MITRE

MTR230084
MITRE TECHNICAL REPORT

Evidence-Based List of Exploratory Questions for Artificial Intelligence Trust Engineering (ELATE)

Author(s):
Stephen L. Dorton
Jeff C. Stanley

April 2024

Dept. No.: L170
Project No.: OVH010.LABS.HAI.J2

Approved for public release. Distribution unlimited 23-00018-2.

Bedford, MA

Abstract

Currently, developing "Trustworthy Artificial Intelligence (AI)" hinges upon top-down, principles-based approaches, which are difficult to put into practice. Thus, we have developed the Evidence-Based List of Exploratory Questions for AI Trust Engineering (ELATE) to help AI developers take a bottom-up approach to developing trustworthy AI. We collected data from multiple sources about incidents where people gained or lost trust in AI "in the wild," and developed a list of items to consider, each including exploratory questions and evidence from real-world examples. We recommend a notional method to apply ELATE in agile development practices; however, we acknowledge more research will be necessary to effectively employ ELATE, and discuss several options for such employment methods.

Record of Changes

Version No.	Date	Description of Change(s)
1	03.02.2023	Initial draft.
1.1	04.06.2023	Cleared for public release/DISTRO A.
1.2	04.10.2024	Minor adjustment to language and formatting of the ELATE items. Reduced three ELATE columns to two (combined the Response and Metrics columns).

Table of Contents

List of Figures

List of Tables

This page intentionally left blank.

1 Introduction

There is a continuing trend of principles-based approaches to developing Artificial Intelligence (AI) that is judged to be better in some way: *Assured* AI, *Ethical* AI, *Responsible* AI, *Trustworthy* AI, etc. These approaches typically take the form of a framework or set of stated principles that, if considered, should generate better AI. The ultimate objective of each of these frameworks typically aligns with supporting human needs and capabilities, something Shneiderman (2022) calls Human-Centered AI. Despite the noble intent, there are noted issues with such principles-based approaches.

First, there is a considerable amount of overlap between these different, seemingly competing frameworks. Blasch et al. (2021) highlights the conceptual overlap of the underlying principles in these approaches, showing how despite the diversity of terminology, they can be distilled to a handful of core principles concerned with making AI more fair, transparent, accountable, trustworthy, etc. Second, and more importantly, such principles-based approaches can be too abstract to be actionable. For example, what would a data science or software engineering team do differently if instructed to "make sure the AI is fair and governable?" Mittelstadt (2019) discusses how these relatively vague principles can make it difficult to come to consensus on how to act on them (e.g., people may have wildly different interpretations of what "fair" means in a given context). Munn (2022) further builds on this argument by highlighting the gap between principles and practice, citing various studies showing a lack of impact of principles on the process or outcome in AI development (e.g., McNamara et al., 2018).

We have taken a broad and inclusive approach to defining AI as "technology that acts intelligently and/or simulates human cognition." This definition is inclusive of various technologies including rule-based expert systems or state machines, as well as statistical machine learning (e.g. supervised, unsupervised, and semi-supervised approaches), and deep learning approaches (e.g. neural networks). We took this inclusive approach for several reasons. First and foremost, splitting hairs to precisely define AI is notorious for stymying efforts to conduct analysis and produce policy, regulation, and guidance on AI (O'Shaughnessy, 2022). Second, we see no downside to supporting trustworthiness of developed technologies, regardless of whether they are "true AI" or not. Thus, there is no penalty for erring on the side of inclusivity in developing tools for trust engineering. Third, there are pragmatic considerations. Namely, the narrower the scope, the less useful this effort is. Additionally, in many cases, we (nor the end users themselves) do not know enough detail about the studied technologies to determine whether they meet some narrower definition of AI.

1.1 Trust & Trustworthy AI

Of all the different frameworks for better AI, we focus particularly on trust or trustworthiness for a variety of reasons. First, trust is critical to adoption: People will work around or refuse to use an AI tool, regardless of its performance, if they do not trust the AI or its outputs (Lee & See, 2004; Dorton & Harper, 2022a). Second, too much or too little trust can adversely affect performance of the human-AI team in completing work (Parasuraman & Riley, 1997; McDermott & ten Brink, 2019).

The most broadly accepted definition of trust within the context of AI, robotics, and autonomy is "*the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability*" (Lee & See, 2004, p. 54). This definition emphasizes two themes from earlier sociological and economic approaches to trust – achieving goals and vulnerability.

Other definitions (e.g., Bhattacharya et al., 1998) emphasize expectancies being met as well as achievement or positive outcomes. Interacting with, sharing resources with, or allowing observation by an AI are behaviors that often (but not always) indicate a trust attitude, because they accept the risk that the AI will not help achieve goals, or worse, act counter to them.

Trust is a multifaceted, complex, and dynamic phenomenon (Chiou & Lee, 2023). There are dozens of human-, system-, and environmental-based factors that affect trust in AI (Schaefer et al., 2016; Dorton & Harper, 2022a). Further, trust in AI is dynamic (i.e., changes over time with interactions), and people can trust specific components within a work system (i.e., technologies or people) while distrusting others (Dorton et al., 2022; Dorton & Harper, 2022b; O'Hear et al., 2022). Ezer et al. (2019) have introduced the concept of trust engineering- the idea that design decisions must explicitly consider trust when developing intelligent systems. Although not under this moniker, others have similarly focused on engineering transparency into intelligent systems to increase user trust (e.g., Lyons et al., 2016).

Because trustworthy AI ideally maximizes benefit and minimizes harm to humans, it has been conflated to some extent with ethical AI. As previously mentioned, frameworks and guidelines for trustworthy AI are often based on reasoning about first principles that are synonymous with other AI frameworks focused on fairness, transparency, and safety; frameworks with limited demonstrable applicability (Fjeld et al., 2020; Blasch et al., 2020). Practitioners who want to engineer appropriate trust and reliance in AI (e.g., Ezer et al., 2019; Lee & See, 2004) have limited means to connect trustworthiness principles to real world trust impacts.

1.2 Inspirations from Naturalistic Decision Making

Given the aforementioned challenges with principles-driven approaches, it seems only appropriate that an alternate approach be empirically driven, focused on studying how humans behave with AI in the environments and contexts in which they accomplish work. The Naturalistic Decision Making (NDM) tradition does just that. The field of NDM began decades ago with a goal of researching how human decision makers actually make decisions in the chaos of complex and dynamic settings with ill-defined goals and tasks (Klein et al., 1993; Klein, 2022). This research tradition has grown over the years not just because of the increased understanding of how people think and work in real-world settings afforded by NDM, but because it also facilitates the engineering of systems to support these cognitive processes (Klein et al. 1993). The NDM approach has proven effective when systems developed on principles from laboratory work have failed (Nemeth & Klein, 2011).

So while our immediate focus is not necessarily on decision making itself, a naturalistic approach is likely to provide insights as to how knowledgeable end users interact and calibrate trust with AI in the context of their work. Recent research has validated this notion, where naturalistic, evidence-based approaches have supported understanding how trust is gained and lost with AI in complex, high-consequence work (e.g., Dorton & Harper, 2022; Dorton, 2022), or more generally, how AI can fail under a variety of scenarios (e.g., Rotner et al., 2021). Such an approach can yield insights and enable development of tools that that are adequately specific and actionable, because they are grounded in the observations that yielded them, with demonstrated impacts.

1.3 Challenges and Objectives

In summary, the trustworthiness of AI is important from a performance and adoption perspective. While there are existing frameworks and guidance for trustworthy AI, and other conceptually

proximal constructs, these kinds of principles-driven guidelines have been criticized for being disconnected from the actual practice of developing AI. Noted challenges to operationalization include (Krijger, 2022; Munn, 2022):

1. Lack of focus on immediate needs and state of the technology,
2. tension among principles and within principles,
3. gaps between principles and the reality of implementation,
4. the impression that existing technical solutions are good enough.

To these we add another challenge:

5. Confusion about which of these principles demonstrably impact human-AI trust, and how.

Our overarching objective is to address these challenges by building an evidence-based (vs principles-based) resource or toolkit to support trust engineering of AI. This tool should be specific enough to be actionable for AI development teams, but not so specific that it is brittle to context. We build upon a format that has been proposed to encourage AI developers to think critically about how humans may gain or lose trust in AI during operational use: exploratory question sets (Dorton 2022). We collected data from various sources and translated these lessons learned from AI "in the wild" into a toolkit to help AI developers consider trust more explicitly. Our objective is not to help *measure* trust, which has been addressed elsewhere (see Kohn et al., 2021 or McDermott & ten Brink, 2019), but to help *develop* AI that is more likely to be measured as trustworthy. Further, we seek to complement existing principles-based approaches- not to replace them.

2 Methods

The overarching method of this research was to explore past incidents where trust was gained or lost in AI. We collected data from three sources: (1) interviews with users of AI in high consequence work domains, (2) case studies or interviews featured in previously published literature, and (3) suitable incidents from the AI Incident Database (McGregor, 2021). This multi-source approach, detailed below, allowed us to incorporate a greater number and range of incidents than if we had limited ourselves to only interviews conducted during this study. From the incidents, we systematically derived a set of exploratory questions to support those who develop and deploy AI and can make decisions that impact the trustworthiness of that AI.

2.1 Data Collection

Our first data source was a set of interviews conducted using an adaptation of the Critical Incident Technique (CIT; Flanagan, 1954). We interviewed a sample (N = 8) of AI users in high-consequence work domains such as national security (i.e., defense and intelligence, n = 4), aviation (n = 3), and healthcare (n = 1). We used a semi-structured interview prompt (Appendix A), which was a streamlined version of the one used by Dorton & Harper (2022), with a few questions removed for being non-diagnostic in previous research, and to focus on the research goals of this effort. The CIT prompt included three phases of questioning: Background and context of use of the AI, the incident where trust was gained or lost, and retrospective thinking. Each interview focused on a single incident and took approximately 30-45 minutes to complete.

The second data source was excerpts, quotes, and case studies from previously published work that used a nearly identical approach to achieve similar goals. That is, we used quotes and events

from previous publications using the CIT to investigate how people gained and lost trust in AI (Dorton & Harper, 2022a; Dorton & Harper, 2021; Dorton, 2022; Dorton & Harper, 2022b; Dorton et al., 2022). This included extracting the considerations (in the form of exploratory questions) that were listed in Dorton (2022), as this effort aims to build upon that research.

The final data source was incidents extracted from the AI Incident Database (McGregor, 2021), a crowdsourced catalog of incidents in which AI systems harmed human stakeholders. We systematically queried and analyzed the full dataset to arrive at 30 incidents to include in this study (this process is shown in Figure 1). We started with a keyword search to identify reports in which trust or related terms were explicitly mentioned; we then looked for quotes from actors and actions indicating a change in trust such as suspending use of a deployed system. We filtered out incidents where we could not identify the specific individual, group, or system to complete the sentence, "[*Human(s)*] lost trust in [*AI system*]." We decided to include one incident (52) in which the driver of an autonomous vehicle died but had previously expressed high trust in the AI. We excluded reports that were referring to a different incident than the one they were mapped to. For example, reports about an Autonomous Vehicle (AV) failure recounted a previous crash involving the death of a pedestrian, which was already reported in another incident (4). Finally, we removed one incident (179) because it was an academic study, not an instance of humans losing trust in AI "in the wild," falling outside the scope of naturalistic inquiry.

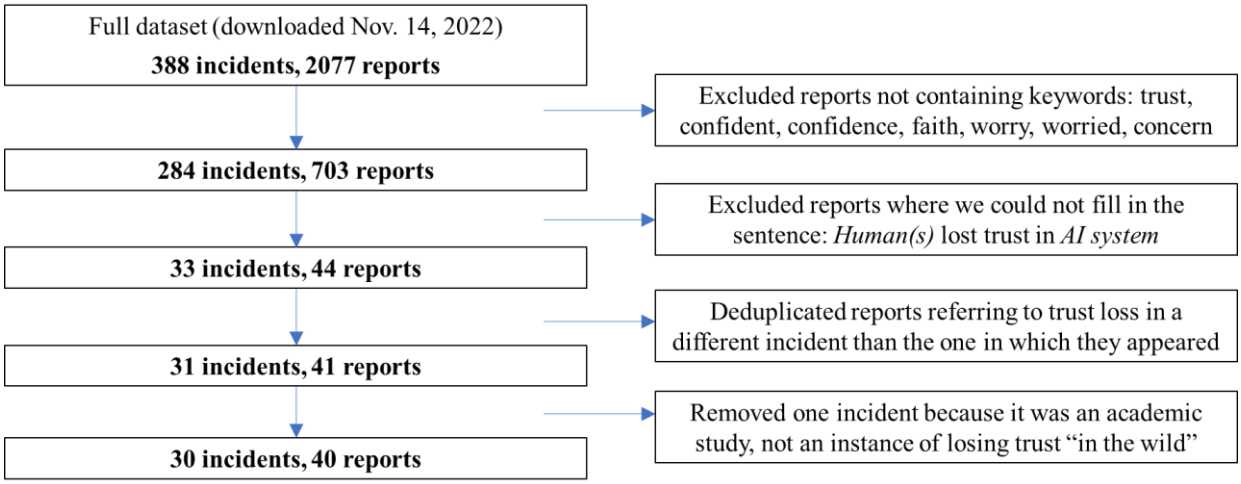

Figure 2-1. Querying and Exclusion Process for Collection from the AI Incidents Database.

2.2 Analysis

At a high level, we examined each incident in the dataset, regardless of source, to answer the following questions:

1. What happened that impacted trust?
2. What were the contributing factors to this happening?
3. Could the development team have done anything to control or mitigate these factors?
4. What questions could the development team have asked themselves, such that the answers to those questions would drive decisions to mitigate (in the cast of lost trust) or ensure (in the case of gained trust) those factors?

While some incidents did not yield any, most incidents yielded one or more factors that could have been addressed by a development team (or were addressed, in the cases of gained trust). Some factors were common across multiple incidents, resulting in a many-to-many relationship between incidents and factors. Each factor yielded one or more exploratory questions to address it. We applied a structured analytic process to translate raw data from incidents into actionable resources for engineers. Below is a step-by-step walkthrough of this process, including an example where relevant.

1. ***Identify Items***: The first step was to identify items of relevance to trust engineering. First, we read through source material to identify factors that contributed to the gain or loss of the participant's trust in the AI. These factors were often explicitly stated, although we also identified factors where they were more implicitly discussed. Each item was given a short, descriptive title.
    - **Example**: From the following passage in an interview transcript, we identified the issue of "Raw Data Inspection," or the need for users to compare AI outputs with raw data to assess AI performance in situ: "*They removed the raw [data] in the new one and it really pissed off a lot of people. They can no longer see the little signals that were used as a cue to see where the AI is missing it. [You know], something's there but it hasn't tripped the computer.*"
2. ***Draft Questions***: For each identified item we developed questions that, if answered, would provide AI developers with insights toward how to prevent (in cases of losing trust) or encourage (in the case of gaining trust) similar outcomes in their system. Most items start with at least one human-centered question to encourage empathy with users and consideration of various contexts, followed by at least one AI-centered question to encourage thinking about how to better design or engineer the AI to engender or avoid outcomes (depending on if the outcomes were good or bad). In many cases we drafted more than two questions to capture nuance or context of the source incident that would otherwise be missed.
    - **Example**: For Raw Data Inspection, we drafted the following questions: "*What kinds of "raw" input data might users rely on to verify or interpret AI outputs? How might this information be available when needed, but hidden when not?*"
3. ***De-Duplicate and Merge***: After extracting items and writing initial questions we removed duplicate items and merged items that were conceptually proximal. In the latter case we considered items to be conceptually proximal if their questions shared similar verbiage, or if we believed they were likely to elicit similar answers. Since one incident could yield multiple items, and items were merged across incidents, this process resulted in a many-to-many relationship between incidents and items. For instance, three different incidents could serve as evidence for the same two items. In practice, we performed this step multiple times, once for each data source before aggregating them and again afterward.
4. ***Organize***: We then collaboratively organized items into "Focus Areas" based on their conceptual similarities. This iterative process involved writing notional definitions for each focus area, and then examining each item to make sure it met the definition. When an item did not meet a definition, we either moved it to a more appropriate focus area, or we modified the definition for the focus area to be inclusive of the items it contained.

- **Example**: The Raw Data Inspection item was placed in the Presenting and Exploring Outputs focus area, defined as: "Assessment of how AI outputs facilitate user decision making."

5. ***Quality Check***: The final step was to ensure that each item complied with a common presentation syntax, which is described in detail in Section 3.1.

Table 1 provides an overview of the number of items initially elicited from each source, and the final number of unique items remaining after completing the aforementioned process. One should note that the step of de-duplication and merging reduced the set of items by 48% of its original size of 101.

Table 2-1. Summary of Items Elicited from Sources

Data Source	# Items Elicited
Interviews (n = 8)	24
Previous work (n = 29)	34
AI Incident Database (n = 30)	43
Total Items	101
Unique Items	**53**

3 Results

In the following sections we describe the results of this effort, which include both an overview of the resultant ELATE, and a notional procedure for employing it in AI development efforts.

3.1 Evidenced-Based List of Exploratory Questions for AI Trust Engineering (ELATE)

Completing the process in Section 2.2 generated ELATE, which is presented in its entirety in Appendix B. This initial version contains 53 items organized into eight different focus areas. As previously mentioned, focus areas are collections of items that are conceptually proximal to each other and fall under a shared theme. Each focus area has a concise name and description of the kinds of items it contains.

Each item follows a syntax, to the extent possible, to ensure a consistency and ease of use. The following are standard elements included with each item:

- **Title**: A short (four words or less), descriptive title of the item.
- **Questions**: A set of at least two questions, which typically includes one human-centered and one AI-centered. Wherever possible, questions were phrased as "How might [user/AI]…?" for consistency and to be as open-ended as possible. However, for some items, questions follow a different wording, format, or ordering.
- **Example**: A short, plain language description of how the underlying factor of the item played a role in an actual incident where trust was gained or lost. In some cases, the example may include a direct quote or excerpt from the relevant interview, publication, or database entry to provide additional context for how trust was gained or lost.

While Appendix B contains the entire set of items, Table 3-1 provides a brief overview of the initial focus areas, the content of the items they contain, their relative size, and an exemplar item from that focus area (examples and quotes are not included for the sake of brevity).

Table 3-1. Summary of Focus Areas and Items in ELATE

Focus Area	# Items	Exemplar Item
User Requirements: Ensuring the AI meets user needs.	**6**	*Capturing Expert Knowledge*: How do users currently make decisions without AI? What information and heuristics do they use? Have these been adequately codified in the AI?
Data, Features, & Models: Aligning the inputs and mechanics of the AI to user expectations.	**8**	*Bias Awareness*: How might users be made aware of biases and default settings in the AI? Which ones may be problematic if unknown?
Controls & Safeguards: Preventing harms to users and others.	**10**	*Bad Faith Actors*: How might a bad faith actor adversely affect the AI, or use the AI to break law or policy? How will users know when bad faith actors are affecting the AI? How might the AI be resilient to such actions?
Presenting & Exploring Outputs: Facilitating user cognition through presentation of outputs.	**8**	*Operationalization of Outputs*: Will users know what to do with the outputs? How will they be aware of the capabilities, limitations, and conditions for which outputs are validated in various scenarios and contexts?
Understanding AI Behavior: Establishing and maintaining a shared understanding between users and the AI.	**4**	*Plan Awareness:* How might users develop and maintain an understanding of the AI's plans? How might the AI make it immediately obvious when it is deviating from plans?
Work System Collaboration: Using the AI in the context of third parties within a work system.	**4**	*Shared Awareness:* How might disparities in AI outputs across different roles cause confusion or conflict? How might the AI promote shared awareness and collaboration across various roles?
Testing, Debugging, & Error Handling: Determining the causes of problems to calibrate and repair trust.	**5**	*Forensic Support:* What heuristics or cues do users and maintainers rely on to troubleshoot issues with the AI? How might the AI help them readily answer the most diagnostic questions after an incident? How might the AI fail to capture a significant irregularity/error in logs or debug outputs?
Sustainment: Maintaining and upgrading the AI to meet user needs after deployment.	**8**	*Update Awareness & Impacts:* How will users become aware of updates to the AI or associated system components, and understand the impact of those updates on observable behavior? How might routine updates to the AI create unexpected behaviors?

3.2 Notional Employment

There are a variety of approaches to employ ELATE; however, barring research into, or evidence of, the pros and cons of each, we default to the simplest and most straightforward approach, leaving room for AI development teams to tailor application based on their processes, practices, and work tempo. An iterative application is likely to be the most productive, given that most software

development projects use agile or agile-hybrid approaches (Raunak & Binkley, 2017). Further, there have been calls for developing human factors methods that dovetail with agile development processes (Dorton et al., 2020), or more generally, methods that are less academic in nature, and compatible with constraints and business practices of technology developers (Bruni, 2022).

Thus, the default approach for putting ELATE into practice is:

1. Employ the template in Appendix B as a living document, kept in a commonly accessible location within the development environment so it can be found and updated regularly.
2. Identify which focus areas (i.e., sets of items) are likely to be most important in the context of the AI being developed. For example, the "Presenting and Exploring Outputs" focus area may be more important for developing AI with a high degree of user interaction, whereas the "Data, Features, and Models" focus area may be more important for AI that controls complex processes with minimal human intervention.
3. Invite all members of the development team, a sample of end users, and other appropriate stakeholders to a brief session (e.g., 30-60 minutes) at the onset or conclusion of each sprint. Ideally, this session would coincide with other sprint activities such as backlog grooming or retrospectives (see Al-Saqqa et al., 2020 for an overview), when the entire team should be accustomed to meeting.
4. Select a focus area and for each item in that focus area, read the questions aloud and have participants respond to the questions (in the center column) with specific examples of how the item could apply to this AI. (In some cases, the item may not be applicable to the AI being developed, and can be marked as such.) Once consensus has been achieved on the response(s) for an item, brainstorm and document measures or milestones (in the right column) for knowing whether the team is adequately addressing or has addressed it.
    - We recommend starting with the "User Requirements" focus area, since it is critical to address these items as early as possible in the development process, and because they have the broadest implications for the development process.
    - After addressing the user requirements focus area, move on to focus areas determined by relative priority in step 2.
5. Repeat steps 3 and 4 throughout the development lifecycle. We recommend continuing to review items even after they have been addressed. The very nature of agile development means that requirements and designs will change based on stakeholder feedback after each sprint, so many items will likely need to be reassessed as the underlying requirements and designs of the AI may change considerably from when the AI was previously addressed.

4 Discussion

While this effort generated a tool to support trust engineering for intelligent systems, there are some methodological and pragmatic limitations worth addressing. First and foremost, we used a relatively uncontrolled method to generate consensus on findings from collected data. Unlike previous studies that we build upon (e.g., Dorton & Harper, 2022), we did not use an independent coding and adjudication approach or generate statistics to formally assess inter-rater agreement of codes (e.g., Cohen, 1960). However, we do not make claims about the relative frequency or impact of different findings, only that each item applies to at least one specific incident in our data. Thus, we believe having two researchers develop concurrence through dialogue rather than through an independent coding process is sufficient to these ends.

Another methodological limitation is that the content in ELATE as presented here is limited to findings from only the incidents considered in this study (i.e., from interviews, literature, and databases). Although we have generated a relatively large sample of incidents from which to extract questions (n = 67), we cannot assert that this sample is exhaustive, nor stratified, across different types of AI nor work domains. However, we find the removal of 48 redundant items from the original set of 101 (a 48% decrease) during the de-duplication process as indicative of achieving a relative level of saturation.

While we believe we have sufficiently argued the merits of a bottom-up, evidence-based, and practice-oriented approach, we also acknowledge that principles-based approaches facilitate certain kinds of discussions. Rather than reject principles and frameworks entirely, we hope to explore how these two approaches may complement one another. For example, such guidelines or frameworks could inform focus areas or approaches to implementation.

ELATE's delivery mechanism needs refinement. First, it is all too easy for AI developers to simply dismiss many legitimate questions by claiming they are not applicable to their contexts, or to "check the box" by providing superficial answers (Dominguez et al., 2021). Second, it does not explicitly recommend what to do about answers to specific questions. That is, this initial approach might nudge AI developers to think critically and to uncover and characterize possible risks, but it falls short of providing concrete guidance for mitigating said risks. In addition to continued collection of incidents to expand the scope of ELATE, future work will identify what kinds of delivery mechanisms and artifacts are most effective, and for whom. In most cases, we envision ELATE as a resource to develop tools to facilitate participatory design or co-design approaches to trustworthy AI, which rely on involvement of end users working with developers (e.g., see Papautsky et al, 2021; or Steen et al., 2011). The following are some ideas for potential delivery mechanisms to be investigated:

- ***Interviews:*** Use exploratory questions to inform interview prompts. Notionally, human-centered questions would inform interviews with users, and AI-centered questions, as well as the users' responses to the human-centered questions, would then inform interviews with members of the development team. While this approach would generate valuable insights, human factors engineers and user experience (UX) professionals would need to perform additional work to make those insights actionable for AI developers.
- ***Targeted PreMortem***: A facilitated brainstorming activity similar to other divergent and convergent activities used in participatory design or design thinking workshops. Notionally, we envision a modified version of the premortem technique (see Klein, 2022; or Bettin et al., 2022), where instead of an open-ended approach, brainstorming is targeted to elicit causes of specific outcomes (e.g., the AI fails because of issues with training data). That is, brainstorming may be targeted on focus areas within the ELATE framework, based on the context of the AI being developed.
- ***Consensus-Building:*** Use exploratory questions associated with items as prompts within the Delphi Method (Helmer, 1967) or Nominal Group Technique (Delbeq & Van de Ven, 1975). Both developers and users (including stakeholders affected by the AI) would answer exploratory questions independently of each other, then share and iterate upon their answers until consensus was formed. This could be done asynchronously through surveys, or synchronously through facilitated sessions. Such a process would facilitate shared expectations across developer and user communities, and help developers understand their own limitations in predicting how users might employ their creations in the wild.

- ***Design Patterns**:* Breaking away from the participatory design tradition, another mechanism may be to develop a set of design patterns based on the findings from various incidents. Design patterns are a way to codify or represent such expert knowledge or lessons learned to inform design (Heer & Agrawala, 2006). However, design patterns are meant to share information about successful implementations (Zhang & Budgen, 2012). Given that the majority of incidents are regarding lost trust (i.e., failed implementations), we may be more likely to experiment with design anti-patterns (e.g., see Mo et al., 2021).

Finally, we recognize ongoing discussions about the utility of trust as a concept in human-AI teaming and the possibility that many stakeholders seeking to improve trust actually want to affect related concepts such as adoption, reliance, and safety (Bolton, 2022). Future work could potentially apply the methods described in this paper to develop tools focused specifically on those phenomena. However, our focus is on trust engineering, and we posit that trust is appropriate and efficient for this kind of tool because of its utility in affecting multiple outcomes of interest. As such, we expect that focusing on one of these related variables (e.g., reliance) would considerably limit the generalizability of the tool.

Despite these potential limitations and need for future work, we believe this is a necessary first step toward exploring naturalistic, bottom-up, evidence-based approaches to trust engineering of AI. We hope to not only continue this line of research and development (particularly in the area of delivery mechanisms), but also to provide a foundation from which the greater community of practice (Government, Industry, Academia, etc.) can close the principles/practice gap to develop more trustworthy AI.

5 References

Al-saqqa, Abdel-Nabi, H., & Sawalha, S. (2020). Agile software development: Methodologies and trends. *International Journal of Interactive Mobile Technologies, 14*(11), 246-270. https://doi.org/10.3991/ijim.v14i11.13269

Bettin, B., Steelman, K., Wallace, C., Pontious, D., & Veinott, E. (2022). Identifying and addressing risks in the early design of a sociotechnical system through premortem. *Proceedings of the 2022 HFES 66th Annual Meeting, 66*(1), 1514-1518. https://doi.org/10.1177/1071181322661307

Bhattacharya, R., Devinney, T.M., & Pillutla, M.M. (1998). A formal model of trust based on outcomes. *Academy of Management Review, 23*(3), 459-472. https://jstor.org/stable/259289

Blasch, E., Sung, J., & Nguyen, T. (2021). Multisource AI scorecard table for system evaluation. *arXiv preprint arXiv:2102.03985*.

Bolton, M.L. (2022). Trust is not a virtue: Why we should not trust trust. *Ergonomics in Design*. https://doi.org/10.1177/10648046221130171

Bruni, S. (2022). Introducing PATI: The pareto analysis for technology insertion – A human-centered methodology to identify and prioritize innovation in complex systems. *Creativity, Innovation, and Entrepreneurship, 31,* 64-74. https://doi.org/10.54941/ahfe1001508

Butterfield, L. D., Borgen, W. A., Amundson, N. E., Maglio, A. T. (2005). Fifty years of the critical incident technique: 1954-2004 and beyond. *Qualitative Research, 5*(4), 475-497. https://doi.org/10.1177/1468794105056924

Chiou, E.K., & Lee, J.D. (2023). Trusting automation: Designing for responsivity and resilience. *Human Factors*. https://doi.org/ 10.1177/00187208211009995

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Delbeq, A.L., & Van de Ven, A.H. (1975). A group process model for problem identification and program planning. *Journal of Applied Behavioral Science, 7*(4), 466-491. https://doi.org/10.1177/002188637100700404

Dominguez, C., Rayo, M., Dorton, S**.**, Morey, D., Naikar, N. & Roth, E. (2021). Cognitive engineering: Will they know our name when we are 40? *Proceedings of the 2021 International Annual Meeting of the Human Factors and Ergonomics Society, 65*(1), 257-261. doi: 10.1177/1071181321651042

Dorton, S.L. (2022). Supradyadic trust in artificial intelligence. *Artificial Intelligence and Social Computing, 28*, 92-100. https://doi.org/10.54941/ahfe1001451

Dorton, S.L., & Harper, S.B. (2022a). A naturalistic investigation of trust, AI, and intelligence work. *Journal of Cognitive Engineering and Decision Making, 16*(4), 222-236. https://doi.org/10.1177/15553434221103718

Dorton, S.L. & Harper, S.B. (2022b). Self-repairing and/or buoyant trust in artificial intelligence. *Proceedings of the HFES 66th International Annual Meeting, 66*(1), 162-166. https://doi.org/10.1177/1071181322661098

Dorton, S.L. & Harper, S. (2021). Trustable AI: A critical challenge for naval intelligence. *Center for International Maritime Security (CIMSEC).* Retrieved from: https://cimsec.org/trustable-ai-a-critical-challenge-for-naval-intelligence/

Dorton, S.L., Harper, S.B., & Neville, K.J. (2022). Adaptations to trust incidents with artificial intelligence. *Proceedings of the HFES 66th International Annual Meeting, 66*(1), 95-99. https://doi.org/10.1177/1071181322661146

Dorton, S.L., Maryeski, L.R., Ogren, L., Dykens, I.T., & Main, A. (2020). A wargame-augmented knowledge elicitation method for the agile development of novel systems. *Systems, 8*(27), 1-15. doi:10.3390/systems8030027

Ezer, N., Bruni, S., Cai, Y., Hepenstal, S.J., Miller, C.A., & Schmorrow, D.D. (2019). Trust engineering for human-AI teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63*(1), 322-326. https://doi.org/10.1177/1071181319631264

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2022). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication No. 2020-1*. http://dx.doi.org/10.2139/ssrn.3518482

Flanagan, J. C. (1954). The Critical Incident Technique. *Psychological Bulletin, 5*, 327-358. http://dx.doi.org/10.1037/h0061470

Heer, J., & Agrawala, M. (2006). Software design patterns for information visualization. *IEEE Transactions on Visualization and Computer Graphics, 12* (5), 853-860. https://doi.org/ 10.1109/TVCG.2006.178

Helmer, O. (1967). Analysis of the future: The delphi method. *The RAND Corporation.* https://apps.dtic.mil/sti/pdfs/AD0649640.pdf

Klein, G. (2022). *Snapshots of the mind*. MIT Press.

Klein, G.A., Orasanu, J., Calderwood, R., & Zsambok, C.E. (1993). *Decision making in action: Models and methods*. Ablex Publishing.

Kohn, S.C., de Visser, E.J., Wiese, E., Lee, Y., & Shaw, T.H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology, 12*(604977), 1-23. https:/doi.org/10.3389/fpsyg.2021.604977

Krijger, J. (2022). Enter the metrics: critical theory and organizational operationalization of AI ethics. *AI & Society, 37*, 1427–1437. https://doi.org/10.1007/s00146-021-01256-3

Lee, J.D., & See, K.A., (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Lyons, J.B., Koltai, K.S., Ho, N.T., Johnson, W.B., Smith, D.E., & Shively, R.J. (2016). Engineering trust in complex automated systems. *Ergonomics in Design, 24*(1), 13-17. https://doi.org/10.1177/1064804615611272

McDermott, P.L. & ten Brink, R.N. (2019). Practical guidance for evaluating calibrated trust. *Proceedings of the Human Factors and Ergonomics Society 2019 Annual Meeting, 63*(1), 362-366. https://doi.org/10.1177/1071181919631379

McGregor, S. (2021). Preventing repeated real world ai failures by cataloging incidents: The AI incident database. *Proceedings of the Thirty-Third Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-21)*. Virtual Conference.

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? *Proceedings of the 26th ACM Joint Meeting on European Software Engineering Conference on the Foundations of Software Engineering*, 729-733. https://doi.org/10.1145/3236024.3264833

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1* (11), 501-507. https://doi.org/10.1038/s42256-019-0114-4

Mo, R., Cai, Y., Kazman, R., Xiao, L., & Feng, Q. (2021). Architecture anti-patterns: Automatically detectable violations of design principles. *IEEE Transactions on Software Engineering, 47*(5), 1008-1028. https://doi.org/10.1109/TSE.2019.2910856

Munn, L. (2022). The uselessness of AI ethics. *AI and Ethics.* https://doi.org/10.1007/s43681-022-00209-w

Nemeth, C., & Klein, G. (2010). The naturalistic decision making perspective. In *Wiley Encyclopedia of Operations Research and Management Science*. SemanticScholar. https://doi.org/10.1002/9780470400531.eorms0410

O'Hear, E. H., Atchley, A., Gholston, S., Weger, K., Mesmer, B., & Tenhundfeld, N. L. (2022). System-wide trust: The impact of an error in a multi-component system. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1777–1781. https://doi.org/10.1177/1071181322661119

O'Shaughnessy, M. (2022). One of the biggest problems in regulating AI is agreeing on a definition. *Carnegie Endowment for International Peace*. https://carnegieendowment.org/2022/10/06/one-of-biggest-problems-in-regulating-ai-is-agreeing-on-definition-pub-88100

Papautsky, E. L., Strouse, R., & Dominguez, C. (2020). Combining cognitive task analysis and participatory design methods to elicit and represent task flows. *Journal of Cognitive Engineering and Decision Making*, *14*(4), 288-301.

Raunak, M.S. & Binkley, D. (2017). Agile and other trends in software engineering. *IEEE 28th Annual Software Technology Conference (STC)*, 1-7. https://doi.org/10.1109/STC.2017.8234457

Rotner, J., Hodge, R., & Danley, L. (2021). Five AI fails and how we can learn from them. *The MITRE Corporation*. https://www.mitre.org/news-insights/publication/five-ai-fails-and-how-we-can-learn-from-them

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

Steen, M., Manschot, M., & De Koning, N. (2011). Benefits of co-design in service design projects. *International Journal of Design*, *5*(2).

Zhang, C., & Budgen, D. (2012). What do we know about the effectiveness of software design patterns? *IEEE Transactions on Software Engineering, 38*(5), 1213-1231. https://doi.org/10.1109/TSE.2011.79

Appendix A Critical Incident Technique Prompt

The following prompt was used to conduct all interviews. Should a need arise to build upon this initial toolkit with domain-specific questions, one may use this prompt to conduct their own interviews, and generate additional questions by following the process outlined in Section 2.2.

Introductory prompt:

- Think of an AI-based technology you work with that you really trust, or do not trust…
- Was there a defining event, or series of events, where you gained or lost trust in that AI?
- Let's talk about that incident.

Background:

1. What should we call this AI? Can you give us a short (3 words or less), descriptive title?
2. Did you have a role in deciding to use this AI?
3. How much, and what kind of training did you receive on it?
4. Was it enough to use the AI effectively?
5. What tasks did the AI perform?
    a. Does it perform the same tasks as you?
    b. Does it perform tasks you used to do?
    c. Does it perform different tasks towards the same goals or outcomes?
6. What was the task frequency (hourly, daily, weekly, monthly, quarterly, yearly, single instance)?
7. What is the task criticality?
    a. 1 = Not critical. If the task is not done correctly or on time there are numerous work-arounds, and/or there is a negligible outcome to life, safety, or mission/operations.
    b. 10 = Highly critical: If the task is not done correctly or on time it will result in death and/or mission/operational failure.

Incident:

8. How long ago was the incident (in months/years)?
9. Was it a single event, or a series of events?
    a. If series, how long was the series of events (e.g. over 3 months during deployment)?
10. Please describe the incident where you gained or lost trust in the AI…
    a. This is the main "question" of the CIT process, and most time should be spent here.
    b. Interviewers should feel free to ask probing or follow-on questions based on what was said by the participant.

11. Was it clear why the AI did what it did?
12. How did the incident change your behavior with the AI? Do you do anything differently now?
13. What was your level of trust (1-10) ***before*** the incident? Why?
    a. 0 = No trust. You assume everything the AI does is wrong or a failure.
    b. 10 = Complete trust. You assume everything the AI does is correct or a success.
14. What was your level of trust (1-10) ***after*** the incident? Why?
    a. 0 = No trust. You assume everything the AI does is wrong or a failure.
    b. 10 = Complete trust. You assume everything the AI does is correct or a success.

Afterwards:

15. What was your level of trust (1-10) in the AI ***now***? Why?
    a. 0 = No trust. You assume everything the AI does is wrong or a failure.
    b. 10 = Complete trust. You assume everything the AI does is correct or a success.
16. (If the incident is about losing trust) What do you wish you would have known that would have mitigated the loss of trust?
17. Choose one of the following based on the nature of the incident:
    a. If they *gained* trust in the incident: What would it take to make you trust the AI at a 0 or 1 (complete lack of trust)? Why?
    b. If they *lost* trust in the incident: What would it take to make you trust the AI at a 10 (complete trust)? Why?

Appendix B ELATE v1.2

The first version of the Evidence-Based List of Resources for AI Trust Engineering (ELATE) is provided in the matrix below. Items are numbered to aid in reference and are organized into focus areas based on their content. The medium by which you employ ELATE is less important than the employment itself. Feel free to use this document, create a spreadsheet, an online collaboration space (e.g., MURAL or Miro), Jira tasks, or whatever medium works best for the needs of your team (developers, engineers, users, stakeholders, etc.). It is more important that this be treated as a living document, where it is accessible such that team members can revisit responses and update the matrix throughout the development lifecycle.

Review each item for not only its set of questions, but also for the examples provided, as they are likely to spur discussion. It is likely that not every item is applicable to the particular AI you are developing; however, we encourage you to resist the urge to discount an item as "Not Applicable" (NA), and to think analogically about how it may pertain to your case.

For example, Item 20 (Access Control) references an issue where violent content was accidentally shown to toddlers when the mature content was snuck into an approved channel for kids. This item should not be viewed as NA simply because a hypothetical AI does not curate content. Analogously, one may consider harms from personal information (e.g., medical, financial, or other personally identifiable information), or even classified information being recommended or presented to those who are not approved to see it. It does not take an undue effort to imagine where a relatively innocuous application like a preventative maintenance prediction system may ingest large amounts of unclassified data, but the outputs may present a compilation risk of being classified when put in aggregate (further, the data at rest in the database may be an issue unto itself). ELATE will be most useful when you have a diverse set of participants from technical and operational backgrounds, and emphasize creativity.

Table B-1. ELATE

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
User Requirements: Ensuring the AI meets user needs.	
1. Stakeholders in Development. How might users, domain experts, and other stakeholders best be involved in the development of the AI (early and often enough)? How might non-users impacted by the AI have a voice in its development? *Example: Involving domain experts in design processes increased the trustworthiness of an AI for intelligence analysis: "We were lucky to have experts or former targeting officers, it was helpful for the development of models" (Dorton & Harper, 2022a).* *Example: Users lost trust in an AI where end users were not involved in the conceptualization and design of the AI: "They missed a huge part… and they did not include the [users] enough… Sometimes it needs to include less technical focus and more involvement of the people who make these decisions every day without computers" (Dorton, 2022).*	
2. Capturing Expert Knowledge. How do users currently make decisions without AI? What information and heuristics do they use? How might these be adequately codified in the AI? *Example: Failure to assess how users think and make decisions in operational contexts resulted in AI outputs that could not be operationalized or acted upon in high-consequence work: "They knew the math behind it… they couldn't translate it for the [users]… if somehow the developers knew what we were trying to achieve [it would have succeeded]" (Dorton, 2022).*	
3. Operational Utility. What do users most need the AI to do, when, and why? How might the AI provide the most utility to users, regardless of algorithmic performance? *Example: Despite high performance when given enough data, users lost trust in a predictive analytics AI because real world operating scenarios could never be able to provide it enough data to make a prediction; therefore, it had no utility outside of laboratory setting: "We made the best DIME and PMESII models ever, they just won't ever have enough data… so [they] cancelled the program." (Dorton & Harper, 2022a).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
4. Workflow. How might users struggle to integrate the AI into their workflows? How might inputs and outputs of the AI map to these workflows? *Example: A medical treatment recommender extracted and aggregated data from various reports; the user would have had more trust if it functioned more as a search engine for browsing relevant articles.*	
5. Legacy Systems & Expectations. How might user experience with legacy systems affect expectations for the AI? How might users evaluate the effectiveness of the AI against some legacy system? What capability must the AI provide so users do not decide to revert to the legacy system? *Example: A user lost trust and stopped using a new content search and curation AI because it did not employ Boolean operators in querying in the same manner as all legacy systems they had used: "The impression I got was that the Boolean words and markers I was using elsewhere were only kind of halfway used in the new system… It took a liberal interpretation of Boolean Logic."*	
6. Usability. How might users find the AI difficult to work with? What efforts should be taken to ensure the AI is intuitive for end users in operational contexts? *Example: Users have reported that they would trust the AI more if it had a more usable interface: "to get me to [trust it more] is not even the algorithm, but the user interface. It would just need to look pretty sleek, user friendly, organized, as opposed to bazillions of buttons and tabs and drop downs in a gonkulator vs a clean and fluid thing."* *Example: A lack of usability prevented a user from repairing trust after an incident with the AI, "I was spending hours per day combing through reports… It was hard to find relevant information, so I used it less frequently" (Dorton & Harper, 2022a).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
Data, Features, & Models: Aligning the inputs and mechanics of the AI to user expectations.	
7. Data Quality. Who will annotate data? What knowledge, skills, or abilities should they have? How might the quality of AI inputs be managed? *Example: Users lost trust in AI because its performance suffered from bad inputs from fallible humans: "The reason I put less faith in this stuff is that humans can train it on bad data with bad features and then say It's gold when the outcomes are bad… you can see how much human assumptions can affect the model's performance" (Dorton & Harper, 2022a).* *Example: Users lost trust in AI because only the least experienced people were available to annotate data to train the system: "I don't think they understood the need for quality. They should have made teams with senior people so they could make sure it was good" (Dorton, 2022).*	
8. Data Recency. How might the data driving the AI be out of date or not representative of the current or desired future world state? How might data be checked against the latest expectations? *Example: A résumé screening AI was trained on past successful candidates, resulting in a bias for resumes from male candidates, based on the old hiring practices they sought to overcome (Incident 37).*	
9. Selective Annotations. Are those providing inputs able to capture their level of confidence, or refuse to provide an annotation when they are unsure? Where might annotation go wrong? *Example: Users lost trust in an entity classification AI because data annotation processes forced inexperienced users to provide an annotation, even when they were unsure: "Analysts were not allowed to rate their confidence for [annotation] or say 'no' or 'I don't know,' they had to make a call" (Dorton, 2022).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
10. Data Priorities. How might users value different kinds of inputs or sources of data (e.g., primary vs. secondary)? How should the AI codify these priorities? *Example: An intelligence user lost trust in an AI because it was providing results with primary and secondary sources of information combined together without distinction: "In intelligence there is a big primary source vs secondary source distinction... [in Open Source Analysis] there's just a lot of punditry going on… which is different than getting a report… So there's that kind of sourcing issue… I'll be honest, I remember trying to go in and filter out against some of the secondary sources that I wasn't getting the response [I desired from the AI]."*	
11. Data Integrity. How might the data driving the AI be compromised? How might padding techniques to increase the dataset change its distribution or integrity? How might this affect user interactions with the AI during use? *Example: Users lost trust in a medical treatment recommender that was trained on data padded with synthetically generated cases that did not sufficiently reflect real cases (Incident 225).*	
12. Reliability. How might users react to different outputs given the same inputs? How might the AI produce such inconsistent or unreliable outputs? *Example: Users gained trust in poorly performing AI because it was at least reliable and could provide them with a means to test if it was working properly: "I guess I've gained confidence because the algorithm consistently gives results that are imperfect" (Dorton & Harper, 2022a).* *Example: UAS pilots cited reliability as an important factor in gaining trust in autonomous flight software: "[To trust it completely?] A million [successful] flights, I guess."*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
13. Bias Awareness. How might users be made aware of biases and default settings in the AI? Which ones may be problematic if unknown? How can the AI make these biases more explicit to users? *Example: A junior military user nearly created an unwarranted emergency because they were unaware that the AI was biased to recommend the most dangerous threat that could not be completely ruled out, regardless of what region of the world they were operating in (Dorton & Harper, 2021).*	
14. Equitable Outcomes. How might some group assert the AI is biased against them? Who is it, what is their argument, and how would you respond? How might the AI be designed or deployed to prevent inequitable outcomes? *Example: Students claimed that test monitoring software had trouble detecting darker skinned students, flagging them for stepping away from the exam more than lighter skinned students (Incident 38).*	
Controls & Safeguards: Preventing harms to users and others.	
15. High-Risk/Low-Frequency. How might a single erroneous output from this AI cause a big problem for someone? How might the AI enable preventing this before it happens, and/or make it easier to reverse the impacts? *Example: Users acted on a contraceptive app's predictions, leading to unwanted pregnancy (Incident 150).*	
16. Low-Risk/High-Frequency. How might numerous errors at scale create an unmanageable workload? How might the AI support error reporting and mitigation at scale? *Example: The Australian government utilized an AI for welfare services that erroneously sent thousands of debt notices (Incident 57).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
17. Human Approvals. What tasks, decisions, or outputs require human approval? Are humans involved at all the appropriate places? *Example: Users gained trust in a targeting AI because it required a human to approve a nominated entity as a target: "A human verifying a nomination gives me much higher confidence than the algorithm feeding itself" (Dorton & Harper, 2022a).*	
18. Human Intervention. How might this AI be overridden, deactivated, or redirected if something goes wrong? How might humans fail to intervene? How might the AI keep their attention when it is most needed? *Example: Autonomous vehicles rely on humans to monitor and take over in emergencies but fail to keep human attention, leading to crashes: "The vigilance required… arguably requires significantly more attention than just driving the vehicle normally" (BBC 2018; Incident 323).* *Example: After an update, Tesla vehicles began braking unpredictably. Users were able to avoid crashes by overriding the brake (Incident 208).*	
19. Interaction Obviousness. How might users accidentally interact with the AI (potentially without knowing it)? How might the AI make obvious when it is doing something that could impact users, and allow them to intervene? *Example: A couple lost trust in a home assistant AI after it misinterpreted their conversation, incorrectly understanding commands to record and send the conversation to a coworker (Incident 361).*	
20. Access Control. Is there any content that any users or third parties should not have access to? How might the AI expose the wrong content to the wrong people? How might we prevent this? *Example: YouTube content service presented disturbing content to toddlers. Parents lost trust and were "horrified to see such content on a site [they] trusted" (Peters 2016; Incident 1).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
21. Adaptability. How might a user, third party, or the operational environment behave in a way the AI is unable to adapt to? How might the AI be designed to adapt to novel situations? *Example: Third parties lost trust after a security robot collided with a toddler who ran across its determined path (Incident 51).*	
22. Proactive Safety. What safety measures are users likely to expect to be engaged, when, and why? How might the AI ensure safety protocols are active when needed? *Example: Users lost trust in a robot after a handler accidentally pressed a button, causing it to crash through a glass wall and injure a third party. Obstacle avoidance was disabled (Incident 217).*	
23. Bad Faith Actors. How might a bad faith actor adversely affect the AI, or use the AI to break law or policy? How will users know when bad faith actors are affecting the AI, and how can it be resilient to such actions? *Example: A stock trading algorithm was used to manipulate the stock market by placing orders in bad faith, causing the "Flash Crash" of 2010 (Incident 28).* *Example: Military users said they would lose trust in an AI if adversaries could interfere with sensor data that the AI relied upon.*	
24. Privacy Protection. How might stakeholder privacy be violated directly or indirectly? Are all stakeholders fully aware when they are being "observed" by an AI, what data are collected, how their data will be used, and who to contact with concerns? *Example: HireView removed facial expression tracking from its interview assessment software after users lost trust and issued complaints about lack of transparency regarding what information was being captured (Incident 95).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
Presenting & Exploring Outputs: Facilitating user cognition through presentation of outputs.	
25. Presentation of Outputs. What kinds of critical decisions will be most important for users to make during AI use? How can the AI provide the right presentation for time pressured, high-consequence decision making? *Example: A military AI provided decision makers with all of the right information, but did not highlight the relevant information to enable a quick and actionable decision: "The problem with this AI in particular is that if used properly it provides a wealth of information, but… you as a human, you have to look through the parameters, figure out what it declared this one a hostile, it's because [Criteria A], [Criteria B]…"* *Example: Physicians lost trust in a clinical reasoning AI, despite its high performance, because outputs were presented in a manner that was unintuitive for medical decision making: "Some of the issues were… how it presented information to the physicians. [COMPANY] had their own idea on how to do that, but …. Maybe the AI was working better than it seemed, but the way you presented it really affected their perceptions."*	
26. Curation of Outputs. How might different use cases or user roles drive expectations for the quantity, filtering, and logical aggregation of outputs? In what scenarios might users want more or less context or explanation of outputs? *Example: A user lost trust in a content search and curation AI because the AI consistently provided too many results, and results that were not relevant to their specific needs: "It brought back too much, and some was garbage. Not just one man's trash is another man's treasure, but like, a report vs an evaluation of a report by a low level line analyst in a remote location, which is not as important to me as the original report… So those things, the inappropriate results and the volume of results got me to lose trust in it and abandon it."*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
27. Operationalization of Outputs. Will users know what to do with the outputs? How will they be aware of the capabilities, limitations, and conditions for which outputs are validated in various scenarios and contexts? *Example: Users lost trust in a Geospatial AI, despite believing its outputs were correct, because they could not determine how to act on the outputs in a high-consequence scenario: "It will give you a GEO Plot with red, yellow, green… and red means… uh… We realized we don't know what it meant… Don't go there ever? Be [extra] alert if you do go there?" (Dorton & Harper, 2022a).*	
28. Explainability of Outputs. How might the users readily determine how (generally) and why (in specific cases) the AI is producing the outputs that it is? *Example: Users gained trust when the AI highlighted, in plain language, what features it was basing its determinations on for a given set of inputs: "You could look at the model and see the words and phrases it sorts for… the model gave you some insights as to why it was flagged. This reinforced that it was capturing the right things" (Dorton & Harper, 2022a).*	
29. Outside Confirmation of Outputs. How might users of varying levels of domain expertise assess the correctness of AI outputs (i.e., AI performance)? What if results are especially counterintuitive? What other sources of information (e.g., people or sensors) might users rely on to confirm AI outputs? Are AI outputs in a format that easily facilitates this confirmation? *Example: Users calibrated their trust in the AI by confirming AI outputs with non-AI sources: "I talked to [another intelligence cell] and they confirmed the threat [the AI identified] was in the area" (Dorton, 2022).*	
30. Calibrated Trust. How might users arrive at too much or too little confidence in the AI's outputs? How might the AI encourage appropriate scrutiny of outputs? *Example: A TikTok user was surprised to learn that she had been seeing both true and fake news stories, including deceptive videos (Incident 185).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
31. Completeness of Outputs. How might users be (un)aware of the exhaustiveness and relative importance of various inputs of the AI? What variables are not accounted for? How might the AI make obvious the limitations of current inputs, and the effects of limitations on outputs? *Example: A user lost trust in an AI-driven dashboard because it misleads analysts by presenting findings based only on variables with readily available data, rather than all variables that are typically considered by humans: "They're important to consider in your decision calculus… you're just leaving out variables because you don't have good data on them… The AI looks impressive, but … they might just trust or accept what it says and not consider what the equation is or what variables are left out."*	
32. Raw Data Inspection. What kinds of "raw" input data might users rely on to verify or interpret AI outputs? How might this information be available when needed, but hidden when not? *Example: Users lost trust and protested when raw data were removed from a display as AI was introduced, since they relied on the combination of raw data and AI annotations to determine trustworthiness of outputs: "They removed the raw [data] in the new one and it really pissed off a lot of people. They can no longer see the little … cue[s] to see where the AI is missing it: [You know] something's there but it hasn't tripped the computer. They actually went back and added [raw data as a real-time check on the AI]."* *Example: After losing trust in the AI, users added steps to their workflow to check every input to the system before running it, which added a considerable amount of workload: "We had to go through every line of data. A lot of human effort was spent cleaning the data. It has influenced how much time the team [now] spends digging to find the right data" (Dorton et al., 2022).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
Understanding AI Behavior: Establishing and maintaining a shared mental model between users and the AI.	
33. Plan Awareness. How might users develop and maintain an understanding of the AI's plans? How might the AI make it immediately obvious when it is deviating from plans? *Example: UAS pilots gained trust in the flight control AI when they had more knowledge of vehicle's planned route, because they could more readily recognize and take over when deviations occurred: "Now the pilot has to know much more in depth about exactly what the flight plan is. That has seemed to help with trust, knowing more about what the expected [route] is. As more incidents occurred, we want the pilot to know exactly what's going to happen."*	
34. Symmetric Feedback. How might users know that the AI is working properly over long durations of use when no outputs are generated? How would users prefer the AI provide positive and negative feedback during use? *Example: Users lost trust in an AI designed for a vigilance task because it did not provide feedback that it was working as designed, but just had not found the entity of interest: "We thought it was broken because it never worked… it was on all the time and didn't spit anything out until it received something. We used it extensively afterwards… I was a lot less nervous when we didn't pick anything up because I trusted the indicators that it was operating correctly." (Dorton & Harper, 2022a; Dorton et al., 2022).*	
35. Probing & Checksums. How might users employ checksums or other methods to probe the AI *in situ* to ensure it is working correctly? How might the AI help facilitate such checksums with minimal effect on operations? *Example: Users altered the value of a certain input parameter as a checksum to make sure the AI was working correctly and that they could trust its outputs: "Yeah the checksum [parameter] is asked on a routine basis [now]… [Analysts] learned to add the [checksum parameter] to the brief if they knew I was the reviewer." (Dorton et al., 2022).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
36. Long-term Recalibration. How might users recalibrate their trust in the AI if they spend extended periods of time not using it? How will they know if the AI's capabilities changed? *Example: Users simply assumed that the AI had improved over time without any supporting evidence, making themselves vulnerable to over-trusting or misusing the AI: "I would have assumed the [AI developers] had done their due diligence and started from the failures that we previously had" (Dorton & Harper, 2022b).*	
Work System Collaboration: Using the AI in the context of third parties within a work system.	
37. Authorities. How might parties with authorities in the AI-enabled workflow hold up time-sensitive operations? What information will they need, from who (people)/what (AI), to fulfill their role in a timely manner? Can the authorities be delegated if it is unreasonable to act at the required speed? *Example: Trust was lost when military pilots had to break the chain of command because the person with authority to give an order did not have the AI-provided information required to give that order fast enough.*	
38. Shared Awareness. How might disparities in AI outputs across different roles cause confusion or conflict? How can the AI promote shared awareness and collaboration across various roles while still tailoring outputs to individual needs? *Example: Military users had different operational pictures generated from a common AI-enabled system, requiring them to verbally coordinate different subsets of data, delaying critical action.*	
39. Third-Party Acceptance of Outputs. How might users or other third parties choose to accept or reject AI outputs, and will those criteria or thresholds be known by others in the broader sociotechnical system? How might the AI facilitate consistency in how people use or disregard outputs? *Example: Users lost trust in an AI when it was not clear whether the AI missed detecting something, or if the AI detected it and other users simply dismissed the detection for a variety of plausible reasons: "[I don't know] whether [the AI] didn't see it, or it saw it, flagged it, and a human disregarded it… A human could have saw it and said 'well [criteria wasn't met] so I won't take action yet'" (Dorton, 2022).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
40. Third-Party Misuse of Outputs. How might less-savvy users or third parties misuse the outputs of the AI? What downstream effects might that cause in the work system? How can the AI mitigate these misuses? *Example: Expert users lost trust in AI because new users pushed AI outputs to superiors as fact, despite the AI only being valid for quick estimates: "I understand what it should be used for… I lost trust in [other] people's use of the tool… We take a month or longer to find the actual [answer], so when our number comes out, the difference with the AI was off. We essentially had to tell people that they had to tell their bosses that they were wrong" (Dorton & Harper, 2022a; Dorton, 2022).*	
Testing, Debugging, & Error Handling: Determining the causes of problems to calibrate and repair trust.	
41. User-Centered Performance. What performance measures matter to users in the context of their work? How might the relevance of performance measures differ across operational contexts? Does the test and evaluation plan focus on the measures most relevant to users? *Example: Intelligence users gained trust in an AI with poor accuracy (e.g. F1) and precision, because recall was critical in the context of their work (i.e. they did not want to miss any real threats): "If it found something for me at all it was a huge positive, because there was such a huge quantity of data that there was probably no way for me to get to it in a practical sense- I had nothing to lose by using the tool" (Dorton & Harper, 2022a).*	
42. Forensic Support. What heuristics or cues do users and maintainers rely on to troubleshoot issues with the AI? How might the AI help them readily answer the most diagnostic questions after an incident? How might the AI fail to capture a significant irregularity or error in logs or debug outputs? *Example: UAS operators mentioned how determining if an event was due to HW or SW failure was a first step, and the importance of logs helping them diagnose issues to repair trust in the AI: "First step is [determining] if it's HW or SW. If it's HW we can fix it [or] complain to the manufacturer. But restoring trust in AI when you don't know what part was broken… even if it's working… trust doesn't come back."*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
43. Recognizable Warnings. What types of issues might be noticed by experts but not novices? How can the AI give more obvious indications and warnings to novice users? *Example: Trust was lost by a senior UAS pilot when only expert UAS pilots deeply familiar with the underlying flight control AI realized that observed flight behaviors were a safety issue, when novices observing the same behaviors did not: "They were just more comfortable with the behavior and maybe that's just because they didn't know the software in and out… lacking knowledge of charting that flight path… to do that they would need a deeper understanding."*	
44. Hardware Errors. What issues might come up with real hardware configurations that do not manifest in a simulator? How can these be tested in a realistic controlled environment? How can the AI recognize and make users aware of hardware errors? *Example: UAS pilots lost trust after a faulty configuration between hardware components caused a software error and erratic UAS behavior, forcing a halt in operations: "Normally there's a setting that the aircraft remembers, if you don't set it, it will hold the last known [hardware configuration]… But instead it was [showing null values for flight controls]… that's what caused it to spiral… It was a bug in the software that forgot what the failsafe setting was, so it would go to [null commands]."*	
45. Graceful Degradation. How might the AI perform poorly or unpredictably if some resource or component it depends on degrades or becomes unavailable (e.g., connectivity, power, sensors or other data streams)? How will the user know if these sources have been interrupted? How might the AI more gracefully degrade during such losses? *Example: Trust was lost when Cruise autonomous vehicles lost connection with the central server, which caused them to stop and block roads until they could be moved manually (Incident 253).* *Example: UAS pilots lost trust in a flight control AI because it did not support graceful degradation of capabilities, and simply crashed the UAS if anything went wrong: "…It can't just halfway work- it has to work. There's [got to be] degraded states where, you know, I can manually fly it through radio… so I'm [just] keeping it airborne vs flying the mission. It has to degrade gracefully."*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
Sustainment: Maintaining and upgrading the AI to meet user needs after deployment.	
46. Recency. How might the AI become outdated for user needs, or for contexts in the current state of the world? How frequently are updates needed (to data, models, etc.)? *Example: Users lost trust in Waze when it directed users in California to drive through wildfires because it could not be updated quickly enough to keep pace with their spread (Incident 22).*	
47. Environmental Robustness. How might deployment to a new target operational environment or use case affect performance? How might the AI need to adapt to changing inputs, operational use cases, or required outputs over its lifecycle? *Example: Trust was lost when a firearms detection system was deployed in a school setting, where it failed to distinguish between firearms and school supplies (Incident 349).* *Example: A UAS pilot said that they would need to see the UAS successfully operate in different environments to trust it completely: "I'd have to see it run in many environments. We always run it in the same field, a semi-controlled environment. We'd want it in a separate field, a desert, etc. We'd run it in many environments [at] different times of year. Different corner cases."* *Example: Users gained trust in a system that classifies threats because it was able to update its models based on new inputs as the threats changed: "The system was able to learn new data… getting data from intelligence on targets that are always changing" (Dorton & Harper 2022a).*	
48. Technical Support. How might users receive technical support during operational use of the AI? What questions and technical issues will most likely need support? *Example: Military users gained trust in an AI because they had near-real time technical support to fix issues that emerged during operational use: "When there is a big failure we know we can get it fixed… the fact that I can yell across the hallway and get answers and fixes quickly is a big factor in my trust." (Dorton, 2022).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
49. Update Impacts & Awareness. How might routine updates create unexpected behaviors? How will users become aware of updates to the AI or other associated components, and understand the impact of those updates on observable behavior? *Example: Users lost trust when, after a Tesla update removing dependence on radar, vehicles began braking suddenly and unexpectedly (Incident 208).* *Example: Pilots noted that they would recalibrate their trust and change workflows after updates were made to UAS flight control AI: "We've flown it many, many times…[But] If we [updated] the software, I'd want to reset my [trust level]."*	
50. Continuous Improvement. What feedback loops exist for users to report issues or desired improvements? Is funding secured and management committed to act on these requests and continuously improve the AI through its expected lifecycle? *Example: Users lost trust in an AI when they realized they had no way to have developers update the AI to meet their needs that were evolving over time: "I only talked to people in my branch – they basically just commiserated… I didn't have any insights on [how to give feedback to] the developers or maintainers" (Dorton, 2022).*	
51. User Adaptations. How might users adapt their workflows based on gaining or losing trust in the AI to accomplish certain functions? How might the AI be designed to support such adaptive workflows? *Example: Signals Intelligence (SIGINT) users gained trust in an AI to do some functions, but lost trust in it to do other functions, so they adopted their team's workflows and responsibilities based on what they trusted the AI to do correctly in a high-consequence workflow, "Yeah we primarily didn't use it for [entity] ID, just for detection as a tipper" (Dorton et al., 2022).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
52. Evolution. How might new users and missions, over longer periods of time, change the underlying mechanics of the AI? How might the AI make others aware when new users and their inputs modify the functionality of the AI? *Example: A network management AI experienced complexity creep and changes to policies and practices as new rules and parameters were added to the system to meet new needs over decades of time: "It's mostly canned since we've been using it since the 80s, so a lot of this is codified. But you get new people doing it and they add new things, new parameters... you're adding more and more people to the network with more and more reporting requirements and parameters."*	
53. Skill Decay. How might the introduction of AI drive skill decay or complacency in users? How can the AI support them in critical situations requiring deep attentive and cognitive resources? *Example: Military users said they might lose trust in a system because of skill decay and complacency as AI is upgraded to take on an increasing proportion of a workflow: "What I'm concerned about… is we are trying to give more and more tasks to the machines because they want to think about other stuff and move up the cognitive ladder, but when shit goes down you need to know how to do it."* *Example: Senior users lost trust in an adopted AI because their junior colleagues were no longer trained on how to do the analysis manually (after the AI was adopted), which meant these junior analysts were not able to recognize when the AI was performing incorrectly: "These abilities of the human specialist in the loop are decreasing… they aren't able to pick out errors." (Dorton 2022).*	

References for ELATE

BBC (30 May 2018). Tesla hit parked police car 'while using Autopilot.' *BBC News.* https://www.bbc.com/news/technology-44300952

Dorton, S.L. (2022). Supradyadic trust in artificial intelligence. *Artificial Intelligence and Social Computing, 28*, 92-100. https://doi.org/10.54941/ahfe1001451

Dorton, S.L., & Harper, S.B. (2022a). A naturalistic investigation of trust, AI, and intelligence work. *Journal of Cognitive Engineering and Decision Making, 16*(4), 222-236. https://doi.org/10.1177/15553434221103718

Dorton, S.L. & Harper, S.B. (2022b). Self-repairing and/or buoyant trust in artificial intelligence. *Proceedings of the HFES 66th International Annual Meeting, 66*(1), 162-166. https://doi.org/10.1177/1071181322661098

Dorton, S.L. & Harper, S. (2021). Trustable AI: A critical challenge for naval intelligence. *Center for International Maritime Security (CIMSEC).* Retrieved from: https://cimsec.org/trustable-ai-a-critical-challenge-for-naval-intelligence/

Dorton, S.L., Harper, S.B., & Neville, K.J. (2022). Adaptations to trust incidents with artificial intelligence. *Proceedings of the HFES 66th International Annual Meeting, 66*(1), 95-99. https://doi.org/10.1177/1071181322661146

Peters, T. (5 August 2016). Moms warn of disturbing video found on YouTube Kids: 'Please be careful.' *Today*. https://www.today.com/parents/moms-warn-disturbing-video-found-youtube-kids-please-be-careful-t101552

All Incidents from: https://incidentdatabase.ai

This page intentionally left blank.