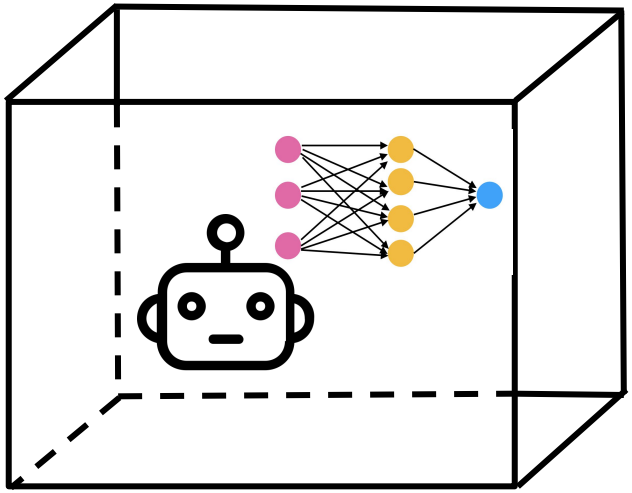
A Situation Awareness-Based Framework for Design and Evaluation of Explainable AI

Lindsay Sanneman and Julie Shah

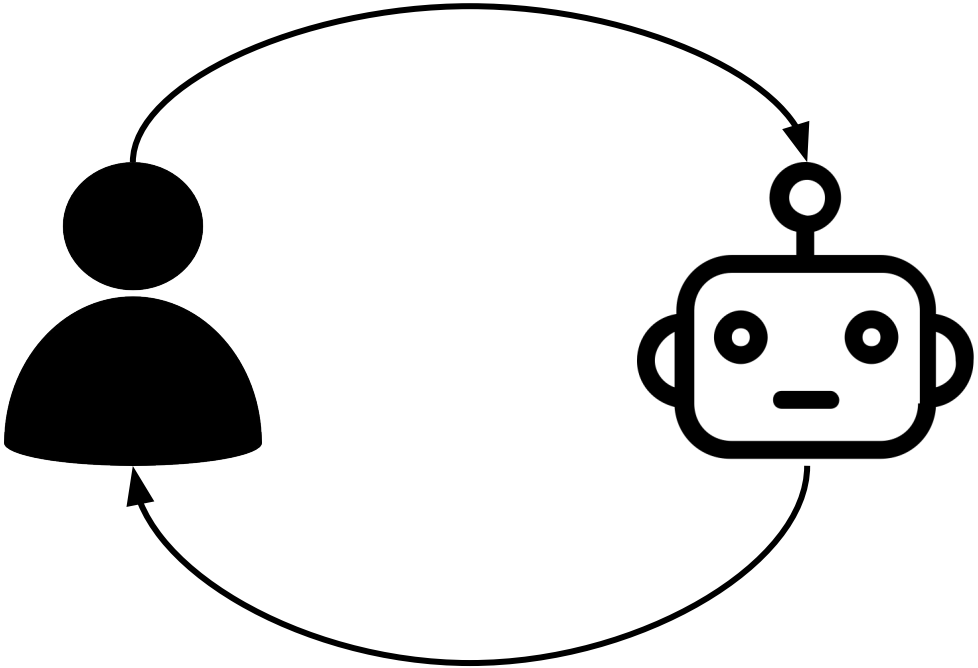
EXplainable, TRansparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS) Workshop
AAMAS 2020

Massachusetts Institute of Technology
Department of Aeronautics and Astronautics
Computer Science and Artificial Intelligence Lab (CSAIL)

Explainable AI (XAI)



Human Factors and Team Performance



- Situation Awareness
- Trust
- Mental Workload

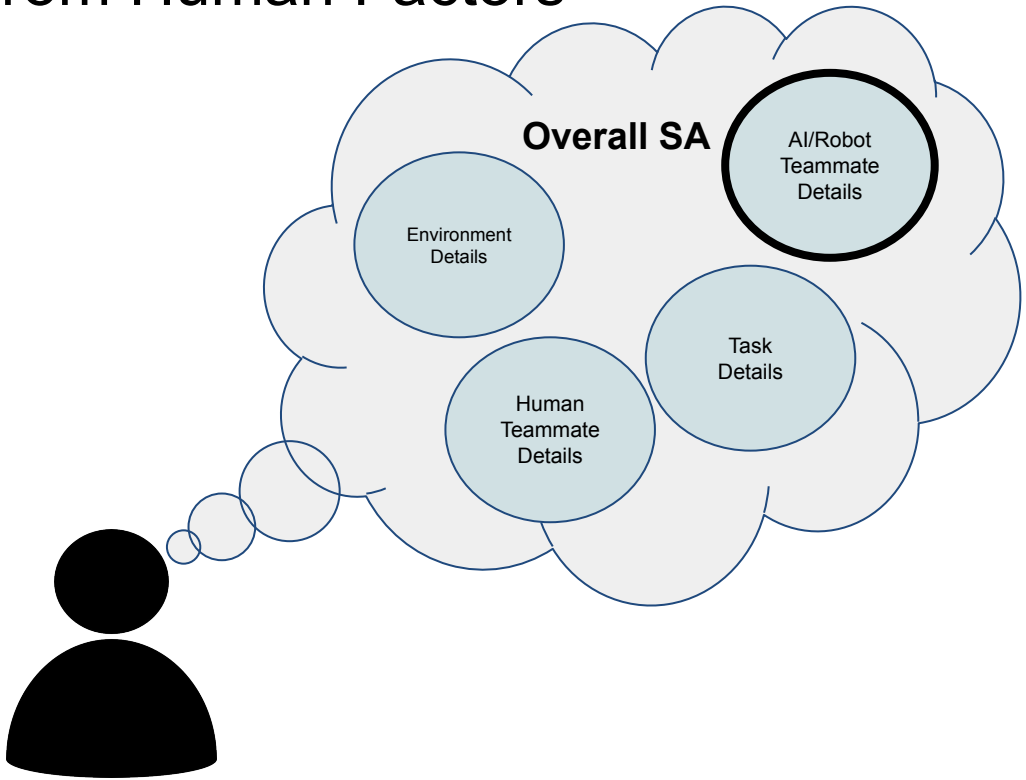
Situation Awareness (SA) from Human Factors

Endsley 1995 Definition[1]:

Level 1
The ***perception*** of elements in the environment within a volume of time and space

Level 2
The ***comprehension*** of their meaning

Level 3
The ***projection*** of their status in the near future



[1] Endsley, Mica R. "Toward a theory of situation awareness in dynamic systems." *Human factors* 37.1 (1995): 32-64.

Contributions

Situation Awareness-Based Framework

Situation Awareness-Based Evaluation Method

Proposed Future Directions for XAI Research

Situation Awareness-Based Levels of XAI Framework

Level 1: XAI for Perception

Level 1 XAI:
XAI for Perception
Explanations of what an AI system did or is doing and the decisions made by the system

Answers "What" Questions[2]

Example Types of Information:
- Input Information
- Output Information

Example Approaches

- Belief-based explanations from explainable Belief-Desire-Intent (BDI) agents
  Broekens et al. 2010, Harbers et al. 2010, Harbers et al. 2011
- Plan information from a planning agent
  Borgo et al. 2018, Chakraborti et al. 2019, Sreedharan et al. 2018
- Prototypes of clusters
  Kim et al. 2014

[2] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence* 267 (2019): 1-38.

Level 2: XAI for Comprehension

Level 2 XAI:
XAI for Comprehension
Explanations of why an AI system acted in a certain way or made a particular decision and what this means in terms of the system's goals

Answers "Why"/"How" Questions[2]

Example Types of Information:
- Model Information

Example Approaches

- Feature or state importance techniques
  Ribeiro et al. 2016, Hayes and Shah 2017, Kim et al. 2017
- Saliency maps
  Adebayo et al. 2018
- Desire-based explanations from explainable Belief-Desire-Intent (BDI) agents
  Broekens et al. 2010, Harbers et al. 2010, Harbers et al. 2011
- Explanations based on objective functions, rewards, or goals
  Dannenhauer et al. 2018, Borgo et al. 2018

[2] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence* 267 (2019): 1-38.

Level 3: XAI for Projection

Level 3 XAI:
XAI for Projection
Explanations of what an AI system will do next, what it would do in a similar scenario, or what would be required for an alternate outcome

Answers "What If" Questions[2]

Example Types of Information:

- Changed Inputs → Outputs
  Forward Reasoning
- Outputs → Required Inputs
  Backward Reasoning
- Effects of Model Changes
- Nominal Next Actions

Example Approaches

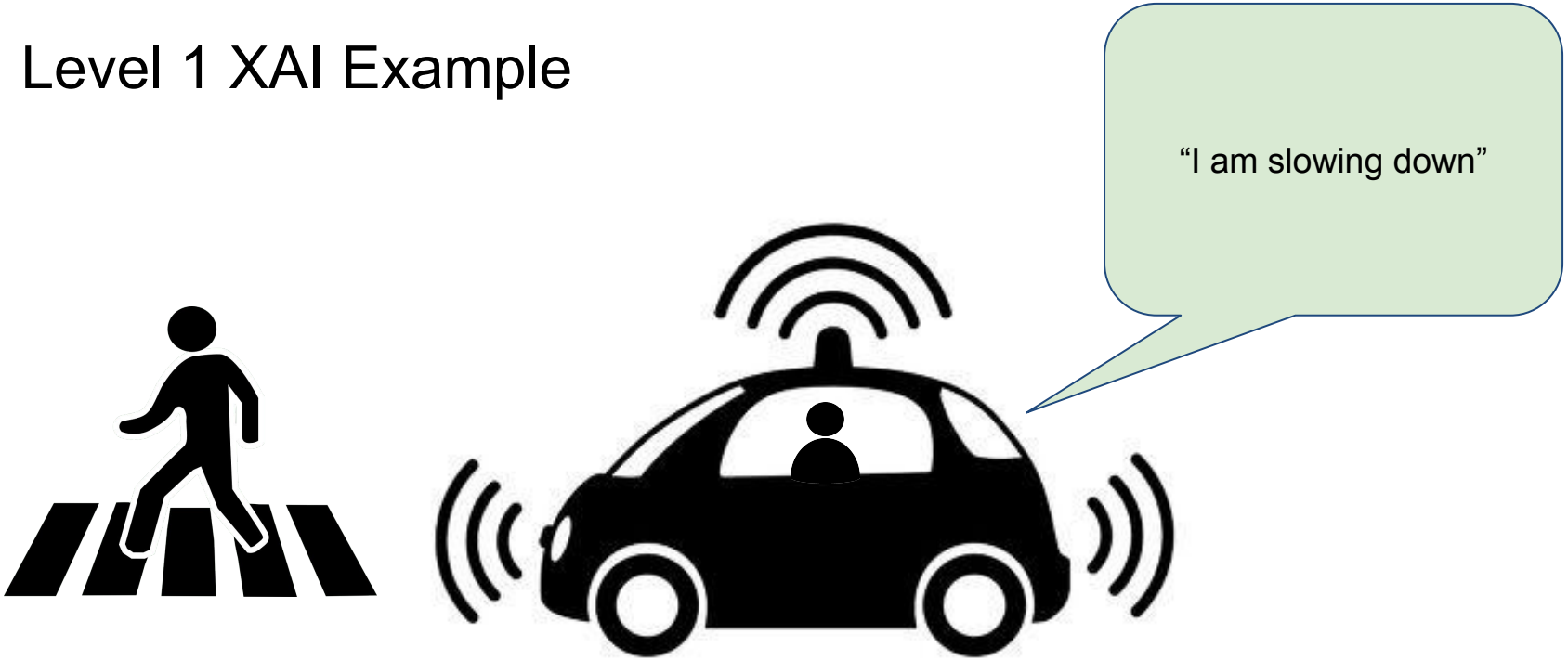
- Failure prediction and misclassification amendment
  Bansal et al. 2014, Marino et al. 2018
- Reasoning about states for forward/backward simulation
  Hayes and Shah 2017
- "Important" traces of agent behavior (state-action pairs)
  Amir and Amir 2018
- Intent-based explanations from explainable Belief-Desire-Intent (BDI) agents
  Broekens et al. 2010, Harbers et al. 2010

[2] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence* 267 (2019): 1-38.

Example Domain: Autonomous Vehicle

- Explainable AI system for vehicle path planner
- Human passenger takes control in off-nominal scenarios

Level 1 XAI Example

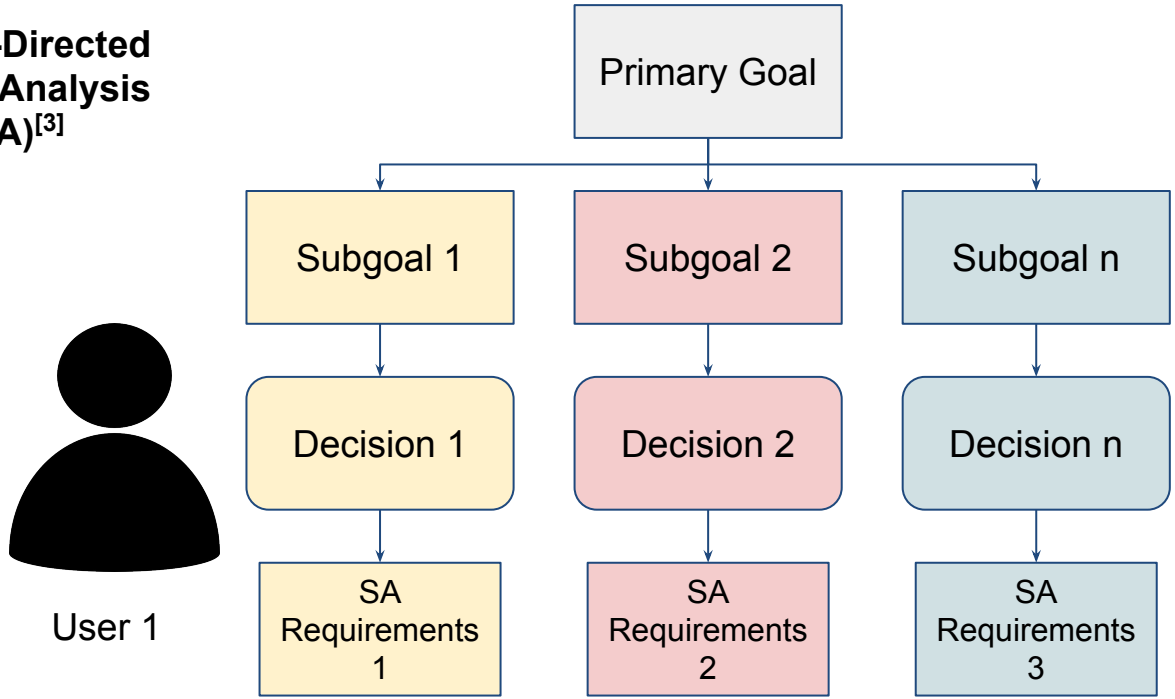


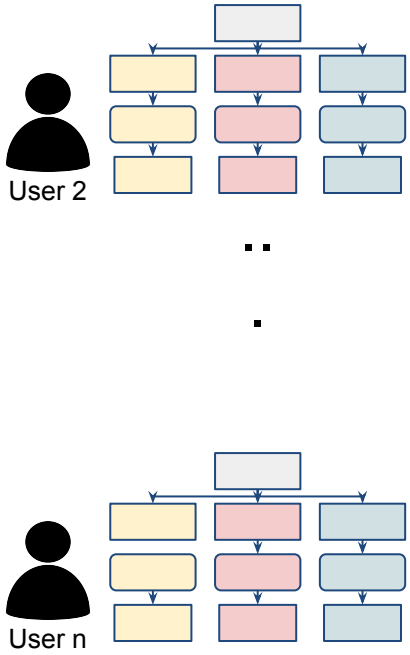
Level 2 XAI Example

Level 3 XAI Example

Determining Human Informational Needs

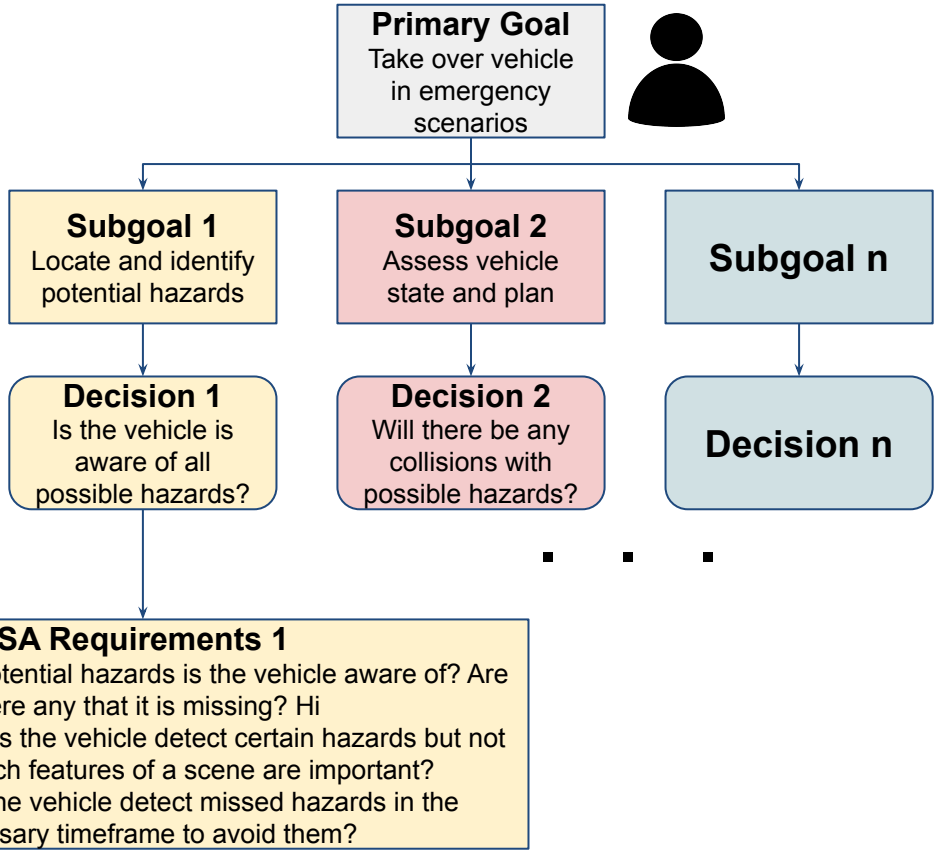
Goal-Directed Task Analysis (GDTA)[3]





[3] Endsley, Mica R. "Direct measurement of situation awareness: Validity and use of SAGAT." *Situational Awareness*. Routledge, 2017. 129-156.

GDTA Example



Evaluating Explanation Quality: A Method for Situation Awareness-Based XAI Assessment

Situation Awareness Global Assessment Technique (SAGAT)[4]

General	For XAI
Enumerate SA requirements using a process such as GDTA.	Enumerate SA requirements ***related to AI behavior*** using a process such as GDTA.
Simulate the scenario of interest, including all relevant human tasks and decisions.	Simulate the scenario of interest, including all relevant human ***and XAI*** tasks and decisions.
Freeze the simulation at randomly selected time, and ask the human questions related to their identified informational needs.	Freeze the simulation at randomly selected time, and ask the human questions related to their identified informational needs ***related to AI behavior***.

SAGAT Question Examples:

- **Level 1** - How many potential hazards is the vehicle missing?
- **Level 2** - Which features are causing the vehicle to miss those hazards?
- **Level 3** - Will the vehicle recognize the hazards before a collision occurs?

[4] Endsley, Mica R. "Situation awareness global assessment technique (SAGAT)." *Proceedings of the IEEE 1988 national aerospace and electronics conference*. IEEE, 1988.

Future Directions

XAI system that addresses all three levels of XAI

Assessment of XAI techniques using the GDTA process and SAGAT test

Enhanced methods for providing user-tailored explanations

Conclusion

Situation Awareness-Based Framework for XAI

Level 1 XAI: XAI for Perception	Level 2 XAI: XAI for Comprehension	Level 3 XAI: XAI for Projection

Situation Awareness-Based Evaluation Method

- Goal Directed Task Analysis (GDTA) for definition of informational needs related to XAI
- Situation Awareness Global Assessment Technique (SAGAT) for evaluation of XAI systems

Proposed Future Directions for XAI