A Naturalistic Investigation of AI, Trust, and Intelligence Analysis: Early Results

Steve Dorton & Sam Harper

VDODHFETAG
29-30 June
"Human Factors of Data Driven Decision Making"

AIRE
authorship & intent recognition environment

HUMAN - AUTONOMY INTERACTION LABORATORY

Agenda

- Introduction
  - Use Case & Challenges
  - Explainable Artificial Intelligence (XAI)
  - Trust
- Methods
  - Critical Incident Technique
  - Participants
  - Dataset
- Results
  - Thematic Analysis
  - Trust Rebound
- Discussion
  - Key Findings
  - Future Work
  - Q&A

Acknowledgements

This work is supported by the US Army Combat Capabilities Development Command (DEVCOM) under Contract No. W56KGU-19-C-0062.

The views, opinions and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other documentation.

This presentation was approved for public release as Distribution A on 14 JUN 2021, Approval #A271

Use Case & Challenges

- Intelligence analysis is:
  - A continuous process that requires collecting, processing, exploiting, and disseminating information to inform decision making [1]
  - A high-stakes domain that is rife with challenges for individual and team cognition, including [2]:
    - A surplus of non-diagnostic information
    - Time-constrained cognition
    - Working across a distributed team
- Artificial Intelligence and Machine Learning (AI/ML) have numerous applications across different disciplines of intelligence analysis [3]
- Tools for Intelligence Analysis are often built without considering the complexity of human cognition, leading to them being [4]:
  - Ineffective
  - Rigged/Gamed
  - Disused

Explainable AI (XAI)

- Explainable AI (XAI): AI that can be easily understood, and a human is able to interpret *why* and *how* the system arrived at a specific decision [5,6]
- A common challenge for XAI is the tradeoff between interpretability and accuracy
- Factors of XAI include:
  - **Justification**: The AI explains why the answer provided is a good answer [7, 8]
  - **Transparency**: The AI explains how the system reached the answer (where decisions are explained in terms, formats, and languages we can understand) [7-9]
  - **Conceptualization**: The AI clarifies the meaning of concepts [7,8]
  - **Learning**: The AI teaches you about the domain [7,8]
  - **Bias**: The AI has verification that decisions made based on the AI system were made fairly and are not based on a biased view of the world [9]

Trust

- Trust: “The attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” [10]
- Users must trust the agent in order to be able to make informed decisions with them
- Users must also have the appropriate level of trust (calibrated trust) in order to effectively use systems [11]
  - Too ***much*** trust: Misuse of a system where the user relies on the system for more than is intended
  - Too ***little*** trust: Disuse of a system where the user does not take advantage of the capabilities of the system

Factors Affecting Trust

- Reputation: The agent has received endorsement or reviews from others [12]
- Usability: The agent is easy to interact with [12]
- Reliability/Predictability: The agent is reliable and consistent and/or predictable in functioning over time [13-15]
- Understandability/ Explainability: The extent to which you are able to understand what the agent is doing, why it is doing it, and how it is doing it [15]
- Security/ Privacy protection: The importance of operational safety and data security to the agent [12]
- Utility: The usefulness of the agent in a task [12]
- Robustness: The agent is able to function under a variety of circumstances [16, 17, 12]
- Goal Congruence: The extent to which the agent's goals align with your own [12]
- Feedback: The agent is able to explicitly give clear feedback on its intended actions [15]
- Errors/False Alarms: Information provided by the agent does not contain errors or false alarms [11]

Methods

Critical Incident Technique

- The Critical Incident Technique (CIT) is a method to collect data on systematically defined criteria from observed incidents [18, 19]
- Prompt to elicit stories (one AI/ML and one human):
  - Think of an AI/ML technology or person you really trust or distrust in the context of intelligence analysis.
  - Why? Was there a defining event (or series of events) that where you gained or lost trust?
- We conducted three passes for each story:
  1. Background Information
  2. Incident Description
  3. Afterwards/Retrospective

Thematic Analysis

- Read through the transcript to identify notional high-level themes in responses
- Two researchers independently coded responses for themes
- Codes were reviewed and differences reconciled
- Resultant themes enabled:
  - Descriptive and inferential statistics (where appropriate)
  - More detailed thematic analysis within high-level themes

Participants

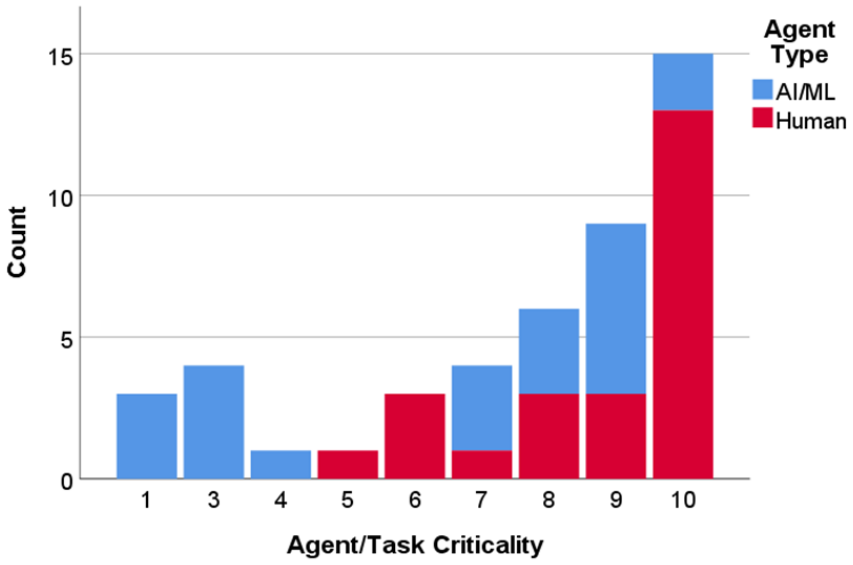
- 29 Intelligence Professionals: Collection, Analysis, Decision Making
- Various backgrounds across IC and DOD: CIA, DIA, NSA, DOS, DCSA, ONI, NASIC, USN, USA, USMC, and USCG
- 563 total years of experience in intelligence (*M* = 20.11, *SD* = 11.26)

Intelligence Discipline	*n*	*M*	*SD*	Sum
All-Source	20	13.18	11.35	346
SIGINT: Signals Intelligence	21	20.11	9.62	324
ELINT: Electronic Intelligence	3	17.33	14.19	52
COMINT: Communications Intelligence	1	25.00	-	25
OSINT: Open Source Intelligence	13	13.38	10.08	174
HUMINT: Human Intelligence	4	10.50	9.88	42
IMINT: Imagery Intelligence	2	11.00	1.41	22
GEOINT: Geospatial Intelligence	3	13.33	10.41	40
MASINT: Measurement & Signature Intelligence	6	15.17	12.75	91
ACINT: Acoustic Intelligence	6	15.00	8.94	90

Dataset

- Generated a total of 55 stories
- Most stories of humans were negative
- Agent/Task criticality was high (*M* = 7.59, *SD* = 2.79)
  - 1 = Redundancies and workarounds are readily available
  - 10 = Results in loss of life or complete mission failure
- AI/ML work was bi-modal, and significantly less critical than human work (*U* = 125.00, *Z* = -.313, *p* < .01)
  - AI/ML often has human redundancies in earlier phases
  - Humans are generally critical path, sometimes with other human redundancies

Valence	Agent Type	
	AI/ML	Human
Gained Trust (+)	12	7
Lost Trust (-)	13	17
General Story (+/-)	6	



Thematic Analysis
Trust Rebound

Results

Thematic Analysis Results

- Eight (8) themes were identified in participant responses
- Each theme can be examined by agent type (AI/ML vs Human)

Theme	Reliability *K* (*p*)	AI/ML (*n* = 27)	Human (*n* = 27)	Total (*N* = 55)
Reputation	.80 (< .01)	7	19	26
Character	.78 (< .01)	5	20	25
Trust by Proxy	.62 (< .01)	11	2	13
Users in Development	.33 (< .01)	10	1	11
Impressions	.20 (> .05)	1	4	5
Trust by Failure	.73 (< .01)	4	0	4
External Stress	.65 (< .01)	0	3	3
Asymmetric Feedback	.85 (< .01)	3	0	3

Thematic Analysis (cont.)

Trust by Proxy (*n* = 11): Trust in AI/ML was affected by human behavior or other involvement with the AI/ML

- Trust lost, usually regarding the provenance of data (*n* = 7)

> *"It's only as good as the information that's entered. There's a lot of people who have accounts… you don't know everyone who is putting info into it."* [CIT 22]
>
> *"They didn't recognize the fact that you can't just have one person [tagging] data- it wasn't the performance of the neural net."* [CIT 50]

- Trust gained, due to human involvement with data (*n* = 2)

> *"A human verifying a nomination gives me much higher confidence than the algorithm feeding itself."* [CIT 29]

- Trust gained in humans because of AI/ML utility (*n* = 2)

> *"… I'd be able to get some sleep because I would be able to trust more junior officers are in better hands, I trust the automation more."* [CIT 36]

Thematic Analysis (cont.)

Users in Development (*n* = 10): Trust was affected by…

- Trust in the AI/ML was gained because end users and SMES were involved in the development process (*n* = 6)

> *"We were lucky to have and experts or former targeting officers, it was helpful for the development of models."* [CIT 55]

- Conversely, trust was lost because end users were not involved in the development process (*n* = 4)

> *"'It was the mathematicians that developed it, and they did **not** include the experts enough. The design requirement input to develop the AI was flawed."* [CIT 10]

Thematic Analysis (cont.)

Reputation (*n* = 7): Reputation of the AI/ML affected trust

- Used to calibrate trust in an AI/ML technology before using it (*n* = 6)
  - A positive reputation increased trust in the AI/ML before using it (*n* = 4), but a negative reputation decreased trust before use (*n* = 2)

> *"I've heard anecdotal stories where people tell me it didn't pick this [event] up and it should have."* [CIT 44]

  - Sometimes a positive reputation was confirmed (*n* = 2) by the experience, but sometimes it was unwarranted (*n* = 2), and trust was lost in the AI/ML

> *"It was really pushed down from high… 'hey you guys should use this all the time- it's got great performance…' so first you really hop into it… but the novelty wore off fast."* [CIT 41]

- Reputation also has a role in regaining trust: In some cases (*n* = 2) the participants noted they would need corroboration from others, in addition to seeing improvements first-hand

Thematic Analysis (cont.)

Character (*n* = 5): A lack of character boosted trust in AI/ML systems and/or highlighted the importance of the user towards mission success

Conversely, trust was lost in humans because of "character flaws" such as selfishness (goal congruence) or dishonesty

"It's a computer program that does what we tell it to do… so we force our own goals on it." [CIT 28]

"The machine is doing to do what it's going to do. With open data part of the problem is how the machine has very little blame, because it's doing what it's told to." [CIT 55]

"I don't believe technology is good or bad, it's really the way I use it." [CIT 50]

Thematic Analysis (cont.)

Trust by Failure (*n* = 4): Participants gained trust in the system after it failed

- Failure helped identify limitations or boundary conditions for the AI/ML technology (*n* = 2)

> *"In a way, [the system] increased my trust because I have a better understanding of what challenges can occur."* [CIT 50]
>
> *"I didn't necessarily lose trust in the system- I learned its limitations."* [CIT 17]

- Failure demonstrated that the system behaved consistently or predictably, if nothing else (*n* = 2)

> *"I guess I've gained confidence because the algorithm consistently gives results that are imperfect."* [CIT 29]
>
> *"Yeah. Abject failure improved my trust in the machine… it demonstrated the machine is doing what I told it to."* [CIT 55]

Thematic Analysis (cont.)

Asymmetric Feedback (*n* = 3): Asymmetric feedback adversely affects trust

- Participants gained trust in AI/ML once positive feedback was finally provided, since no negative feedback was regularly provided (*n* = 2)

> *"We thought it was broken because it never worked… When you were underway it was on all the time and didn't spit anything out until it received something."* [CIT 53]

- One participant cited a similar issue with an ELINT-based system

> *"They won't regain confidence because you won't know if you got saved– you don't know that you didn't get blown up,"* [CIT 18]

- However, there was an overall lack of feedback in both directions that affected trust

> *"Conversely, if you get blown up you don't know if [the AI/ML] broke or if it's a different [threat feature] [than it is currently programmed to look for]."* [CIT 18]

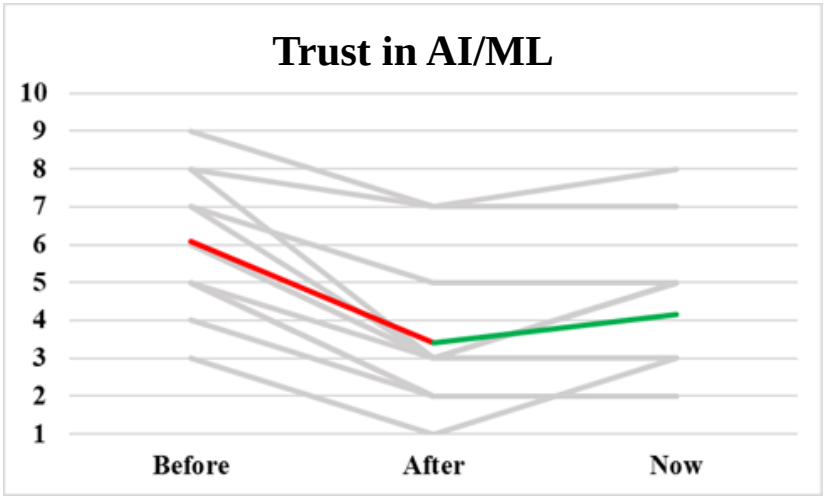
Trust Rebound Results

Trust Rebound Results

- Asked about trust at three points in time: Before (T_B), After (T_A), and Now (T_N) on a 1-10 scale, where:
  - 1 = No trust- you assume everything the AI/ML does is wrong or a failure.
  - 10 = Complete trust – you assume everything the AI/ML does is right or successful.
- Noticed that trust "rebounded" after losing it in an incident, despite having no further interactions
- Created additional measures:
  - Loss (L) = T_A-T_B
  - Rebound (R) = T_N-T_A
  - Rebound Ratio (RR) = R/L
- Conducted thematic analysis of rationales to find *why*

Trust Rebound (cont.)

- Rebound occurred in 5 of 12 incidents (41.7%) where trust was lost in AI/ML
- Developed hypothesis: $T_N > T_A$
- Descriptive statistics showed marked increase, despite low frequency of occurrence:
  - Mean T_B = 6.08 (*SD* = 1.78)
  - Mean T_A = 3.42 (*SD* = 1.93)
  - Mean T_N = 4.17 (*SD* = 1.99)
  - Mean Loss = 2.67 (*SD* = 1.07)
  - Mean Rebound = 0.75 (*SD* = 0.97)
  - Mean Rebound Ratio = .30 (*SD* = 0.40)



There was a significant rebound effect (n = 12, Z = -2.12, p < .05)

Trust Rebound (cont.)

- But *why* does trust rebound?
- Conducted thematic analysis of cases with rebound (*n* = 5)

Theme	*n*	Example Rationale(s)
Assumed Improvement	3	CIT 8: "I would have assumed [the AI/ML developers] had done their due diligence and started from the failures that we previously had." CIT 33: "Algorithms and technology being what it is, is always advancing and becoming more reliable over time… with trial and error. I lost trust at that time, but realize it's a work in progress."
Knowledge of Improvement	1	CIT 26: "It's got to prove that it's better, but there have been a lot of [interface] changes and it provides feedback now…" *CIT 40: "The newer systems [have new AI/ML feature]… There's data we have now that he didn't have back then. He probably wouldn't have made the [bad] call. He'd be a better operator off the bat."
Self-Maturation	1	CIT 31: "I would most likely understand how that engine works and I know [how to run] queries better, so I would have direct access to it and be able to control it."

Discussion

Key Findings

- AI/ML has a bifurcated role in intelligence (all-in or w/ human redundancy)
- Trust in AI/ML is difficult to disentangle from user behavior (i.e. trust by proxy)
  - Data entry/provenance
  - Misuse/disuse/abuse
- Involving end users and domain experts is critical to increasing trust in AI/ML systems
- Reputation plays a large role in calibrating trust in AI/ML
- Trust in AI/ML rebounds, even with little/no further interaction with the system

Future Work

- Conduct primary analysis for factors of trust
  - Descriptive statistics and qualitative analysis
  - Regression (sample size permitting)
- Analyze magnitude by theme (loss/gain)
- Analyze survey data
  - Propensity to Trust Technology (PTT) survey
  - Ranked priorities for XAI
  - Ranked priorities for trust factors
    - AI/ML
    - Human
- Extract design seeds

What questions do you want answered?

How would you approach this?

References

1. Clark, R.M. (2014). *Intelligence Collection*. Los Angeles, CA: CQ Press.
2. Trent, S.A., Patterson, E.S., & Woods, D.D. (2007). Challenges for cognition in intelligence analysis. *Journal of Cognitive Engineering and Decision Making, 1*(1), 75-97.
3. Lee, M., Valisetty, R., Breuer, A., Kirk, K., Panneton, B., & Brown, S. (2018). *Current and future applications of machine learning for the US Army* (Report No. ARL-TR-8345). Aberdeen Proving Ground, MD: US Army Research Laboratory.
4. Moon, B.M. & Hoffman, R.R (2005). How might "transformational" technologies and concepts be barriers to sensemaking in intelligence analysis, *Proceedings of the Seventh International Naturalistic Decision Making Conference*, J.M.C. Schraagen (Ed.), Amsterdam, The Netherlands, June 2005.
5. Volz, V., Marjchrzak, K., & Preuss, M. (2018). A social science-based approach to explanations for (game) AI. *2018 IEEE Conference on Computational Intelligence and Games (CIG),* 474-481.
6. Michael, N. (2019). *Trustworthy AI - Why Does It Matter?* National Defense. https://www.nationaldefensemagazine.org/articles/2019/11/19/trustworthy-ai-why-does-it-matter
7. Roth-Berghofer, T. R., & Cassens, J. (2005). Mapping goals and kinds of explanations to the knowledge containers of case-based reasoning systems. *ICCBR 2005, 3630*, 451–464.
8. Sørmo, F., Cassens, J., & Aamodt, A. (2005). Explanation in case-based reasoning – perspectives and goals. *Artificial Intelligence Review, 24*(2), 109–143.
9. Hagras, H. (2018). Toward human-understandable, explainable AI. In *Computer, 51*(9), 28-36. doi: 10.1109/MC.2018.3620965.
10. Lee, J., & See, K.A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors, 46*(1), 50-80 .
11. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434.
12. Siau, K. & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal, 31*, 2.
13. Muir, B. M. (1994) Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37*(11), 1905-1922, DOI: 10.1080/00140139408964957
14. Holmes, J. & Rempel, J. (1985). Trust in close relationships. *Journal of Personality and Social Psychology, 49*. 10.1037//0022-3514.49.1.95.
15. Balfe, N., Sharples, S., & Wilson, J. R. (2018). Understanding is key: An analysis of factors pertaining to trust in a real-world automation system. *Human Factors, 60*(4), 477–495.
16. Sheridan, T. B. (1999). *Human supervisory control*. In Sage, A. P., Rouse, W. B. (Eds.), Handbook of systems engineering and management (pp. 645–690). New York, NY: Wiley & Sons.
17. Woods, D. D. (1996). *Decomposing automation: Apparent simplicity, real complexity*. In Parasuraman, R., Mouloua, M. (Eds.), Automation technology and human performance (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.
18. Flanagan, J.C. (1954). The Critical Incident Technique. *Psychological Bulletin, 5*, 327-358. doi: http://dx.doi.org/10.1037/h0061470
19. Rosala, M. (2020). *The Critical Incident Technique in UX*. Nielsen Norman Group. https://www.nngroup.com/articles/critical-incident-technique/

Steve Dorton
Director, Human-Autonomy Interaction Laboratory
Sonalysts, Inc.
sdorton@sonalysts.com

Q & A