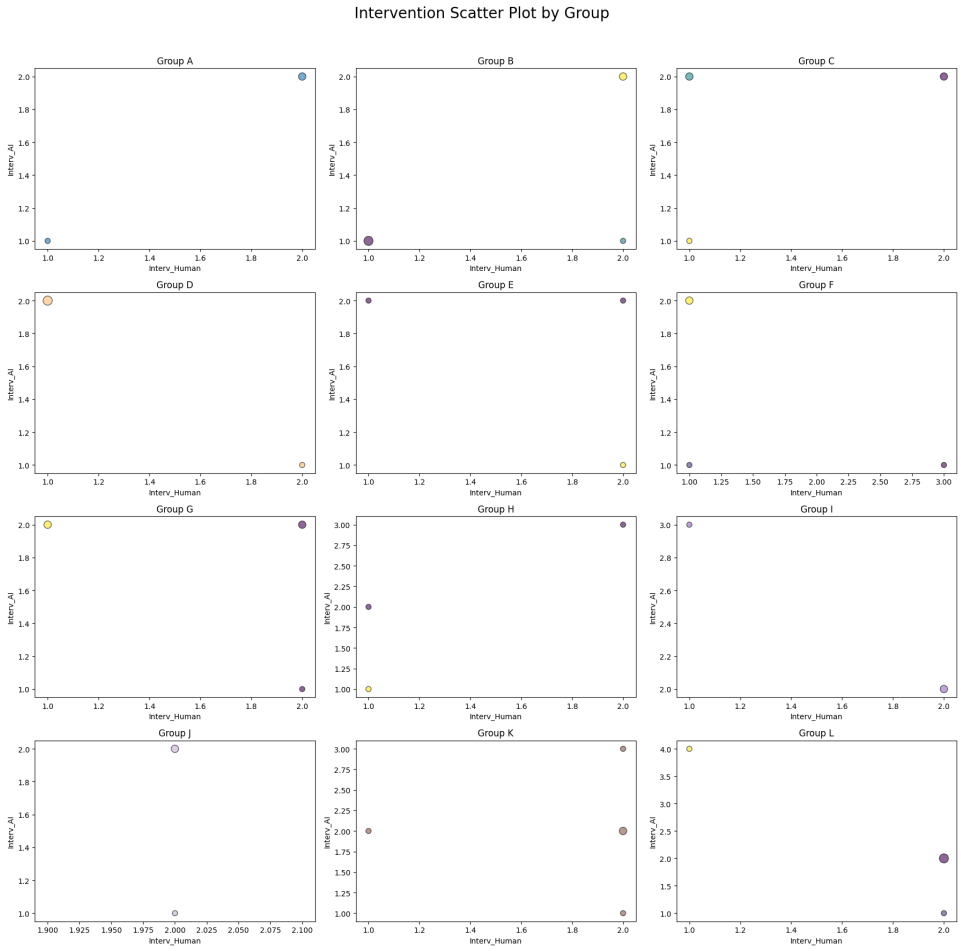
[9] I. Kamara, R. Leenes, K. Stuurman, and v. J. Boom, "The cybersecurity certification landscape in the netherlands after the union cybersecurity act," *Tilburg Institute for Law, Technology, and Society*, 2020.

[10] NCSC, "What is a ncsc advisory," 2023. [Online]. Available: https://www.ncsc.nl/documenten/publicaties/2019/juli/02/wat-is-een-ncsc-beveiligingsadvies

[11] ——, "Inschalingsmatrix," 2023. [Online]. Available: https://www.ncsc.nl/documenten/publicaties/2019/juli/02/inschalingsmatrix

APPENDIX

[OPTIONAL] You can see possible causes of confounding and interaction effectshere in the scatter plot in Figure 8



Describe here the key insight of the diagram. This diagram must be of course regenerated with your data from gcolab "Intervention Scatter Plot by Group".

Fig. 8: REGENERATE THIS SCATTER PLOT WITH THE REAL DATA