Trustworthy Autonomous Systems Through Social Explainable AI

Balint Gyevnar

Supervised by: Chris Lucas, Shay Cohen, Stefano Albrecht

In Doctoral Consortium of the 26th European Conference on Artificial Intelligence (ECAI 2023)

THE UNIVERSITY *of* EDINBURGH informatics

Autonomous Agents Research Group

NLP UKRI CENTRE FOR DOCTORAL TRAINING

UK Research and Innovation

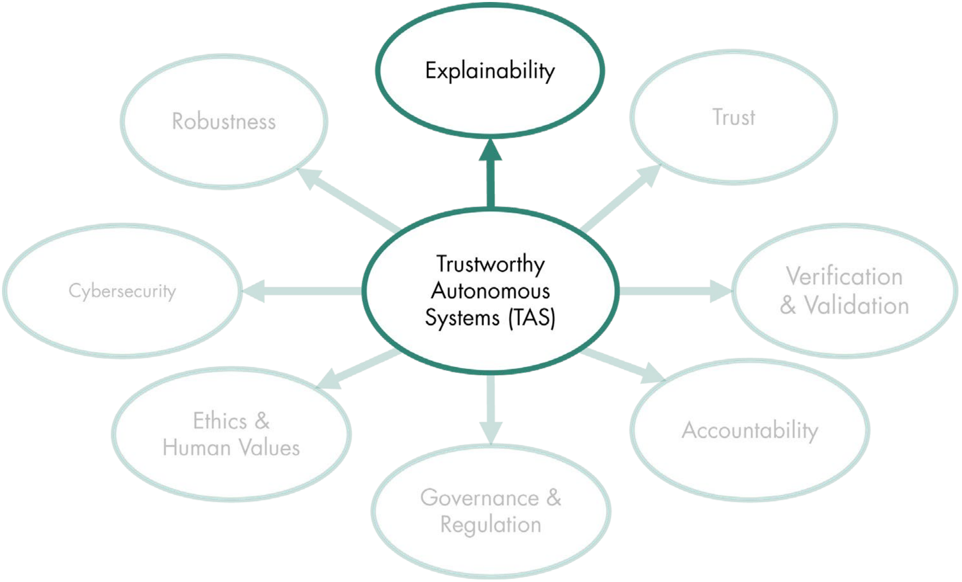
Trustworthy Autonomous Systems:

- Multi-faceted and cross-disciplinary

- How to contest decisions?
- Give informed consent?
- Ask for explanations?

- Explainability → Restore agency

- **Social explanations**

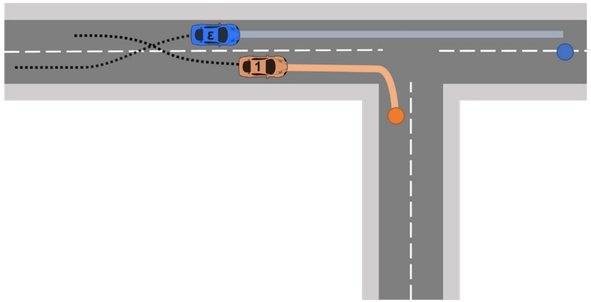


Social Explainable AI:

- In terms of a causal chain of events
- Involving contrast cases
- Addressing human cognitive biases
- Building gradual knowledge through conversations

CEMA:

- Causal Explanation in Multi-Agent systems
- Autonomous driving



User **Agent**

Why did you change lanes?

It decreases the time to reach the goal.

Why does it decrease the time to the goal?

Because vehicle 1 was slower than us.

Why was it slower?

It was decelerating and turning right.

What if it hadn't changed lanes before?

We would have gone straight.

Causal Explanations for Sequential Decision-Making in Multi-Agent Systems. **Balint Gyevnar**, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, Stefano V. Albrecht; IJCAI 2023 Workshop on Explainable Artificial Intelligence, 2023. - https://arxiv.org/abs/2302.10809

Evaluation of CEMA:

- Causal explanations in complex scenarios with many queries.
- Works for large number of agents:
  - Tested with up to 20 agents;
- Two-stage human evaluation for goodness of explanations:
  - Comparison against human-written baseline from participants;

What is left?

- Evaluation in more domains:
  - E.g., grid-worlds, Pacman, SMAC.
- Integration with state-of-the art NLP:
  - Support fluent conversation from query to response.
- Cognitive state tracking:
  - Theory of mind modelling for better targeted explanations

More on my website gbalint.me