Artificial Intelligence Threat Reporting and Incident Response System

D.2.7 IRIS evaluation and impact assessment

Project Title:	Artificial Intelligence Threat Reporting and Incident Response System
Project Acronym:	IRIS
Deliverable Identifier:	2.7
Deliverable Due Date:	31/08/2023
Deliverable Submission Date:	15/09/2023
Deliverable Version:	v1.0
Main author(s) and Organization:	Nikolas Kapsalis (KEMEA)
Work Package:	WP2 - System co-design
Task:	System Evaluation and Assessment
Dissemination Level:	PU: Public

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 101021727. Content reflects only the authors' view and European Commission is not responsible for any use that may be made of the information it contains.

Quality Control

	Name	Organisation	Date
Editors	Lorena Volpini, Carmela Occhipinti	CEL	08/06/2023
	Nikolas Kapsalis	KEMEA	08/06/2023
Peer Review 1	Gonçalo Cadete	INOV	15/09/2023
Peer Review 2	Susana Zarzosa	ATOS	14/09/2023
Submitted by (Project Coordinator)	Gonçalo Cadete	INOV	15/09/2023

Contributors

Organisation
CEL
KEMEA

Document History

Version	Date	Modification	Partner
v0.1	01/06/2023	ToC and Introduction	KEMEA
v0.2	31/08/2023	1st version	KEMEA, CEL
V0.3	11/09/2023	Final full draft	KEMEA
v0.4	14/09/2023	Reviewers' feedback	ATOS, INOV
V0.5	14/09/2023	Address reviewers' feedback	CEL
v1.0	15/09/2023	Final editing	INOV

Legal Disclaimer

Contents

List of Figures

List of Tables

Table 1. Relation to other project documents

List of Abbreviations and Acronyms

Abbreviation/ Acronym	Meaning
AI	Artificial Intelligence
CE	Capacity Enabling
CEL	CYBERETHICSLAB Srls
CERT	Computer Emergency Response Team
CISO	Chief Information Security Officer
DPA	Data Protection and Accountability
Expected Systemic Change	ESC
GDPR	General Data Protection Regulation
HiL	Human in the Loop
IoT	Internet of Things
ITW	Institutional Trustworthiness
LK	Likeability
PBC	Perceived Behaviour Control
PEU	Perceived Ease of Use
PR	Perceived Risks
PU	Perceived Usefulness
PUC	Pilot Use Case
RL	Reliability
SAT	Social Acceptance of Technology
SIW	Stakeholders and Industrial Workshop
SME	Small and Medium Enterprise
TR	Transparency
USC	User Perceived Certainty
UX	User eXperience

Executive Summary

This deliverable, a result of task T2.6 "System Evaluation and Assessment", aims at providing a qualitative and quantitative evaluation and assessment of the IRIS technology, based on the Social Acceptance of Technology (SAT) methodological framework, which was described in D2.4 "Human factors for co-design methodology". To this end, the methodology described in D2.4 was employed to measure the Social Acceptance of the IRIS technology and tools to a network of stakeholders in order to capture different perceptions and perspectives. Furthermore, the defined qualitative and quantitative techniques developed in the context of T2.4 "Human factors for co-design of effective cross-border threat intelligence sharing" were applied to **assess the social acceptance of the IRIS technology** by the main stakeholders. In particular, the qualitative – employing focus groups and interviews with highly qualified practitioners and quantitative – employing the questionnaire presented in D2.4 - assessment that is described in the present Deliverable enabled the project members to investigate how the IRIS technology and tools are perceived by practitioners. Furthermore, the analysis of the data gathered and the questionnaire results serve as a basis for refinements and improvements of the IRIS platform and tools.

The feedback received by the relevant stakeholders, employing the aforementioned qualitative and quantitative methods were then analysed in order to extract lessons from it and provide feedback to other stakeholders other than the IRIS' practitioners, such as for instance the project pilots.

1 INTRODUCTION

1.1 Deliverable Purpose

Within the context of project task T2.6, this deliverable aims at describing the application of the IRIS methodology for social acceptance assessment of the IRIS platform and tools. The social acceptance is carried out through the application of the methodology which was defined and described in the context of Task 2.4, and that takes into account both qualitative and quantitative aspects.

1.2 Relation to other project activities

As previously mentioned, the present document is strongly related to the activities that were carried out in Task T2.4, given that in the present deliverable the results of the application of the methodology described in Deliverable 2.4 are presented. Furthermore, the feedback and lessons learned described in the present Deliverable will be employed in the pilot planning, execution and evaluation, in the context of WP7. It also connects with Deliverable 2.3, "Ethics and data protection requirements specification", where the ethics requirements that the project needs to follow in the development phase were defined.

#ID	Deliverable name	Deliverable description	Submission Date/Deadline
D2.3	Ethics and data protection requirement specification	It aims to specify the requirements related to ethics, data protection and secure sharing of data.	Month 8
D2.4	Human factors for co-design methodology	It aims at providing the **human factors** that will be taken into account in the process of co-designing the IRIS technology	Month 16

Table 1. Relation to other project documents

1.3 Document structure

Section	Title	Brief Summary
1	Introduction	It provides a brief explanation of the aim of the present deliverable and its structure.
2	IRIS Social Acceptance Methodology Results	This section describes the validation results of the IRIS Social Acceptance Methodology
3	IRIS Social Acceptance Assessment	This section describes the IRIS social acceptance assessment results
4	Lessons Learned and Feedback	This section describes the lessons learned from the IRIS social acceptance assessment and describes enhancements and improvements to the IRIS platform and tools
5	Conclusions	This section concludes the document

Table 2. Document Structure

2 IRIS SOCIAL ACCEPTANCE METHODOLOGY APPLICATION

2.1 Assessment of the IRIS Social Acceptance methodology

The IRIS methodology for social acceptance assessment, described in D2.4, was at first applied through an in-person engagement of experts. The stakeholders engaged for this assessment were three Chief Information Security Officers (CISOs) from Romania, each one from a different industry sector, namely energy, banking and SME sectors. The interlocutors from the energy sector and the banking sector are CERTs, while the one from the SME sector operates different organisations, among which many CERTs.

A CISO, or Chief Information Security Officer, is a senior-level executive responsible for the information security of an organization. CISOs are responsible for developing and implementing security policies and procedures, managing security staff, and overseeing the security of the organization's information systems. Therefore, they can offer valuable perspective on potential risk and challenges, and they can provide recommendations on how to mitigate them. Their input and overall involvement in the IRIS project hold substantial value due to their diverse backgrounds. This stems from the fact that they hail from various sectors, enabling them to contribute distinct perspectives that are closely attuned to the specific security requirements and strategies of their respective industries.

For these reasons, their contribution has been considered relevant and valuable in this phase of research and development of the IRIS system.

The experts group engagement was held during the Security and Defence 2023 Conference (RISE-SD) held in Rhodes from May 29 to May 31, 2023, as described below. Following a comprehensive demonstration of the IRIS platform and its tools, during the II Stakeholder and Industrial Workshop demonstration (which took place during the aforementioned event) the experts were encouraged to share their perspectives on the IRIS platform. This was done to gain a well-informed assessment of potential enhancements and improvements that could be made to the platform.

2.1.1 Validation of SAT Methodology

The procedure outlined in Task 2.4, which details the methodology for assessing the social acceptance of technology developed for the IRIS platform, was initially put into practice at the RISE-SD. At this event, the SAT-based methodology for evaluating the social acceptance of IRIS was introduced to three Chief Information Security Officers (CISOs). A live demonstration of the IRIS tools was conducted during a face-to-face meeting,

followed by soliciting their feedback through two focused group sessions and the completion of the questionnaire specified in D2.4. The questionnaire was completed online via the EU survey platform. The method, as tailored to IRIS platform, was being applied for the first time, serving as a validation of its efficacy and suitability in the given context.

3 IRIS SOCIAL ACCEPTANCE ASSESSMENT AND RESULTS

3.1 Qualitative Assessment

IRIS methodology for social acceptance assessment was applied in-person during the II SIW meeting held in Rhodes, with the engagement of three cybersecurity experts, as stated above.

Following a comprehensive demonstration of the IRIS platform and its tools, the experts were encouraged to share their perspectives on the IRIS platform. This was done to gain a well-informed assessment of potential enhancements and improvements that could be made to the platform.

3.1.1 Qualitative assessment Methodology

At this stage, the SAT assessment qualitative component methodology included the demonstrations of IRIS technology (narrative approach combined with video demos) focusing on IRIS Platform Use in a smart city (Pilot Use Case 1), the enhanced MeliCERTes ecosystem and the Data Protection and Accountability component.

Following the demonstrations, the workshop proceeded with a focus group discussion coordinated and moderated by CEL.

WHO: experts to assess the IRIS technology

WHAT: expectations/perception of experts on the IRIS Platform

HOW: semi-structured group interview, plus discussion

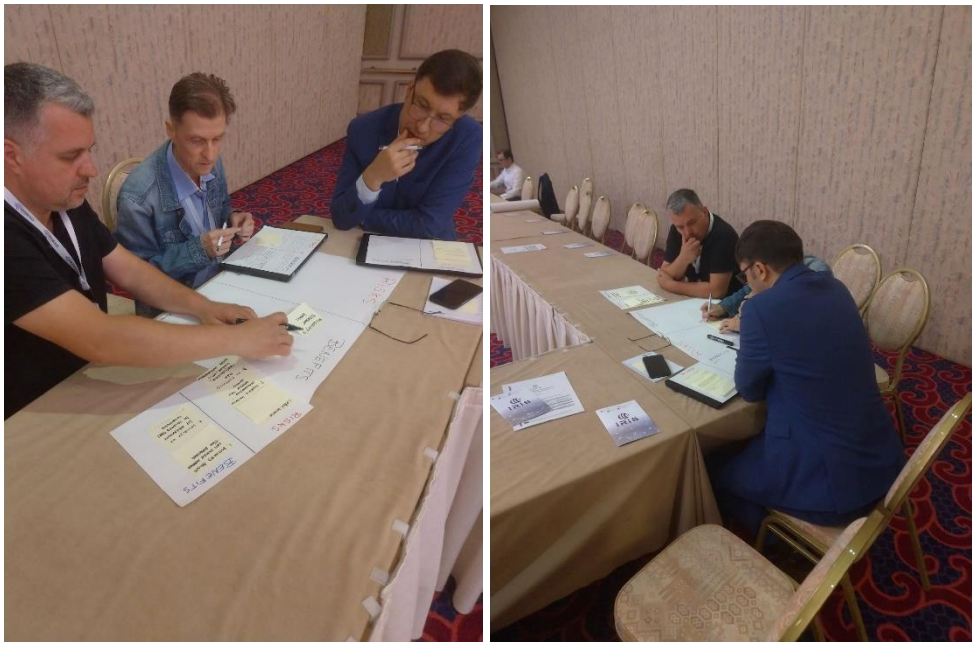

Figure 1 - Discussion group, session 1

The research was conducted by deepening two of the Social Acceptance of Technology clusters, called "bubbles" (see D2.4), namely the User eXperience (UX) bubble and the Value Impact bubble. The focus group was divided in two sub-sessions. (1) During the first session SAT methodology (cf. D2.4) was presented to participants and the discussion was focussed on risks and benefits of IRIS technology (UX factor). This point was considered relevant for other projects activities as well. Hence, at the end of this session, the results here reported were presented to the rest of II SIW participants.

Figure 2 - Discussion group session 1 methodology

The second session, focussing on the "Value Impact" factor of Social Acceptance included a brief presentation on the notion of value, the idea of values embedded in technology (Flanagan et al., 2018, van de Poel, 2020b) and the specific values relevant to cybersecurity technology (Christensen et al., 2020). The participants were asked to identify the values embedded in IRIS technology, according to their perception, and then discuss if these values would harmonise or conflict with other values crucial to stakeholders (end-users, society in general, specific social groups).

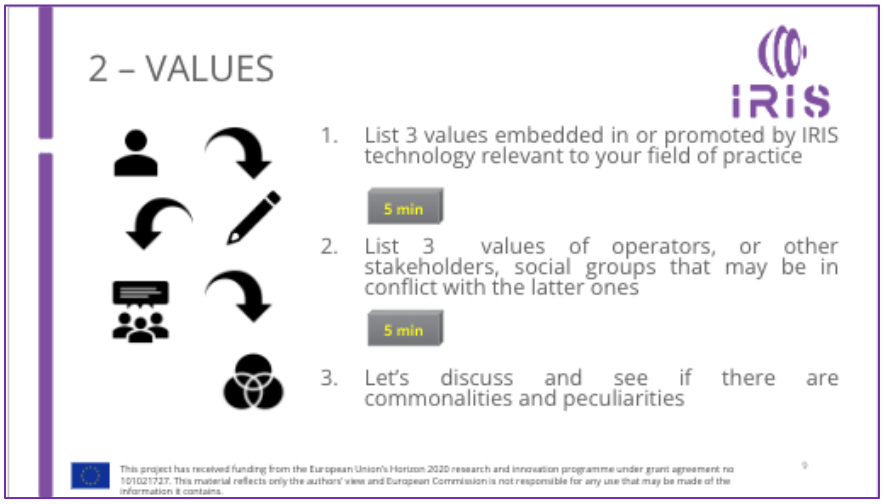

Figure 3 - Discussion Group session 2 methodology

The notion of "value" adopted in this phase is the one referred by the EU project CANVAS (https://canvas-project.eu/), also conceptualised in the literature on the ethics of cybersecurity. According to this body of knowledge, value embodies the idea of something good and desirable, also providing individuals with some form of orientation on how to behave (Christen et al., 2020). The values on which the group discussed were the ones recognised as relevant to cybersecurity: security, privacy, fairness and accountability. However, instead of taking them as simple and discrete entities they were here thought of as clusters (van de Poel, 2020), as showed in Figure 4.

Figure 4 - Value Clusters

This approach views values as fluid and context-dependent. It allows discussing values as they emerge from different situations (such as use cases or hypotheses) and different perspectives (such as those of different stakeholders or social actors). This makes it easier to analyse how values interact among each other and become relevant in different contexts.

One classic demonstration of these variable interactions refers to the value of security. In fact, one may hardly argue that "security" in itself is not a value to be pursued. But in considering concrete applications and contexts, one may see that there are different types of security relevant to specific situations or applications and social groups. A famous example is the encryption dilemma: a government has tech companies to build backdoors into encrypted messaging for national security against terrorists. However, this compromises personal security and privacy, showing that in the trade-off between preventing attacks and safeguarding individual rights different notions of security as a value come into play. The take-away message is that depending on the application and the context of use, different interactions and even conflicts among values may arise.

It is worth noting that the research activity carried out in the second session about values has required participants to think in terms of values embedded in technology and values spread among stakeholder groups (operators, professionals, social groups and society in general) which resulted to be unusual in their professional practice. For this reason, the discussion group's dynamic had been conceived and planned in a flexible way: when

needed CEL researchers would intervene through questions able to stimulate participants' reflection and further discussion.

This kind of research methods allowed the collection of data which are less structured than those resulting from surveys employing scaling techniques. On the other hand, they allow a hermeneutic approach to discursive interactions and content analysis of linguistic material which are better suited procedures to grasp the semantic connections structuring interlocutors' perceptions and organising values (Rositi, 1993). Moreover, they allow the gathering of stakeholders' feedback which is useful for the project, also beyond the scope of social acceptance research. For this reason, the outcomes of both sessions have been shared by presenting them in a face-to-face session to the partners participating at the II stakeholders and industrial workshop.

3.1.2 Qualitative assessment Results

3.1.2.1 Session 1 - Risks and benefits

Regarding the risks, the research participants identified risks or areas of concern related to different aspects of IRIS technology.

The first one relates to the use of open-source software. Its inherent nature presents potential security challenges. The transparency of the source code allows for a comprehensive understanding of the system's development and functionality, thereby making it more susceptible to the discovery and exploitation of vulnerabilities by malicious actors.

The second area of concern relates to potential liabilities that arise after the system's release. Specifically, it focuses on maintenance, support, improvement, and regulatory compliance following the system's implementation (or the lack thereof) and possible liability generated by it. It is perceived as crucial to establish a comprehensive maintenance and support framework for the system. This includes regular updates, bug fixes, and performance optimizations. Continuous improvement efforts should also be prioritized to enhance system functionality and address evolving user needs and to maintain a high level of system reliability. Ensuring regulatory compliance was also mentioned in this second area of concern: regulatory frameworks are subject to change, and organizations must remain vigilant to adapt their systems to guarantee compliance.

The third area of concern was the privacy of IoT users. This concern was motivated by the fact that much data that the platform will ingest could be personal data or confidential data -- for example data collected by IoT devices (e.g. security camera IP) that would be exchanged between CERTs. This concern must be taken into consideration, given that the system is due to be used in many countries. In the light of GDPR one must think of potential risks of identification, re-identification and/or singling out. Moreover, there are constraints privacy and confidentiality related, beyond GDPR, which are country-specific and sector-specific.

Although they were asked to identify only three risks, other areas of concern related to AI have been brought forward to the discussion and are worth of mention.

One of these is AI Vector Attacks. These attacks occur by poisoning of training data to undermine the effectiveness of threat detection and incident response mechanisms. The manipulation of training data can result in compromised system performance, potentially leading to serious security breaches. According to the interlocutors, addressing and mitigating the risks associated with AI vector attacks is crucial for maintaining robust and reliable system functionality.

In addition to AI vector attacks, interlocutors raised concerns about the ethical implications of AI implementation. These concerns encompass various ethical dimensions associated with the utilization of AI systems. Stakeholders expressed apprehension regarding the operational-level deployment of AI-generated incident response due to issues surrounding accountability, transparency, and potential biases. The stakeholders consistently emphasized the importance of AI systems being compliant with relevant regulations. This underscores the need to establish a framework that ensures adherence to legal and ethical standards governing AI implementation. By aligning AI systems with regulatory requirements, organizations can enhance trust, mitigate risks, and foster responsible AI practices. This area of concern was also the one in which the stakeholder from the electrical sector claimed some peculiarities regarding critical infrastructures namely, legal aspects related to sensitive information, and how information is used, which varies according to each country's law.

Figure 5 - Risks and benefit discussion results

In regard of the benefits, the fact of IRIS being based on Open-Source software, previously listed among the concerns, was also considered from the point of view of advantages: beyond lower development costs, by having the source code open, contributions for improvement can make the solution stronger.

Another benefit emphasized by the participants was the fact of IRIS solution being applicable to different domains (upon improvement).

The automated process was included among the benefits because it reduces human interference and human error as well as increases cost efficiency. While there are initial implementation costs, in the long run, automation can lead to cost savings by optimizing resource utilisation (e.g. human resources).

A positive contribution to European security is also perceived as afforded by the fact of the platform being collaborative at national and international level. Something that should be more clearly understood is the data sharing mechanisms and processes and related data protection and security.

As a last input on benefits, peculiar to strategic infrastructure and particularly the energy sector, the IRIS solution was considered as useful for the security of IoT devices in isolated areas (e.g. security cameras, sensors).

The table below resumes the risks and benefits as emerged in the discussion session.

Risks	Benefits
Use of Open-Source software	Use of Open-Source software
After system's release liability	Applicability to different domains
AI vector attacks	Automated process
Privacy	Collaboration and interoperability
Ethical and Legal Aspects	Security of IoT in isolated locations
Ethics of AI	

Table 3. IRIS solution risks and benefits

3.1.2.2 Session 2 – The impact on values

In this session the discussion objective was to identify the values embedded in and directly enabled by IRIS technology, while also trying to prefigurate potential tensions with the values prevalent among users, and/or stakeholders (e.g. social groups).

The exploration of values inherent in technology and their potential conflicts with meaningful societal values was prefaced by normative considerations concerning the values that should underpin an acceptable cybersecurity solution. Among the values presumed as implicit were security, cybersecurity and accountability. Additionally, participants acknowledged the significance of innovation. The underlying notion is that, in the eyes of the stakeholders, the viability and adoption of technology hinge on demonstrating that the solution introduces novel attributes while also streamlining human effort. The reduction of human effort is interpreted here from an economic standpoint. This perspective spurred a discourse on how such an endeavour might engender tensions and subsequently generate costs due to apprehensions among operators about the

prospect of being replaced. Engaging with the viewpoints of the participants revealed that this apprehension was primarily associated with the integration of AI within the IRIS solution. According to their vantage point, AI represents a substantial advancement in terms of operational efficiency. Nevertheless, a prevailing belief among them is that a cybersecurity solution entirely driven by AI remains a remote possibility - at least for the time being. This sentiment was underscored by specific reservations voiced in relation to the understanding of AI's role in generating alerts (e.g., threat detection) without autonomously executing subsequent actions. In tandem, they accentuated the intrinsic value of human creativity within incident response procedures, all the while accentuating concerns regarding the potential for occupational displacement within the sector.

A noteworthy progression of this discussion circled back to a facet underscored during the session on risks and benefits: that of aligning with AI regulations. From the perspective of the CISOs, despite the multifaceted considerations surrounding AI in cybersecurity solutions, a framework for AI compliance is imperative as a foundational element for system design. Indeed, many functionalities may prove advantageous for operators, but they could entail actions that run counter to legal and managerial dimensions. From this discussion, a tension among the smartness of innovation and trust in the process is observable among our conversation partners. Quite interestingly, from a researcher's point of view, this tension has not been consciously acknowledged in each participant analysis but emerged in a second moment from the content analysis of the discussion recordings.

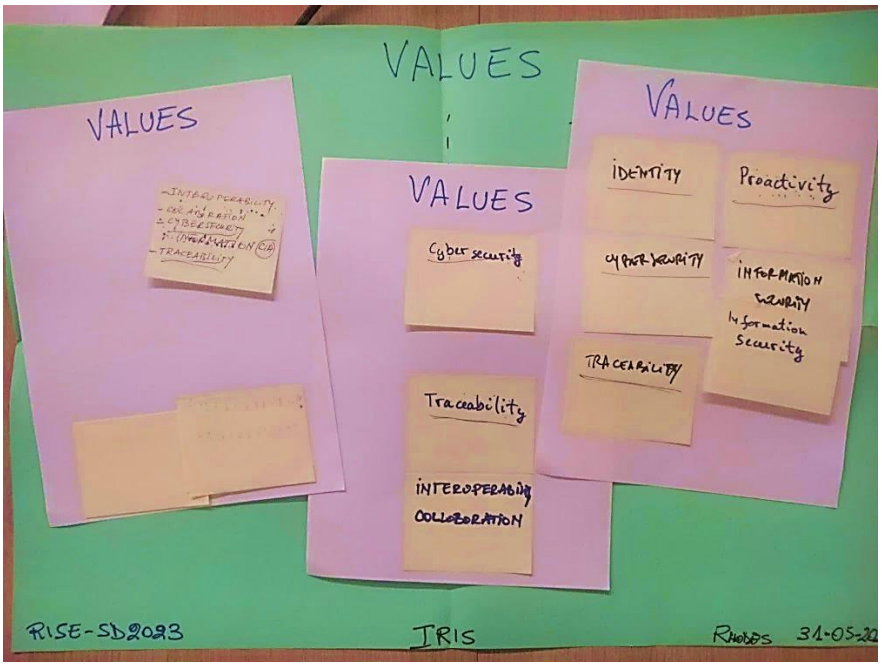

Figure 6 - Identification of values embedded in IRIS solution

The values recognised as embedded in and enabled by IRIS technology, are in general from the security cluster (cf. Figure 4), also implying **cybersecurity** as a specific kind of **security** – here recognised as involving **safety** too. Other values were identified as specific

to IRIS technology: **interoperability** and **proactivity** were recognised as IRIS-specific values enabled by the platform. Interoperability because it is conceived and designed as a platform to allow the **sharing** of cyber-attacks notice, as well as incident-response policies among different actors (e.g. EU National CERTs), thus also embedding proactivity in technology. Another value identified as important in IRIS comes from the **accountability** cluster and it is the **traceability** promoted by the DPA component demonstrated to the stakeholders. In fact, accountability gains relevance together with the awareness of the fact that governments or companies can potentially harm others (citizens, companies, public services...) when taking cybersecurity measures. There is also at least one other dimension: citizens and consumers are growingly dependent on companies and governments for the secure storage of their personal data (e.g. the banking sector, the public administration). This also require traceability of cybersecurity solutions which are intended to handle cyber-threats and cyber-attacks. DPA component developed for auditing by using blockchain technology guarantee the traceability of automatic and semi-automatic incident response. Smartness of innovation was also mentioned as a business value in IRIS research and development.

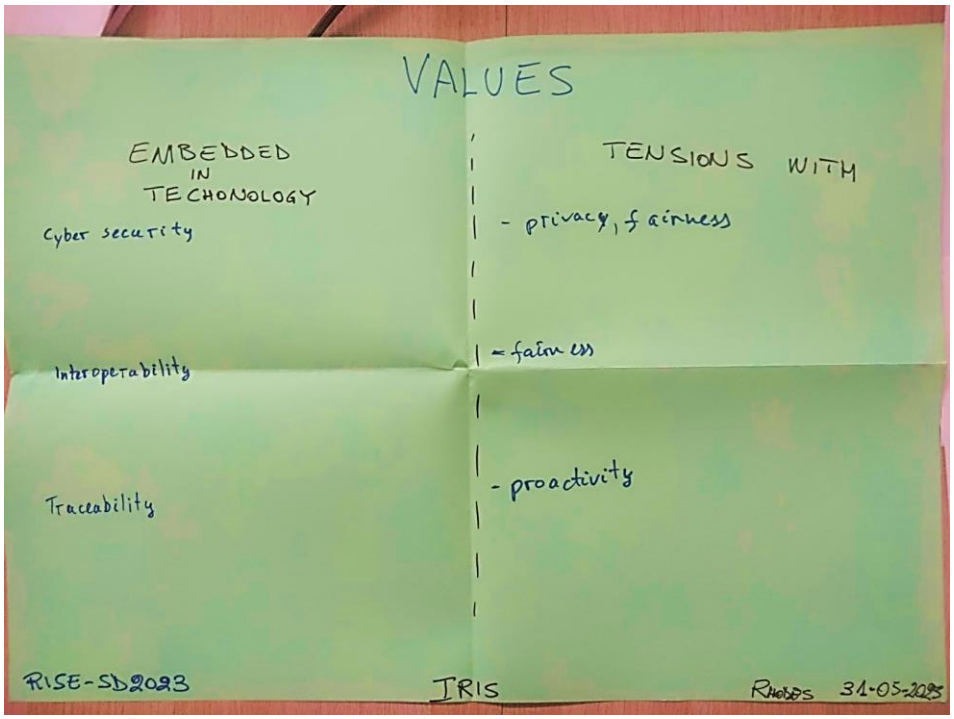

Figure 7 - Session 2: potential value tensions

Regarding potential conflict among IRIS solutions' embodied values and relevant societal values, tensions have been discussed among the following:

- Cybersecurity – Privacy & fairness
- Interoperability – Fairness (Privacy & Confidentiality)
- Traceability – Proactivity

Cybersecurity - Privacy & Fairness: Among IRIS solution objectives there is to enhance cybersecurity by allowing sharing critical information on threats and incident response policies across borders. However, this sharing can conflict with the values of privacy and fairness. Questions arise about the fairness of sharing information without knowing the extent of its impact on individuals' privacy and on data confidentiality (interestingly, during the activity we noticed an overlapping among the idea of privacy as an individual right and the property of confidentiality of data among tech experts). Balancing cybersecurity needs with safeguarding data confidentiality and individuals' privacy rights becomes a challenge.

Interoperability - Fairness: The data exchange between different cybersecurity actors to bolster cybersecurity can clash with concerns about fairness, according our discussion participants. Even though here the accent is on fairness, together with this value, privacy and confidentiality were mentioned as concerns. Sharing information, from their viewpoint, raises concerns about the type of data being shared and the level of confidentiality maintained. Achieving interoperability while ensuring that shared data remains private and confidential still presents a complex trade-off.

Traceability - Proactivity: The tension between traceability and proactivity arises from the need for timely response to cyber incidents. While established procedures provide a traceable framework for accountability, they might not always align with the need for swift and creative responses. Sometimes, the formal procedures may prove time-consuming or outdated. This point can be resumed by the need of balancing traceable processes with the urgency of proactive actions related to the potential tension between following established procedures for accountability and the need for creative, efficient responses that may challenge operators' accountability.

In conclusion, the intricate interplay of values such as cybersecurity, privacy, fairness, interoperability, traceability, and proactivity in an automated incident response platform highlights the need for careful navigation and balance among these factors to achieve effective cross-border collaboration while upholding essential ethical and legal principles.

3.2 Quantitative assessment

For the quantitative component of the research, the questionnaire is designed with the following rationale: questions based on a *likert scale* (1-5) and will have three response options, two aimed at investigating a certain content, and the other as a response check - sometimes in a negative form - to ascertain the respondent's attention, also allowing to weigh the answers given to the previous two questions. The questions presented to the external experts are presented in Annex A.

In order to participate in this survey, it is essential to have experienced a demonstration of the technology and to be familiar with well-defined use cases. Currently, the project has primarily adopted a narrative approach for the demonstrations due to the software not being fully developed at this stage. Consequently, the survey was distributed to a limited

group of 3 respondents through an in-person engagement. The subsequent analysis holds value not only as feedback on the demonstrations but also as a means to validate the survey methodology, providing insights into whether the planned questions effectively address the objectives of the methodology. It is important to note that due to the small number of respondents, the results do not possess statistical significance.

As thoroughly described in D2.4, the IRIS social acceptance methodology takes into account four human factors for the acceptance of technology, namely the User Experience, Value Impact, Perceived Trustworthiness and Social Disruptiveness, which are then divided into several factors, namely the Perceived Usefulness (PU), Perceived Ease of Use (PEU), Likeability (LK), Reliability (RL), Perceived Behaviour Control (PBC) and Human in the Loop (HiL), Capacity Enabling (CE), Transparency (TR), User Perceived Certainty (SC), Perceived Risks (PR), Institutional Trustworthiness (ITW) and Expected Systemic Change (ESC).

3.2.1 Quantitative assessment Results

3.2.1.1 Perceived Usefulness Results

The evaluation results regarding the Perceived Usefulness (PU) in the working sphere are depicted in Figure 8. As readily observed from Figure 8, IRIS is perceived as potentially useful in their working sphere, as the responses were either neutral or positive.

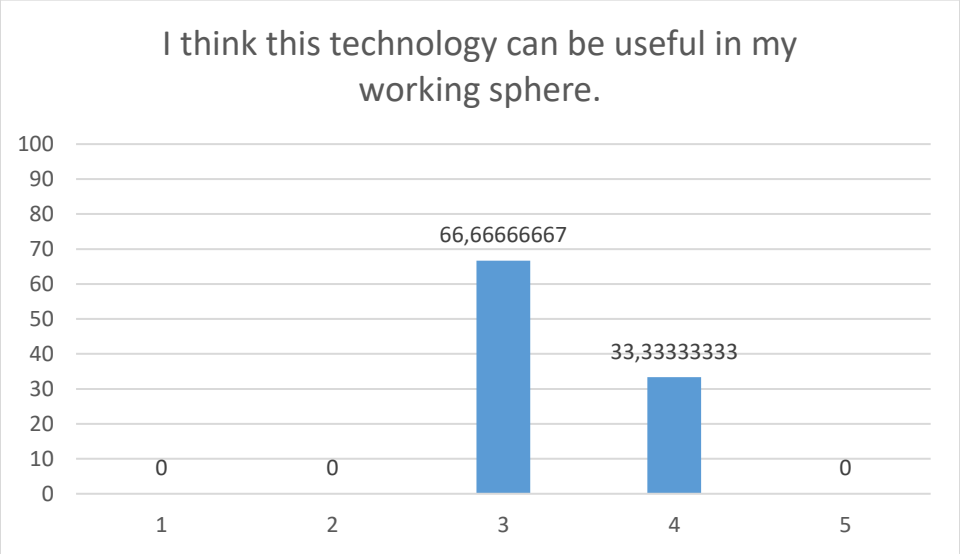

Figure 8. Perceived Usefulness in the working sphere.

The perceived usefulness of the IRIS technology in the daily life of the respondents is shown in Figure 9. As readily observed from it, the IRIS technology and tools is mostly considered neutral regarding its usefulness in daily lives of the respondents.

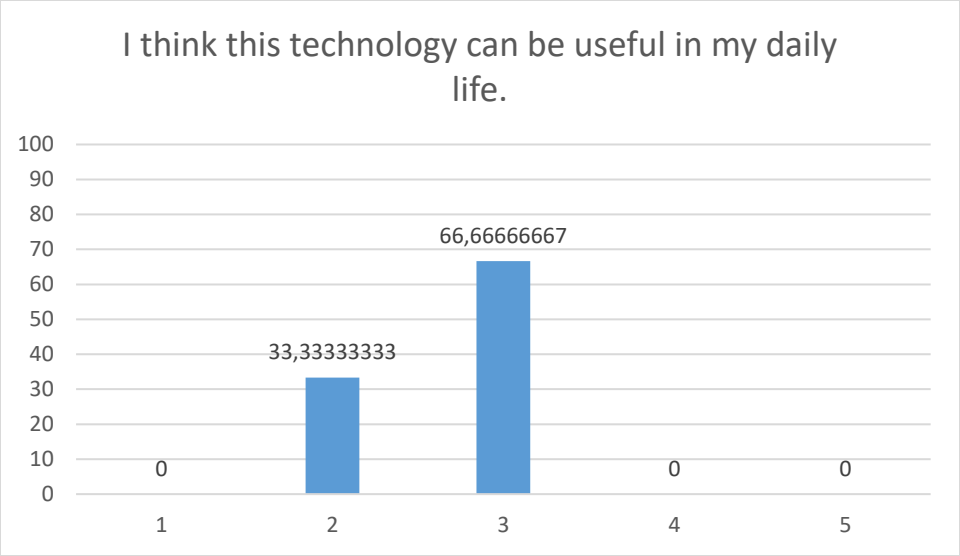

Figure 9. Perceived Usefulness in the daily life.

The results depicted in Figure 9 are in accordance with the results for the "trick question" results depicted in Figure 10. As readily observed from Figure 10, the majority of the respondents lean towards a neutral approach regarding the usefulness of the IRIS technology.

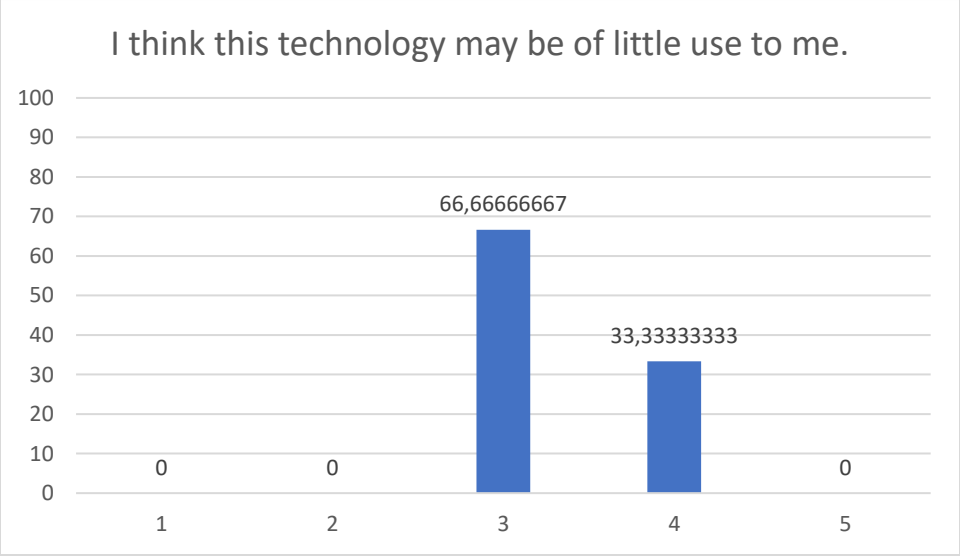

Figure 10. Perceived Uselessness of the IRIS technology

3.2.1.2 Perceived Ease of Use Results

The IRIS technology is very well perceived regarding its ease of use, as the majority of the responses agree that it is intuitive, as depicted in Figure 11.

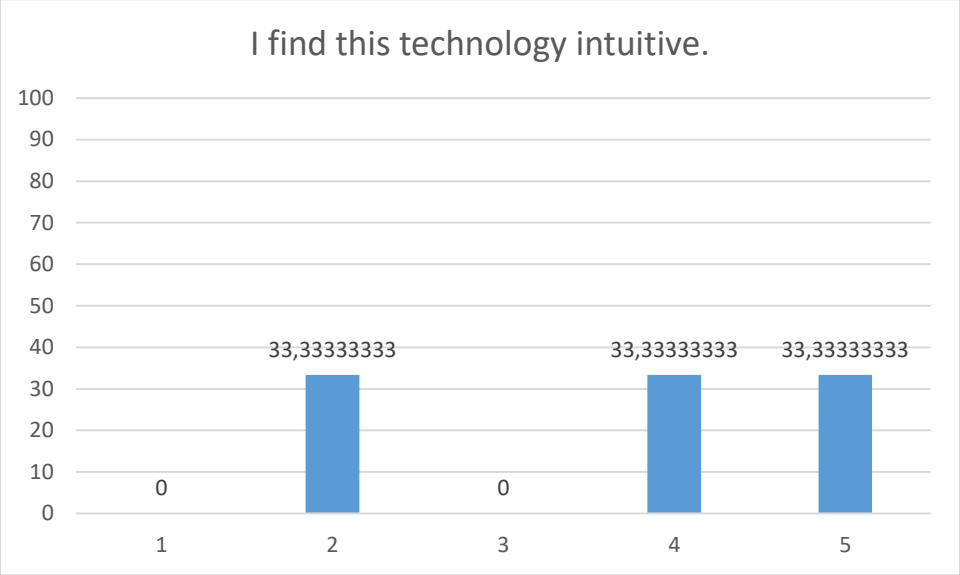

Figure 11. Perceived Ease of Use results.

Accordingly, the IRIS technology is perceived as easy to learn, as shown in Figure 12, since although the majority of the respondents again lean towards neutral answers, there is no negative feedback.

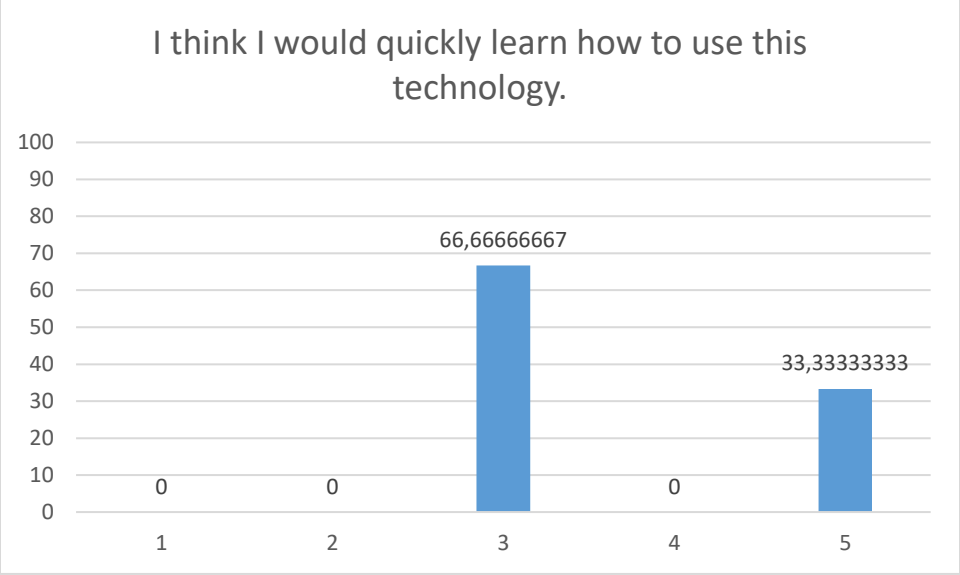

Figure 12. Perceived Ease of Use regarding the learning capabilities of IRIS.

The aforementioned results are further corroborated by the results depicted in Figure 13, from which it can be observed that 66.67% of the respondents strongly believe that it is not hard to understand how the IRIS technology is working.

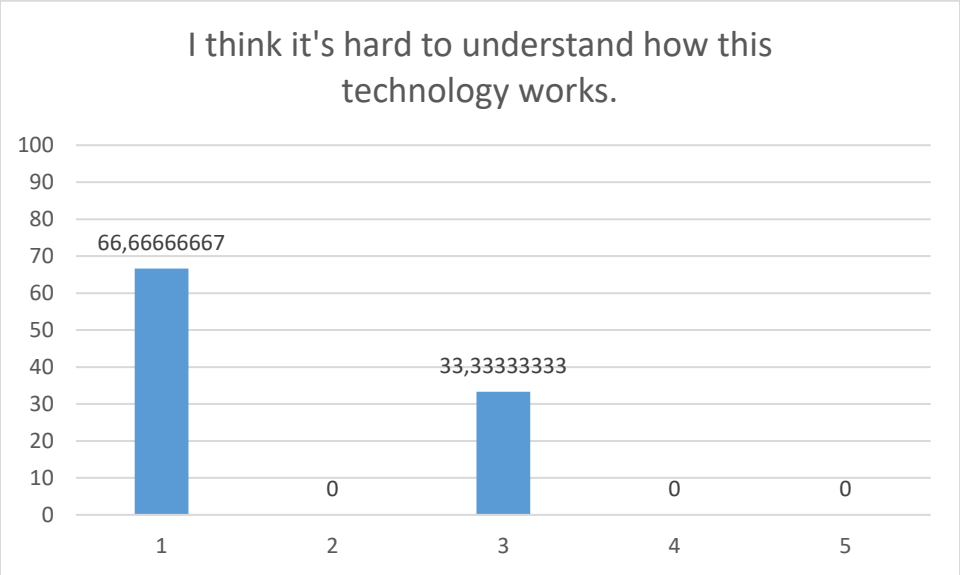

Figure 13. Perceived Ease of Use results ("trick question").

3.2.1.3 Likeability Results

The results regarding the potential adoption of the IRIS technology are depicted in Figure 14, from where it can be readily observed that all the respondents are neutral towards adopting the IRIS technology.

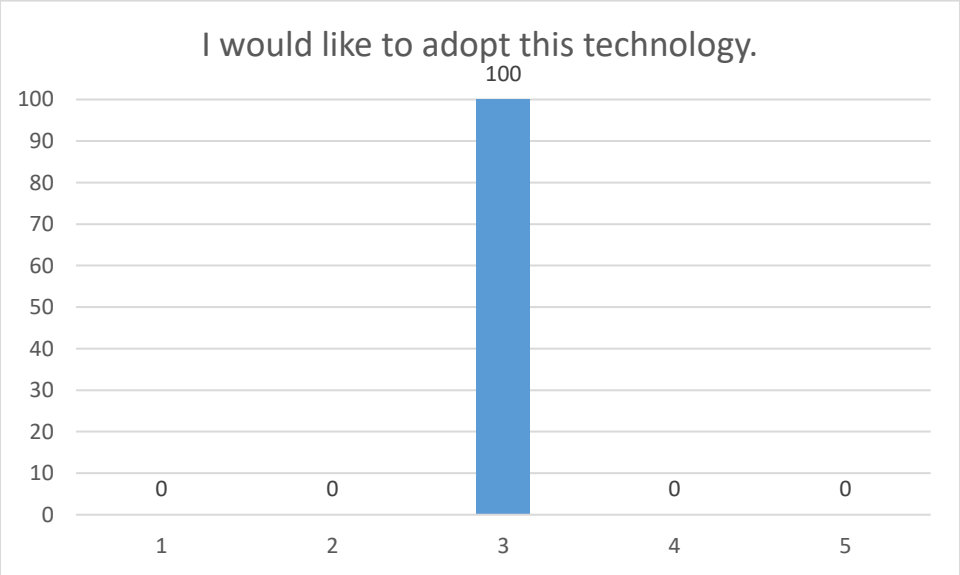

Figure 14. Potential adoption of the IRIS technology results.

Identical results regarding the likeability of the IRIS technology are depicted in Figure 15, thus indicating that the evaluators lean towards neutrality regarding the question if the IRIS technology is smart and nice.

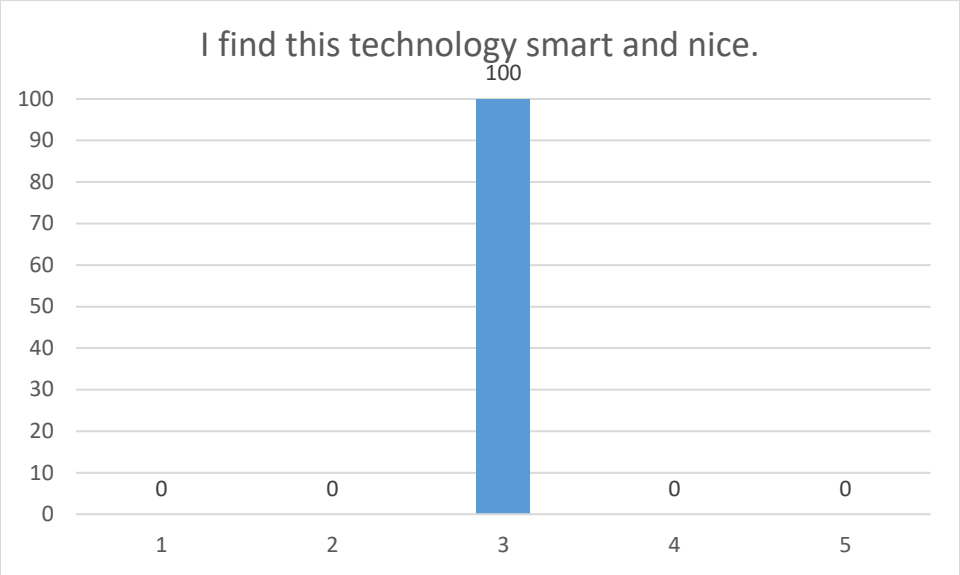

Figure 15. Likeability of the IRIS platforms evaluation results.

However, the evaluators strongly provided positive feedback regarding how pleasant the IRIS technology is, as shown in Figure 16, since they find the IRIS technology pleasant.

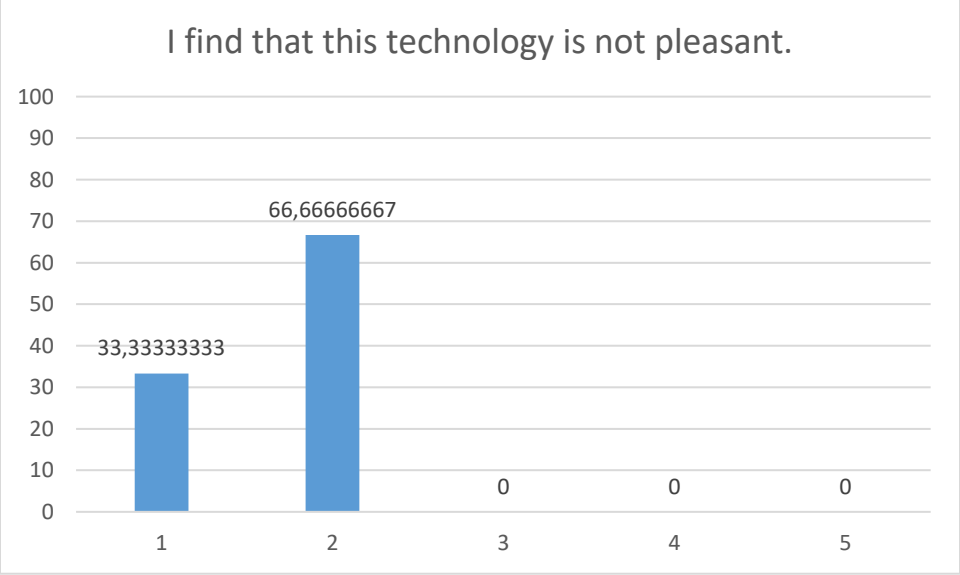

Figure 16. Likeability evaluation results ("trick question").

3.2.1.4 Reliability Results

Regarding the reliability evaluation, all respondents do not agree or disagree regarding the IRIS ability to work as it is supposed, as shown in Figure 17.

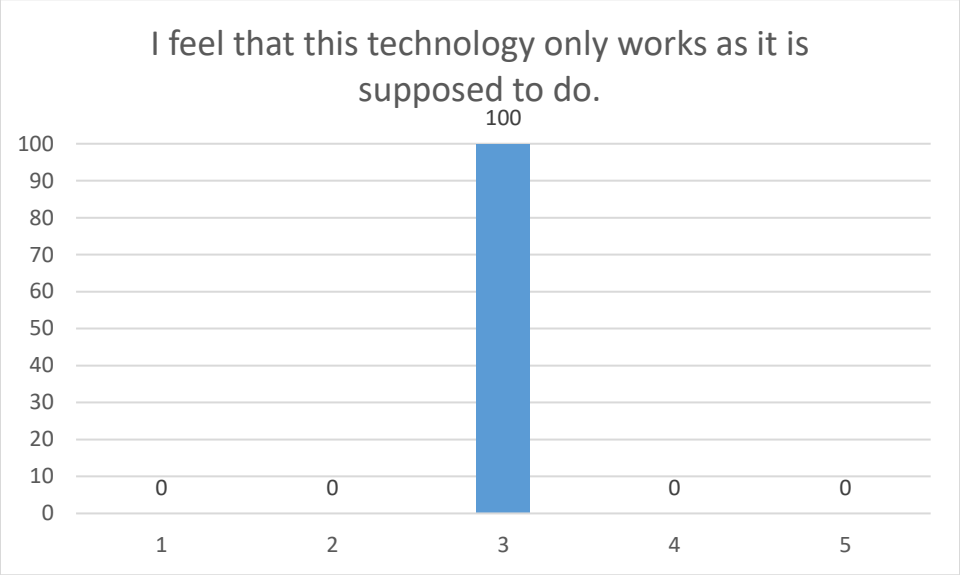

Figure 17. Reliability evaluation results.

Accordingly, as depicted in Figure 18, the respondents provided mostly neutral feedback regarding their perceived relying in the IRIS tools without worries.

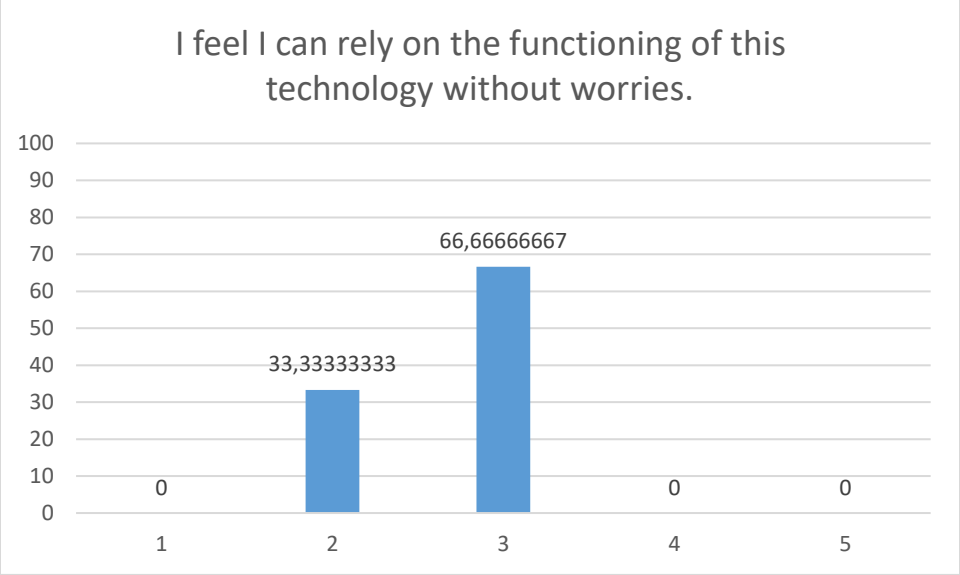

Figure 18. Reliability evaluation results.

The results depicted in Figure 17 and Figure 18 are in accordance with the results depicted in Figure 19, from where it can be observed that the evaluators do not believe that the employment of the IRIS tools gives them the sense of working randomly.

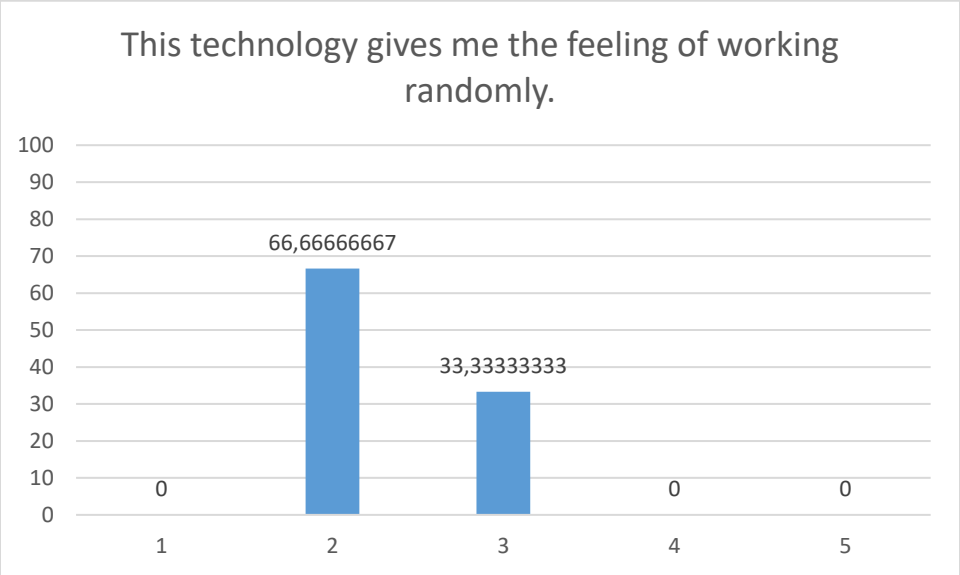

Figure 19. Reliability evaluation results ("trick question").

3.2.1.5 Perceived Behaviour Control and Human in the Loop results

Regarding the PBC and HiL evaluation results, as shown in Figure 20, the evaluators are indecisive regarding the fact that they fully control the IRIS technology.

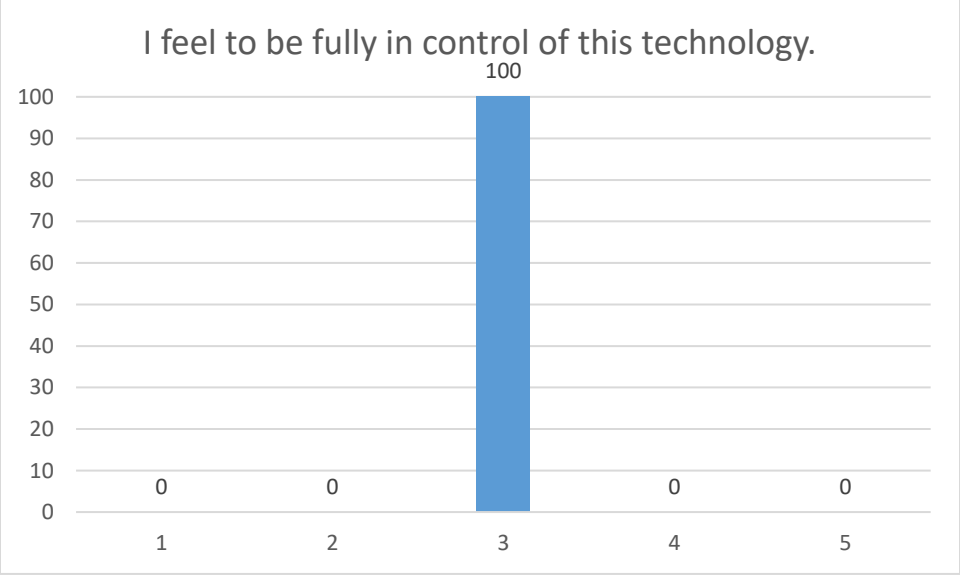

Figure 20. PBC evaluation results.

Accordingly, the evaluators are indecisive regarding their comfort in using the IRIS tools and technology, as shown from the results depicted in Figure 21.

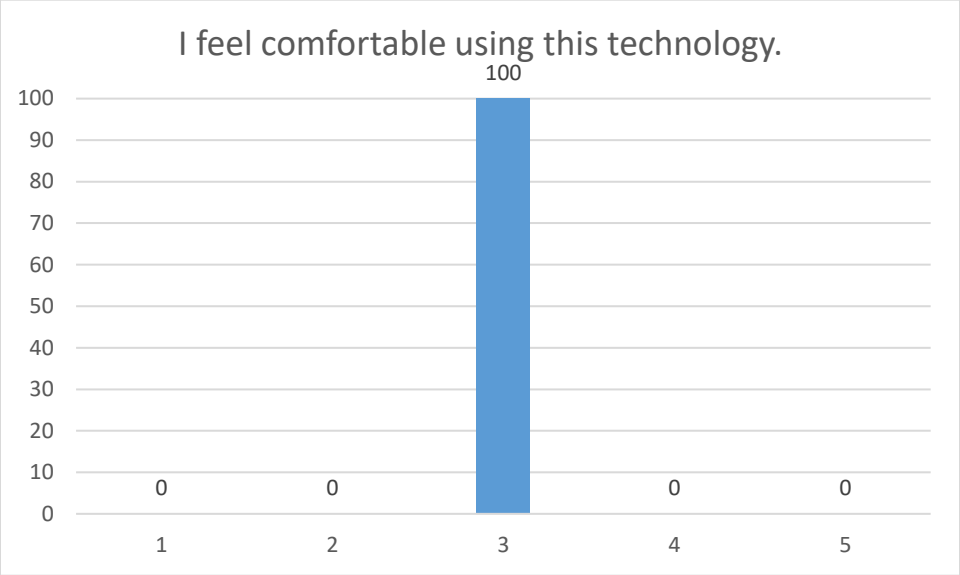

Figure 21. PBC evaluation results.

However, the aforementioned results are not in complete convergence with the results derived from Figure 22. As readily observed from Figure 22, one respondent is in strong disagreement with the feeling that the effects of the IRIS technology are beyond the evaluator's control.

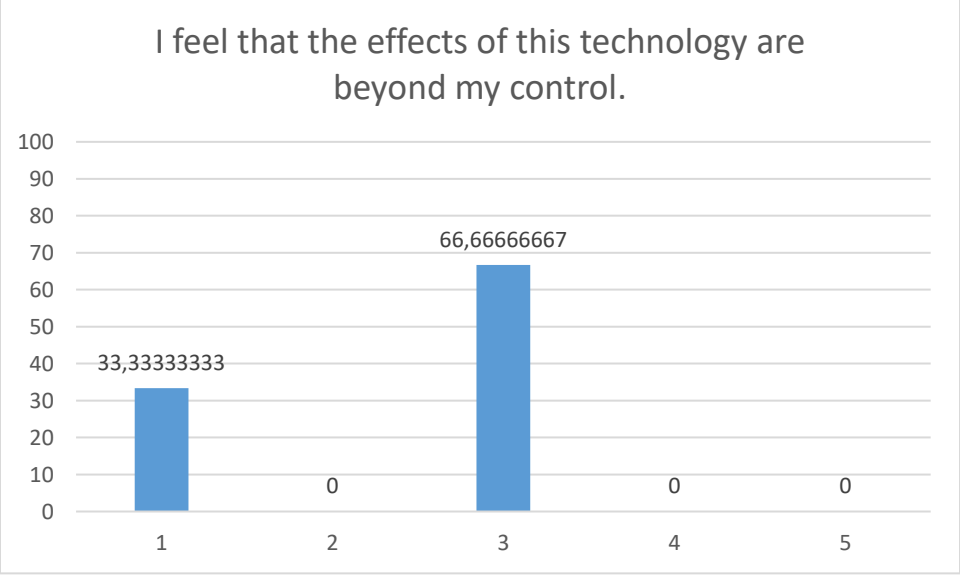

Figure 22. PBC evaluation results ("trick question") .

3.2.1.6 Capacity Enabling results

Regarding the Capacity Enabling Human Factor, the evaluators provided positive feedback regarding their perception of the ability of IRIS to provide them with a sense of ability and efficacy, as depicted in Figure 23.

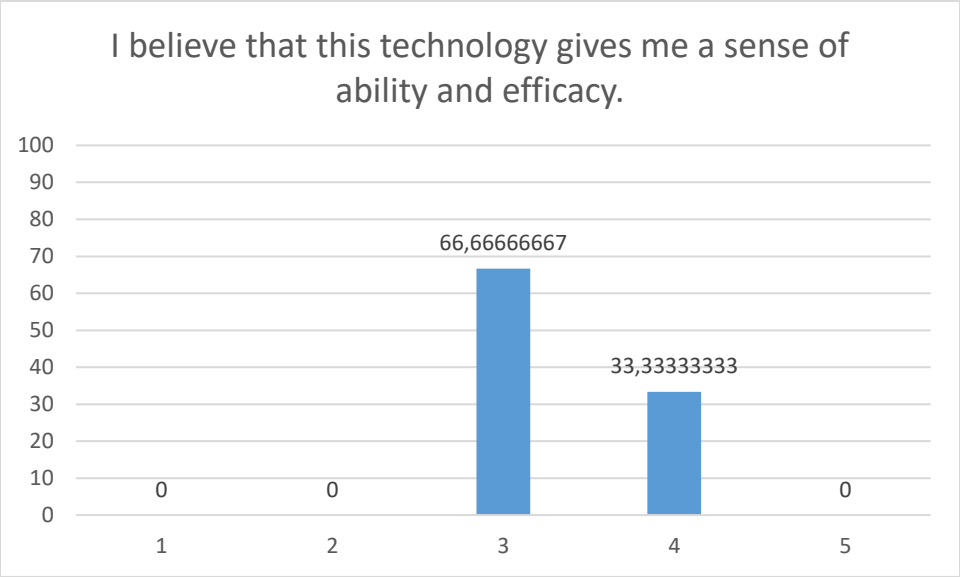

Figure 23. Capacity Enabling evaluation results.

Accordingly, the evaluators provided positive feedback in their perception of the fact that the employment of IRIS tools and technology will enable them to achieve their goals, as depicted in Figure 24.

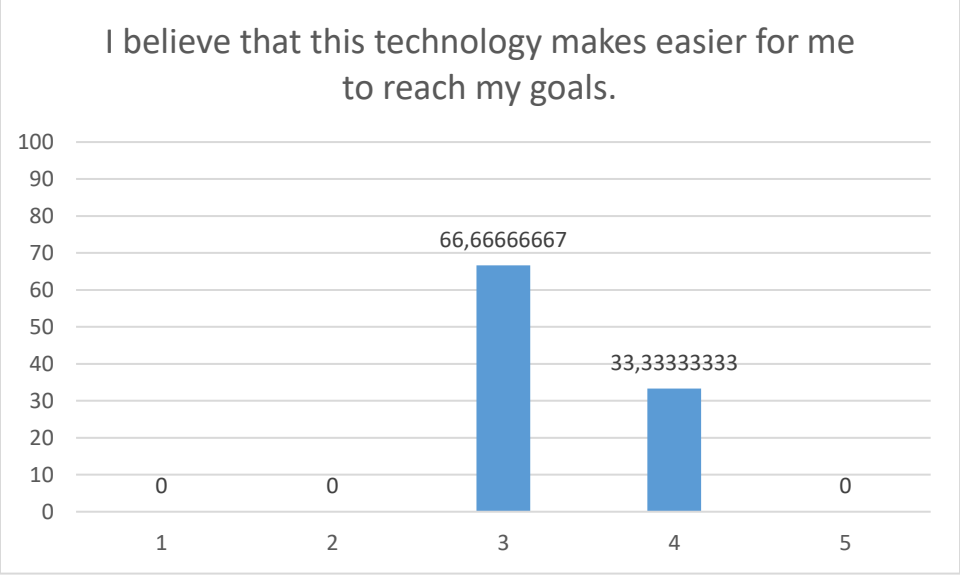

Figure 24. Capacity Enabling evaluation results.

The aforementioned results are further corroborated by the results depicted in Figure 25, as the evaluators believe that the IRIS technology helps them to improve their abilities.

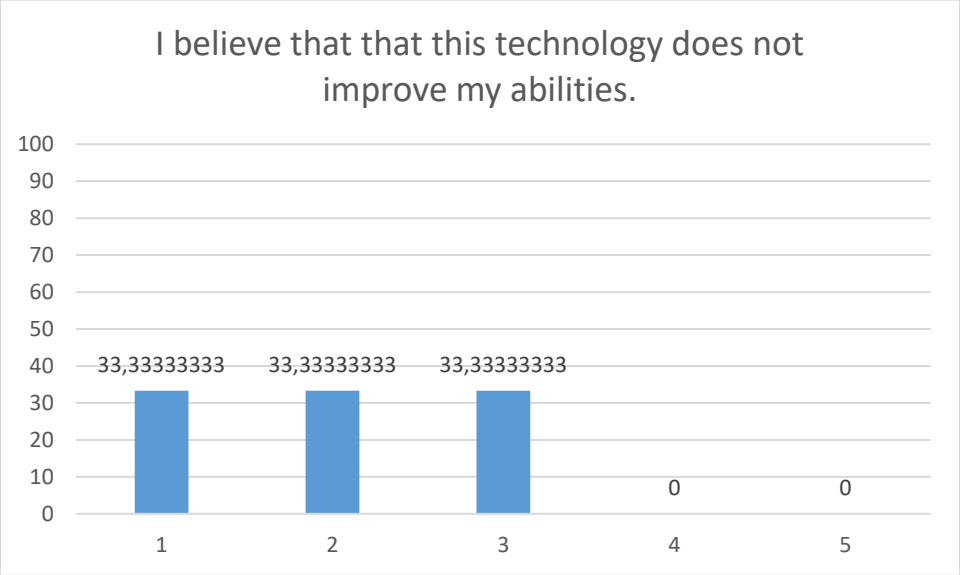

Figure 25. Capacity Enabling evaluation results ("trick question")

3.2.1.7 Transparency results

The IRIS technology achieves significant results regarding the Transparency Human Factor, as 66.67% of the evaluators believe that IRIS is understandable, as clearly shown in Figure 26.

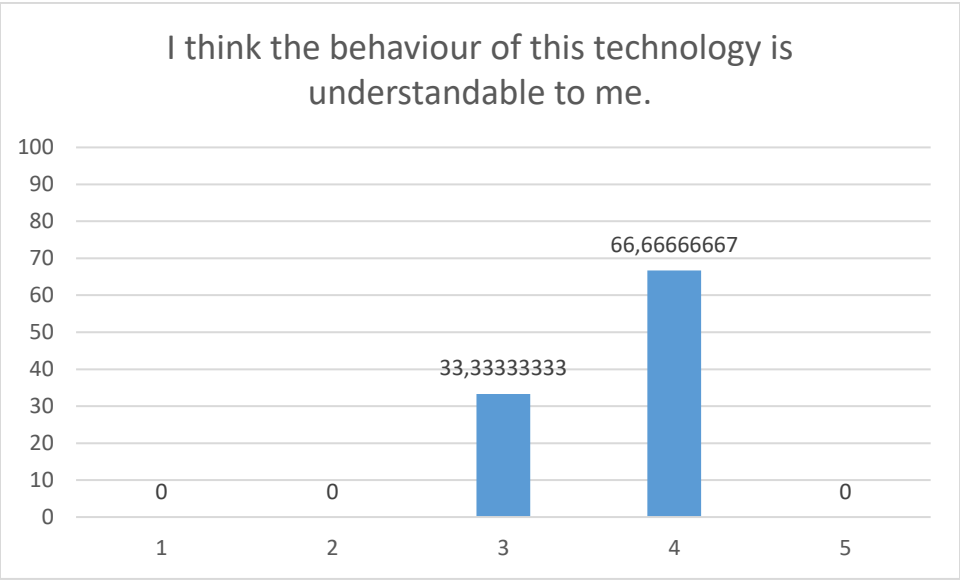

Figure 26. Transparency evaluation results.

This may not be attributed to the IRIS documentation, as the evaluators provided neutral feedback about the documentation of IRIS, as depicted in Figure 27. However, the fact that the documentation of the IRIS tools is not yet readily available to be presented to the evaluators, and thus this Human Factor must be revisited after the documentation is made available to the end users.

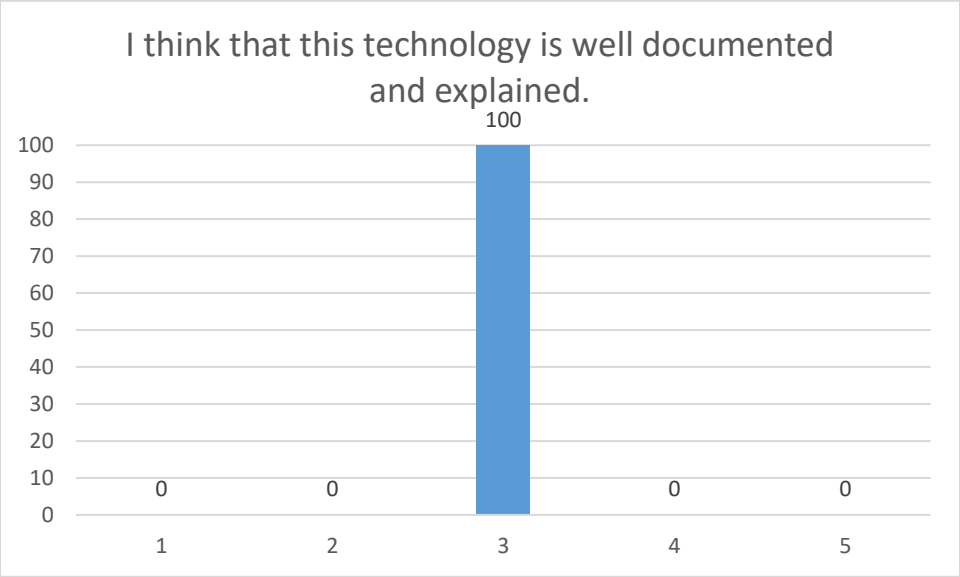

Figure 27. Transparency evaluation results.

However, the IRIS technology is not obscure to the evaluators, as readily observed from Figure 28. The results regarding the Transparency Human Factor are expected to improve after the completion of the documentation.

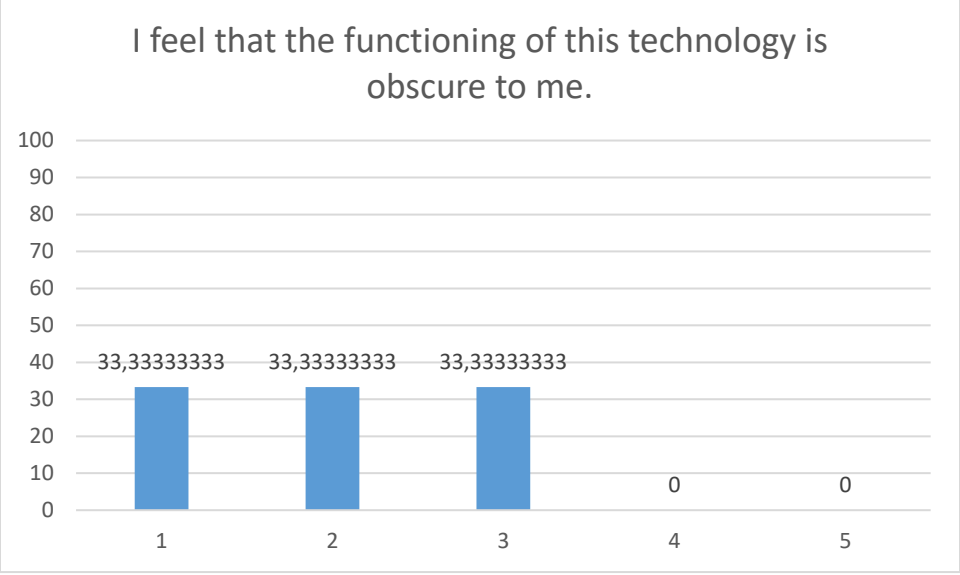

Figure 28. Transparency evaluation results ("trick question").

3.2.1.8 User Perceived Certainty results

The evaluation results regarding the User Perceived Certainty Human Factor, depicted in Figure 29, show that the respondents again provided neutral feedback regarding their understanding of the operational details of IRIS. This Human Factor is also expected to be improved after the documentation is made available to the end users.

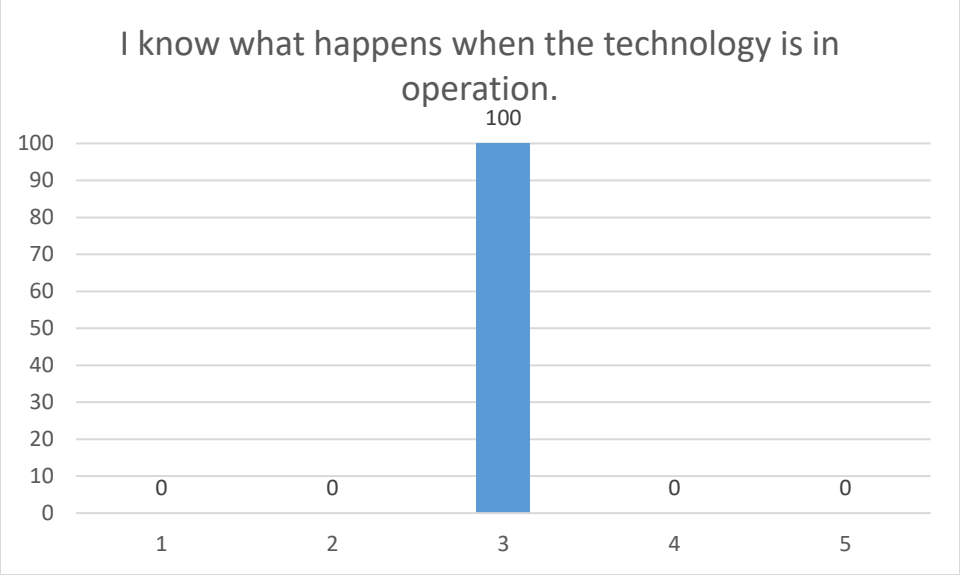

Figure 29. User Perceived Certainty evaluation results.

Similar results can be drawn by readily observing Figure 30, where the responses about the users' ability to predict the effects of the IRIS technology and tools both on themselves and their external environment is depicted.

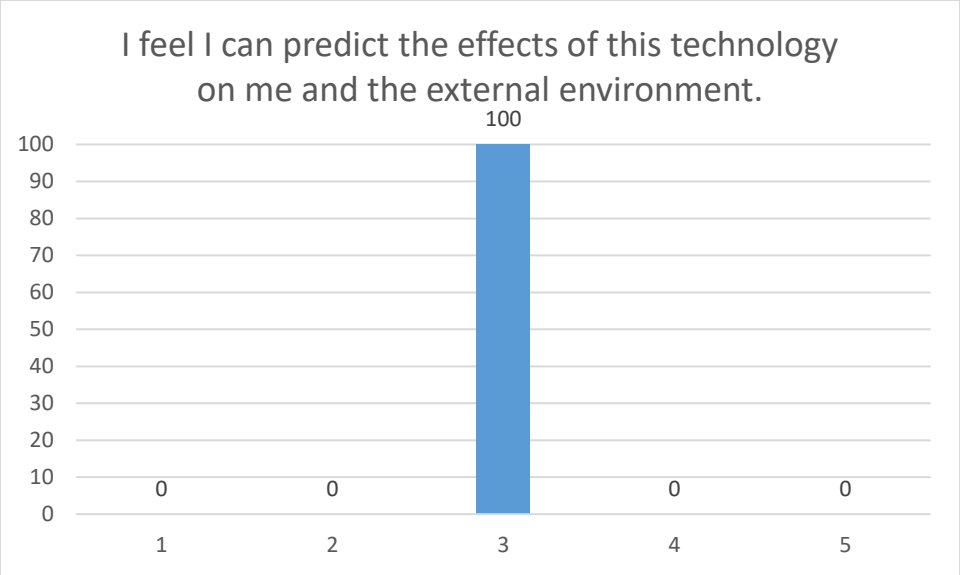

Figure 30. User Perceived Certainty evaluation results.

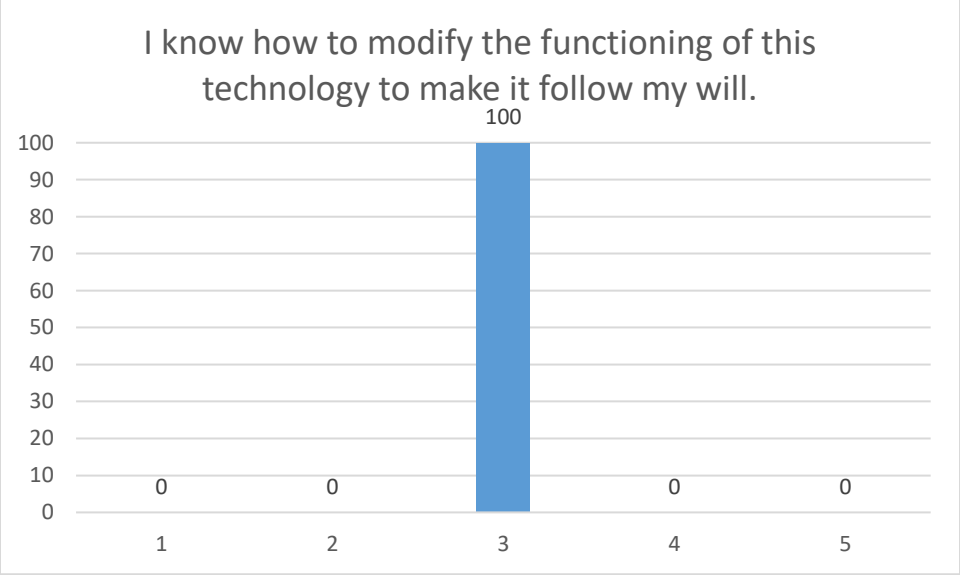

Figure 31. User Perceived Certainty evaluation results ("trick question").

The results presented in 3.2.1.8 are further corroborated by the results depicted in Figure 31, where the respondents are indecisive regarding their ability to modify the IRIS functioning to better serve their purposes. As already mentioned, IRIS will attain better results in this Human Factor, by providing more detailed documentation to the end users.

3.2.1.9 Perceived Risks results

The overall evaluation of the Perceived Risks Human Factor, depicted in Figure 32, is neutral regarding the risks associated with the employment of the IRIS technology and tools.

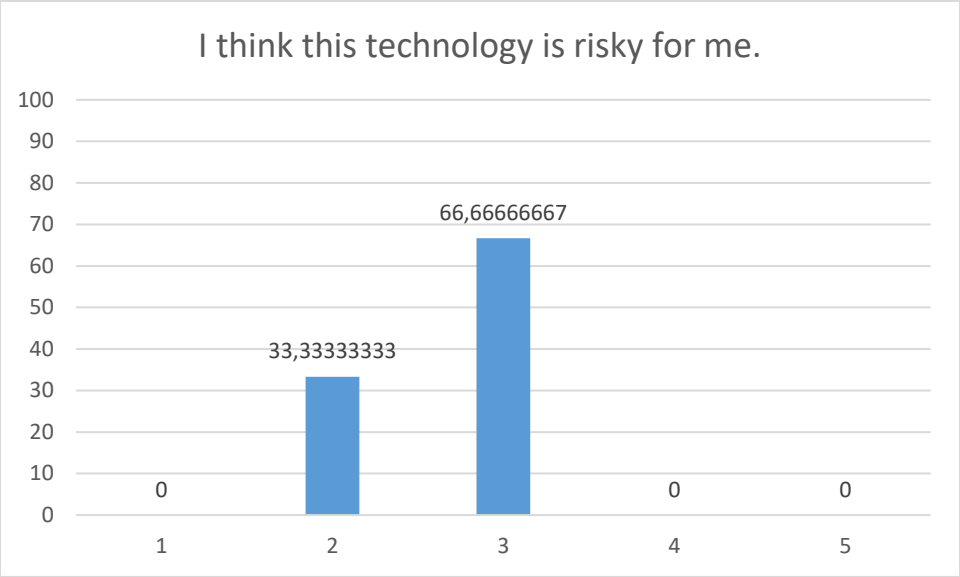

Figure 32. Perceived Risks evaluation results.

The results depicted in Figure 33 are in accordance with the perceived risks associated with the use of the IRIS technology, as the evaluators do not believe that the IRIS technology can harm someone or something.

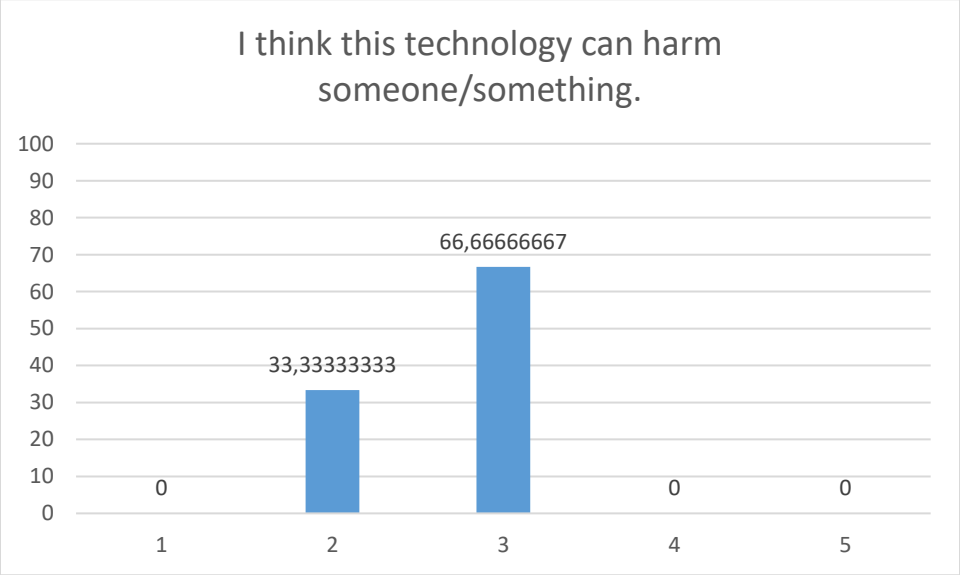

Figure 33. Perceived Risks evaluation results.

Finally, as readily observed from Figure 34, the evaluators are neutral regarding whether the impact associated with the use of IRIS tools is relevant to them.

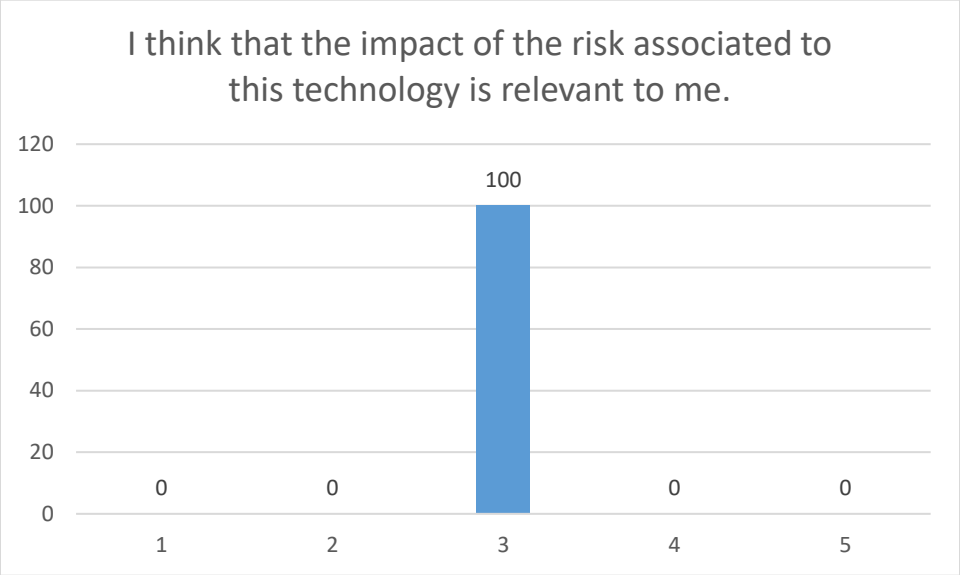

Figure 34. Perceived Risks evaluation results ("trick question")

3.2.1.10 Institutional Trustworthiness results

The Institutional Trustworthiness evaluation results are shown in Figure 35. As readily observed from it, the respondents provided neutral feedback regarding the perceived trust of the regulatory body in charge of controlling the IRIS technology.

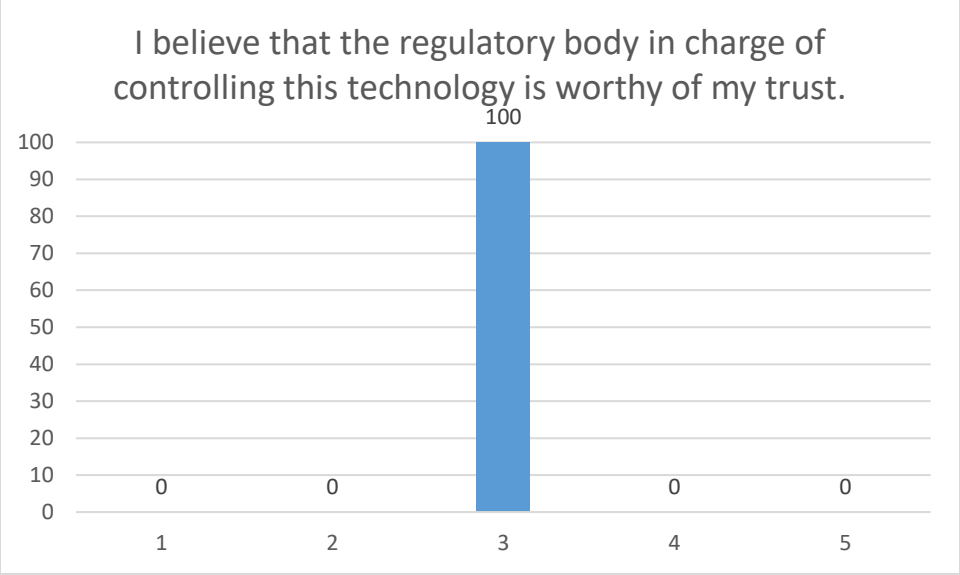

Figure 35. Institutional Trustworthiness evaluation results.

Similar results are yielded about the manufacturers of the IRIS technology, as readily observed from Figure 36.

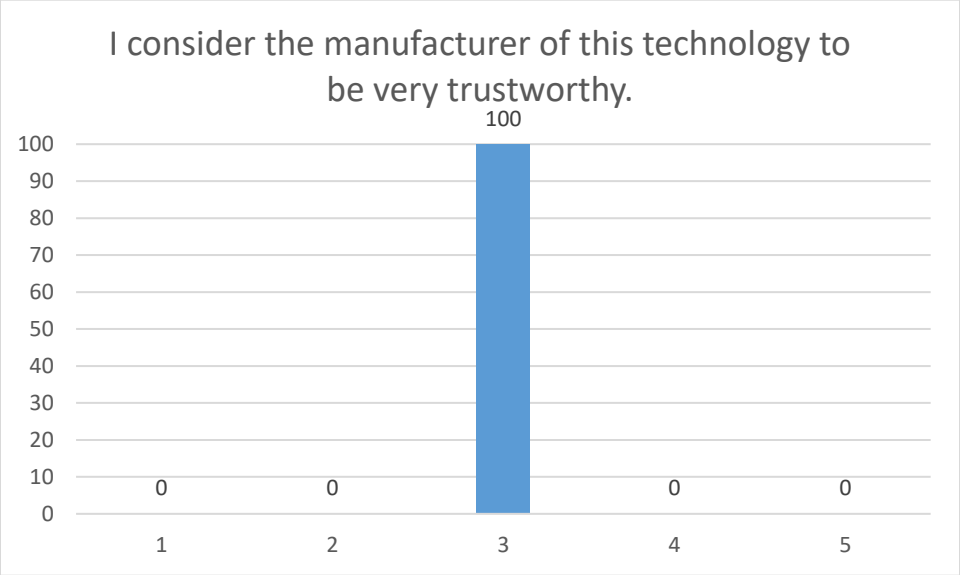

Figure 36. Institutional Trustworthiness evaluation results.

The aforementioned results are in partial agreement with the results depicted in Figure 37, from where it can be derived that the regulatory bodies and manufacturers of the IRIS technology are perceived as trustworthy by some respondents.

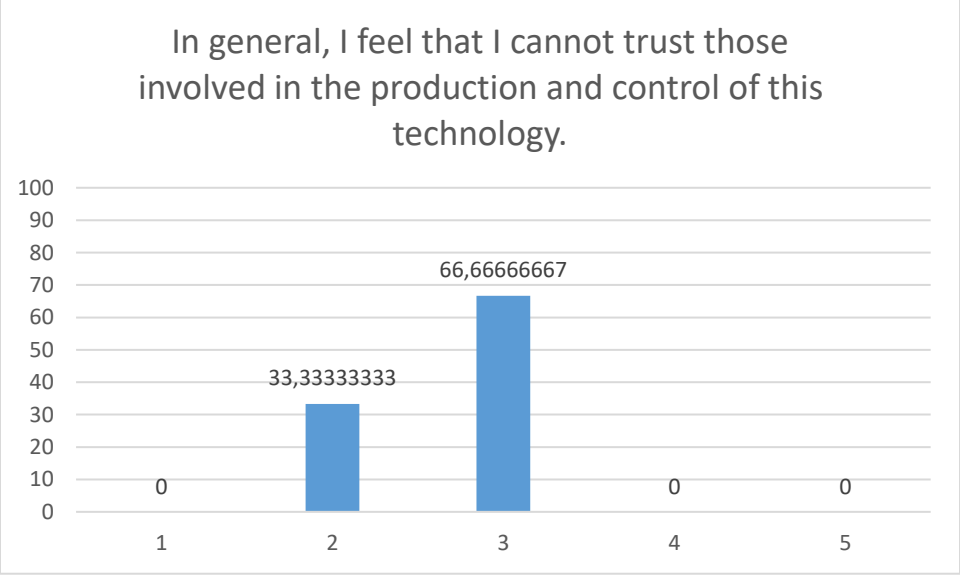

Figure 37. Institutional Trustworthiness evaluation results ("trick question").

3.2.1.11 Expected Systemic Change results

The evaluators provided neutral feedback regarding the fact of IRIS dealing with cyberthreats, as shown in Figure 38.

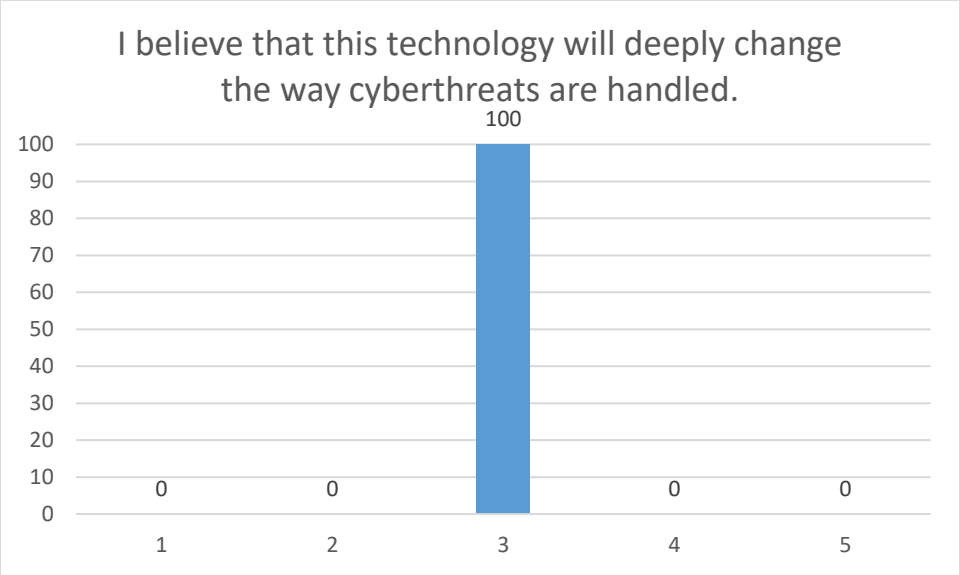

Figure 38. Expected Systemic Change evaluation results.

Accordingly, as depicted in Figure 39 and Figure 40 the evaluation results clearly indicate that the evaluators are mostly indecisive about the question regarding the IRIS technology and tools having a long-term impact in the cybersecurity landscape, but however some positive feedback was also provided.

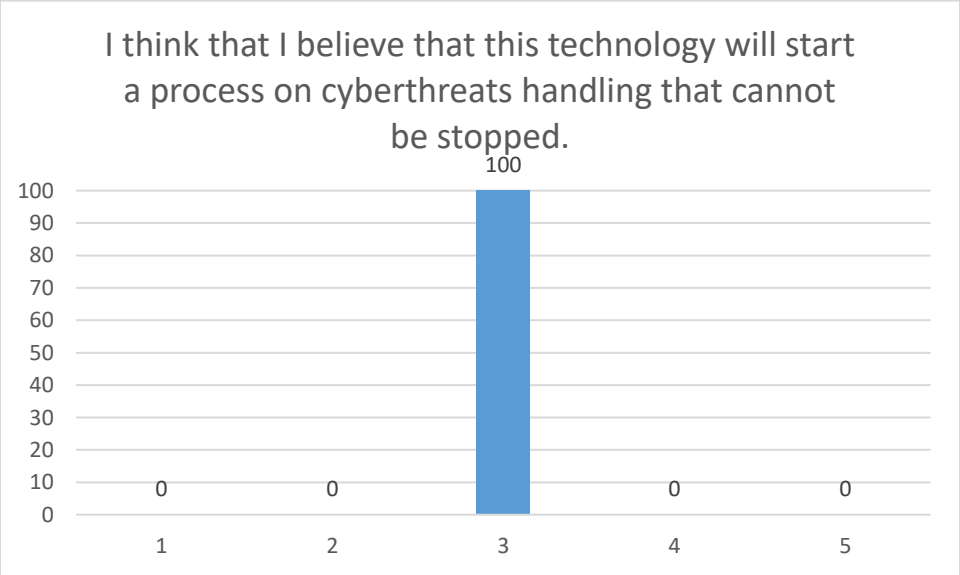

Figure 39. Expected Systemic Change evaluation results.

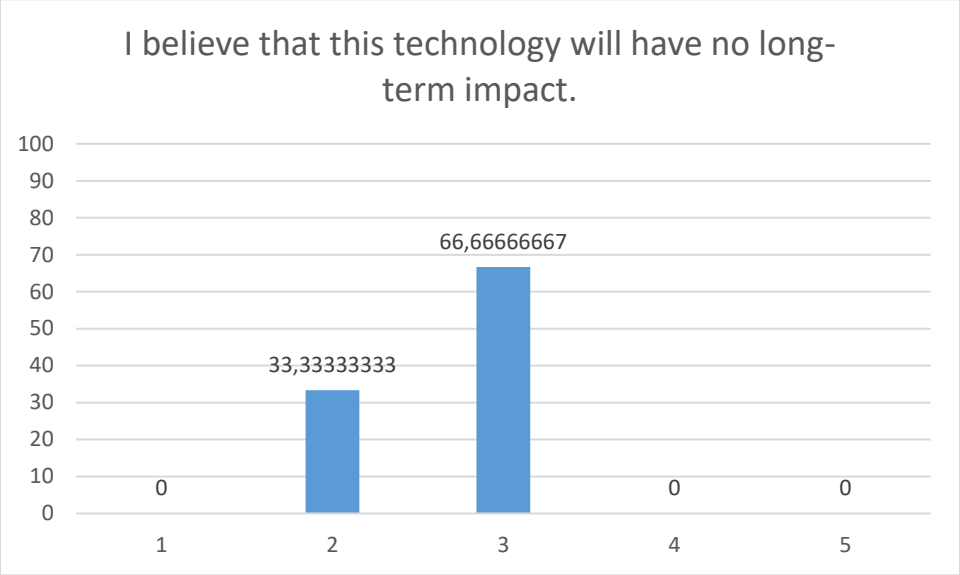

Figure 40. Systemic Change evaluation results ("trick question")

For a more comprehensive analysis the mean of the positive questions (excluding the "trick questions") are tabulated in Table 4.

Table 4. Evaluation results of the positive questions classified per Human Factor Area.

#	Human Factor	Mean and Median Values
HFA1 User Experience		
HFA1.1	**Perceived Usefulness (PU)**	**Mode:** 3 **Overall result**: Practitioners are indecisive about the IRIS perceived usefulness.
HFA1.2	**Perceived Ease of Use (PEU)**	**Mode:** 3, 5 **Overall result**: Practitioners perceive the IRIS technology and tools as easy to understand and use.
HFA1.3	**Likeability (LK)**	**Mode:** 3 **Overall result**: Practitioners are indecisive about the IRIS likeability.
HFA1.4	**Reliability (RL)**	**Mode:** 3 **Overall result**: Practitioners are indecisive about the IRIS reliability.

#	Human Factor	Mean and Median Values
HFA2 Value Impact		
HFA2.1	**Perceived Behaviour Control (PBC) - Human in the loop (HiL)**	**Mode:** 3 **Overall result**: Practitioners are indecisive about the IRIS Behaviour Control.
HFA2.2	**Capacity enabling (CE)**	**Mode**: 3 **Overall result**: Practitioners are indecisive whether their abilities are increased by the technology.
HFA3 Perceived Trustworthiness		
HFA3.1	**Transparency (TR)**	**Mode**: 4 **Overall result**: Practitioners are indecisive about the IRIS transparency.
HFA3.2	**User Perceived Certainty (SC)**	**Mode**: 3 **Overall result**: Practitioners are indecisive about the IRIS Perceived Certainty.
HFA3.3	**Perceived Risks (PR)**	**Mode:** 3 **Overall result**: Practitioners are indecisive about the IRIS Perceived Risks.
HFA3.4	**Institutional Trustworthiness (ITW)**	**Mode:** 3 **Overall result**: Practitioners are indecisive about the Institutional Trustworthiness associated with the IRIS tools and technology.
HFA4 Social Disruptiveness		
HFA4.1	**Expected systemic change (ESC)**	**Mode:** 3 **Overall result**: Practitioners are indecisive about the Expected Systemic Change associated with the IRIS tools and technology.

4 LESSONS LEARNT AND FEEDBACK

As already stated, the main Human Factor Areas that are taken into consideration are User Experience, Value Impact, Perceived Trustworthiness and Social Disruptiveness, which are then further analysed by taking into account several Human Factors, namely Perceived Usefulness (PU), Perceived Ease of Use (PEU), Likeability (LK), Reliability (RL), Perceived Behaviour Control (PBC) - Human in the loop (HiL), Capacity enabling (CE), Transparency (TR), User Perceived Certainty (SC), Perceived Risks (PR), Institutional Trustworthiness (ITW) and Expected systemic change (ESC).

The evaluation and assessment results clearly show that the practitioners provided mostly neutral feedback in almost all the aforementioned Human Factors.

However, if these results are read together with the qualitative data resulting from observations drawn during the focus group, some of the answers given to the survey may become more interpretable and useful for future work. For example, during the focus group some participants asked clarifications on the software and on details regarding the use of AI. Some of these questions remained unanswered because of the technology maturity level and because of technical partners not being present at the session. Hence, some neutral answers (score 3) could be related to this, to the level of clarity of demonstration also linked to the technology maturity level.

Additionally, during the discussion, one of the topics that emerged pertained to the ethics of AI and, more broadly, the trustworthiness of AI in incident response. Notably, one respondent expressed scepticism about the overall reliability of AI (not just in IRIS). This individual displayed a keen interest in gaining a deeper understanding of AI's role within IRIS, seeking additional information to inform their assessment. Although the demonstration session at this stage did not delve extensively into AI, it may prove beneficial to incorporate a more comprehensive exploration of AI's role in future demonstration sessions. This research phase also offers some useful experience regarding the research process itself and the method.

In relation to the survey, it would have been advantageous to present the questions to the respondents before the demonstrations. This approach could have allowed them to seek any necessary clarifications, enabling more informed and accurate responses to the survey.

Furthermore, during this workshop, the survey was completed following the focus group discussions on UX and the factors impacting value. In this regard, it might have been beneficial to conduct the survey *beforehand*, so as to allow a quick group discussion of results and *then* engage in the focus group. This approach, particularly, could have enhanced the validation of the questionnaire.

Regarding the survey questions, in more than one section of the survey, some "trick questions" remained unanswered. It may be possible that these questions confused the

respondent. Either these questions should be better formulated, or their function should be explained to the respondents.

The focus group took place in a spacious conference room and was attended by the project coordinator and other members of the IRIS consortium. Although the focus group was allowed sufficient privacy for the discussion, it is beneficial that focus groups are conducted in a dedicated, private room with the presence of researchers. This allows for less distractions, intrusions, better quality of recordings and an atmosphere more suitable to research practice.

Moreover, the participants to this focus group all were Romanian native speakers, while the researchers that conducted the focus group do not understand Romanian. Although the focus group was conducted in English, during some focus group activities, the interlocutors exchanged ideas among them while speaking in Romanian. This is pretty normal, in such circumstances, but from a behaviour point of view it created two distinct areas of interaction among the group. One private, among the participants, and one more public, with the researchers, in English. Some very useful qualitative data in focus groups come exactly from the observation of participants while they act and talk spontaneously, rather than when they know that they are observed and listened to or they speak in public. So, in future workshops strategies to have the participants speaking in English the whole time should be implemented.

Other lessons that may be drawn from the evaluation results presented in the previous sections are outlined below:

- Demonstrations of technical artifacts should be better organised, with a stronger focus on effectively communicating the project's outputs and the benefits associated with their use to the final users
- In future phases of the project, it is advisable to provide more favourable conditions for the organisation of focus groups (regarding the space, time allowed).
- A more successful and optimistic dissemination and communication campaign may be beneficial. To this end, the IRIS consortium is planning a Stakeholders and Industrial Workshop, which will enable the engagement of several stakeholders. Furthermore, the IRIS tools will be in a more mature state, thus enabling the stakeholders to attain better understanding of the IRIS tools.
- The IRIS tools developers must provide detailed information and documentation about their tool’s usage, in order to improve the way the practitioners perceive the risks associated with the usage of the IRIS technology.

5 CONCLUSIONS

The present deliverable presents and describes the evaluation and assessment results of the Social Acceptance of Technology of the IRIS platform and tools, aiming at understanding the potential barriers hindering the acceptance of the IRIS platform and tools.

The evaluation of the human factors presented in the present deliverable are crucial in relation to AI development as well as in relation to cybersecurity and information sharing.

The present deliverable presents and demonstrates how the methodological framework described in D2.4 was finalized and employed to evaluate and assess the SAT of the IRIS platform and tools.

The application of the aforementioned methodological framework enables to conduct a quantitative and qualitative evaluation and analysis to extract meaningful lessons and feedback which will then be used throughout the project's lifetime to improve the potential adoption of the IRIS technology by relevant stakeholders.

In terms of social acceptance, even if the results of the quantitative component are not significant in statistical terms, they were valuable to validate the methodology and make improvements in the questionnaire.

Regarding the qualitative component, the feedback on benefits and risks was deemed as useful by the consortium and informed productive discussions about requirements and design. The session on value impact showed there are no significant conflicts of values specific to IRIS. In fact, most of the discussion referred to concerns that regard cybersecurity in itself. It is important to stress that this is not the final evaluation and the results of this first stage will be deepened in the next iteration of Social Acceptance of Technology assessment to contribute to pilots' evaluation (WP7 and WP8), to refine the IRIS communication plan, and contribute to final recommendations.

6 REFERENCES

[1] Christen, M., Gordijn, B., & Loi, M. (Eds.). (2020). *The Ethics of Cybersecurity*. Springer Nature. https://doi.org/10.1007/978-3-030-29053-5

[2] Flanagan, M., Howe, D. C., & Nissenbaum, H. (2008). Embodying Values in Technology: Theory and Practice. In J. Van Den Hoven & J. Weckert (Eds.), *Information Technology and Moral Philosophy* (1st ed., pp. 322–353). Cambridge University Press. https://doi.org/10.1017/CBO9780511498725.017

[3] van de Poel, I. (2020a). Core Values and Value Conflicts in Cybersecurity: Beyond Privacy Versus Security. In M. Christen, B. Gordijn, & M. Loi (Eds.), *The Ethics of Cybersecurity* (pp. 45–71). Springer International Publishing. https://doi.org/10.1007/978-3-030-29053-5_3

[4] van de Poel, I. (2020b). Embedding Values in Artificial Intelligence (AI) Systems. *Minds and Machines*, *30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

ANNEX A. QUESTIONNAIRE PROVIDED TO THE EXTERNAL EXPERTS

- Is your organisation a CERT or a CSIRT?
- Does your organisation support MeliCERTes platform?
- I think this technology can be useful in my working sphere.
- I think this technology can be useful in my daily life.
- I think this technology may be of little use to me.
- I find this technology intuitive.
- I think I would quickly learn how to use this technology.
- I think it's hard to understand how this technology works.
- I would like to adopt this technology.
- I find this technology smart and nice.
- I find that this technology is not pleasant.
- I feel that this technology only works as it is supposed to do.
- I feel I can rely on the functioning of this technology without worries.
- This technology gives me the feeling of working randomly.
- I feel to be fully in control of this technology.
- I feel comfortable using this technology.
- I feel that the effects of this technology are beyond my control.
- I believe that this technology gives me a sense of ability and efficacy.
- I believe that this technology makes easier for me to reach my goals.
- I believe that that this technology does not improve my abilities.
- I think the behaviour of this technology is understandable to me.
- I think that this technology is well documented and explained.
- I feel that the functioning of this technology is obscure to me.
- I know what happens when the technology is in operation.
- I feel I can predict the effects of this technology on me and the external environment.
- I know how to modify the functioning of this technology to make it follow my will.

- I think this technology is risky for me.
- I think this technology can harm someone/something.
- I think that the impact of the risk associated to this technology is relevant to me.
- I believe that the regulatory body in charge of controlling this technology is worthy of my trust.
- I consider the manufacturer of this technology to be very trustworthy.
- In general, I feel that I cannot trust those involved in the production and control of this technology.

- I believe that this technology will deeply change the way cyberthreats are handled.

- I think that I believe that this technology will start a process on cyberthreats handling that cannot be stopped.
- I believe that this technology will have no long-term impact.
- Do you have any comments on previous questions/answers or suggestions about this questionnaire?