Human-Intelligent Systems Integration
https://doi.org/10.1007/s42454-020-00021-z

RESEARCH ARTICLE

Explainable artificial intelligence (XAI) interactively working with humans as a junior cyber analyst

Eric Holder[1] · Ning Wang[2]

Received: 9 July 2020 / Accepted: 4 December 2020

Abstract
There are many applications where artificial intelligence (AI) can add a benefit, but this benefit may not be fully realized, if the human cannot understand and interact with the output as required by their context. Allowing AI to explain its decisions can potentially mitigate this issue. To develop effective explainable AI methods to support this need, we need to understand both what the human needs for decision-making, as well as what information the AI has and can make available. This paper presents an example case of capturing those requirements. We explore how an operational planner (senior human analyst) for a cyber protection team could use a junior analyst virtual agent to scour, analyze, and present the data available on vulnerabilities and incidents on both the target systems as well as similar systems. We explore the interactions required to understand these outputs and to integrate additional knowledge held by the human. This is an exemplar case for integrating XAI into the real-world bi-directional workflow: the senior analyst needs to be able to understand the junior analysts results, particularly the assumptions and implications, in order to create a plan and brief it up the command chain. He or she may have further questions, or analysis needs to achieve this understanding. The application is the junior analyst agent and senior human analysts working together to create this understanding of threats, vulnerabilities, incidents, likely future attacks, and counteractions on the mission relevant cyber terrain that their unit has been assigned a mission on.

Keywords Human agent interaction · Explainable artificial intelligence · Transparency · Situation awareness · Cyber · Human factors

1 Introduction

As the DoD pushes towards the unified platform to conduct joint cyber operations, we need to ensure that cyber operators from across the services can work together to successfully execute missions. This will require effective communication and collaboration predicated upon a shared understanding of the cyber environment. The future operating environment will likely include input and output from AI and other intelligent agents in the workflow that will play a key role in this required understanding.

The focus of this overall project, sponsored by the ASD (R&E) Cyber Security and Applied Research Program, is on effective information sharing and visualization regarding the cyberspace components of mission command at combatant commands (CCMDs). The primary foci are on tools for providing information and visualization support to CCMD cyber protection teams (CPTs), mission owners, and network owners for the identification of cyber key terrain (KT-C) and mission relevant terrain-cyber (MRT-C) and the planning and tasks related to protecting and utilizing this key terrain (HQ USINDOPACOM 2019; Raymond et al. 2014).

The CCMD represents the ideal test case to examine the information sharing and information visualization constructs as multiple services come together and have to coordinate with each other, as well as other mission-relevant partners. For example, the CCMD might be run by the Army, while utilizing networks and systems owned by Navy and industry partners, and supported by an Air Force CPT. Each of these services and stakeholders brings in unique perspectives,

✉ Eric Holder
eric.w.holder4.civ@mail.mil

Ning Wang
nwang@act.edu

[1] Combat Capabilities Development Command, US Army Research Laboratory, Ft. Huachuca, AZ 85613, USA

[2] Institute for Creative Technologies, University of Southern California, Marina Del Rey, CA 32826, USA

Published online: 06 January 2021

techniques, and terminology that can impact both the information sharing and understanding across the team. Within the CCMD workflow dynamic, the planning process for CPT missions was chosen as the focus area as this is the stage where the stakeholders have to come together to ensure understanding of the specific, and overall, mission and how that relates to the terrain and threat landscape to produce the most effective plan to achieve the mission's intent.

Through extensive knowledge acquisition (KA) activities, including documentation review and more than 20 interviews with various CPT members and other cyber stakeholders, it was consistently explained to us that there were petabytes of data out there that could be valuable, but no one had the time or personnel to analyze it. One of these data sources was the incident data available in large databases but would need to be related to system components, threats, and vulnerabilities to be useful along with the requirement to understand the analysis and its implications and assumptions in order to create a plan and brief it. Therefore, we identified a value-added use case for artificial intelligence (AI) within the CCMD CPT planning process as part of the toolkit to support the planning process by mining and combining available data on the target and similar systems. The applied use case required applying explainable AI (XAI) methods to explain the results and assumptions of the AI analysis and get the planner up to speed on the target system in terms of what has happened (incidents, vulnerabilities, threat presence), likely follow on adversary activities and where to monitor, harden, or counteract those. This paper will explain the development of this use case and the information requirements for an AI-driven junior cyber analyst.

2 Background for XAI

Artificial intelligence (AI) is at the core of the future of Army technology. The US Department of Defense (DoD) is investing $2 billion to create human-like AI to be the Soldiers' "partners in problem-solving" (Walker 2018). Cyber security analysis tools will be powered by state-of-the-art AI. Such AI will work alongside the cyber operators on and off the battlefield. For many future use cases, including cyber operations, the understanding of the decisions of the AI and the rationale behind such decisions can be key to the success of the man-machine team. However, the complexity and the "black-box" nature of many AI algorithms create a barrier for establishing such understanding within their human counterparts. Without such understanding, a human interacting with such an AI system is likely to fall into the pitfall of misuse or disuse of the automation (Parasuraman and Riley 1997). For an AI to play an effective role in many human-machine teams, it must make its decisions understood by its human counterparts.

Early work in explainable AI (XAI) focused on generating explanations of expert decisions within rule-based and logic-based AI systems, not addressing the quantitative nature of much of the AI used today (Swartout and Moore 1993; Johnson 1994; van Lent et al. 2004; Core et al. 2006). More recent work on agent-based XAI used Markov decision processes (MDPs), the completely observable subclass of partially observable MDPs (POMDPs) (Elizalde et al. 2008; Dodson et al. 2011; Khan et al. 2011). The author's work on agent-based XAI was the first to develop the algorithms to automatically generate explanations based on POMDPs (Pynadath et al. 2016). More recently, as machine learning (ML) systems become more prevalent in our everyday life, there has been a surge of research into making their decision-making more transparent (Ribeiro et al. 2016; Hendricks et al. 2016; Guo et al. 2018). Some of these efforts have taken the approach of incorporating human-interpretable models, such as AND-OR trees (Si and Zhu 2013) or Bayesian networks (Shih et al. 2018), into the ML process. To address the needs to make AI more transparent, DARPA created an explainable AI program in 2016. In a review of the program mid-way through its 4-year course, the program director, Dave Gunning, has pointed out that state-of-the-art XAI research has enjoyed success in explaining the decisions behind, for example, recommendation systems for image recognition (Hendricks et al. 2016; Chang et al. 2018) and AI playing video games (Koul et al. 2018). However, generating explanations for automation with ML will remain a significant challenge for years to come, because of the fundamentally higher level of complexity of the AI decisions for automation relative to those for simpler recommendation systems (Gunning 2019). An AI-driven cyber analyst agent, as a decision-support automation, still presents a challenge to making its decision process transparent, given the state-of-the-art XAI research.

3 Background for CCMD cyber operations planner

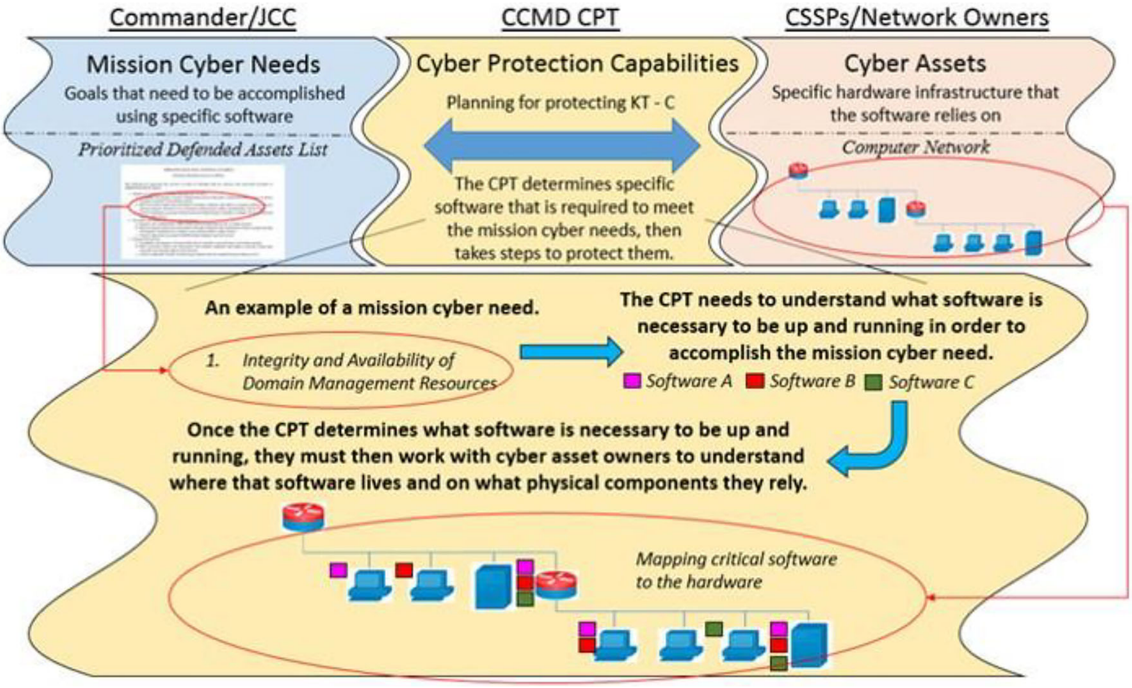
Cyber operations planners receive various forms of guidance for planning their missions. This can include high level joint doctrine, such as that provided by US Cyber Command (Allen 2015; USCYBECOMMAND 2020), guidance provided by service commands (e.g., Army Cyber Command), unit standard operating procedures, and lessons learned from on the job experience. Through our KA activities (literature review, interviews, observations) across various CCMDs and CPTs, it was clear that there is much variation between CCMDs and CPTs on the missions and planning process, largely in terms of the mission scope and specificity provided (e.g., vague tasking like go do cyber to specific tasking such as look at this one component); the organization of the planning responsibilities and experience levels (e.g., variations in rank and use of civilian deputies); the materials provided (e.g., old vs up-to-date network maps vs nothing); and the timeframe for both planning and response (e.g., 1- to 2-year plans vs quick reaction requirements). For example, CPTs supporting CCMDs in active combat regions often receive more quick reaction force taskings with less formality and detail, whereas others may receive missions a year or more out in advance.

It is also important to note that the current set-up is very unique in that a strategic level entity, CCMD, is tasking a tactical level unit (CPT) often bypassing the official operational level planning entity or steps. There is nowhere else that we have observed in the DoD planning process where this happens. Various approaches, both formal and informal, have been implemented to help fill this gap, but the overall result is that 2 types of planning activities have been required at the CPT operations planner level with various levels of support. The first type is the more internal and traditional military decision-making processes, and the second type is a more external planning piece to scope the mission and to organize the stakeholders, accesses, points of contact, schedule, and related tasks.

Across all CPTs, there is an initial step where the CPT operations planner, in conjunction with mission owner (typically representatives of the CCMD's Joint Staff and Joint Cyber Center, namely, the J-3 (operations cell) or J-6 (communication systems)) that provide tasking and clarification to the CPTs supporting the CCMD, and network owners need to get on the same page (see Fig. 1). The goal is for the planner to (1) understand and identify the KT-C/MRT-C which could be identified as task critical assets, systems, or other more or less detailed entities that describe what systems and components are essential to the mission; (2) understand the KT-C/MRT-C in terms of how data flows and supports the mission, as well as security and sensors already in place; (3) arrange access to both specific data and system components as required, especially if building a virtual training environment; (4) be able to pass this understanding along (hand-off) to the actual CPT mission team who will be doing the mission, and (5) in some cases help identify courses of action that best support the mission goals. Most planners reported that at this early stage, this is primarily an artifact and discussion-based review of the terrain, without yet getting hands on the actual network or pulling data. The initial information available can often be out of date or incorrect, and as the mission proceeds, a better understanding of the KT-C/MRT-C can be gained. As seen in the use case, this will be important for the user being able to update or edit the materials input into the process, as well as the analysis done by AI. The core point is that the better the CPT operational planner can understand the network, its status and vulnerabilities, the threats on it, and what these threats are likely to do at this early stage of planning, the better the planner can prepare the CPT mission team for their mission to go on that network and perform defensive actions that matter.

4 Developing the use case: methods and results

The methods and results here are discussed sequentially although there was significant iteration, interaction, and updating in parallel as the research proceeded.


Commander/JCC
CCMD CPT
CSSPs/Network Owners
Mission Cyber Needs
Goals that need to be accomplished using specific software
Prioritized Defended Assets List
Cyber Protection Capabilities
Planning for protecting KT - C
Cyber Assets
Specific hardware infrastructure that the software relies on
Computer Network
1. Integrity and Availability of Domain Management Resources
🟪 Software A
🟥 Software B
🟩 Software C
Mapping critical software to the hardware

The CPT determines specific software that is required to meet the mission cyber needs, then takes steps to protect them.

An example of a mission cyber need.

The CPT needs to understand what software is necessary to be up and running in order to accomplish the mission cyber need.

Once the CPT determines what software is necessary to be up and running, they must then work with cyber asset owners to understand where that software lives and on what physical components they rely.

Fig. 1 Stakeholders involved in CCMD CPT tasking

4.1 Identifying the use case

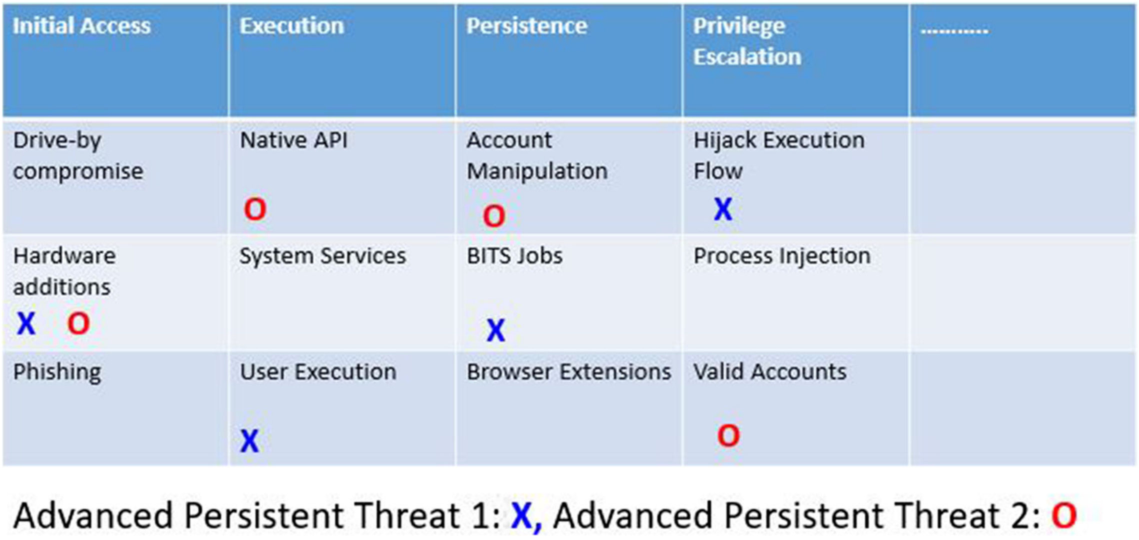
During the course of our KA activities, including literature review, interviews, and observations, in addition to learning about the planning workflow, information requirements, and constraints, we were also looking for areas where the advantages of AI could be combined with the needs of the operators for AI support and transparency to create a value-added application of XAI. A consistent theme across the KA activities was that there were massive amounts of data stockpiled that may contain valuable insights, but CPTs lacked the manpower and pressing priority to force analysis. The ability to analyze large amounts of data is a strength of AI compared to humans (O'Leary 2013). Further, AI is good at identifying patterns in data (Bishop 2006). It was clear from the interviews that cyber analysts and operators would have to explain their findings in terms of assumptions, sources, logic, and limitations and also conduct further analysis integrating additional data available to the humans. These are also characteristics that define a need for any AI applied to be explainable and bi-directional.

The input from subject matter experts really helped guide the use case. One discussion that was particularly informative described an idea for using the MITRE ATT&CK® model (MITRE 2018). The ATT&CK model lays out the steps of cyber attacks into a kill chain (initial access, execution, persistence, privilege escalation, defense evasion, credential access, discovery, lateral movement, collection, command and control, exfiltration, impact) with known techniques that are used at each step of the chain to accomplish that goal. For threats that are known, primarily advanced persistent threats (APTs), the model provides patterns of the techniques used by each specific threat at each step based on their history of attacks (see Fig. 2 for a simple example). The idea presented was to analyze the patterns of (APTs) in terms of the model's kill chain and how to combine that with system vulnerabilities to identify likely adversary courses of action (COAs) on a network or system of interest. This means that if using the ATT&CK model to see what threats typically do, you can see what they are currently doing on your system and predict what technique(s) they are likely to use next and, based on your system's vulnerabilities, the likely location on your network where they will try to do that. This also would help support incident response actions (what to harden to prevent further threat actions) and attribution of attacks and incidents. The SME provided a white paper to support this via personal communication.

The second was an interview with an instructor at the Intelligence Center of Excellence, Ft. Huachuca. She discussed the workflow process between junior and senior analysts and how those outputs are used. This highlighted the need for explainable outcomes and the concept of keeping the human as the senior analyst and using AI as the junior analyst agent to do the legwork and produce usable outputs to the senior analyst. The human senior analyst then uses the results to brief further up and down the chain of command. The senior analyst is going to have to explain and field questions on the results and suggestions he or she is making to the commander and will need to be able to provide those justifications and supporting details. This also included a need for the ability to let the human senior analyst drill down into the results provided by the junior analyst and "retask" the junior analyst to answer additional questions.

These inputs were combined with the rest of the KA insights to produce the following use case:

At the combatant command (CCMD) level, there will be a large number of networks within their area of operations that support the CCMD missions to various degrees. For each network and component, there are a stockpile of incident



Initial Access	Execution	Persistence	Privilege Escalation	....
Drive-by compromise	Native API O	Account Manipulation O	Hijack Execution Flow X	
Hardware additions X O	System Services	BITS Jobs X	Process Injection	
Phishing	User Execution X	Browser Extensions	Valid Accounts O	

Advanced Persistent Threat 1: X, Advanced Persistent Threat 2: O

Fig. 2 MITRE ATT&CK simplified threat pattern example

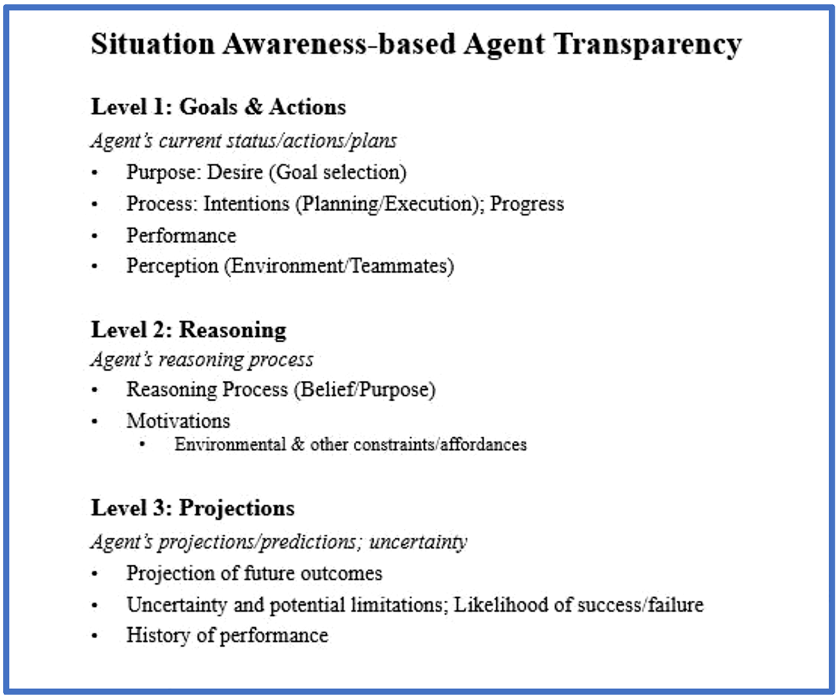
reports (e.g., via the Joint Incident Management System (JIMS) and other systems), some of which are investigated fully and others not. There are also databases of known exploits and vulnerabilities (such as the common vulnerabilities and exposures (CVE) database, Nessus, or the Air Force's Genesis tool). Threat models that include attack patterns and tactics, techniques and procedures (TTPs), and the signatures left behind, such as those based on the MITRE ATT&CK model, are also available. When assigned a mission, the cyber protection team (CPT) planner first identifies the networks and systems involved in that mission, the KT-C or MRT-C. These networks, systems, and components represent the base of the workflow for a junior cyber analyst agent.

An XAI-driven junior cyber analyst agent could be used to allow the human operator to identify the systems and components of interest (e.g., pulling directly from available data or manually entering). Then, the XAI-driven junior analyst agent would pull relevant data concerning exploits and vulnerabilities and the incident reports from the target system and similar systems/components to look for patterns that identify past attacks, likely attack patterns and next steps in attacks (adversary courses of action) and the threat actors present, along with the logic of these recommendations and additional recommendations on required actions (hardening, sensors or monitoring). AI should be able to "see" much more in terms of patterns of connections between large databases (e.g., incidents-JIMS, threat-ATT&CK, and vulnerabilities-CVE) than a human operator and do this much faster. The tool should provide an interface to allow modifications and drill downs into the analysis providing an interactive, bi-directional, component to include additional knowledge and further analyze and "what if" the results. This would also support coordination between the CPT operational planner and the CPT's intelligence group and with CYBERCOM as needed to integrate updated information (new threats or patterns, vulnerabilities, etc.) into the analysis. This will allow the senior analyst to plan what the CPT mission team should do next to either protect the mission, gather more information (sensors, from network owner, etc.), and also facilitate discussions with the mission owner and network owners.

4.2 Identifying the bi-directional information requirements

After identifying a use case and the need for XAI, the next step was to identify the information requirements (input, processing, and output) for both the AI and the human, along with the types of interactions that might be required to support this. The situation-awareness agent transparency (SAT) model (Chen et al. 2014) provided a base approach to accomplish these goals. The SAT model breaks out the situation awareness (SA) requirements for the human based on the outputs of the intelligent agent. See Fig. 3 for an overview of the SAT model. Chen and colleagues (Chen et al. 2014, 2018; Chen and Barnes 2015) define transparency in terms of understanding the internal underpinnings of the intelligent agent's (IA) courses of action (COAs). The SAT model defines the agent's suggested COAs as comprising three transparency levels (L): the agent's perception of its plan (L1), its logic (L2), and its predicted outcomes and their perceived likelihood (L3). SAT

Fig. 3 SAT model reproduced with permission from Chen et al. (2018)



Situation Awareness-based Agent Transparency

Level 1: Goals & Actions

Agent's current status/actions/plans

- Purpose: Desire (Goal selection)
- Process: Intentions (Planning/Execution); Progress
- Performance
- Perception (Environment/Teammates)

Level 2: Reasoning

Agent's reasoning process

- Reasoning Process (Belief/Purpose)
- Motivations
  - Environmental & other constraints/affordances

Level 3: Projections

Agent's projections/predictions; uncertainty

- Projection of future outcomes
- Uncertainty and potential limitations; Likelihood of success/failure
- History of performance

Table 1 Intelligent agent's (IA) level 1 SA, logic, and human interactions

Level 1: data used in analysis (perceptual pieces)	Reasoning logic (filtering)	Human interaction[1]
- Incident reports: # of incident reports included (perhaps excluded)	Commonalities: What components of target system were matched (hardware, software, network components, etc.), why included or excluded	Drill down to examine what was included and why. Include additional, or exclude, categories/classes of reports or weight them
- System types and components: analyzed/included	Similarity: Why systems were included (similarity factors)	Drill down on factors to understand what was included and why, and include additional, or exclude, systems and related data or weight them
- Vulnerabilities and exploits: other databases and data samples included (vulnerabilities, exploits)	Relevance: Why databases and data included/excluded (relevance to target system)	Drill down on factors to understand and include additional, or exclude, databases or data in analysis or weight them
- Known threats (patterns, signatures) considered	Relevance to protected mission: Why threats included/excluded	Drill down on factors to understand and include additional, or exclude, threats (patterns) in analysis or weight them. Modify/update threat patterns based on new intelligence not in model

[1] Note that many of these choices could be made before running the analysis (what to include/exclude) and what parts of the target system are the focus of the planning. These could then be filtered in and out when reviewing the analysis

is similar to Endsley's (2015) original SA model but derived from the IA's output perspective. The SAT model has been tested in three diverse military paradigms, showing improved calibration of trust and performance (reduced misuse and disuse of autonomously generated COAs) for an agent that conducted perimeter defense (Mercado et al. 2016), infantry support (Selkowitz et al. 2016), and convoy route planning (Wright et al. 2016).

We modified the SAT analysis to break down the workflow process of the IA into 3 levels of SA (1 = perception and base data, 2 = reasoning and assessment of system's current state, 3 = projecting to the future state) and then map in parallel how the human would be presented with, and interact with, that output to achieve their levels 1–3 SA (see Fig. 4). In our application the XAI was not just producing COAs so we had to capture additional elements of the analysis and output steps. See Tables 1, 2, 3, and 4 for an overview of the process and the output breaking down the IA's information requirements by the 3 levels of SA, along with documenting likely uncertainties in the analysis. It was important to map out the IA's inputs, reasoning, and outputs to support the future task of identifying the best AI methods to apply that could both effectively do that processing and yet provide the output and additional information (explanations) that the human required to be able to use this output for their levels 1–3 SA and to support their tasks and workflow. This task of identifying the best AI methods has not yet been finished in this project but will be in the coming months.

Table 2 IA's level 2 SA, logic, and human interactions

Level 2: assessment of system (key terrain) and current state	Reasoning logic	Interaction
- Target system vulnerabilities: identified vulnerabilities and weak points in target system based on system status and past incidents on target and similar systems, and vulnerabilities and known exploits database	Vulnerabilities prioritization reasoning: based on what data and factors-e.g., used by most threats, known vulnerability, exploits available, patched, etc., why similar systems were relevant	Filter for likelihood of exploitation by Threat N-Z, impact of patches, redundancies, etc. The results can also alert CPT of questions to be answered by network owner (e.g., was this patch, STIG, etc. implemented) as it might be unknown to the AI
- Relevant incidents: overlays or lists of past incidents on target system and similar systems and key components attacked/compromised and patterns	Explain any relevance logic, especially if system components have changed or differences between target and non-target system	Filter by date, severity, systems included, etc.
- Threats: overlays/list of likely threats working on system based on past attack patterns on target and similar systems and area of operations	Relate patterns to threats and target system components and highlight the factors that differentiate between threats	Filter or sort incident by parameters (IP, components, systems included) combine across attacks to match to larger patterns (e.g., heat map), conduct if/then analysis on adding or removing components to see if it changes the threat picture

Table 3 IA's level 3 SA, logic, and human interactions

Level 3: future state: recommendations and predictions	Reasoning logic	Interaction
- Likely threat courses of action on target system: scaled by likelihood based on each threat, based on MITRE ATT&CK patterns or other	COAs matched to threat(s) and how determined/matched from past attacks to target system (e.g., how an attack step such as file and directory discover applies to target system)	Compare COAs and rank on likelihood/severity/etc.
- Named areas of interest: to monitor (place sensors) and behaviors to look for to differentiate between COAS	Overlay the threat attack patterns and highlight how this NAI differentiates between threats (e.g., Threat X enters via the router, where Threat Y uses URL viruses)	Overlay and compare threat patterns, edit patterns as required, save results

By identifying the inputs, processing, outputs, reasoning, and possible interactions, we were prepared to examine how those interactions might be portrayed to the human user and what AI techniques might support this. At this stage, we were not able to identify in detail the level 3 SA requirements (projecting into the future) for human users as the AI had already made its predictions to the human as the human's level 1 input. Level 3 information is not always required, but it was left as something to track and add, if needed, as the work proceeded.

Table 1 captures the level 1 (perceptual) requirements of the junior cyber analyst agent. For an AI, this was translated into what data from the databases (incidents, vulnerabilities, threat) it would use in the analysis (at level 2) and the explainable aspects of this focused on why those data items were selected or not. This is particularly important for data where there is matching required, such as how target systems and components were matched with similar systems for analysis or what vulnerabilities apply to the target system. This is a step where the human analyst would be expected to drill back down into that data and filter, add, or weight different aspects of the data to get the analysis and information he or she requires. It is expected that the human might also have information or intelligence that is not already captured in the system and be able to use that to impact the analysis. This could include updating things like the network map to better identify the components or interest, or knowing about new vulnerabilities or vulnerabilities that have been patched, or emerging threats and threat models to load into the XAI-driven junior cyber analyst agent.

Table 2 captures the level 2 (reasoning) done by the junior cyber analyst agent. This is the base analysis, taking the data included from level 1 and determining what does that mean to the target system in terms of incidents, vulnerabilities, and threats. This is answering where is the system vulnerable based on the system make-up and components and known vulnerabilities found in the database. This includes analyzing the incidents on your target system components, as well as incidents on similar systems, to see where your system is getting attacked and how, and if, this matches other similar systems or if there might be incidents you are missing. The threats are analyzed by looking at the incidents on your system in comparison to known patterns by threat actors to help determine who is active on your system. This information is then combined to look for patterns and help identify the priority system components to look at (e.g., vulnerability is present on your system, has been exploited and is known to be used by the threat actors seen on your system). The explainability of the junior cyber analyst should provide insight into why vulnerabilities were prioritized, what made incidents relevant or not (e.g., changes over time), and highlight the factors that differentiated between different threats. The human is expected to interact by filtering and sorting this information, exploring the impact of vulnerabilities and patches, and conducting if/then analyses to look at the impact on the threat picture.

Table 3 captures the level 3 (projection into the future) done by the junior cyber analyst agent. The core of this analysis is to go beyond the current state of the target system to identify predictions for what is likely to happen next. This includes looking at the known threat patterns and their current known activities to identify likely next steps (courses of

Table 4 Modeling uncertainties, logic, and human interactions

Uncertainties	Logic	Interaction
- Uncertainties based on deception to imply other threat or decoy from main attack, data limits (timeframes included), how close of a match to prior patterns, new attack patterns (similarity), if target system has implemented various cybersecurity measures	Portray information to allow user to know where the analysis is most/least certain and identify the key factors that, if changed, would impact the results (e.g., adding security measures, updates, new threat, etc.)	Drill down to understand uncertainties. Rerun model to "what if" changes to key factors (e.g., with new threat model or including systems that have been patched for comparison)

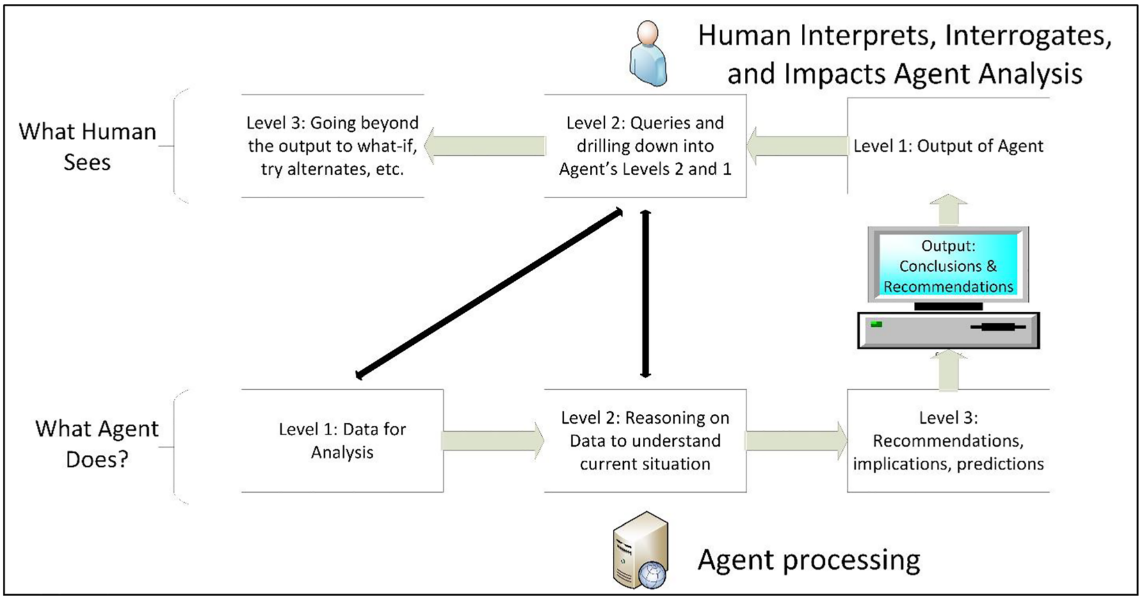

Human Interprets, Interrogates, and Impacts Agent Analysis
What Human Sees
Level 3: Going beyond the output to what-if, try alternates, etc.
Level 2: Queries and drilling down into Agent's Levels 2 and 1
Level 1: Output of Agent
Output: Conclusions & Recommendations
What Agent Does?
Level 1: Data for Analysis
Level 2: Reasoning on Data to understand current situation
Level 3: Recommendations, implications, predictions
Agent processing

Fig. 4 Modified SAT model as applied to our use case

action) and based on those threats' preferred techniques and your system make-up determine where those next steps are likely to happen. The explainable components of the AI include transparency into how it came to these conclusions in terms of references to the threat models and how that was matched to the target system. The human will want to compare different possible attack patterns to help determine where the CPT mission team will need to focus and also identify ways to confirm or deny certain threats are present when there is ambiguity. The human might also want to edit the parameters of this analysis based on additional information he or she has concerning threats and their adaptations. The human planner needs to evaluate and produce enemy courses of action and rank them by likelihood and risk and this output will help support that, for example, identifying named areas of interest, which would be where to look for techniques, or the targeting of specific system components, that would differentiate between suspected threat actors.

Table 4 separately addresses uncertainty information in the analysis. This has been included in prior work on using the SAT model (Mercado et al. 2016), and making uncertainty information available has shown an impact on performance. Therefore, we want to make sure it was addressed in the information and explanations available. The primary uncertainties center on helping identify confidence in any of the results provided, any assumptions made, and taking into account factors such as data limitations, the degree of match between systems or threat profiles, and adversary deception that might attempt to hide their presence or activities. In terms of explanations, this keys on making the human aware of where the analysis results are most, or least, certain and why. The human could filter the analysis on parameters such as confidence or drill down to understand the uncertainties. The cyber threat world is constantly adapting so their will consistently be a degree of mismatch between past behaviors and techniques and future techniques, and the AI and human have to work together to understand and interpret those. There can also be uncertainties in the analyses on factors, such as which vulnerabilities have been patched and to what degree, and this is something that the human operator may be able to obtain information on and add to alter the analysis. It can also provide the human operator the ability to what if the analysis even if ground truth is not known to answer questions such as what if this vulnerability exists or is removed or if the threat tries this instead.

4.3 Information portrayals

The next step, after reviewing the SAT model content with operational SMEs, was to turn that mapping into information portrayal examples to examine how the human user would receive information from the junior cyber analyst and then interact with it to build his or her SA. These were mocked up using Microsoft Visio to represent user interface content and options and to allow walk-through sessions with SMEs. As you can see, these were mapped fairly consistently onto the SAT model outputs produced in the prior step (Tables 1, 2, 3, and 4). Some examples of user interfaces, interactions, and the reasoning are provided in the following to show the basic process.

As depicted conceptually in the model, the human SA building starts with the output from the junior cyber analyst. In reality, a human may have already interacted with the system to define the key terrain (systems and components of interest on the target and similar systems) and input some of the settings, but this step of starting with the XAI output works as a good conceptual starting point for describing the flow.

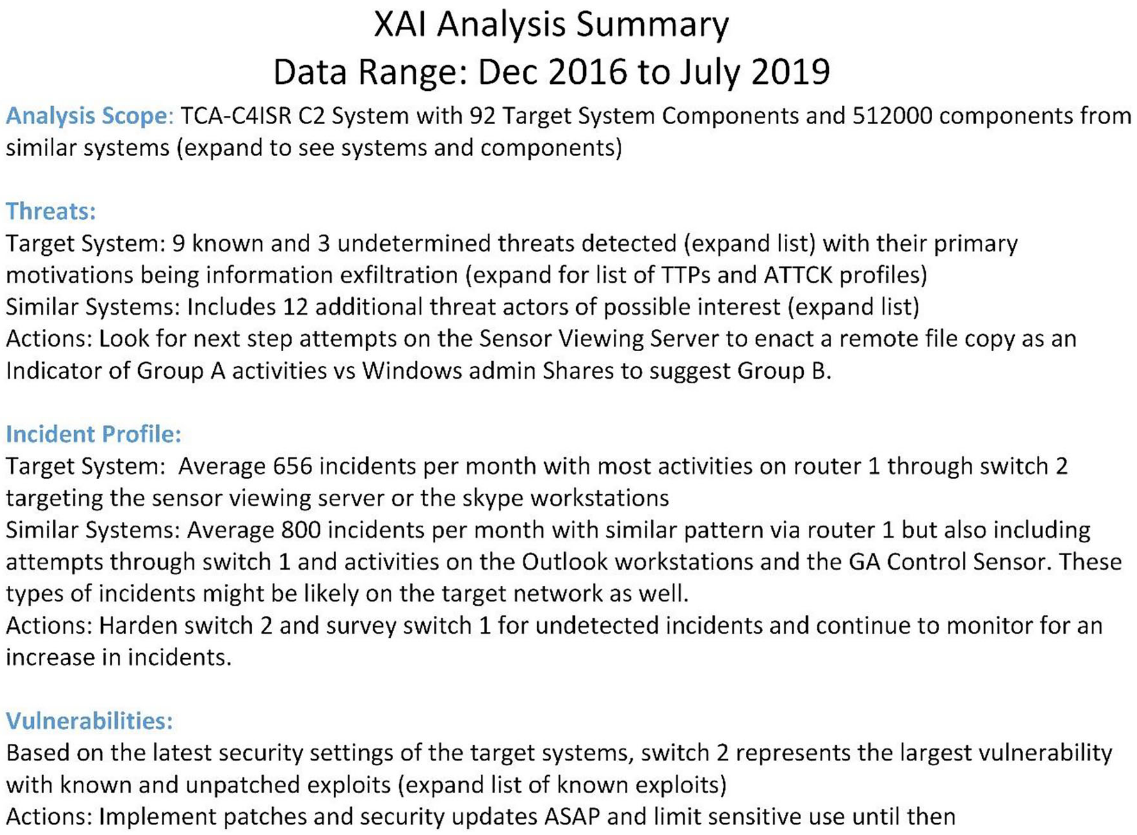
The first output screen, see Fig. 5, consists of a summary of the analysis conducted. This is envisioned to provide an overview of what was included and what was found and give the human analyst a starting point to branch off into more detailed analyses as warranted by their task and information needs. A summary of scope defines at a high level what was included in the analysis in terms of the target and similar systems. The results for each component of the analysis, threats, attacks/incidents, and vulnerabilities, are summarized for the target system and similar systems and then augmented with the junior cyber analyst's recommended actions. This gives an overview of the junior cyber analyst agent's levels 1, 2, and 3 SA and outputs.

Figure 6 presents a mock-up portrayal of incidents on the target system with larger circles indicating more incidents. The analyst could click on the circles to drill down into the incidents for information concerning the threat, TTPs, and vulnerability exploited. This portrayal also shows the various filter options available: date, vulnerabilities, target, or similar systems and by known threats and stages in the attack cycle.

Figure 7 provides a view if the analyst wants to see both the target and similar systems simultaneously with red and yellow combining to make orange for common incidents. A non-color coding method such as patterns is also being explored as are other alternate presentations. The goal was to allow the analyst to see similarities but also differences as a way to learn from other systems to identify incidents that were perhaps missed by prior analyses or represent foreshadowing of future incidents and attacks.

Figure 8 shows how vulnerability information could be shown as well, in this case overlaid on the incident information. The vulnerability portrayal (3-level triangle) was created as a simple way to show 3 levels of vulnerability (high, medium, and low). It was not a standard portrayal, but we needed a representation that could be shown simultaneously with the incidents. The levels were envisioned to map onto a combination of ease of use by adversaries and amount of damage (risk) if exploited. The plan was to confirm the presentation and levels with SMEs in a follow-on review step. In the example shown, you can see a drill down onto a vulnerability to gather more information. The date slider in combination with the vulnerability portrayal was seen as one way to explore how incidents and vulnerabilities were related to system updates and patches. This can allow the senior human analyst to understand the history but also filter out attack vectors (vulnerabilities) that are no longer relevant.

Figure 9 portrays an example user interface for modifying the analysis and parameters. This screen allows for loading or editing a network map, modifying the threat list, and managing what is considered a similar system of interest. It was not



XAI Analysis Summary
Data Range: Dec 2016 to July 2019

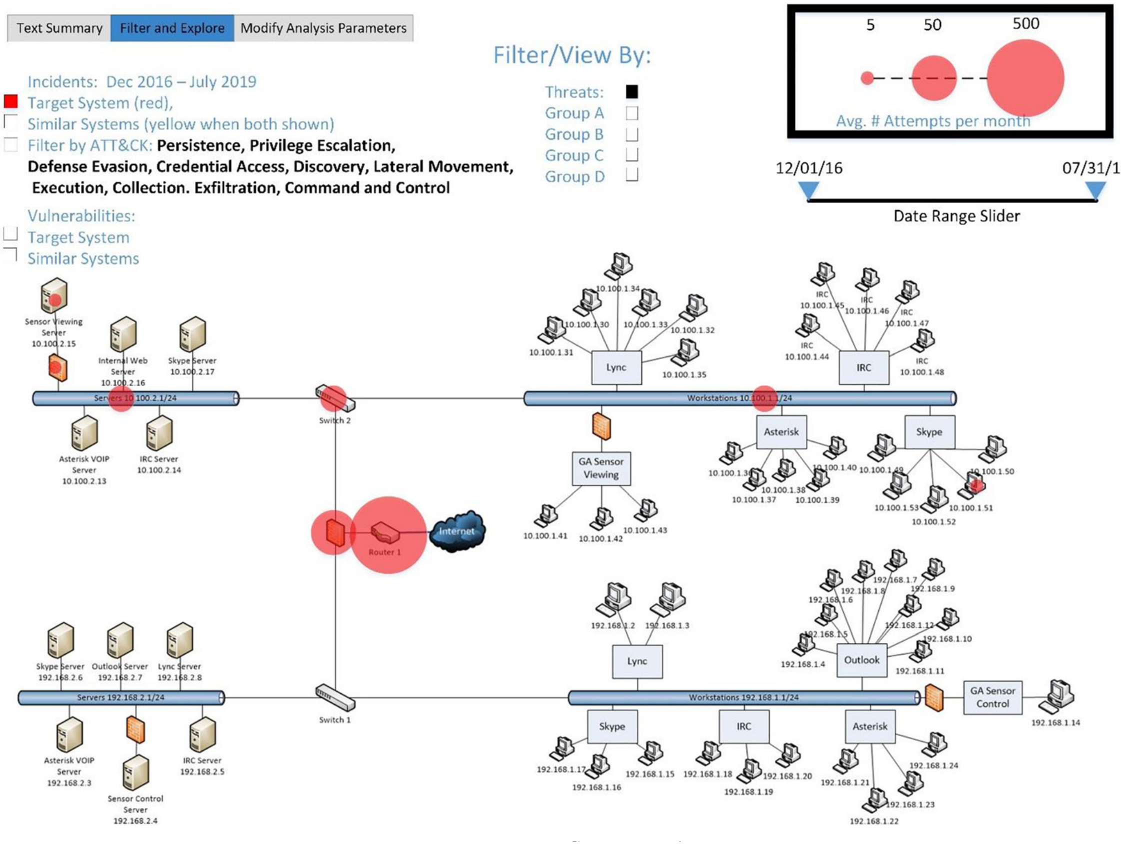
Analysis Scope: TCA-C4ISR C2 System with 92 Target System Components and 512000 components from similar systems (expand to see systems and components)

Threats:
Target System: 9 known and 3 undetermined threats detected (expand list) with their primary motivations being information exfiltration (expand for list of TTPs and ATTCK profiles)
Similar Systems: Includes 12 additional threat actors of possible interest (expand list)
Actions: Look for next step attempts on the Sensor Viewing Server to enact a remote file copy as an Indicator of Group A activities vs Windows admin Shares to suggest Group B.

Incident Profile:
Target System: Average 656 incidents per month with most activities on router 1 through switch 2 targeting the sensor viewing server or the skype workstations
Similar Systems: Average 800 incidents per month with similar pattern via router 1 but also including attempts through switch 1 and activities on the Outlook workstations and the GA Control Sensor. These types of incidents might be likely on the target network as well.
Actions: Harden switch 2 and survey switch 1 for undetected incidents and continue to monitor for an increase in incidents.

Vulnerabilities:
Based on the latest security settings of the target systems, switch 2 represents the largest vulnerability with known and unpatched exploits (expand list of known exploits)
Actions: Implement patches and security updates ASAP and limit sensitive use until then

Fig. 5 XAI junior analyst analysis summary screen



Incidents: Dec 2016 – July 2019

🟥 Target System (red),

□ Similar Systems (yellow when both shown)

□ Filter by ATT&CK: **Persistence, Privilege Escalation, Defense Evasion, Credential Access, Discovery, Lateral Movement, Execution, Collection. Exfiltration, Command and Control**

Vulnerabilities:

□ Target System

□ Similar Systems

Threats: ■

Group A □

Group B □

Group C □

Group D □

Fig. 6 Target system incidents

felt that modification needed to be restricted to just this screen. For instance, circling areas of interest or excluding aspects could be provided on any screen. This portrayal provided the ability to discuss those with SMEs in one place though if not covered elsewhere. As noted before, these portrayals were not meant to be functioning user interfaces but rather a means to get SMEs to talk through how such a system would be used and the information requirements. In this case, one of the primary goals was to drill into how "similar" would be defined in the interviews that followed.

5 Matching information needs to AI methods

As networks of hosts continue to grow in both size and criticality to operations, evaluating their vulnerability to attacks becomes increasingly more important to automate. The human-automation interaction in the cyber application described here centers around the vulnerability and threat analysis based on the MITRE ATT&CK model. Specifically, the research utilizes AI methods to analyze the past cyber incident reports for the target network and similar networks based on the publicly known adversary tactics and techniques described in the MITRE ATT&CK repository and known vulnerabilities. As of June 2020, the MITRE ATT&CK repository provides a total of 174 techniques belonging to 15 preattack tactics and 266 techniques belonging to 12 post-exploit tactics (MITRE, 2020). A tactic is a behavior that supports a strategic goal; a technique is a possible method of executing a tactic. Each technique is performed through various procedures. A sequence of techniques from different tactics used for an attack is called a TTP (tactics, techniques, procedures) chain. The combination of MITRE ATT&CK techniques in a TTP chain represents various attack scenarios that can be composed in an attack graph (Jha et al. 2002). The ATT&CK dataset can be used to construct attack graphs that are associated with different known threat groups to represent their standard TTPs.

Attack graphs can serve as a basis for detection, defense, and forensic analysis. All systems will exhibit at least some vulnerabilities, and many of these are known and/or discovered and reported. When evaluating the security of a network, it is not enough to consider the presence or absence of isolated vulnerabilities. A large network builds upon multiple platforms and diverse software packages and supports several modes of connectivity. The assessment of the vulnerability of a network of hosts should consider the effects of interactions of local vulnerabilities and include global vulnerabilities introduced by interconnections. Scanning tools can determine the vulnerabilities of individual hosts. An attack graph can

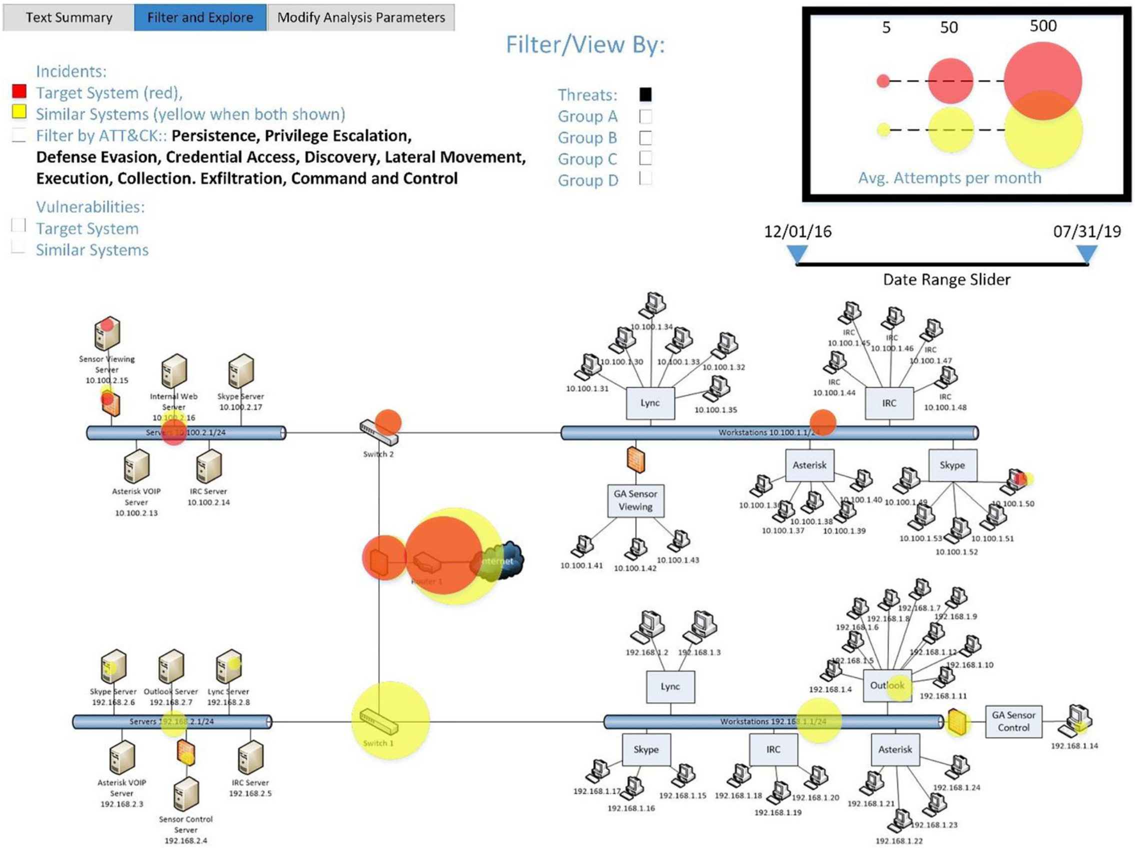
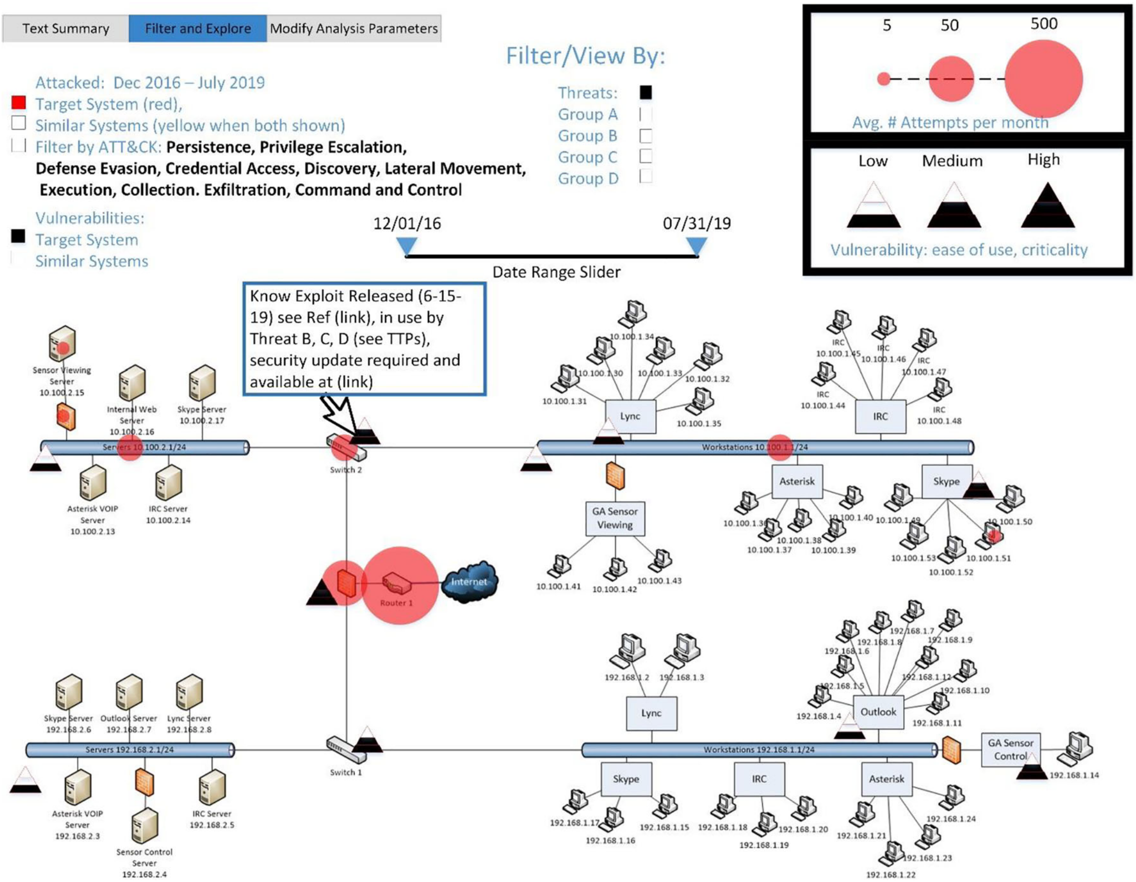


Fig. 7 Incidents on target and similar systems

express the local vulnerability information along with other information about the network, such as connectivity between hosts. Each path in an attack graph is a series of exploits that lead to an undesirable state (e.g., a state where an intruder has obtained administrative access to a critical host). An attack graph is a succinct representation of all paths through a system that end in a state where an intruder has successfully achieved his goal.

The combination of MITRE ATT&CK techniques in a TTP chain represented in an attack graph captures the life cycle of an attack. Lockheed Martin first described the cyber attack life cycle as the cyber kill chain that composes of seven stages: reconnaissance, weaponize, deliver, exploit, control, execute, and maintain (Hutchins et al. 2011). The tactics in ATT&CK follow this life cycle as well. The 15 tactic categories for preattack were derived from the first two stages (recon and weaponize), and the 11 post-exploit tactic categories within ATT&CK were derived from the later stages (exploit, control, maintain, and execute) of a seven-stage cyber attack life cycle (MITRE, 2020). An attack sequence would involve at least one technique per tactic, and a completed (post-exploit) attack sequence would be built by moving from left (initial access) of the ATT&CK matrix to right (command and control). It is possible for multiple techniques to be used for one tactic. For example, a well-known attack group APT28 might try both spearphishing attachment and spearphishing link technique as initial access tactics. It is not necessary for an attacker to use all 11 post-exploit tactics. Rather, the attacker will likely use the minimum number of tactics to achieve the objective, as it is more efficient and provides less chance of discovery. For example, APT28 may perform initial access to the credentials of an administrative assistant using a spearphishing link technique delivered through an email. Once they have the admin's credentials, APT28 can look for documents through file and directory discovery in the discovery stage. If the data APT28 is after is in a folder to which the admin also has access, then there is no need to go through the privilege escalation phase. In the end, APT28 could use various techniques in the collection phase, such as data from local system, to download files to their own machine.

The cyber incident report data can represent different scenarios in the attack graphs built using the ATT&CK model. While the US Military has a stockpile of reports and data on cyber incidents, we choose to not use the real incident reports dataset for the research of this project, due to the security restrictions and sensitive nature of such data. Alternatively, we created a simulated dataset based on the structure of the cyber incident report used in the military. Each incident report describes the technical specification of the target system, techniques used by the adversary, detection methods, impact on


Text Summary
Filter and Explore
Modify Analysis Parameters
Filter/View By:
Attacked: Dec 2016 – July 2019
Target System (red),
Similar Systems (yellow when both shown)
Filter by ATT&CK: Persistence, Privilege Escalation, Defense Evasion, Credential Access, Discovery, Lateral Movement, Execution, Collection. Exfiltration, Command and Control
Threats:
Group A
Group B
Group C
Group D
Vulnerabilities:
Target System
Similar Systems
12/01/16
07/31/19
Date Range Slider
5
50
500
Avg. # Attempts per month
Low
Medium
High
Vulnerability: ease of use, criticality
Sensor Viewing Server 10.100.2.15
Internal Web Server 10.100.2.16
Skype Server 10.100.2.17
Servers 10.100.2.1/24
Asterisk VOIP Server 10.100.2.13
IRC Server 10.100.2.14
Switch 2
Router 1
Internet
Lync
IRC
Workstations 10.100.1.1/24
Asterisk
Skype
GA Sensor Viewing
Skype Server 192.168.2.6
Outlook Server 192.168.2.7
Lync Server 192.168.2.8
Servers 192.168.2.1/24
Asterisk VOIP Server 192.168.2.3
Sensor Control Server 192.168.2.4
IRC Server 192.168.2.5
Switch 1
Workstations 192.168.1.1/24
Outlook
GA Sensor Control
192.168.1.14

Know Exploit Released (6-15-19) see Ref (link), in use by Threat B, C, D (see TTPs), security update required and available at (link)

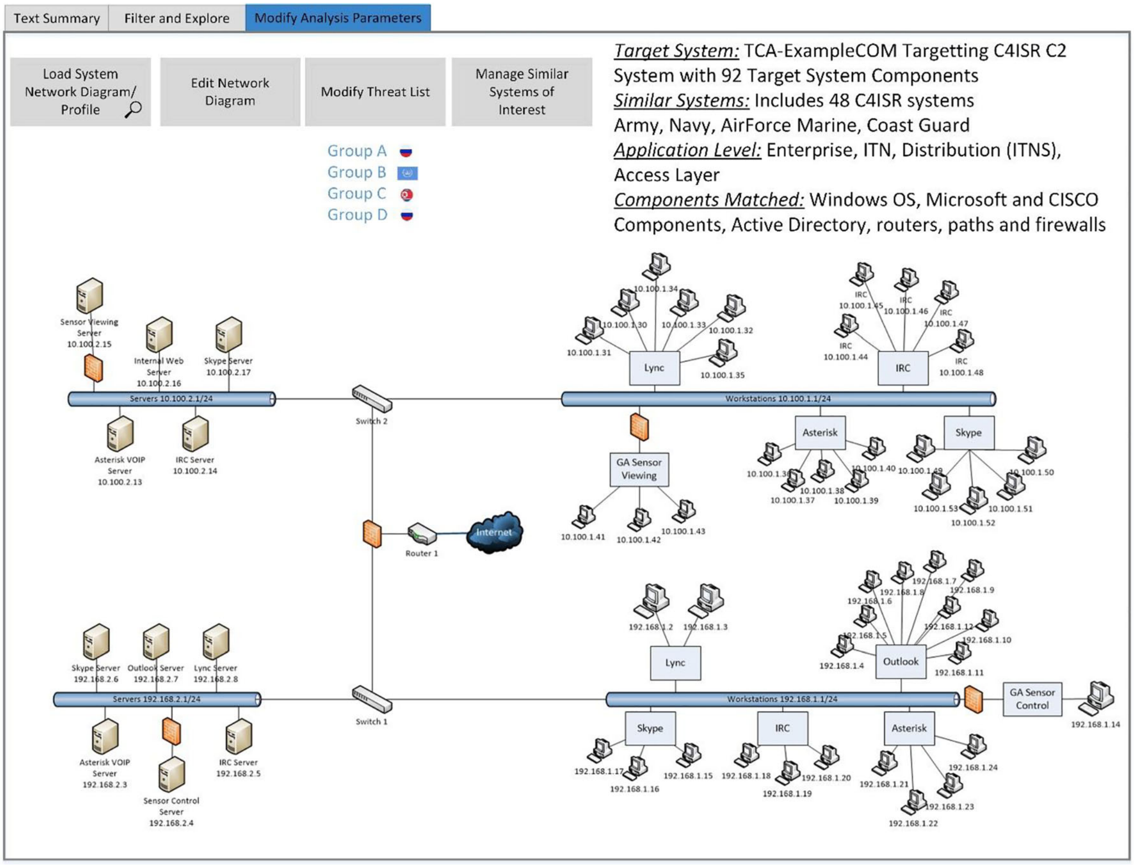
Fig. 8 Incident and vulnerabilities combined

the target system and the mission, and other technical and non-technical information related to the incident. For example, some non-technical information includes military categorization of the attack, date and time of incident and reporting, and impact on the operational unit and the mission. The technical details of the incident can be simulated using the data categories from the ATT&CK repository. For example, the "technique, tool, or exploit used" by the adversary can map onto the techniques, such as power shell and network sniffing, described in the MITRE ATT&CK repository. The "root causes", "method of detection", and "mitigation strategies" can be simulated using the "software", "detection", and "mitigations" subfield in the description of a technique in the ATT&CK matrix.

The goal of cyber incident data analysis is to understand (1) who might be behind the attack and (2) what might happen next if the vulnerability is left unaddressed. AI can be used to automate such processes. For example, a simple method to infer the identity of the adversary is through probabilistic inference. In the ATT&CK model, each group is known for using a subset of techniques within each tactic. Each technique is achieved via software, which is developed or frequently used by certain groups. Thus, given a technique used during an incident in the dataset, there is a Technique ➔ Group probability distribution, based on the description of the ATT&CK matrix. This distribution describes the likelihood that a specific group(s) is behind the attack. Such probabilities update as more techniques are revealed through the sequence of attack incidents in this time series incident data. Alternatively, simple heuristics can be used to nominate the most likely group behind the attack, e.g., the timing and overlap of techniques identified in the incident reports and the techniques used by an adversary group.

Alternatively, the identity of the adversary in the past attack can be inferred through policies derived from supervised learning on the ATT&CK dataset. Using the tactics as features, techniques as feature values, and the adversary as label, identifying the adversary becomes a classification task for machine learning.

Upon identifying the adversary, we can thus infer possible next steps in the attack, if the vulnerability is left unaddressed. Another method to understand what might happen next to the target network is to analyze what commonly happens together in an attack based on the behavior of the adversary described in the MITRE ATT&CK repository. The analysis of commonly associated steps can be achieved using unsupervised machine learning on incident reports, such as clustering (Al-Shaer et al. 2020).



Target System: TCA-ExampleCOM Targetting C4ISR C2 System with 92 Target System Components
Similar Systems: Includes 48 C4ISR systems Army, Navy, AirForce Marine, Coast Guard
Application Level: Enterprise, ITN, Distribution (ITNS), Access Layer
Components Matched: Windows OS, Microsoft and CISCO Components, Active Directory, routers, paths and firewalls

Fig. 9 Set-up or modify analysis

One of the next steps is to develop explanations in both verbal and visual form on how AI arrives at the decision of who was behind the attack and what might happen next. Some methods are easily understandable to humans and can be used to generate explanations, such as Bayesian network, decision trees, and clustering. The attack graphs are already good explanations of patterns that can be matched to known threat groups. We are currently developing a simulation testbed, where a human subject plays the role of a senior analyst, to perform network vulnerability analysis and implement mitigation strategies. Data from human subject studies with the simulation testbed on how such explanations can facilitate cyber operator's decision-making will inform the efficacy of the explanations and the best techniques to use.

6 Operator feedback on portrayals and concept

To date, we have conducted 3 SME review sessions on the portrayal concepts that were produced and workflows, including 10 SMEs. For each session, the SMEs were walked through an overview of the project and use case and then a series of screen shots of the portrayals. SMEs were encouraged to discuss how they would interact with the "features" of each portrayal, how it would support or conflict with their workflow for different missions, additional information required, data sources, and suggested improvements. The feedback is currently being used to update the portrayals and build out the concept of use for such a system as the project moves forward.

Overall the feedback was very positive on the concept as a way to improve SA. Several other tools and programs to integrate with were identified as well. Some of the feedback provided was specific to threats or applications and cannot be included in an unclassified publication or were outside of the use case provided but examples of high-level feedback relevant to the use case are provided.

Across the SMEs, the need to edit or adapt the information was a consistent request and speaks to the need for a bi-directional analysis tool. For example, known, historical vulnerabilities were seen to be exploitable by any knowledgeable threat (e.g., can be bought and used). Therefore, adding updated vulnerability information, as well as adversary TTPs for exploiting systems, the senior human analyst could obtain from their intelligence sources; colleagues or learned/war-gamed by their offensive cyber counterparts would add value. Similarly the network maps were frequently seen to be out of date and in need of updating or editing based on new information. Being able to compare the discrepancies between various network maps or representations would support analysis as well. Another aspect to be able to add/edit or filter would be patterns that are known to be benign (e.g., this incident happens every time the system is started up). This was seen as a way to weed out a lot of false alarms and noise in the data. The need for the tool to facilitate communications with other stakeholders, such as the network owners, was also consistently brought up. This included examples such as getting feedback from the network owners on the unknowns identified. Therefore, the ability to share or export the analysis should also be a core component.

The discussion of similar systems highlighted that this will be very context and mission dependent and provided other examples of sorting/matching categories. The function/architecture of the system was one of those categories, for instance, between enterprise level down to access layer characterizations could have a large impact on what the senior analyst is looking at and concerned with. It was felt that the threat actors and TTPS can be specific to these categories. Providing an automatic filtering of sub-options by the choice of higher level options (e.g., operating system) was also suggested.

The use of the threat functionality (e.g., filter by threats) was also discussed and suggested as more of a filter to use while viewing the incident and vulnerability portrayals. There were times when filtering on the stages of the threat attack models would be relevant to the CPT tasking or stage of their mission as well, for example, if they wanted to see all the options available for the next step in an attack while securing that CCMD's MRT-C. Another suggestion was to augment the portrayal screen with a list of attacks by the threats selected and the ability to search through and drill down on those. The ability to also select a section or component of the target system to focus the analysis of threats on was also seen as a way to support this analysis.

The feedback from these sessions and additional sessions will be used to update the portrayals, the overall use case, and to integrate these into a concept of operations document that captures this information in a way that supports integrating user requirements into system development and acquisition processes.

7 Summary discussion and future activities

This paper discusses an example and method for developing the concept for an XAI-driven junior cyber analyst based on an understanding of the information requirements of both humans and AI components in terms of the work context and workflow. This approach was not only very helpful and useful but might be required in order to develop future systems that humans can use, in particular for systems where the human stakeholders are not able to work with blackbox outputs from intelligent agents. This is common in the military for analysts that have to brief up the chain of command and explain why they are making the predictions or recommendations that they are. These analysts cannot use AI inputs that they cannot explain or justify. This has benefits to software developers as well to be able to understand exactly what needs to be explained and why, rather than trying to explain everything. This should allow informed decisions on which AI approaches to use and how to implement them.

The results to date will be integrated into a concept of operations' document that captures this information in a way that supports integrating user requirements into system development and acquisition processes. The results from human studies analyses on the explainability of the XAI will inform recommendations on the AI techniques to use and how to provide the explanations. This tool will be demonstrated as a prototype mock-up along with a larger work aid being developed by the project to support the cyber operations planner.

Authors' contributions Yes both contributed significantly.

Funding Open Access funding enabled and organized by Projekt DEAL. The research was supported by the Office of the Assistant Secretary of Defense Research and Engineering by their Cyber Security and Applied Research Program.

Compliance with ethical standards

Ethics approval N/A

Consent to participate N/A

Consent for publication Approved by ARL Form 1 process.

Availability of data and material N/A

Code availability N/A

References

Allen, G. (2015) Cyber protection team (CPT) crew operations manual. US Cyber Command ACT-P 2015

Al-Shaer R, Spring J, Christou E (2020) Learning the associations of MITRE ATT&CK adversarial techniques. Retrieved December 15, 2020 from: https://www.researchgate.net/publication/

341149123_Learning_the_Associations_of_MITRE_ATTCK_Adversarial_Techniques

Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin

Chang, C. H., Creager, E., Goldenberg, A., & Duvenaud, D. (2018). Explaining image classifiers by adaptive dropout and generative in-filling. arXiv preprint arXiv:1807.08024

Chen JYC, Lakhmani SG, Stowers K, Selkowitz AR, Wright JL, Barnes M (2018) Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theor Issues Ergon Sci 19(3):259–282

Chen JYC, Procci K, Boyce M, Wright J, Garcia A, Barnes MJ (2014). Situation awareness-based agent transparency. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2014 Apr. Report No.: ARL-TR-6905

Chen JYC, Barnes MJ (2015). Agent transparency for human-agent teaming effectiveness. Proceedings of the IEEE conference on Systems, Man, and Cybernetics (SMC); 2015 Oct; Hong Kong

Core M, Traum D, Lane HC, Swartout W, Gratch J, van Lent M, Marsella S (2006) Teaching negotiation skills through practice and reflection with virtual humans. Simulation 82(11):685–701

Dodson T, Mattei N, Goldsmith J (2011) A natural language argumentation interface for explanation generation in Markov decision processes. In: In international conference on algorithmic decision theory. Springer, Berlin, pp 42–55

Elizalde, F., Sucar, L. E., Luque, M., Diez, J., & Reyes, A. (2008). Policy explanation in factored Markov decision processes. In proceedings of the 4th European workshop on probabilistic graphical models (PGM 2008) (pp. 97-104)

Endsley MR (2015) Situation awareness misconceptions and misunderstandings. J Cognit Eng Decis-making 9:4–15

Gunning, Dave. (2019). Keynote presented at the ACM IUI 2019. Retrieved from: http://iui.acm.org/2019/keynotes.html#gunning-abstract

Guo, W., Mu, D., Xu, J., Su, P., Wang, G., & Xing, X. (2018, October). Lemna: explaining deep learning based security applications. In proceedings of the 2018 ACM SIGSAC conference on computer and communications security (pp. 364-379). ACM

Hendricks LA, Akata Z, Rohrbach M, Donahue J, Schiele B, Darrell T (2016, October) Generating visual explanations. In: In European conference on computer vision. Springer, Cham, pp 3–19

HQ USINDOPACOM (2019). Mission relevant terrain—cyber (MRT-C) campaign update. Briefing

Hutchins EM, Cloppert MJ, Amin RM (2011) Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. Retrieved December 15, 2020 from https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/LM-White-Paper-Intel-Driven-Defense.pdf

Jha S, Sheyner O, Wing J (2002) Two formal analysis of attack grpahs. Retrieved Dec. 15, 2020 from http://www.cs.cmu.edu/wing/publications/Jha-Wing02.pdf

Johnson, W. L. (1994). Agents that learn to explain themselves. In Proc. of the 12th National Conference on artificial intelligence (AAAI)

Khan O, Poupart P, Black J, Sucar LE, Morales EF, Hoey J (2011) Automatically generated explanations for Markov decision processes. In: Decision Theory Models for Applications in Artificial Intelligence: Concepts and Solutions, pp 144–163

Koul, A., Greydanus, S., & Fern, A. (2018). Learning finite state representations of recurrent policy networks. arXiv preprint arXiv:1811.12530

Mercado J, Rupp M, Chen J, Barber D. Procci K, Barnes M. Intelligent agent transparency in human-agent teaming for multi-UxV management. Hum Factors 2016;58(3):401–415

Mitre Corporation. (2018, April 13). Adversarial tactics, techniques & common knowledge. Retrieved from Mitre.org:https://attack.mitre.org/wiki/Main_Page

O'Leary DE (2013) Artificial intelligence and big data. IEEE Intell Syst 28(2):96–99

Parasuraman R, Riley V (1997) Humans and automation: use, misuse, disuse, abuse. Hum Factors 39(2):230–253

Pynadath DV, Rosoff H, John RS (2016, May) Semi-automated construction of decision-theoretic models of human behavior. In: In proceedings of the 2016 international conference on autonomous agents & multiagent systems, pp 891–899

Raymond D, Conti G, Cross T, Nowatkowski M (2014) Key terrain in cyberspace: seeking the high ground. In: Proceedings of the 6th international conference on cyber conflict. Estonia, Tallin

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: explaining the predictions of any classifier. In proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). ACM

Selkowitz A, Lakhmani S, Larios C, Chen JYC (2016). Agent transparency and the autonomous squad member. Presented at the 2016 International Annual Meeting of the Human Factors and Ergonomics Society; 2016 Sep 19–23; Washington DC

Shih, A., Choi, A., & Darwiche, A. (2018). A symbolic approach to explaining Bayesian network classifiers. arXiv preprint arXiv:1805.03364

Si Z, Zhu SC (2013) Learning and-or templates for object recognition and detection. IEEE Trans Pattern Anal Mach Intell 35(9):2189–2205

Swartout WR, Moore JD (1993) Explanation in second generation expert systems. In: In second generation expert systems. Springer, Berlin, pp 543–585

Walker, S. (2018). Closing remarks presented at DARPA D60 symposium

Wright JL, Chen JYC, Barnes MJ, Hancock PA (2016). The effect of agent reasoning transparency on automation bias: an analysis of performance and decision time. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2016

US Cyber Command (2020). Cyber warfare publication 3–33.4: cyber protection team (cpt) organization, functions and employment (28 January 2020)

van Lent, M., Fisher, W. & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. Proc. of the 16th conference on innovative applications of artificial intelligence (IAAI)