ELATE v1.2

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.

Check for updates to the ELATE framework at https://www.mitre.org/news-insights/publication/evidence-based-list-exploratory-questions-ai-trust-engineering-elate.

Approved for public release. Distribution unlimited 24-00712-1.

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).

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."*	
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).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
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).*	
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).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
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).*	
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).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
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."*	
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).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
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).*	
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."*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
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).*	
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).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
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).*	
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.*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
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).*	
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).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
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."*	
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]."*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
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."*	
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).*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
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).*	
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]."*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
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).*	
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."*	

Questions	Response *Respond to the questions in as much detail as possible. How will you know you've addressed this? By when?*
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.