TU Delft

Delft University of Technology

Document Version
Final published version

Citation (APA)
van Gerwen, S., Constantino Torres, J. E., Roothaert, R., Weerheijm, B., Wagner, B., Pavlin, G., Klievink, B., Schlobach, S., Tuma, K., & Massacci, F. (2024). To Know What You Do Not Know: Challenges for Explainable AI for Security and Threat Intelligence. In T. Sipola, J. Alatalo, M. Wolfmayr, & T. Kokkonen (Eds.), *Artificial Intelligence for Security : Enhancing Protection in a Changing World* (pp. 55-83). Springer. https://doi.org/10.1007/978-3-031-57452-8_4

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

To Know What You Do Not Know: Challenges for Explainable AI for Security and Threat Intelligence

Sarah van Gerwen, Jorge Constantino, Ritten Roothaert, Brecht Weerheijm, Ben Wagner, Gregor Pavlin, Bram Klievink, Stefan Schlobach, Katja Tuma, and Fabio Massacci

1 Introduction

Threat intelligence (TI) builds upon many, sometimes unknown or unreliable sources and must operate under operational and legal constraints that cannot be interpreted by a single automated system. In a threat intelligence hybrid workflow (TIHW), human analysts and machines powered by artificial intelligence (AI) cooperate [12, 92]. Analysts routinely assemble findings derived from data generated by machines or assembled by other human analysts. These findings must often be assembled from data that analysts are not able to share or even have access to. Yet, TI must also be actionable to be useful for planning an intervention (such as apprehension of suspect of a cyberattack). Actionable information is relevant, timely, accurate, complete, and ingestible [79]. These properties are difficult to assert when the data itself cannot be accessed and/or when all the sources cannot be trusted.

S. van Gerwen (✉) · R. Roothaert · S. Schlobach · K. Tuma
Vrije Universiteit, Amsterdam, Netherlands
e-mail: s.a.m.van.gerwen@vu.nl

J. Constantino · B. Wagner
Delft University of Technology, Delft, Netherlands

B. Weerheijm · B. Klievink
Leiden University the Hague, Den Haag, Netherlands

G. Pavlin
Thales Research and Technology, Delft, Netherlands

F. Massacci
Vrije Universiteit, Amsterdam, Netherlands

University of Trento, Trento, Italy
e-mail: fabio.massacci@ieee.org

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

T. Sipola et al. (eds.), *Artificial Intelligence for Security*,
https://doi.org/10.1007/978-3-031-57452-8_4

Consequently, at each step of TIHW, the analyst must revise this limited and uncertain information and recommend actions. For example, they must choose which explanation among many is more likely or ask for additional yet proportional investigations. Further down the line of the TIHW, part of the gathered intelligence may be used to determine the proportionality of interventions, an investigation conducted by oversight bodies [57]. For example, in the Netherlands, the Review Committee on the Intelligence and Security Services (CTIVD) investigates the lawfulness of conduct by the General Intelligence and Security Service (AIVD) and the Military Intelligence and Security Service (MIVD) and in 2022 reported that in case of cable interception (report no. 75), the duty of care had been insufficiently implemented.[1]

Yet, intelligence sharing remains challenging due to fear of negative publicity, legal constraints, quality issues, and prevalence of other uncertainties [81, 92] despite decades of research [104]. In addition, AI-powered solutions and human experts will always have biases [2, 54, 106] or imperfect models [78] which may further contribute to the overall uncertainty of the gathered intelligence and assembled findings.

The goal of this chapter is to discuss the emerging socio-technical implications and technical challenges in the formalization and quantification of uncertainty within threat intelligence. We will start with a description of the situation at hand (cf. Sect. 2), discussing of related work (cf. Sect. 3) in the threat intelligence environment. Thereafter, we will discuss socio-technical challenges within the legal, societal, and organizational field (cf. Sect. 4). Afterward, in Sect. 5, the technical challenges with regard to the formalization and empirical evaluation in the TIHW are presented. Finally, we close the chapter with an overview of the bigger picture.

2 The Problem of Threat Intelligence

Threat intelligence relies on analysts to bring information together and distill actionable intelligence from it [76, 109, 111]. Thus, what is distilled as actionable is a product of human decision-making. Expert judgment enables analysts to make decisions in real time in a different way than novices [75]. The intuitive form of expert judgment relies on the ability to make predictions about the environment and the possibilities to learn about the commonalities within the environment [53]. In other words, expert judgment is contingent on the situational awareness of the expert.

Figure 1 illustrates the process of decision-making in a threat intelligence scenario. The new information (green in Fig. 1) shows the meta-level information about the intelligence that could be useful for decision-making but is not always known (or available).

[1] https://english.ctivd.nl/investigations.

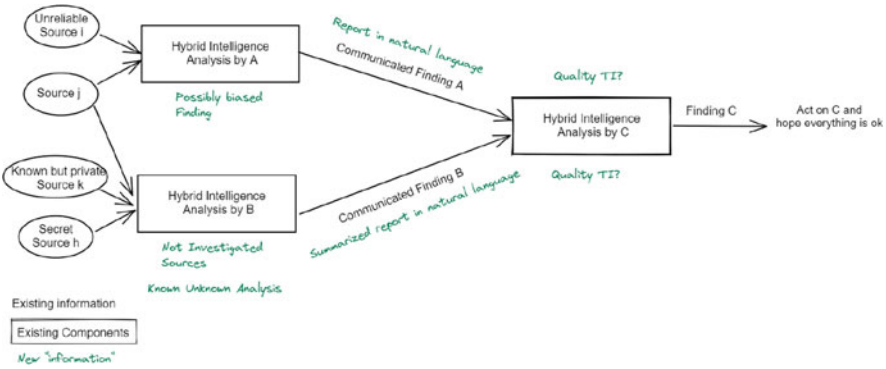

Fig. 1 The threat intelligence process and problem. *When A and B communicate their findings to C, it is hard for C to formally evaluate the uncertainty of the initial sources where A and B's findings are based upon*

Table 1 Overview of recently (2018–2023) researched biases and uncertainties in cybersecurity from the perspective of the analyst/defender

Uncertainty/bias	Description	Ref	Regarding
Overconfidence	Predictions are too certain or too uncertain given the actual performance	[33]	Managing a cyber-physical environment under threat.
Primacy bias	The first item in a series has the best recall	[38]	Attribution of cyber operations
Seizing and freezing	To combat cognitive dissonance, seizing shows a predisposition to information that confirms existing beliefs and freezing shows the refusal to adjust judgments to maintain beliefs	[39]	Attribution of degradative cyber operations
False sense of validation	When there is a perception of a human in the loop, action is undertaken on less and incomplete information	[107]	Use of AI in cyber conflicts
Information-pooling bias	In a team, information that is known by most members is more likely to be shared than information that is unique to an individual	[83]	Incident correlation in cybersecurity threat detection

Table 1 shows an overview of recently researched biases and reasoning with uncertainty (approximately 2018–2023) within the field of cybersecurity from the position of the analyst/defender. Furthermore, time pressures [46], height of stakes of decisions [50], secrecy [57], and a range of complexity are associated with intelligence problems [70]. Achieving situational awareness is difficult, because these characteristics make it hard to know, understand, and make predictions about the environment [53].

2.1 Adding Artificial Intelligence to the Equation

When using AI to obtain threat intelligence, we can speak of a double-edged sword. On the one hand, large datasets can be collected and processed to extract potentially useful information for analysts. On the other hand, this collection process can be expensive [89], invasive [13], and biased or unfair [78]. Furthermore, the sheer amount of information in combination with its dubious provenance and varying quality could result in increased confusion [92] or no actionable intelligence.

For example, in case studies [80], data on the entire Washington DC was collected, and AI-based techniques were used to predict criminal events with high accuracy. Yet, the predictions were not tactically actionable as the predicted hot spot areas amounted to the whole pedestrian area downtown. In contrast, systems with lower accuracy (e.g., 30%) but taking into account "awareness information" on uncertainty were considered more actionable by officers planning for Improvised Explosive Devices detection in Iraq [50].

2.2 Key Issues

Key issues in the field of threat intelligence are categorized as follows:

1. **Socio-technical context: legal and societal elements**—Providing threat intelligence in a genuine democratic society cannot focus solely on maximizing the quality of threat intelligence from a technical perspective. Threat intelligence needs to be accompanied by constitutional safeguards such as providing reliable and robust oversight systems and advancing privacy rights to ensure societal trust in security and intelligence operations. Engaging with this value multiplicity around threat intelligence is crucial to understanding the *legal and societal restrictions*, as well as the wider societal context. As noted by Laura Carlsen [14], "there can be no security without human rights."
2. **Socio-technical context: organizational elements**—Showing how risk-based decisions feed into professional practices and generate knowledge despite limited and constrained sharing [8] is not a trivial task. A structural change in work reflects an *organizational change.* In practice, organizational changes and learning only happen through its members [85], for the good or the bad in their social context [40]. Impact can only be achieved through concrete outcomes in the change of employees' daily work. Organizational learning is studied in environments where knowledge sharing is not heavily restricted [4]. The heavily restricted and sometimes misunderstood nature of threat intelligence sharing [104] results in different challenges viewed through this organizational lens.
3. **Uncertainty of information**—Sources and methods often have significant uncertainties and biases. Analysts are aware of these limitations, but uncertainty is yet to be captured and quantified within the threat intelligence workflow. The need arises for a formalization of uncertainty that can be both machine-readable and human-interpretable.

This undertaking is difficult due to the conditions of decision-making and information sharing in threat intelligence, a heavily restricted and possible deceptive environment [104]. Uncertainty can arise from multiple factors including potential observation errors, imprecision, communication errors, ambiguity, and unknown credibility of the sources. Quantifying and *formalizing uncertainties* can stem from qualitative concepts. These formalizations are distilled from a plethora of heterogeneous sources and need to be relevant in the decision-making process. This type of formalization results in specific challenges.

4. **Empirical evaluation**—The complex conditions of decision-making in threat intelligence workflows makes conducting *empirical validation* extremely challenging. Methodologies from existing literature (as further explained in Sect. 3.7) cannot be directly applied in the context of threat intelligence. For example, with respect to cybersecurity, existing experimental methodologies assume that the individual inputs to a method under scrutiny come from known sources, are complete and correct [59, 91, 97]. This conflicts with the reality of a threat intelligence workflow.

3 Related Work

In this section, we will discuss the related literature on information fusion, vocabularies for threat intelligence, uncertainty representation and reasoning, human judgment and communication of uncertainty, and experimental methodologies for threat intelligence.

3.1 Information Fusion

Information fusion systems extract actionable information from numerous sources [16]. The task of a threat intelligence analyst can be seen as an information fusion task. The first applications of information fusion combined simple sensory data in situations where the physical model was well-understood [67]. With the progression of AI, fusion systems are enabled to incorporate models that learn and adapt to complicated environments [67, 88]. These environments are characterized by data that comes from heterogeneous sources including (but not limited to) sensors, processes relying on learned models, and humans. One example is the DPIF platform where data and information sharing is supported during TIHW processes [100].

When information fusion systems increase in complexity (e.g., increase in amount of heterogeneous sources or overall scale), it also becomes increasingly difficult to make accurate inferences and support effective decision-making [67]. One key challenge within this difficulty is the role of uncertainty [16, 67, 100].

If uncertainty is not considered properly, fusion processes may deliver "underconfident, overconfident and/or incorrect results" [100].

3.2 Vocabularies for Threat Intelligence

Vocabularies exist for standardizing cyber threat intelligence. See Tounsi and Rais [92] for an overview of standards used for TI representation and sharing. Two of the most used vocabularies are STIX™ (Structured Threat Information eXpression) [20] and MITRE ATT&CK™ [22]. STIX™ is a structured language and serialization format that encompasses domain objects like attack pattern and campaign. It also includes relationship objects. MITRE ATT&CK™ is a vocabulary that is concerned with adversary tactics and real-world techniques. It encompasses a plethora of different techniques ranging from defense evasion to credential access. STIX™ and MITRE ATT&CK™ both provide detailed information that can be used to build knowledge graphs for cyber threat intelligence. However, what these vocabularies are missing is information about uncertainty.

3.3 Uncertainty Representation

The need for standardization of uncertainty representation has long been recognized [60]. However, creating a unified vocabulary applicable across multiple domains is difficult as the requirements for such a vocabulary may vary across these domains. As a result, attempts at creating a standardized vocabulary retain some level of domain-specificity (see Table 2).

Uncertainty within the context of information fusion is often discussed using the distinction between aleatoric and epistemic uncertainty. Aleatoric uncertainty arises from the variability in outcome due to randomness [45]. Thus, this type of uncertainty lies within the modeled environment [101]. Epistemic uncertainty refers to a lack of knowledge. This type of uncertainty refers to the epistemic state of an agent instead of random phenomenon [45]. Therefore, it lies outside of the modeled environment and might be mitigated by querying for additional information [101].

In De Villiers et al. [101], this distinction is used to help make uncertainty explicit during the information fusion process. Uncertainty can be factorized to include potential observation errors, imprecision of measurement, communication errors, ambiguity, unknown credibility of sources, and many more. Due to these different factors and categories, explicitness helps in reasoning about involved uncertainties throughout the fusion process. Whether these uncertainties have a one-to-one translation in the current domain remains to be seen.

The URREF ontology [23] is a work in progress and provides an overview of the potential sources of uncertainty. It provides the opportunity to explore the boundaries of the (information fusion) system that one is building. For a more

Table 2 An overview of previous attempts at creating a vocabulary regarding uncertainty representation from various domains

Domain	Relation to uncertainty	Example
Decision sciences	While the domain of decision sciences is incredibly diverse, it is rarely the case that conclusions can be drawn with full certainty; experiments are conducted in controlled environments, and models are developed based on imperfect data. Therefore, proper communication of findings requires reporting the uncertainties associated with those findings	[32, 87]
Earth system science	Earth systems highlight a different challenge when representing uncertainty: spatiotemporal scaling. Atmospheric models can be reasonably accurate when considering a daily global model output but may fail to provide any usable insights on lower scales. Alternatively, small-scale population estimates might not generalize to large-scale ecosystems. This, along with the stochastic nature of ecological processes, measurement error, and human judgment, is an important source of uncertainty within earth systems	[86, 110]
Database management	Data and uncertainty are closely intertwined. Uncertainty would not exist without data, and most, if not all, forms of data is to some degree uncertain. In the context of database management, this means that protocols are needed on how to combine the uncertainty information when merging data sets. These protocols not only depend on the type of uncertainty but also on the domains from which the data originates	[61]
Information fusion	When developing an information fusion system, the uncertainties associated with the fused information determine which fusion method can be applied. Therefore, a careful evaluation of those uncertainties is needed at the start of the development process	[23]

complete account of representing uncertainty in decision-making, we refer the interested reader to the survey of Keith and Ahner [56].

3.4 Reporting Uncertainty in Threat Intelligence

The most basic and most important task of actionable threat intelligence is reporting this intelligence to decision-makers. Reporting uncertainty successfully with respect to threat intelligence has been deemed an important area of study [65]. This is the case, because there is an effect of the way uncertainty is represented on decision-making [29].

Natural language in the form of linguistic categories (e.g., “Likely” or “Probable”) is often used to represent uncertainty within threat intelligence. An example is the Admiralty Code used in several intelligence organizations [46]. Research has shown [65] that this way of conveying uncertainty is often ineffective due

to differences in the way these categories are interpreted by humans, even when definitions are set.

A possible alternative would be using numerical representations [27]. One of the reasons why numerical representations have yet to be implemented is the fear of more risky decision-making by analysts. Indeed, Friedman et al. [35] found that less experienced analysts were overconfident in their decision-making when using numerical representations of uncertainty. However, the same study also found that, generally, numerical representations were actually associated with less risky decision-making and more accurate predictions. These findings suggest that experience might overcome this overestimation [35].

Another possibility is that linguistic categories may contain more than just information of probability [18]. Collins and Mandel [18] proposed that linguistic categories could be used during deliberate argumentation, while numeric categories were most suitable in situations where clear probability estimations were required. A combination of the two formats was also investigated, although no significant difference was found in performance of numerical representation and the combined representation [66]. However, both formats were more effective than just linguistic categories when it came to probability estimation.

Graphical and visual representations for uncertainty representation have also been researched [77]. For example, in a dynamic decision-making missile-defense game, support was found for using graphical representations to increase efficiency even in combinations with numerical representations [7]. For a more complete overview of visual representations, see [77]. Within the context of cyber threat intelligence in a national security environment, as far as the authors are aware, it is not clear how these different representations feed in to professional practices.

3.5 Human Judgment and Bias Under Uncertainty

Decision-making in threat intelligence includes decisions about security risks which are made in face of uncertainty [6], leaving space for subjective and possibly biased judgment [49]. This type of uncertainty is a key element in socio-technical systems.

Judgment under uncertainty has been the object of study for decades [34, 54, 71]. In the past, the focus has been on looking for reasoning shortcuts (heuristics) and bias within human decision-making. This research is built upon by theories of Bayesian inference (the framing of a problem leads humans to view new information in accordance with prior beliefs) [108] and in opposition and digitization (reasoning with uncertain information leads humans to favor the most certain information and ignoring other information [52]). In addition, the dual processing theory expands the existing framework (humans use two systems in the decision-making process, one for quick automatic judgments and one for deliberative, slow, and complex calculations) [108]. For an overview of biases and debiasing techniques within the general field of decision and risk analysis, see Montibeller and Von Winterfeldt [71], and cybersecurity, see Johnson et al. [51].

However, even though the previously mentioned research on known biases (see Table 1), the influence of AI, and team-level heuristics exists, research in reasoning with bias and uncertainty is sparse and scattered over many categories in cybersecurity. Furthermore, the population of intelligence analysts differs from other populations. Especially in the world of cyber professionals, due to heterogeneity in work roles and skill sets, findings are population specific [68]. Whether the same biases and reasoning techniques are relevant within the current field remain to be investigated.

3.6 Existing Approaches to Help Elicit Expert Knowledge in Threat Intelligence

Analysts are highly trained experts that use intuitive decision-making to deal with situations [75]. Okoli et al. [75] put forward that the consensus in literature is that experts are able to make quicker and often better decisions because they are able to use their existing knowledge to assess the situation at hand with the usage of schema. These schema (i.e., strong memory networks) allow experts to have a perceptual advantage even when events unfold in real time [75].

Expert knowledge elicitation techniques are techniques that try to improve the quality of expert judgment with respect to debiasing and reasoning under uncertainty [28]. To that end, structured analytic techniques (SATs) have been previously used in the domain of threat intelligence. SATs are techniques that systematically and transparently aim to externalize internal thought processes [106]. SATs are often not well researched [65], and the few existing studies have shown mixed results [17, 106].

In cases where the SATs have been implemented, an important challenge is to either update these techniques to handle cyber threats and measure their efficacy [17] or develop new ways to make human reasoning transparent and enable explainability. In the field of general decision and risk analysis, see Dias, Morton, and Quigley for an overview [28] of expert knowledge elicitation techniques.

Another approach to make judgments less biased is the use of coherentization and aggregation of judgments [55]. Here, multiple numerical judgments from different analysts are first made coherent with respect to specific statistical assumptions (e.g., probabilities must add up to 1) and, thereafter, aggregated into one prediction. However, this approach has not been researched with respect to realistic cyber threat intelligence scenarios.

In intelligence analysis, meta-information is already used to aid decisions and judgments under uncertainty. The standard NATO Standardization Agreement (STANAG) 2511 captures two qualitative categories [65], although their exact implementation is not uniform between different intelligence agencies [46]. In general, source reliability can be seen as a confidence level based on historical performance. Information credibility captures the extent to which a new piece of

information is consistent with the current reporting [46]. However, these factors are often not enough to provide the necessary uncertainty information. For example, source reliability can vary tremendously in different situations. The current categories do not provide a way to make this distinction explicit [46].

3.7 Existing Experimental Procedures and Methods

Existing approaches assume that the individual inputs to a method under scrutiny come from known sources, are complete, and are correct [59, 91, 97]. Representations of security and malicious threats (e.g., attack and defense trees, data flow diagrams, petri-nets, etc. [95]) are compared by either observing the quality of the representation (compared to a baseline model) [31] or by observing some measure of the analysis output (e.g., the precision of the identified security threats [97] or complexity of generating all attack paths [44]).

Due to deception, the variability in sources, and incompleteness (automatically generated), cyber threat intelligence can only be actionable if it takes the quality of the information into account. To underline this point, Ranade et al. [84] generated fake cyber threat intelligence and observed that experts would consider both the deceptive TI and the authentic TI as equally true. Without an explicit discussion on the quality of TI, these effects remain unnoticed.

Decision support systems can also be used in cybersecurity to assist analysts in their job [11, 37]. For instance, decision support systems can be used to aid optimizing cyber forensic investigations [73], cybersecurity threat and incident management [98], and the manual assessments of proportionality in military cyber operations [64]. The evaluation of decision support systems mainly focused on conducting user studies and evaluating transparency (i.e., explain how the system works) and trust (increase user confidence in the system) of the decision support system [74].

Methods in explainable AI (xAI) [43] are certainly interesting to investigate for the purpose of evaluating persuasiveness. In a recent survey on the evaluation of xAI systems, Nauta et al. [72] found that only one fifth of the analyzed papers evaluated their findings with users. In addition, Dalvi et al. [26] claim that within cyber threat intelligence, multiple xAI implementations to help understand AI algorithms do not actually agree with each other on their explanation. This is no surprise because there is no standard correct or best explanation when it comes to measuring explanations in many scenarios [74].

4 Socio-technical Challenges

In this section, challenges with respect to the socio-technical context of a TIHW are discussed. In a multidisciplinary approach, legal, societal, and organizational

matters are discussed. For each of the identified challenges, ideas are proposed to handle them.

4.1 Identification and Operationalization of Key Societal and Legal Elements

Governments and their agents are expected to follow the law, particularly in sensitive constitutionally protected matters such as privacy, while ensuring that other key rights are also safeguarded [48]. In the larger governance context, political decision-makers need to take these challenges into account to monitor, oversee, and evaluate responsible intelligence and security agencies accordingly. For instance, passing certain laws to facilitate the work of security and intelligence services may support ethical decision-making or would nudge security and intelligence operations to fall into a web of unethical practices [58]. These considerations contribute to strengthening citizens' trust toward security and intelligence operations, particularly where the provision of data or information is needed [24].

At the same time, intelligence and security agencies need to ensure that their practices remain within the boundaries of the law and the core of public service integrity [58]. In this environment, agencies can implement innovations, and government can realign their model in a way that supports confidence in public service by putting the well-being of constituents first above all [21].

The identification and implementation of key elements is hard. For example, transparency has the ability to obscure (e.g., showing so much data to distract from the central information) [3], can encourage "seeing" over understanding (i.e., being able to look inside a system is not enough, one also needs to be able to interact with them in a broader social context to have an actual understanding) [3], and can be used to create an atmosphere of transparency in the public eye instead of fostering accountability within an organization [1].

Socio-technical challenge Sect. 4.1: It is unclear what and how legal and societal constraints can be implemented in a TIHW
The identification and implementation of key legal and societal elements is necessary to ensure that intelligence and security agencies' practices are lawful and enable trust in citizens. Currently, it is not known how legal and societal constraints need to be incorporated in professional hybrid threat intelligence practices.

We are in the process of identifying key societal and legal elements of the TIHW. The workflow needs to follow the standard principle of necessity established under international human rights law: establishing an objective goal deemed to be

necessary in order to protect a legitimate interest (e.g., specific target affecting national security) [57].

Having established necessity, we can then address the proportionality test, which requires the workflow to establish a justification balancing the methods to be utilized against the intended goal. For instance, a TIHW should conduct impact assessments to determine proportionality and assess whether the workflow causes a chilling effect on citizens. In addressing a subsidiary analysis, the workflow will be established in an environment where its framework is regulated by transparency.

It is first necessary to develop a framework based on international best practices of the socio-technical conditions and constraints for risk-based decisions in intelligence communities such as necessity and proportionality. How do comparable threat intelligence procedures take place in other countries and governmental contexts?

Second, an in-depth juridical analysis of relevant laws and regulations should be conducted to inform the design of the TIHW. This can also help understand how international best practices can be integrated into existing workflows and how these could be operationalized in practice.

Idea Sect. 4.1: Develop a framework based on international best practices and relevant laws and regulations
Conducting research in relevant laws and regulations and international best practices can show how key legal and social elements can be integrated into existing workflows and be operationalized in practice.

4.2 Future Proofing Data Protection and Human Right Safeguards

Technology changes rapidly. For example, big tech companies currently try to construct methods for identifying deepfakes while internet enthusiasts and state-sponsored disinformation campaigns keep finding new systems to fool these detectors [25]. This arms race leaves regulations to be outdated.

Even in a perfect world where all limitations would have concrete definitions that were agreed upon, the issue remains that not all limitations can be satisfied at all times. What is perceived by some as an easy solution, for example, mass surveillance of communications, may be the most harmful approach from a privacy or digital rights perspective. Furthermore, these trade-offs also encompass an economical aspect. Although security (e.g., finding threats) is the goal, analyst's time is often considered more important in the decision-making process, since this time is expensive and scarce [12]. To address these limitations in a way that has a better chance of working in practice over a substantial amount of time remains a challenge.

Socio-technical challenge Sect. 4.2: Technology changes rapidly, and limitations cannot be satisfied at the same time
Limitations of the system have to be addressed in a way that has a better chance of working in practice over a substantial amount of time.

Future proofing efficacy and safeguards should include both considering incoming international legal frameworks such as Council of Europe Convention 108+ [30] and responding to existing societal challenges such as the engagement of commercial Open Source Intelligence (OSINT) which might be caught by the future enactment of the European Artificial Intelligence Act [19].

Idea Sect. 4.2: Incoming international legal frameworks and existing societal challenges should be addressed in a TIHW
Future proofing data protection and human rights safeguards should be taken into account to make the workflow not only relevant now but also robust in the future.

4.3 Studying Organizations with Restricted Information Sharing

The amount of information expands rapidly, and no one institution or vendor can hold it all. Bouwman [12] found that, when looking at two commercial providers of threat intelligence, there was minimal overlap in the indicator feeds, even in cases of identical threat actors. Next to processes between organizations, information sharing is also influenced by internal organizational processes. Domains (e.g., military, national security) within threat intelligence work in compartments to limit information flow [58].

Decision-makers, analysts, and field personnel do not have access to the same information. In environments where information sharing is restricted due to need-to-know policies, the sharing of healthy behavior toward regard for human dignity and ethics may also be difficult. Due to compartmentalization, sometimes best practices are difficult to transfer to other compartments. Even when there is a transfer of colleagues between compartments, there is a risk of mixing between colleagues upholding "positive" culture and colleagues caught in "negative" work culture [58]. Instead of a chain of restricted information sharing, decision-makers and analysts have the opportunity to act as autonomous agents [99].

Furthermore, individual motives play in a role in the information sharing as well. Found information can, depending on the significance of said information, result in increased "status" once shared with a superior [102]. The higher the position of

the superior in the organizational hierarchy, the greater the potential "status" gain. Sharing information with peers or superiors lower in the organizational hierarchy could result in them taking credit, leaving the original finder with a diminished "status" gain.

Socio-technical challenge Sect. 4.3: Analyzing/improving information sharing in a restricted environment is difficult
External organizational processes, internal organizational processes, and psychological factors influence the degree of information sharing. It is unclear how these elements feed in to organizational practices within a restricted environment.

An integrative view of the three levels of organizational change is necessary to get a better understanding of the impact of change during the implementation of a system. These levels consist of an operational level (learning of the actual professionals themselves), a tactical level (learning from a management perspective), and a strategical level (learning in the perspective of organizational compliance and regulations). In addition, this integration helps in pinpointing effects that would otherwise stay unseen. This view calls for feedback loops both within the hybrid intelligence pipeline and across these levels, within and between organizations.

A potential solution and opportunity for empirical work could be the implementation of an acceptable (by the organization and the employees) variation on the nice-to-know code among different compartments; this may facilitate the transfer and fostering of values and international principles such as proportionality.

Idea Sect. 4.3: Integrating the operational, tactical, and strategic levels of organizational change with a clear distinction of processes between organizations and those within organizations
Taking into consideration the internal processes of an organization (individual motives, team dynamics, and organizational aversion to change), and the external processes (between organizations, in a broader societal context), using empirical evaluation, can help pinpointing unforeseen effects.

5 Technical and Experimental Challenges

In this section, technical and experimental challenges of formalizing uncertainty within a TIHW are discussed. For each of the identified challenges, ideas are proposed to handle them.

5.1 Representing Uncertainty Stemming from Systems, Humans, and Situations

As discussed in Sect. 3.4, uncertainty within threat intelligence workflows is mainly conveyed in natural language between people. Standardized natural language formats are used in certain domains. For example, within the defense domain, the standard NATO Standardization Agreement (STANAG) 2511 incorporates linguistic labels to communicate source reliability and information credibility [65]. However, such uncertainty representation is typically not suited for machine-based processing. Since these uncertainties often stem from qualitative concepts, it can be challenging to translate them into representations that quantify the uncertainty for the machine-based processing and vice versa. Furthermore, the threat intelligence workflow is hybrid, meaning that the uncertainties themselves will not only stem from abstractions and errors in systems and data but also from the process of human decision-making. In addition, to provide a basis for accountability in the larger societal context, uncertainty information should be available along the chain of communication.

Technical challenge Sect. 5.1: Representing and tracking uncertainty for actors in TIHW is complicated due to qualitative sources of uncertainty
Uncertainty will not only arise from abstractions and errors in systems and data but also from the process of human decision-making. The uncertainty information should be available along the chain of communication. Representing uncertainty stemming from qualitative concepts is challenging.

Uncertainty can be factorized by a plethora of elements. Due to these different factors and categories, explicitness helps in reasoning about involved uncertainties throughout the process.

A possible extension of the URREF ontology [23], introduced in Sect. 3.3, could serve as a basis to start representing uncertainty within the threat intelligence workflow. It provides the opportunity to explore the boundaries of the (information fusion) system that one is building. It also has the expressiveness to incorporate the current NATO-STANAG 2511 standard [9]. It should be considered a checklist, forcing any information system developer to thoroughly analyze the information fusion pipeline and make adjustments where necessary.

For uncertainty to be used not only between two agents who are in immediate connection to one another but also along the chain of communication, it is necessary that uncertainty provenance is tracked. A second proposed framework, handling provenance tracking, is the PROV data model [41]. This model can be used to structure the information in a knowledge graph (KG), making a distinction between entities (things that contain information), activities (the process that produced the information), and agents (persons/software/machines responsible for the taken

actions). If applied within a TIHW, it allows for the construction of a provenance trail of information, providing insights into origin of the information.

Idea Sect. 5.1: Making uncertainty explicit by expanding and combining existing ontologies
Uncertainty can be factorized by a plethora of elements. Due to these different factors and categories, explicitness helps in reasoning about involved uncertainties throughout the fusion process. A possible extension of the URREF ontology [23] in combination with the PROV ontology [41] could serve as a basis of representing these uncertainties.

5.2 Formal Reasoning with Uncertainty

When it comes to providing explanations that agree with human understanding, uncertainty representation is not enough. A specific type of formal uncertainty reasoning that can reflect abductive inference is necessary [69].

With respect to formal reasoning, two directions can be distinguished. These directions are forward and backward reasoning; see Fig. 2. Backward reasoning, in the current context, is about recreating trails and possibly gathering more information to demonstrate proof of the proportionality and subsidiarity of actions for each TIHW component and for the entire TIHW. Forward reasoning, in the current context, could be used in building an actionable strategy with minimal uncertainty. The challenge at hand is that there are no such reasoning tasks that minimize the uncertainty on process level for threat intelligence and also, with respect to proportionality and other legal and societal constraints, in a domain agnostic way.

Fig. 2 Forward and backward reasoning. *Forward and Backward Reasoning: Forward reasoning can be thought of as a form of what-if reasoning where the reasoning starts from the information and moves forward. Backward reasoning can be thought of as an evaluation where the reasoning starts from the decision and moves backward. The green color represents the respective reasoning paths*

Technical challenge Sect. 5.2: Formalizing reasoning tasks to minimize uncertainty is challenging
How to formalize reasoning tasks that minimize the uncertainty on process level, e.g., by backward reasoning ("what-if") or forward reasoning ("why")?

A third framework is needed, one that could unify the uncertainty overview with formalized reasoning about uncertainties. Which to choose is not a trivial choice. This depends on the complexity of reasoning capabilities, and the richness of the uncertainty overview. The combination of uncertainty representation, provenance tracking, and reasoning/causal inference is necessary.

Past approaches in uncertainty reasoning use fuzzy logic [10], epistemic logic [5], Markov network/processes [90], probabilistic logic [42], Bayesian networks [15], and Dempster-Shafer theory [112]. However, automated reasoning over uncertainty is very complex. The choice of reasoning method is dependent on the way the uncertainty is represented (e.g., in a qualitative format [105]). The combination of the URREF ontology [23] and PROV data model [41] with the intent to reason about uncertainty is particularly difficult. The foundations of both the URREF ontology and PROV data model are based on Boolean statements, either something is true or it is not. When dealing with uncertainties, the assigned value lies most often somewhere in the gray area in between.

Idea Sect. 5.2: Combine uncertainty representations with a provenance framework and reasoning/causal inference
To tackle provenance tracking and enable forward and backward reasoning within the workflow, the URREF ontology can be combined with the PROV framework [41] and integrated with a third framework (uncertainty reasoner) to unify the uncertainty overview with formalized reasoning about uncertainties.

5.3 *Experimental Methods Aware of Uncertainty*

Designing methods to evaluate the correctness of a threat intelligence decisions regarding a course of action or event likelihood is not easy, because the lack of ground truth is persistent in threat intelligence. In controlled experimentation, one possible solution is to curate the ground truth manually [93, 96], but this is not always feasible when analyzing large number of threat intelligence sources [62]. In previous work, there exists an optimal choice [55], preferences were measured

[47, 63], or expert judgment was used for validation [64]. These measurements were sometimes in a qualitative format, e.g., interviews [82, 103], and in other cases, they were quantitative, e.g., optimal likelihood estimations [55]. However, it is not clear how incomplete and uncertain threat intelligence information should be treated in the ground truth.

When it comes to expert judgment, the situation becomes complex. On the one hand, analysts are highly trained individuals in high-risk decision-making [75]. The combination of training and experience often leads to "intuitive" decision-making. This behavior is rarely seen in non-experts [75]. On the other hand, decisions about security risks may be affected by biased judgment [6, 49]. Uncertainty and bias are key elements of each socio-technical system. Minimization is not the ultimate goal. However, existing experimental methods lack protocols that can effectively and systematically measure human bias in threat intelligence decision-making [94].

Technical challenge Sect. 5.3: Existing empirical protocols for THIW validation have to be adapted to incorporate the human factor
What existing empirical protocols and measures can be adapted to quantify measures of uncertainties including human bias in TIHW?

The property of the exchanged information in the TIHW is that from an analyst's (or study participant's) point of view, the information may (or may not) be aggregated, incomplete, inaccurate, unreliable, and/or censored. And yet, a sound and convincing explanation (with at least partial traces in the model) for a minimal intervention (i.e., proportionality) must be possible. The current landscape of empirical methods does not cater for investigating such aspects of decision support systems. Important is to measure the human effects. Qualitatively and quantitatively evaluating the entire intelligence pipeline thus calls for novel protocols, measures, and controls to be developed.

See Table 3 for an overview of methods used in recent (2018–2023) research on bias and uncertainties in cybersecurity from the perspective of the analyst/defender (for background information on the research in question, see Table 1). Internal validation of surveys, questions, and other methods were rare. External validation of these methods was most often checked with a group of experts or participants [33, 38, 39]. These findings suggest that there is a need for more internal and external validation methods.

Validation in isolation is not insightful enough. A validation methodology has to be adapted to effectively assess heterogeneous systems with both AI, human, and unknown components.

Table 3 Overview of methods used in recent (2018–2023) research on bias and uncertainties in cybersecurity from the perspective of the analyst/defender

Uncertainty/bias	Ref	Type of study	Measures
Overconfidence	[33]	Cyber game	Argumentation and self-confidence via coded transcripts of verbal discussion
Primacy bias	[38]	Vignette	Attribution via survey on confidence levels
Seizing and Freezing	[39]	Vignette	Attribution via survey on confidence levels and coded justification
False sense of validation	[107]	Vignette	Machine preference via survey on confidence levels and selecting decisions
Information-pooling bias	[83]	Synthetic task environment (i.e. less focus on realism and more on the cognitive task at hand) experiment	Team collaboration and information pooling were measured via coded transcripts of the verbal discussions

Idea Sect. 5.3: Validating effectiveness of a human-based decision-making process (such as TIHW) calls out for human-in-the-loop experimental protocols
To this aim, new experimental protocols must be specifically designed to measure human effects. For instance, similar protocols outlined in [94] could be retrofitted to the domain of threat intelligence.

5.4 Computing with Objects of Evaluation to Measure Their Quality May Not Be Possible

Since direct computation over unknown (or uncertain) values is not possible, the evaluation should take as input meta-information rather than the object of evaluation. So the key question is not whether say an AI image recognition tool works with 80 or 90% of accuracy, but rather, which representation of such uncertainty is actionable for the user. However, methodologies for threat modeling and analysis and the protocols used for their evaluation require the user to specify the sources of security relevant components and the locations where such information is not allowed to flow. Therefore, existing methodologies [31, 44, 93, 95] cannot be directly carried over to evaluate the appropriateness of alternative suggestions by the TIHW, such as an alternative plan of intervention in the presence of a terrorist threat by requesting input from a new source.

Technical challenge Sect. 5.4: Measuring meta-information about objects of evaluation is necessary instead of measuring the object level
How can meta-information about objects of evaluation be measured and under what conditions are these measurements valid?

Since computation with the object of evaluation itself is not always possible, we need to make use of the meta-information that is available (e.g., timestamp, type of device, etc.) to define and compute new measures of quality. As put forward in Zibak, Sauerwein, and Simpson, data quality in threat intelligence has not been properly empirically investigated [113]. To achieve this, the first step is to investigate what type of meta-information is available from the field.

Confounding factors should be balanced within these measurements. For example, a THIW relies on AI modules, which can be symbolic modules explicitly taking uncertainty into account, or sub-symbolic modules (ML-like). For the latter, several studies exist on estimating and propagating uncertainty on the output of, e.g., deep learning models (see, e.g., the popular dropout method [36]), but there is no protocol to propagate the effect of hybrid errors of the next TIHW component.

Idea Sect. 5.4: Validate new measures to quantify meta-information about objects of evaluation
Since computation with the object of evaluation itself is not always possible, we need to make use of the meta-information that is available (e.g., timestamp, type of device, etc.) to define and compute new measures of quality. To achieve this, the first step is to investigate what type of meta-information is available from the field.

6 The Bigger Picture

In this chapter, we discussed the interplay between complex conditions and trade-offs between security and legal, societal, and organizational restrictions that make decision-making under uncertainty a challenging endeavor. Table 4 shows an overview of the illustrated challenges.

In the quest to achieve efficiency and effectiveness in threat intelligence, security and intelligence agencies are implementing AI-powered solutions to find actionable information to aid them in decision-making during uncertainty. However, one must remember that these AI tools to help deal with uncertainty in threat intelligence can end up being a double-edged sword. The development of a threat intelligence hybrid

Table 4 Overview of socio-technical, technical, and experimental challenges discussed in this chapter

Category	Section	Challenge	Idea
Socio-technical	Section 4.1	It is unclear what and how legal and societal constraints can be implemented in a TIHW	Develop a framework based on international best practices and relevant laws and regulations
	Section 4.2	Technology changes rapidly and constraints can not be satisfied at the same time	Incoming international legal frameworks and existing societal challenges should be addressed in a TIHW
	Section 4.3	Analyzing/improving information sharing in a restricted environment is difficult	Integrating the operational, tactical and strategic levels of organizational change with a clear distinction of processes between organizations and those within organizations
Technical	Section 5.1	Representing and tracking uncertainty for actors in TIHW is complicated due to qualitative sources of uncertainty	Making uncertainty explicit by expanding and combining existing ontologies
	Section 5.2	Formalizing reasoning tasks to minimize uncertainty is challenging	Combine uncertainty representations with a provenance framework and reasoning/causal inference
Experimental	Section 5.3	Existing empirical protocols for THIW validation have to be adapted to incorporate the human factor	Validating effectiveness of a human-based decision making process (such as TIHW) calls out for human-in-the-loop experimental protocols
	Section 5.4	Measuring meta-information about objects of evaluation is necessary instead of measuring the object level	Validate new measures to quantify meta-information about objects of evaluation

workflow (TIHW) is not an exception. Uncertainty will likely arise due to communication errors, ambiguity, and unknown credibility of the sources/provenance.

Challenges arise when creating a robust system that advances the embedding of regard for citizens' fundamental rights and responding to efficiency and support of user autonomy to enable intelligence agencies to arrive at the best possible decisions. Achieving this fine point is essential in a democratic society, because it develops societal trust in security and intelligence operations.

Despite developing a system that meets the requirements mentioned in this paper, we also need a path forward in security and intelligence operations to transfer this knowledge within their agencies or organizations. We recommend applying a lens based on international standardized legal principles, such as proportionality and necessity, during human AI interactions or evaluations in the absence of ground truth. Thus, the relationship between developing an AI system and having regard for societal and legal matters are not far from each other.

Uncertainty in a TIHW stems from quantitative as well as qualitative sources. This makes the formalization of uncertainty hard. In addition, uncertainty representation has to be machine-readable, as well as human understandable. Therefore, uncertainty representation should enable reasoning according to abductive inference. Representation and reasoning methods for uncertainty that capture these conditions have not been constructed with respect to threat intelligence.

AI augmented socio-technical systems for threat intelligence must respond to relevance, timeliness, accuracy, completeness, and ingestibility. A TIHW evaluation will require investigating the persuasiveness (e.g., to the oversight body), efficiency (helps analysts make decisions faster), and debugging (helps analysts identify when something is wrong and explore "what-if" scenarios) of the explanations, for which appropriate measures are to this day less explored.

The validation methodology for a TIHW has to holistically incorporate AI, human, and unknown components. In addition, confound-aware methods that measure the meta-level instead of the object-level of a TIHW are necessary. Validation methodologies within threat intelligence that satisfy these requirements have not been thoroughly investigated.

We hope to stimulate discussion and further research in the community by illustrating these challenges and possible ways to answer them.

Acknowledgments We are thankful to Sarah Giest and Iris Cohen for their valuable feedback. This work was funded by the *Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO)* under the HEWSTI Project under grant no. 14261.

Contributions SVG (Sects. 2, 3, 4, 5, 6, Figs. 1, 2, Tables 1, 3, 4), JC (Sects. 4 and 6), RR (Sects. 5.1, 5.2, Fig. 2, Table 2), B Weerheijm (Sect. 4.3), B Wagner (Sect. 4), GP (Sects. 5.1 and 5.2), BK (Sect. 4), SS (Sects. 5.1 and 5.2), KT (Sects. 1, 2, 5.3, 5.4, Fig. 1), and FM (Sects. 1, 2, 6, Fig. 1) have conceived the presented ideas and contributed to their corresponding sections in the manuscript.

References

1. Albu, O.B., Flyverbom, M.: Organizational transparency: conceptualizations, conditions, and consequences. Business Soc. **58**(2), 268–297 (2019). https://doi.org/10.1177/0007650316659851
2. Alexander, P.: Exploring bias and accountability in military artificial intelligence. 7 LSE Law Review, pp. 396–405 (2022)
3. Ananny, M., Crawford, K.: Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc. **20**(3), 973–989 (2018). https://doi.org/10.1177/1461444816676645
4. Argote, L., Miron-Spektor, E.: Organizational learning: from experience to knowledge. Organiz. Sci. **22**(5), 1123–1137 (2011). https://doi.org/10.1287/orsc.1100.0621
5. Banerjee, M., Dubois, D.: A simple logic for reasoning about incomplete knowledge. Int. J. Approx. Reason. **55**(2), 639–653 (2014). https://doi.org/10.1016/j.ijar.2013.11.003. https://www.sciencedirect.com/science/article/pii/S0888613X13002478
6. Bier, V.: The role of decision analysis in risk analysis: a retrospective. Risk Analy. **40**(S1), 2207–2217 (2020)
7. Bisantz, A.M., Cao, D., Jenkins, M., Pennathur, P.R., Farry, M., Roth, E., Potter, S.S., Pfautz, J.: Comparing uncertainty visualizations for a dynamic decision-making task. J. Cogn. Eng. Decis. Making **5**(3), 277–293 (2011). https://doi.org/10.1177/1555343411415793
8. Blagden, D.: The flawed promise of national security risk assessment: nine lessons from the british approach. Intell. Nat. Secur. **33**, 716–736 (2018)
9. Blasch, E., Laskey, K., Jousselme, A., Dragos, V., Costa, P., Dezert, J.: URREF reliability versus credibility in information fusion (stanag 2511). In: Proceedings of the 16th International Conference on Information Fusion, FUSION 2013 (2013)
10. Bobillo, F., Straccia, U.: Fuzzydl: an expressive fuzzy description logic reasoner. In: 2008 IEEE International Conference on Fuzzy Systems (IEEE World Congress on Computational Intelligence), pp. 923–930 (2008). https://doi.org/10.1109/FUZZY.2008.4630480
11. Bohanec, M.: Decision support. In: Mladenić, D., Lavrač, N., Bohanec, M., Moyle, S. (eds.) Data Mining and Decision Support, vol. 745. The Springer International Series in Engineering and Computer Science. Springer, Berlin (2003). https://doi.org/10.1007/978-1-4615-0286-9_3
12. Bouwman, X., Griffioen, H., Egbers, J., Doerr, C., Klievink, B., van Eeten, M.: A different cup of TI? The added value of commercial threat intelligence. In: 29th USENIX Security Symposium (USENIX Security 20), pp. 433–450 (2020)
13. Brown, I., Korff, D.: Terrorism and the proportionality of internet surveillance. Eur. J. Criminol. **6**, 119–134 (2009)
14. Carlsen, L.: Mexico's false dilemma: human rights or security. Nw. J. Hum. Rts **10**(3), 145–135 (2012)
15. Carvalho, R.N., Laskey, K.B., Costa, P.C.G.: PR-OWL – a language for defining probabilistic ontologies. Int. J. Approx. Reason. **91**, 56–79 (2017). https://doi.org/10.1016/j.ijar.2017.08.011. https://www.sciencedirect.com/science/article/pii/S0888613X17301044
16. Catano, V., Gauger, J.: Information fusion: Intelligence centers and intelligence analysis. In: Goldenberg, I., Soeters, J., Dean, W.H. (eds.) Information Sharing in Military Operations, pp. 17–34. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-42819-2_2
17. Claver, A., van de Meeberg, H.M.: Devil's advocacy within dutch military intelligence (2008–2020): an effective instrument for quality assurance? Intell. Nat. Secur. **36**(6), 849–862 (2021). https://doi.org/10.1080/02684527.2021.1946951
18. Collins, R.N., Mandel, D.R.: Cultivating credibility with probability words and numbers. Judg. Decis. Making **14**(6), 683–695 (2019). https://doi.org/10.1017/S1930297500005404
19. Commission, E.: Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (2021). COM(2021), 206 final, 2021/0106 (COD)

20. Committee, C.T.I.T.: Introduction to stix. https://oasis-open.github.io/cti-documentation/stix/intro.html (2023). Accessed 15 Jun 2023
21. Constantino, J.: Exploring article 14 of the eu ai proposal: Human in the loop challenges when overseeing high-risk ai systems in public service organisations. Amsterdam Law Forum **14**(3), 17 (2022)
22. Corporation, T.M.: Mitre att&ck. https://attack.mitre.org/ (2023). Accessed 15 Jun 2023
23. Costa, P., Jousselme, A.L., Laskey, K.B., Blasch, E., Dragos, V., Ziegler, J., de Villiers, P., Pavlin, G.: Urref: uncertainty representation and reasoning evaluation framework for information fusion. J. Adv. Inf. Fusion **13**(2), 137–157 (2018)
24. Court, T.H.D.: Njcm et al. v. the dutch state (2020). https://uitspraken.rechtspraak.nl/#!/details?id=ECLI:NL:RBDHA:2020:865 (2020). ECLI: NL: RBDHA: 2020:865 (NL) and ECLI:NL:RBDHA:2020:1878 (EN) (SyRI): [6.5]
25. Dagar, D., Vishwakarma, D.K.: A literature review and perspectives in deepfakes: generation, detection, and application. Int. J. Multimed Inf. Retr. **11**, 219–289 (2022). https://doi-org.vu-nl.idm.oclc.org/10.1007/s13735-022-00241-w
26. Dalvi, A., Siddavatam, I., Patel, A., Panchal, A., Kazi, F., Bhirud, S.: Predicting attribute effectiveness using biased databases. In: 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), pp. 1–8 (2021). https://doi.org/10.1109/SMARTGENCON51891.2021.9645789
27. Dhami, M.K., Mandel, D.R.: Words or numbers? Communicating probability in intelligence analysis. Amer. Psychol. **76**(3), 549–560 (2021). https://doi.org/10.1037/amp0000637
28. Dias, L.C., Morton, A., Quigley, J.: Elicitation. The Science and Art of Structuring Judgement. International Series in Operations Research & Management Science, vol. 261. Springer, Berlin (2018)
29. Durbach, I.N., Stewart, T.J.: An experimental study of the effect of uncertainty representation on decision making. Eur. J. Oper. Res. **214**, 380–392 (2011). https://doi.org/10.1016/j.ejor.2011.04.021
30. Council of Europe: The convention for the protection of individuals with regard to automatic processing of personal data (cets no. 108). https://www.coe.int/en/web/data-protection/convention108-and-protocol (1981). Accessed 18 Jun 2023
31. Eades III, H., Gadyatskaya, O.: Graphical models for security. In: 7th International Workshop, GraMSec 2020 (2020)
32. Fischhoff, B., Davis, A.L.: Communicating scientific uncertainty. Proc. Natl. Acad. Sci. **111**(Supplement_4), 13664–13671 (2014). https://doi.org/10.1073/pnas.1317504111. https://www.pnas.org/doi/abs/10.1073/pnas.1317504111
33. Frey, S., Rashid, A., Anthonysamy, P., Pinto-Albuquerque, M., Naqvi, S.A.: The good, the bad and the ugly: a study of security decisions in a cyber-physical systems game. IEEE Trans. Softw. Eng. **45**(5), 521–536 (2019). https://doi.org/10.1109/TSE.2017.2782813
34. Friedman, J.A., Zeckhauser, R.: Uncertainty in intelligence. Intell. Natl. Secur. **27**(6), 824–847 (2012). https://doi.org/10.1080/02684527.2012.708275
35. Friedman, J.A., Lerner, J.S., Zeckhauser, R.: Behavioral consequences of probabilistic precision: experimental evidence from national security professionals. Int. Organiz. **71**(4), 803–826 (2017). https://doi.org/10.1017/S0020818317000352
36. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, pp. 1050–1059. JMLR.org (2016)
37. Garae, J., Ko, R.: Visualization and data provenance trends in decision support for cybersecurity. In: Carrascosa, I.P., Kalutarage, H., Huang, Y. (eds.) Data Analytics and Decision Support for Cybersecurity. Springer, Berlin (2017). https://doi.org/10.1007/978-3-319-59439-2_9
38. Gomez, M.: Sound the alarm! updating beliefs and degradative cyber operations. Eur. J. Int. Secur. **4**(2), 190–208 (2019). https://doi.org/10.1017/eis.2019.2
39. Gomez, M.A.: Past behavior and future judgements: seizing and freezing in response to cyber operations. J. Cybersecur. **5** (2019). https://doi.org/10.1093/cybsec/tyz012

40. Gonin, M., Palazzo, G., Hoffrage, U.: Neither bad apple nor bad barrel: how the societal context impacts unethical behavior in organizations. Busin. Ethics Eur. Rev. **21**(1), 31–46 (2012). https://doi.org/10.1111/j.1467-8608.2011.01643.x
41. Groth, P., Moreau, L.: An overview of the prov family of documents. W3C Working Group Note (2013). http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
42. Henderson, T.C., Simmons, R., Sacharny, D., Mitiche, A., Fan, X.: A probabilistic logic for multi-source heterogeneous information fusion. In: 2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Daegu, Korea (South), pp. 530–535 (2017). https://doi.org/10.1109/MFI.2017.8170375
43. Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W.: Explainable AI methods – a brief overview. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.R., Samek, W. (eds.) xxAI - Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science, vol. 13200. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-04083-2_2
44. Hong, J.B., Kim, D.S., Chung, C.J., Huang, D.: A survey on the usability and practical applications of graphical security models. Comput. Sci. Rev. **26**, 1–16 (2017)
45. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach. Learn. **110**, 457–506 (2021)
46. Irwin, D., Mandel, D.R.: Improving information evaluation for intelligence production. Intell. Natl. Secur. **34**(4), 503–525 (2019). https://doi.org/10.1080/02684527.2019.1569343
47. Irwin, D., Mandel, D.R.: Communicating uncertainty in national security intelligence: expert and nonexpert interpretations of and preferences for verbal and numeric formats. Risk Analysis (2022). https://doi.org/10.1111/risa.14009
48. Janssen, M., der Hoven, J.V.: Big and open linked data (bold) in government: a challenge to transparency and privacy? Govern. Inf. Quart. **32**, 363–368 (2015)
49. Jaspersen, J.G., Montibeller, G.: Probability elicitation under severe time pressure: a rank-based method. Risk Analy. **35**(7), 1317–1335 (2015)
50. Jensen, M.A.: Intelligence failures: what are they really and what do we do about them? Intell. Natl. Secur. **27**(2), 261–282 (2012). https://doi.org/10.1080/02684527.2012.661646
51. Johnson, C.K., Gutzwiller, R.S., Ferguson-Walter, K.J., Fugate, S.J.: A cyber-relevant table of decision making biases and their definitions. Technical Report (2020). https://doi.org/10.13140/RG.2.2.14891.87846
52. Johnson, S.G.B., Merchant, T., Keil, F.C.: Belief digitization: do we treat uncertainty as probabilities or as bits? J. Exper. Psychol. General **149**, 1417–1434 (2020). https://doi.org/10.1037/xge0000720
53. Kahneman, D., Klein, G.: Conditions for intuitive expertise: a failure to disagree. Amer. Psychol. **64**(6), 515–526 (2009). https://doi.org/10.1037/a0016755
54. Kahneman, D., Slovic, P., Tversky, A. (eds.): Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge (1982). https://doi.org/10.1017/CBO9780511809477
55. Karvetski, C.W., Mandel, D.R., Irwin, D.: Improving probability judgment in intelligence analysis: from structured analysis to statistical aggregation. Risk Analy. **40**(5), 1040–1057 (2020). https://doi.org/10.1111/risa.13443
56. Keith, A.J., Ahner, D.K.: A survey of decision making and optimization under uncertainty. Ann. Oper. Res. **300**, 319–353 (2021). https://doi.org/10.1007/s10479-019-03431-8
57. Korff, D., Wagner, B., Powles, J.E., Avila, R., Buermeyer, U.: Boundaries of law: exploring transparency, accountability, and oversight of government surveillance regimes. Cybersecurity (2017)
58. Kowalski, M.: Ethics of Counterterrorism. Boom uitgevers Amsterdam (2017)
59. Labunets, K., Massacci, F., Paci, F.: On the equivalence between graphical and tabular representations for security risk assessment. In: Proceeings of the REFSQ'2016, pp. 191–208 (2017)
60. Laskey, K.J., Laskey, K.B., Costa, P.C.G., Kokar, M.M., Martin, T., Lukasiewicz, T.: Uncertainty reasoning for the world wide web. W3C Incubator Group Report (2008). https://www.w3.org/2005/Incubator/urw3/XGR-urw3-20080331/

61. Li, Y., Chen, J., Feng, L.: Dealing with uncertainty: a survey of theories and practices. IEEE Trans. Knowl. Data Eng. **25**(11), 2463–2482 (2012)
62. Li, V.G., Dunn, M., Pearce, P., McCoy, D., Voelker, G.M., Savage, S.: Reading the tea leaves: A comparative analysis of threat intelligence. In: 28th USENIX Security Symposium (USENIX Security 19), pp. 851–867. USENIX Association, Santa Clara (2019). https://www.usenix.org/conference/usenixsecurity19/presentation/li
63. Logg, J.M., Minson, J.A., Moore, D.A.: Algorithm appreciation: people prefer algorithmic to human judgment. Organiz. Behavior Human Decis. Proc. **151**, 90–103 (2019). https://doi.org/10.1016/j.obhdp.2018.12.005
64. Maathuis, C., Pieters, W., van den Berg, J.: Decision support model for effects estimation and proportionality assessment for targeting in cyber operations. Defence Technol. **17**(2), 352–374 (2021). https://doi.org/10.1016/j.dt.2020.04.007
65. Mandel, D.R.: Assessment and communication of uncertainty in intelligence to support decision-making. NATO STO TECHNICAL REPORT, TR-SAS-114 (2020)
66. Mandel, D.R., Irwin, D.: Facilitating sender-receiver agreement in communicated probabilities: is it best to use words, numbers or both? Judg. Decis. Making **16**(2), 363–393 (2021). https://doi.org/10.1017/S1930297500008603
67. Marlin, B.M., Abdelzaher†, T., Ciocarlie, G., Cobb, A.D., Dennison, M., Jalaian, B., Kaplan, L., Raber, T., Raglin, A., Sharma, P.K., Srivastava, M., Trout, T., Vadera, M.P., Wigness, M.: On uncertainty and robustness in large-scale intelligent data fusion systems. In: IEEE Second International Conference on Cognitive Machine Intelligence (CogMI), pp. 82–91 (2020). https://doi.org/10.1109/CogMI50398.2020.00020
68. Maymí, F.J., Thomson, R.: Human-machine teaming and cyberspace. In: Schmorrow, D., Fidopiastis, C. (eds.) Augmented Cognition: Intelligent Technologies, vol. 10915. Springer, Berlin (2018). https://doi.org/10.1007/978-3-319-91470-1_25
69. Medianovskyi, K., Pietarinen, A.V.: On explainable ai and abductive inference. Philosophies **7**(2), 35 (2022). https://doi.org/10.3390/philosophies7020035
70. Menkveld, C.: Understanding the complexity of intelligence problems. Intell. Natl. Secur. **36**(5), 621–641 (2020). https://doi.org/10.1080/02684527.2021.1881865
71. Montibeller, G., von Winterfeldt, D.: Individual and group biases in value and uncertainty judgments. In: Dias, L.C., Morton, A., Quigley, J. (eds.) Elicitation: The Science and Art of Structuring Judgement, vol. 261, pp. 377–392. Springer, Cham (2018)
72. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. ACM Comput. Surv. **55** (2023). https://doi.org/10.1145/3583558
73. Nisioti, A., Loukas, G., Laszka, A., Panaousis, E.: Data-driven decision support for optimizing cyber forensic investigations. IEEE Trans. Inf. Forens. Secur. **16**, 2397–2412 (2021). https://doi.org/10.1109/TIFS.2021.3054966
74. Nunes, I., Jannach, D.: A systematic review and taxonomy of explanations in decision support and recommender systems. User Model. User-Adapted Interact. **27**(3), 393–444 (2017)
75. Okoli, J.O., Weller, G., Watt, J.: Information processing and intuitive decision-making on the fireground: towards a model of expert intuition. Cogn. Tech. Work **18**, 89–103 (2016). https://doi.org/10.1007/s10111-015-0348-9
76. OTAN, N.: Automation in the intelligence cycle (2020). https://www.sto.nato.int/Lists/STONewsArchive/displaynewsitem.aspx?ID=552. Accessed 11 April 2023
77. Padilla, L., Kay, M., Hullman, J.: Uncertainty visualization. In: Piegorsch, W., Levine, R., Zhang, H., Lee, T. (eds.) Computational Statistics in Data Science, pp. 405–421. Wiley, Hoboken (2022)
78. Pagano, T.P., Loureiro, R.B., Lisboa, F.V.N., Cruz, G.O.R., Peixoto, R.M., de Sousa Guimarães, G.A., dos Santos, L.L., Araujo, M.M., Cruz, M., de Oliveira, E.L.S., Winkler, I., Nascimento, E.G.S.: Bias and unfairness in machine learning models: A systematic literature review. arXiv:2202.08176 (2022). https://doi.org/10.48550/arXiv.2202.08176

79. Pawlinski, P., Jaroszewski, P., Kijewski, P., Siewierski, L., Jacewicz, P., Zielony, P., Zuber, R.: Actionable information for security incident response. European Union Agency for Network and Information Security (2014)
80. Perry, W.L., McInnis, B., Price, C.C., Smith, S.C., Hollywoon, J.S.: Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. Rand Corporation, Santa Monica (2013)
81. Petersen, K.L., Tjalve, V.S.: Intelligence expertise in the age of information sharing: public–private 'collection' and its challenges to democratic control and accountability. Intell. Natl. Secur. **33**(1), 21–35 (2018). https://doi.org/10.1080/02684527.2017.1316956
82. Prabhudesai, S., Yang, L., Asthana, S., Huan, X., Liao, Q.V., Banovic, N.: Understanding uncertainty: How lay decision-makers perceive and interpret uncertainty in human-AI decision making. In: Proceedings of the 28th International Conference on Intelligent User Interfaces, pp. 379–396. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3581641.3584033
83. Rajivan, P., Cooke, N.J.: Information-pooling bias in collaborative security incident correlation analysis. Human Factors **60**, 626–639 (2018). https://doi.org/10.1177/0018720818769249
84. Ranade, P., Piplai, A., Mittal, S., Joshi, A., Finin, T.: Generating fake cyber threat intelligence using transformer-based models. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–9 (2021). https://doi.org/10.1109/IJCNN52387.2021.9534192
85. Reagans, R., Argote, L., Brooks, D.: Individual experience and experience working together: predicting learning rates from knowing who knows what and knowing how to work together. Manag. Sci. **51**(6), 869–881 (2005). https://doi.org/10.1287/mnsc.1050.0366
86. Regan, H.M., Colyvan, M., Burgman, M.A.: A taxonomy and treatment of uncertainty for ecology and conservation biology. Ecol. Appl. **12**(2), 618–628 (2002). https://doi.org/10.1890/1051-0761(2002)012[0618:ATATOU]2.0.CO;2
87. Rona-Tas, A., Cornuéjols, A., Blanchemanche, S., Duroy, A., Martin, C.: Enlisting supervised machine learning in mapping scientific uncertainty expressed in food risk analysis. Sociol. Methods Res. **48**(3), 608–641 (2019)
88. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson, Saddle River (2020)
89. Slayton, R.: What is the cyber offense-defense balance? Conceptions, causes, and assessment. Int. Secur. **41**(3), 72–109 (2017). https://doi.org/10.1162/ISEC_a_00267
90. Snidaro, L., Visentini, I., Bryan, K.: Fusing uncertain knowledge and evidence for maritime situational awareness via markov logic networks. Inf. Fusion **21**, 159–172 (2015). https://doi.org/10.1016/j.inffus.2013.03.004. https://www.sciencedirect.com/science/article/pii/S1566253513000523
91. Stevens, R., Votipka, D., Redmiles, E.M., Ahern, C., Sweeney, P., Mazurek, M.L.: The battle for New York: A case study of applied digital threat modeling at the enterprise level. In: 27th USENIX Security Symposium, pp. 621–63 (2018)
92. Tounsi, W., Rais, H.: A survey on technical threat intelligence in the age of sophisticated cyber attacks. Comput. Secur. **72**, 212–233 (2018). https://doi.org/10.1016/j.cose.2017.09.001
93. Tuma, K., Scandariato, R.: Two architectural threat analysis techniques compared. In: Software Architecture: 12th European Conference on Software Architecture, ECSA 2018, Madrid, Spain, September 24–28, 2018, Proceedings 12, pp. 347–363. Springer, Berlin (2018)
94. Tuma, K., Van Der Lee, R.: The role of diversity in cybersecurity risk analysis: An experimental plan. In: 3rd Workshop on Gender Equality, Diversity, and Inclusion in Software Engineering, GEICSE 2022, pp. 12–18. Institute of Electrical and Electronics Engineers (2022)
95. Tuma, K., Calikli, G., Scandariato, R.: Threat analysis of software systems: a systematic literature review. J. Syst. Softw. **144**, 275–294 (2018)

96. Tuma, K., Sion, L., Scandariato, R., Yskout, K.: Automating the early detection of security design flaws. In: Proceedings of the 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems, pp. 332–342 (2020)
97. Tuma, K., Sandberg, C., Thorsson, U., Widman, M., Herpel, T., Scandariato, R.: Finding security threats that matter: Two industrial case studies. J. Syst. Softw. **179**, 111003 (2021)
98. van der Kleij, R., Schraagen, J.M., Cadet, B., Young, H.: Developing decision support for cybersecurity threat and incident managers. Comput. Secur. **113**, 102535 (2022). https://doi.org/10.1016/j.cose.2021.102535
99. van der Voort, H., Klievink, A., Arnaboldi, M., Meijer, A.: Rationality and politics of algorithms. will the promise of big data survive the dynamics of public decision making? Govern. Inf. Quart. **36**(1), 27–38 (2019). https://doi.org/10.1016/j.giq.2018.10.011
100. Villiers, J.P.D., Laskey, J.P., Jousselme, A., Blasch, E., de Waal, A., Pavlin, G., Costa, P.: Uncertainty representation, quantification and evaluation for data and information fusion. In 2015 18th International Conference on Information Fusion. IEEE, pp. 50–57 (2015)
101. Villiers, J.P.D., Pavlin, G., Jousselme, A., Maskell, S., de Waal, A., Laskey, K., Costa, P., Blasch, E.: Uncertainty representation and evaluation for modeling and decision-making in information fusion. J. Adv. Inf. Fusion **13**, 198–215 (2018)
102. Vogel, K.M., Reid, G., Kampe, C., Jones, P.: The impact of ai on intelligence analysis: tackling issues of collaboration, algorithmic transparency, accountability, and management. Intell. Natl. Secur. **36**(6), 827–848 (2021). https://doi.org/10.1080/02684527.2021.1946952
103. Waardenburg, L., Sergeeva, A., Huysman, M.: Hotspots and blind spots. In: Schultze, U., Aanestad, M., Mähring, M., Østerlund, C., Riemer, K. (eds.) Living with Monsters? Social Implications of Algorithmic Phenomena, Hybrid Agency, and the Performativity of Technology, pp. 96–109. Springer International Publishing, Cham (2018)
104. Wagner, T.D., Mahbub, K., Palomar, E., Abdallah, A.E.: Cyber threat intelligence sharing: survey and research directions. Comput. Secur. **87** (2019). https://doi.org/10.1016/j.cose.2019.101589
105. Wei, L., Du, H., Mahesar, Q.A., Al Ammari, K., Magee, D.R., Clarke, B., Dimitrova, V., Gunn, D., Entwisle, D., Reeves, H., Cohn, A.G.: A decision support system for urban infrastructure inter-asset management employing domain ontologies and qualitative uncertainty-based reasoning. Expert Syst. Appl. **158**, 113461 (2020). https://doi.org/10.1016/j.eswa.2020.113461
106. Whitesmith, M.: The efficacy of ach in mitigating serial position effects and confirmation bias in an intelligence analysis scenario. Intell. Natl. Secur. **34**(2), 225–242 (2019). https://doi.org/10.1080/02684527.2018.1534640
107. Whyte, C.: Learning to trust skynet: Interfacing with artificial intelligence in cyberspace. Contempor. Secur. Policy **44**(2), 308–344 (2023). https://doi.org/10.1080/13523260.2023.2180882
108. Willingham, D.T., Riener, C.: Cognition: The Thinking Animal, 4th edn. Cambridge University Press, Cambridge (2019). https://doi.org/10.1017/9781316271988
109. Wirtz, J.J.: The sources and methods of intelligence studies. In: Johnson, L.K. (ed.) The Oxford Handbook of National Security Intelligence. Oxford University Press, Oxford (2010). https://doi.org/10.1093/oxfordhb/9780195375886.003.0004
110. Wu, J., Li, H.: Uncertainty analysis in ecological studies: An overview. In: Wu, J., Jones, K.B., Li, H., Loucks, O.L. (eds.) Scaling and Uncertainty Analysis in Ecology, pp. 45–66. Springer Netherlands, Dordrecht (2006). https://doi.org/10.1007/1-4020-4663-4_3
111. Xiong, W., Lagerström, R.: Threat modeling – a systematic literature review. Comput. Secur. **84**, 53–69 (2019). https://doi.org/10.1016/j.cose.2019.03.010

112. Zhao, K., Li, L., Chen, Z., Sun, R., Yuan, G., Li, J.: A survey: optimization and applications of evidence fusion algorithm based on dempster–shafer theory. Appl. Soft Comput. **124**, 109075 (2022). https://www.sciencedirect.com/science/article/pii/S1568494622003696. https://doi.org/10.1016/j.asoc.2022.109075
113. Zibak, A., Sauerwein, C., Simpson, A.C.: Threat intelligence quality dimensions for research and practice. Digital Threats Res. Practice **3**(4), 44 (2022). https://doi.org/10.1145/3484202