Toward a Causal Modeling Approach for Trust-Based Interventions in Human-Autonomy Teams

Putting AI in the Critical Loop: Assured Trust and Autonomy in Human-Machine Teams

Anthony L. Baker[1], Daniel E. Forster[1], Ray E. Reichenberg[2], Catherine E. Neubauer[1], Sean M. Fitzhugh[1], Andrea Krausman[1]

[1]DEVCOM Army Research Laboratory; [2]Southern New Hampshire University

Aberdeen Proving Ground, MD, USA

21 MAR 2022, AAAI Spring Symposium

Approved for public release/distribution unlimited

1

Approved for public release/distribution unlimited

TODAY'S TALK

1. Introduction
2. Trust measurement in human-autonomy teams
3. A model human-autonomy teaming scenario
4. Overview of causal modeling
5. Causal modeling in context
6. Conclusions & next steps

Approved for public release/distribution unlimited

2

INTRODUCTION

- Appropriate calibration of trust in HAT is important to enable effective technology use and interaction
- Maintaining calibration takes 3 steps:
  a. Defining relevant constructs
  b. Using appropriate trust measures
  c. Intervening when trust is outside optimal limits

Approved for public release/distribution unlimited

INTRODUCTION

- Appropriate calibration of trust in HAT is important to enable effective technology use and interaction
- Maintaining calibration takes 3 steps:
  a. Defining relevant constructs
  b. Using appropriate trust measures
  c. Intervening when trust is outside optimal limits
- Multi-method trust measurement approaches are key to more fully understanding HAT trust

INTRODUCTION

- Appropriate calibration of trust in HAT is important to enable effective technology use and interaction
- Maintaining calibration takes 3 steps:
  a. Defining relevant constructs
  b. Using appropriate trust measures
  c. Intervening when trust is outside optimal limits
- We need a similar holistic understanding of the team's trust-related factors

5

INTRODUCTION

- Appropriate calibration of trust in HAT is important to enable effective technology use and interaction
- Maintaining calibration takes 3 steps:
  a. Defining relevant constructs
  b. Using appropriate trust measures
  c. Intervening when trust is outside optimal limits
- **We propose a causal analysis approach to understanding the effects of trust interventions**

6

Trust Measurement in HAT

7

TRUST MEASUREMENT IN HAT

- Trust is one's willingness to be vulnerable to others
  - indicated by prosocial actions and feeling safe, expressing beliefs, etc.
- Team trust is an emergent property
  - Arises and evolves as a function of teammate interactions, which in turn affect individual teammate trust

8

DEFINING TRUST

- Usually defined as traits (relatively stable) or states (dynamic) (Mooradian et al. 2006)
- Different types have distinct antecedents & indicators
  - Affect-based, cognitive-based, swift trust, etc.
- Researchers must make clear decisions about what "trust" means and how it is situated in a causal network

MEASURING TRUST

- Critical to choose right measures for the context
  - Subjective is common, but has limitations
- A multi-modal approach is key to most effectively and accurately portray team trust (Schaefer et al. 2019; Krausman et al. in press)

MULTI-MODAL MEASUREMENT

TIME

Surveys: sporadic measurement over time

Communication: near-continuous measurement over time

Task data: variable measurement over time (task dependent)

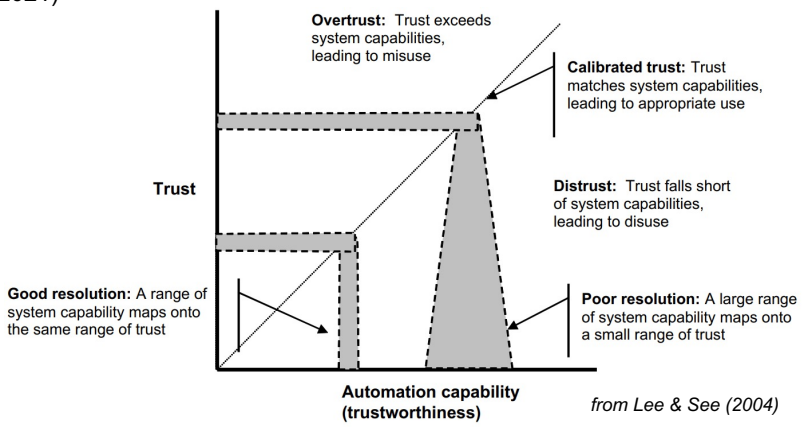
Sensor data: continuous measurement over time (eye tracking, HRV, facial expression)

MEASURING TRUST

- When trust is appropriately measured, we can then determine if it is too high/low for a context given the team's performance level (Muir, 1987)
- This is trust calibration

TRUST CALIBRATION

- Calibration is complex, due to several influential factors such as prior history, personality, and expertise (Wagner & Robinette, 2021)



from Lee & See (2004)

TRUST CALIBRATION

- Overtrust
  - Too much trust for a given degree of automation capability
  - Complacency, less awareness of activities and actions of teammates, potentially leading to costly errors if mistakes aren't swiftly perceived and corrected

TRUST CALIBRATION

- Undertrust
  - Excessive monitoring of teammates -> excessive workload, directing attention away from key tasks (de Visser et al., 2019; De Jong et al., 2016)
  - Team fails to take full advantage of team member skills and expertise

INTERVENTIONS

- If we find that trust is miscalibrated, we can look to developing interventions that promote trust calibration

Intervention process:

- Understand what variables to intervene on (trust is multifaceted)

INTERVENTIONS

- If we find that trust is miscalibrated, we can look to developing interventions that promote trust calibration

Intervention process:

- Understand what variables to intervene on (trust is multifaceted)
- Be capable of shaping those variables

17

17

Approved for public release/distribution unlimited

INTERVENTIONS

- If we find that trust is miscalibrated, we can look to developing interventions that promote trust calibration

Intervention process:

- Understand what variables to intervene on (trust is multifaceted)
- Be capable of shaping those variables
- Then, assuming appropriate measures & measurement were used…

18

18

INTERVENTIONS

- If we find that trust is miscalibrated, we can look to developing interventions that promote trust calibration

Intervention process:

- Understand what variables to intervene on (trust is multifaceted)
- Be capable of shaping those variables
- Then, assuming appropriate measures & measurement were used…
- You can move trust in the desired direction

INTERVENTIONS

- But HAT is highly complex, so evaluating an intervention means accounting for moderators, mediators, covariates, confounds, etc.

- To address these challenges, we argue that causal modeling approaches (e.g. Bayesian Networks or Structural Equation Modeling) are ideally suited for such scenarios.

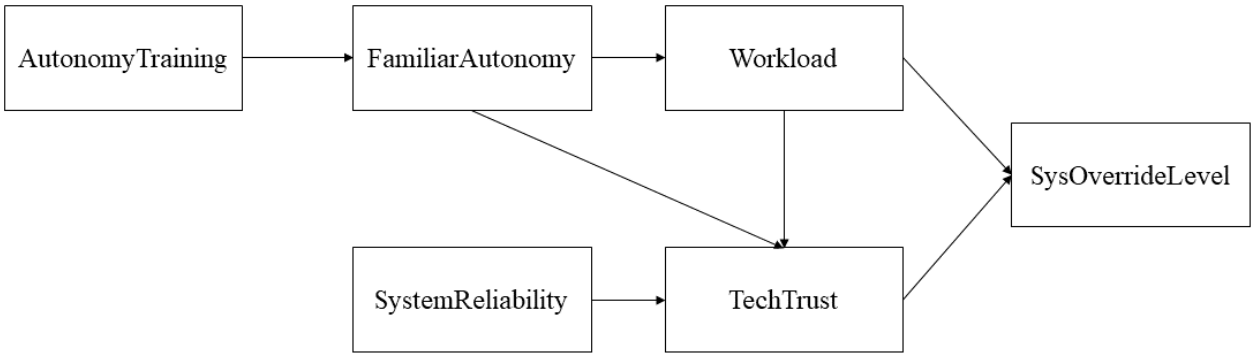
Model Scenario

EXAMPLE SCENARIO

- Imagine a human driver supported by an autonomous sensor package
  - LIDAR for obstacle detection, GPS for route-planning, decision system that sends optimized routes to human

EXAMPLE SCENARIO

- Conceptual causal model of relevant factors



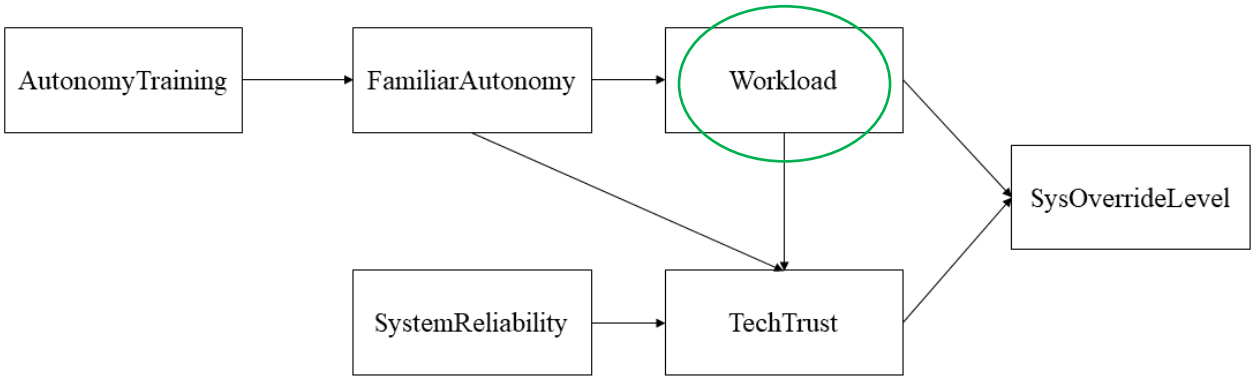
*not meant to be theoretically complete

23

Approved for public release/distribution unlimited

EXAMPLE SCENARIO

- Conceptual causal model of relevant factors

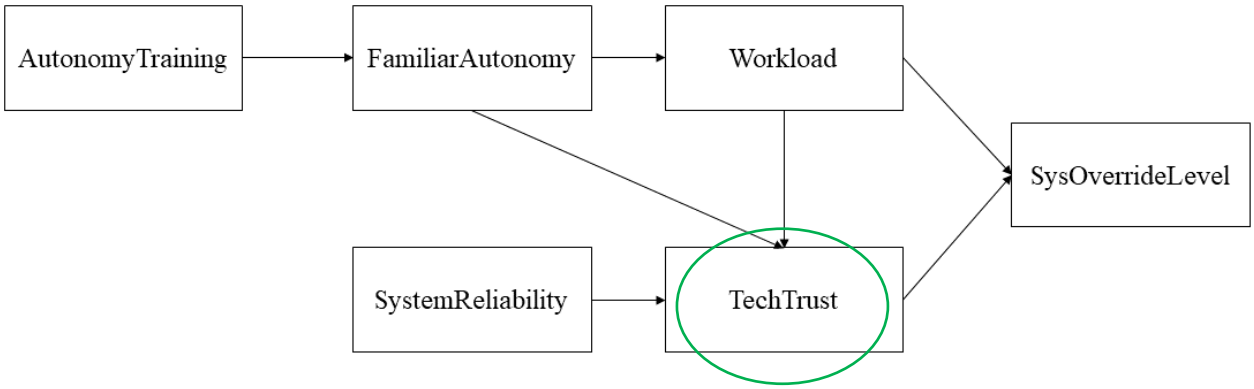


Approved for public release/distribution unlimited

24

EXAMPLE SCENARIO

- Conceptual causal model of relevant factors

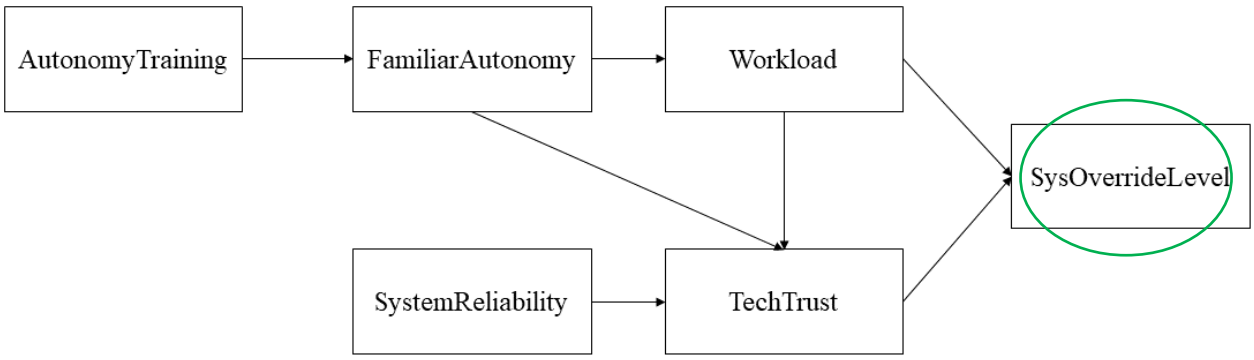


25

Approved for public release/distribution unlimited

EXAMPLE SCENARIO

- Conceptual causal model of relevant factors

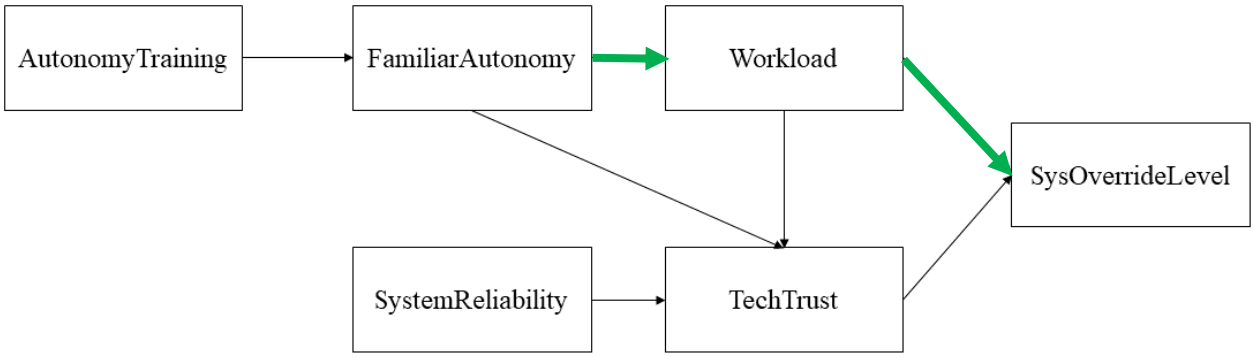


26

EXAMPLE SCENARIO

- Conceptual causal model of relevant factors

Also includes mediation effects



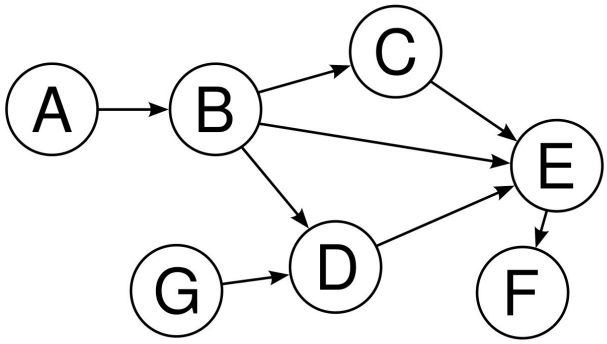
Overview: Causal Modeling

CAUSAL MODELING

- Explaining causal mechanisms that underlie observable phenomena
  - Several frameworks can be used for this; we will focus on Bayesian networks (BNs) and structural equation models (SEM)

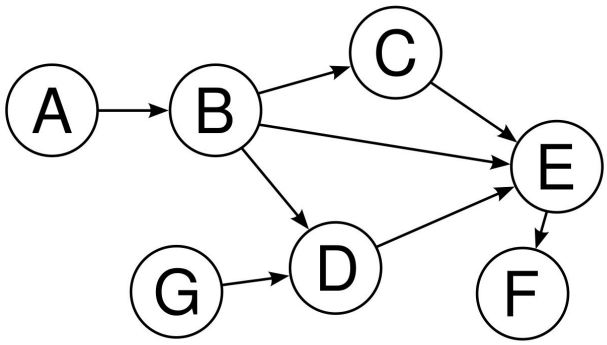
BAYESIAN NETWORKS

- Multivariate distribution of discrete variables, commonly depicted as a directed acyclic graph (DAG)
  - Variables represented as “nodes” in the graph, with “edges” between the variables



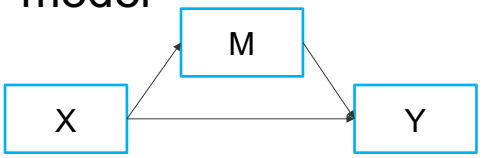
BAYESIAN NETWORKS

- Multivariate distribution of discrete variables, commonly depicted as a directed acyclic graph (DAG)
  - Variables represented as “nodes” in the graph, with “edges” between the variables
  - These edges are directed (i.e., single-headed arrows) and define the structure of the network
  - Expresses the dependence & conditional independence assumptions in the model for the joint distribution

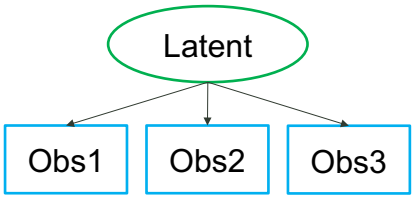


STRUCTURAL EQUATION MODELS

- SEM is a framework in which causal and correlational relationships among observed and latent variables can be specified and simultaneously evaluated
- Two parts: structural model and measurement model



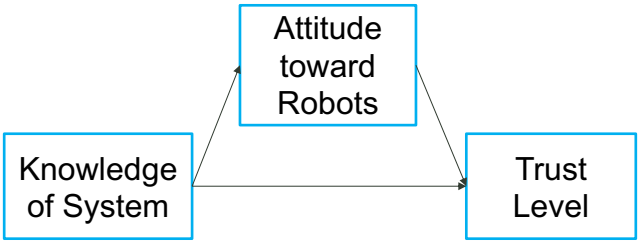
structural



measurement

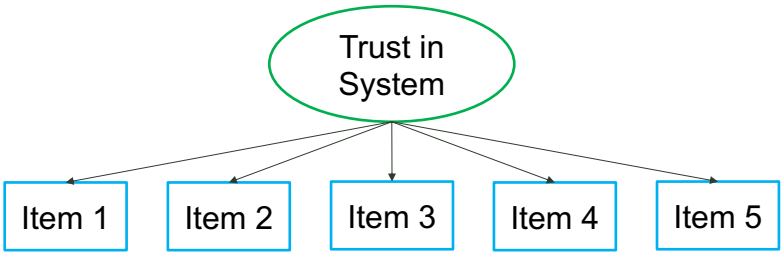
STRUCTURAL EQUATION MODELS

- Structural model
  - A set of equations that defines the causal relationships between the focal variables, which are often latent



STRUCTURAL EQUATION MODELS

- Measurement model
  - Another set of equations that defines or identifies these variables



CAUSAL MODELING

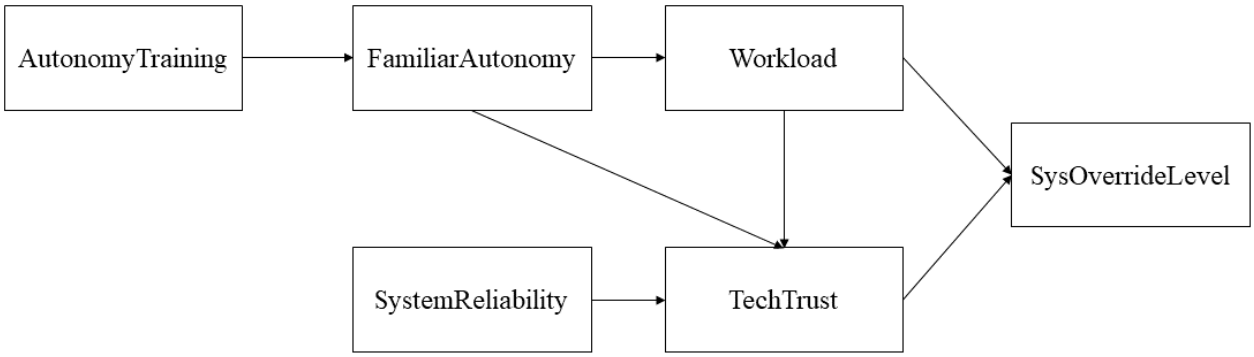
- In HAT, causal modeling can help formalize varied interactions between antecedents, mediators, moderators, confounds, and other key factors that influence trust
- This can offer stronger justifications for intervention designs aimed at HAT trust calibration

35

Causal Modeling in Context

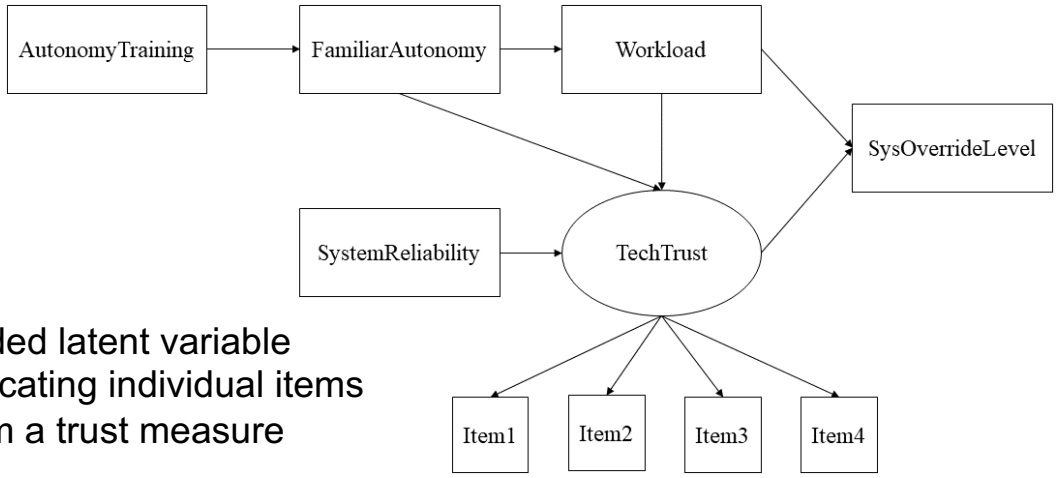
36

SEM VS. BN, IN CONTEXT



EXAMPLE PATH MODEL

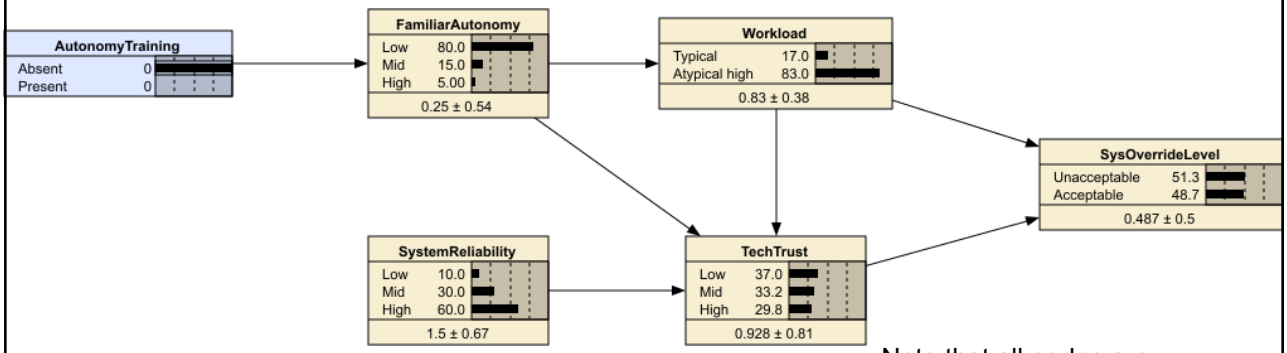
- Can incorporate additional information



Added latent variable indicating individual items from a trust measure

EXAMPLE BN

- Path model's hypotheses as a BN, using Netica software

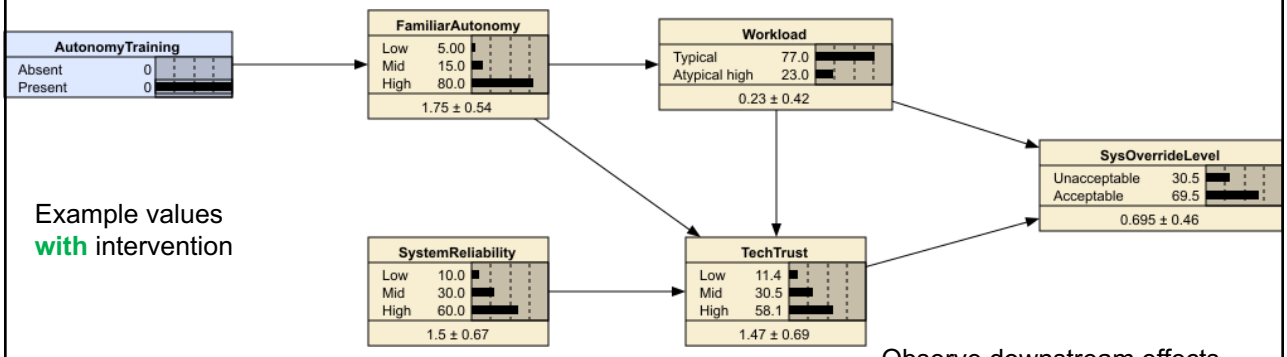


Note that all nodes are categorical, as required in BN

39

EXAMPLE BN

- Path model's hypotheses as a BN, using Netica software

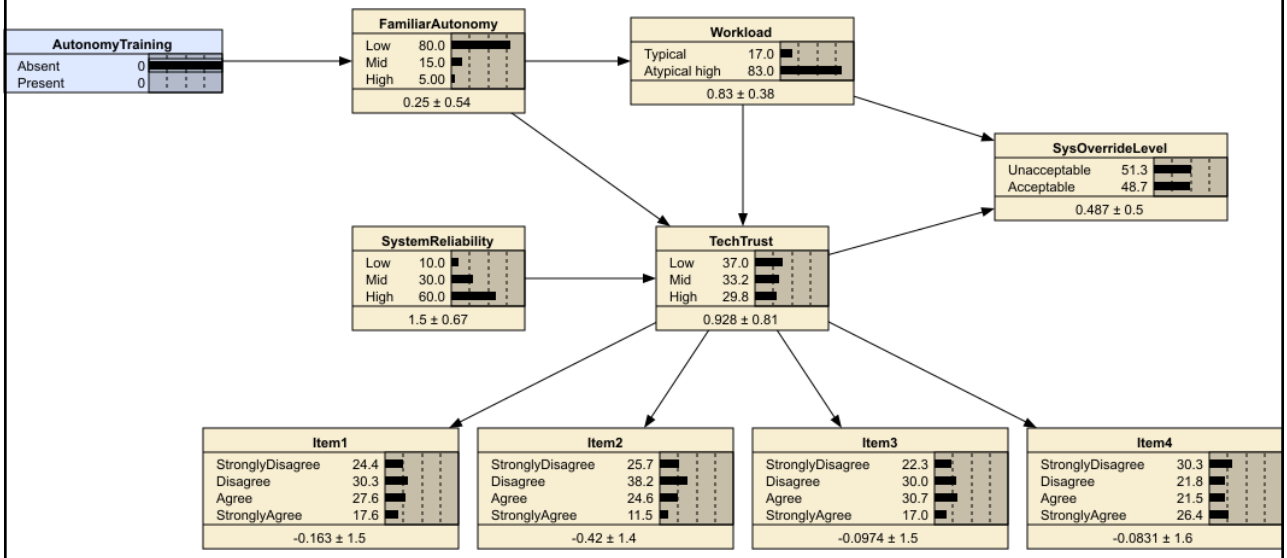


Example values **with** intervention

Observe downstream effects

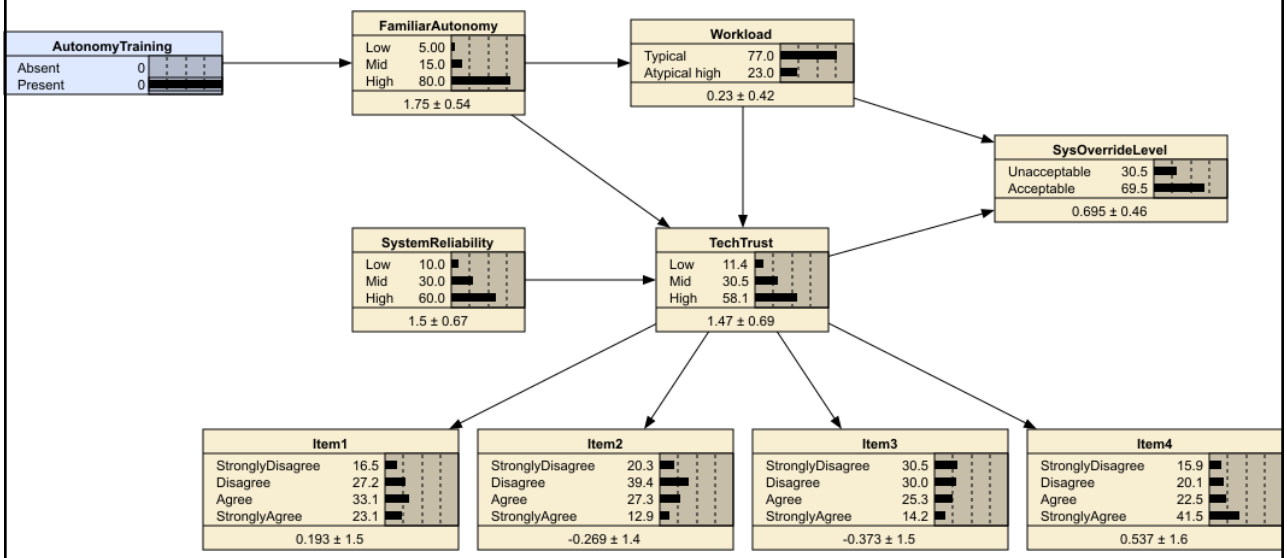
40

EXAMPLE BN WITH ITEMS



41

EXAMPLE BN WITH ITEMS



42

Conclusions & Next Steps

43

CONCLUSIONS

- We presented a causal analysis approach to develop a more comprehensive understanding of the effects of interventions on human-autonomy team trust
- Causal modeling approaches will help researchers and practitioners evaluate interventions and identify novel intervention targets suited for a given context, which can include changes in autonomy behavior, improving communication and transparency elements, providing after-action reviews, and so on

44

NEXT STEPS

- Continue to identify & validate measures of trust that can be used in a multi-method manner for HAT (Krausman et al. in press; Schaefer et al. 2019)
- Evaluate & build on causal modeling approach using live datasets
- Together, these directions will converge on more effective maintenance of trust calibration in complex, dynamic HAT of the near future

45

Thank you!

References

- de Jong, B.A., Dirks, K.T., and Gillespie, N. (2016). Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of Applied Psychology, 101*(8), 1134-1150
- de Visser, E., Peeters, M.M., Jung, M., and Kohn, S., Shaw, T.H., Pak, R., & Neerincx, M.A. (2019). Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics, 12*, 459-478.
- Krausman, A., Neubauer, C., Forster, D., Lakhmani, S., Baker, A., Fitzhugh, S., … & Schaefer, K. (in press). Team Trust Measurement in Human-Autonomy Teams. *ACM Transactions on Human-Robot Interaction.*
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.
- Mooradian, T., Renzl, B., and Matzler, K. (2006). Who Trusts? Personality, Trust and Knowledge Sharing. *Management Learning*, *37*(4), 523-540.
- Muir, B.M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*(5-6), 527- 539
- Schaefer, K. E., Baker, A. L., Brewer, R. W., Patton, D., Canady, J., & Metcalfe, J. S. (2019). Assessing multi-agent human-autonomy teams: US Army Robotic Wingman gunnery operations. In *Micro-and Nanotechnology Sensors, Systems, and Applications XI* (Vol. 10982, p. 109822B). International Society for Optics and Photonics.
- Wagner, A.R., & Robinette, P. (2021). *An explanation is not an excuse: Trust calibration in an age of transparent robots.* In J Lyons and C. Nam (eds.), Trust in Human-Robot Interaction: Research and Applications. Elsevier.

46