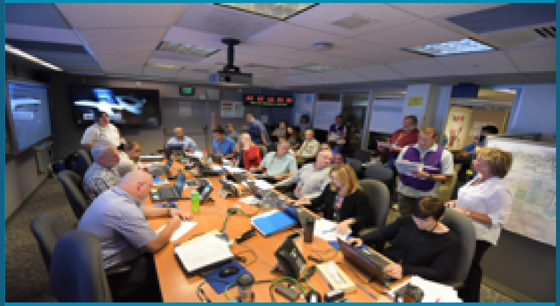
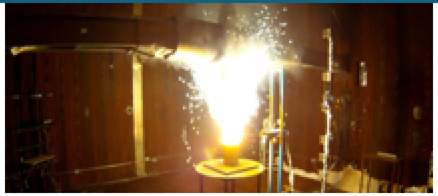
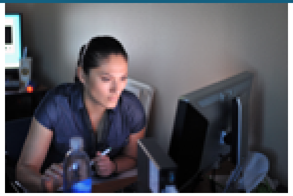
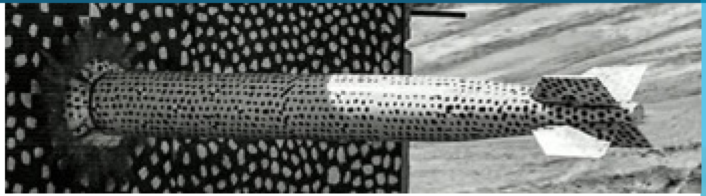
SAND2020-12976PE

Research Needs for Trusted Analytics in National Security Settings









David Stracuzzi and Ann Speed

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Overview

- Bottom Line Up Front
- The Prevailing Hypothesis
  …And Why Its Wrong
- A Motivating Example
- Important Definitions
- Challenges in Measuring Trust
- An Experimental Perspective on Trusted Analytics
- Research Gaps
- Principles for Trusted Analytics Research

Bottom Line

Drivers

- Analytics and AI are here to stay in national security domains
- Complexity and opacity of models raise questions about appropriate use:
  - How do we achieve it?
  - How do we measure it?
- Many gaps in current academic literature, commercial applications
  - Mission contexts often violate laboratory assumptions
  - Mission consequences often more severe than laboratory or commercial applications
  - Ground truth often presents a special challenge in national security domains

Bottom Line

Goal

- Establish principles to guide future research in trusted analytics
  - Trust is not the goal – we want analytics that improve decision making and are correctly used
  - Application domain expertise needs to be well represented during development
  - Mission applications need to be rooted in theory of ML/AI/data science

Why Are Trusted Analytics So Challenging?

- **Setting**: Mitigating human inadequacy
  - Capacity – Too much data from too many sources
  - Time – Maintain situation awareness and decision making as dictated by application
  - Bias and Error – Reduce unjustified assumptions (perspective) and thinking errors
- **Constraints**:
  - Analysts and end-users have expertise not captured by analytics
  - Analysts and end-users may lack expertise in computational analytic methods
  - Ground-truth limitations
- **Outcomes**:
  - Failure to establish appropriate trust in analytics can make mission performance worse

Prevailing Hypothesis

"People don't use analytics because they don't trust them"

Analytic developers response:

If we:

- Produce higher-quality solutions,
- Provide more information about our methods, or
- Explain how our methods made predictions…

Then:

- People will necessarily trust and use our analytics, and
- The analytics will always be beneficial.

Trust is not so simple. Developing trusted analytics less so

Motivating Example: Coronavirus Testing

- **Virus tests are similar to detection algorithms**
  - Black boxes that perform specific tasks with hard-to-estimate performance
  - Binary output despite complex false positive and false negative rates

- **COVID-19**
  - Wide range of symptoms (weak indicators)
  - Asymptomatic cases (hidden patterns)
  - Lack of comprehensive testing (can’t measure everything we’d like)

Motivating Example: Coronavirus Testing

- **Decision-making challenge:**
  - Task does not operate in a vacuum – many other possible diagnoses
  - Test can <u>augment</u> or <u>supplant</u> physician judgement
  - How to incorporate test results appropriately with:
    - Other tests?
    - Patient background?
    - Patient symptoms (or lack thereof)?
    - Exposure level?
    - Physician background knowledge and experience?

What constitutes a well-calibrated decision?

Important Definitions

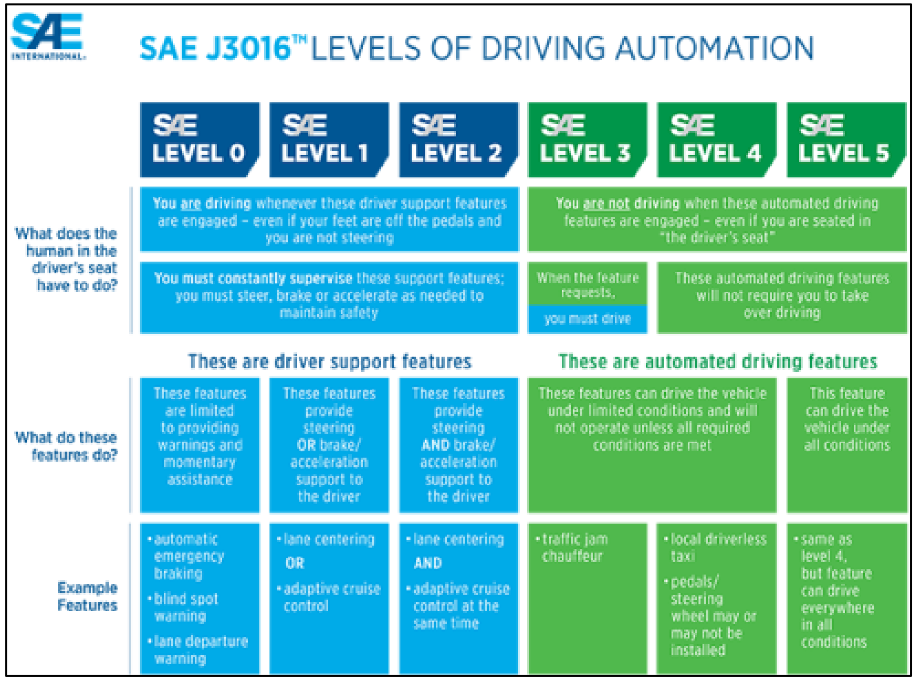
- **Analytic**: any computational method that connects data with decisions
  - Focus on data-driven analytics
  - These tend to be correlational
  - Distinguished from simulation models (causal)

- **Automation**: technology that selects data, transforms information, makes decisions, or controls processes
  - Includes AI/ML/stats models
  - Large, relevant literatures

HIGH
10. The computer decides everything, acts autonomously, ignoring the human.
9. informs the human only if it, the computer, decides to
8. informs the human only if asked, or
7. executes automatically, then necessarily informs the human, and
6. allows the human a restricted time to veto before automatic execution, or
5. executes that suggestion if the human approves, or
4. suggests one alternative
3. narrows the selection down to a few, or
2. The computer offers a complete set of decision/action alternatives, or
1. The computer offers no assistance: human must take all decisions and actions.

LOW

Levels of Automation, from Parasuraman, Sheridan & Wickens, 2000.



Important Definitions

- **Trust**: measurement of the <u>user</u>, defined in terms of subjective and objective measures
  - Subjective – individual's reported level
  - Objective – comparison of human decision to analytic recommendation
  - Frequent dissociation between subjective and objective measures
  - Not binary – lies on a continuum
  - Influenced by decision environment

- **Trustworthiness**: measurement or property of the <u>analytic</u>
  - Degree to which analytic in general, or prediction in particular, should be relied upon
  - Focus area of AI/ML/stats literature (though often confused with trust)
  - Proposed metrics include:
    - Predictive Performance (such as accuracy)
    - Uncertainty Measurements
    - Model Transparency and Explainability
    - Anthropomorphism
  - Impact of most analytic properties on human trust not well established

Important Definitions

- **Trusted Analytic**: Necessary (maybe not sufficient) conditions:
  - Analytic should demonstrate
    - Validity – an established connection between metric and user trust
    - Properties that are relevant to the application
  - Demonstrated trust and use of analytic
    - Measured subjectively and objectively
    - Uncalibrated – an important research waypoint
  - Demonstrated <u>appropriate</u> trust and use
    - Use needs to be calibrated relative to analytic performance
    - Complex and technically challenging for mission applications

Challenges in Measuring Trust

- Analytics are imperfect predictors
- "Proper use" means correctly accounting for the chance that they are incorrect
  - Ideal decision = Bayes optimal
  - Means any decision error is due only to noise in the data/information
- **Case study:** COVID diagnosis and trust in virus test
  - **Available information**:
    COVID test results, patient symptoms, patient history,
    background knowledge of other diseases
  - **Suppose:** we know the probabilistic relationships among these based on observations
  - **Then:** we can measure the difference between ideal and doctor predictions

Challenges in Measuring Trust

Case Study: Why might the doctor differ from optimal?

- The doctor might…
  - Over/under weight COVID test – improper trust
  - Observe sample that yields different relationships among symptoms, tests, and conditions – disagreement (good trust?)
  - Improperly weight certain symptoms – bad decision criteria (good trust?)
  - Incorporate irrelevant information – bad decision criteria (good trust?)
  - Weights information correctly, but reasons incorrectly – good trust, thinking error
- In general, we **can't distinguish** among these (hard to know doctor's relative weighting)
- In many cases, the **probabilistic relationships are also unknowable** or weakly estimated
- Sometimes the test changes – COVID test may not pick up new mutations equally

Challenges in Measuring Trust

Bottom Line: Ground-truth for calibration entails more than just known theoretical relationships or desired outputs.

We need to know the decision-maker's internal evaluation function and background knowledge

An Experimental Perspective on Trusted Analytics

How Analytics Indicate Trustworthiness (Key Independent Variables)

Correctness of Analytic

- Does analytic answer the question of interest?
- Does the data speak to the question of interest?
- Does analytic respect domain constraints and background knowledge?
- Is it consistent with known examples or expected behavior?

Performance of Analytic

- generalize to new examples?
- Does it distinguish between guessing and reliable predictions?
- Does it meet other performance criteria (speed, efficiency, etc)?

How Users Demonstrate Trust in Analytics (Key Dependent Variables)

Subjective Ratings of Trust

- Does the user indicate they trust the analytic?
- Does the user know when the analytic can

Behavioral Reliance on Analytic

- Do subjective trust ratings and calibration predict actual use?
- Does analytic use correspond to its reliability?

Decision Quality

- Does the analytic improve decision quality?

Additional moderating variables further complicate the picture.

Example Moderating Variables

- User Past Experience with Analytics
- Analytic Anthropomorphism
- User Propensity to Trust
- Autonomy Level of Analytic
- Decision Context (workload, consequence, task difficulty, time limits, etc.)
- User Ability to Perform Task Alone

Research Gaps

- **Strategic Gaps**
  - Generalizability of laboratory research
    to national security environments
    - Differences in consequences
    - Lack of ground truth in national security situations
    - Theoretical models may hide nuances that
      drive mission applications
  - Methodological issues in human subjects studies
  - Lack of theoretical framework of trustworthiness and trust
  - Temporal characteristics of trust

- **Focused Gaps**
  - Trustworthiness characteristics of analytics that engender appropriate trust
  - User, task, and environment characteristics that influence willingness to trust
  - Adversarial conditions
    - Detection of adversarial manipulation
    - Potential vulnerability around manipulating analytics to report overconfidence
  - When to automate and to what level

Principles for Research in Trusted Analytics

Trust is not the goal

- Trust is a mediator – people unlikely to use analytics they don't trust

HOWEVER

- Just because they trust it doesn't mean they trust it appropriately
- Just because they trust it appropriately doesn't mean they use it appropriately
- Just because they trust and use it appropriately, doesn't mean all effects are positive, and
- Just because they say they don't trust it doesn't mean it doesn't impact their behavior (explicitly or implicitly)!

The goals are appropriate analytic use and

Improved mission performance

Finally, users trust and use your analytic appropriately…. *There are still risks!*

Appropriately Calibrated Trust and Use of Analytic

Complacency

- Complacency can lead to over-reliance – especially with highly reliable systems, when user is multitasking, or under high workload
- System errors become harder to detect, thus harder to correct

Lack of Situation Awareness

- Limited ability to recognize when analytic is wrong
- Limited ability to diagnose source of analytic error
- Limited ability to effectively take over for the system being driven by the analytic

Subsequent loss and possible repair of trust

- Computational systems punished more severely than humans for same errors (perfect automation schema)
  - Evidence of individual differences
  - Error rates and types
  - Explanations of errors by automation seems to mitigate effect

Principles for Research in Trusted Analytics

- **Incorporate relevant technical and domain expertise in development process**
  - Mission expertise and background knowledge
  - AI, statistics, computing, and mathematics
  - Experimental psychology and/or human factors
- **ML/AI applications built on theoretical foundation**
  - Methods with well-understood strengths and weakness calibrated to application
  - Avoid poorly characterized, ad-hoc approaches
- **Intentional analytic design to include and respect:**
  - Relevant domain expertise
  - Human user needs
  - Anticipated trust-use pitfalls

Summary

1. **Trusted Analytics is not well-defined in the literature**
   - Intersection of computer science, statistics, human factors, psychology, cognitive science
   - Communities tend to ignore each other
2. **As a field trusted analytics lacks strong theoretical and empirical foundations**
   - Theory of factors that influence trust and how they interact with analytic properties is needed
   - Experimental methodology needs to be strengthened
3. **Several factors to consider in developing trusted analytics**
   - Trustworthiness (according to some metric) does not imply trust
   - Trust does not imply appropriate use
   - Properties of the user, task, and application environment influence trust