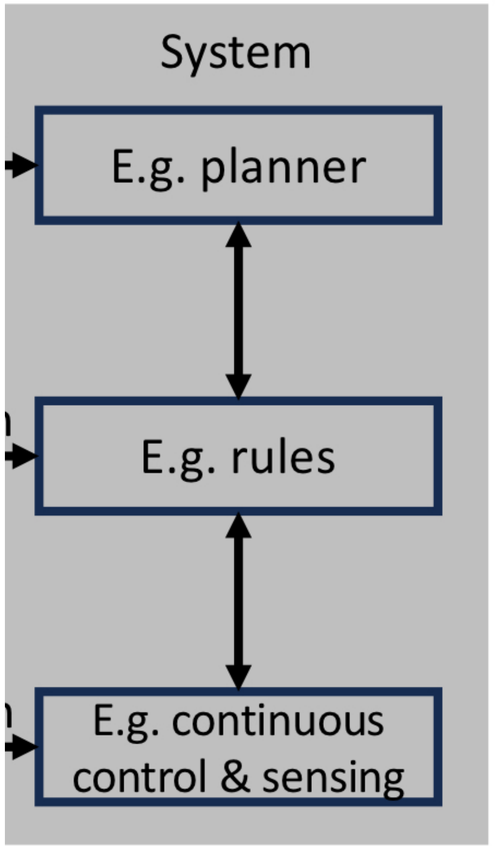
Towards Engineering Explainable Autonomous Systems

Michael Winikoff

EMAS 2024, Auckland, NZ

Motivation …

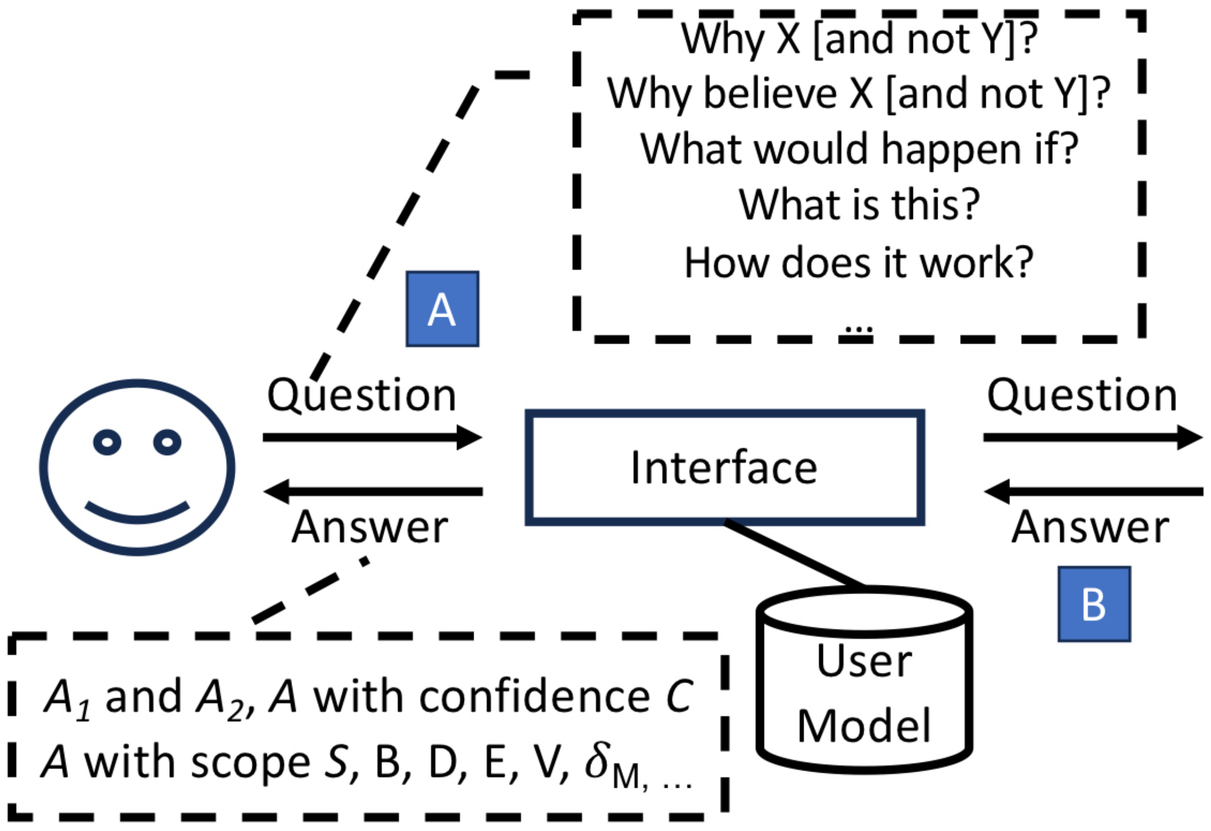
- Explainability important (e.g. calibrating trust, enhancing transparency & understanding)
- Hence lots of work on XAI
- BUT work is typically on techniques for explaining **individual components** (typically machine learning)
- This is **necessary**, but **insufficient**: systems often have multiple components, and need to explain such systems!

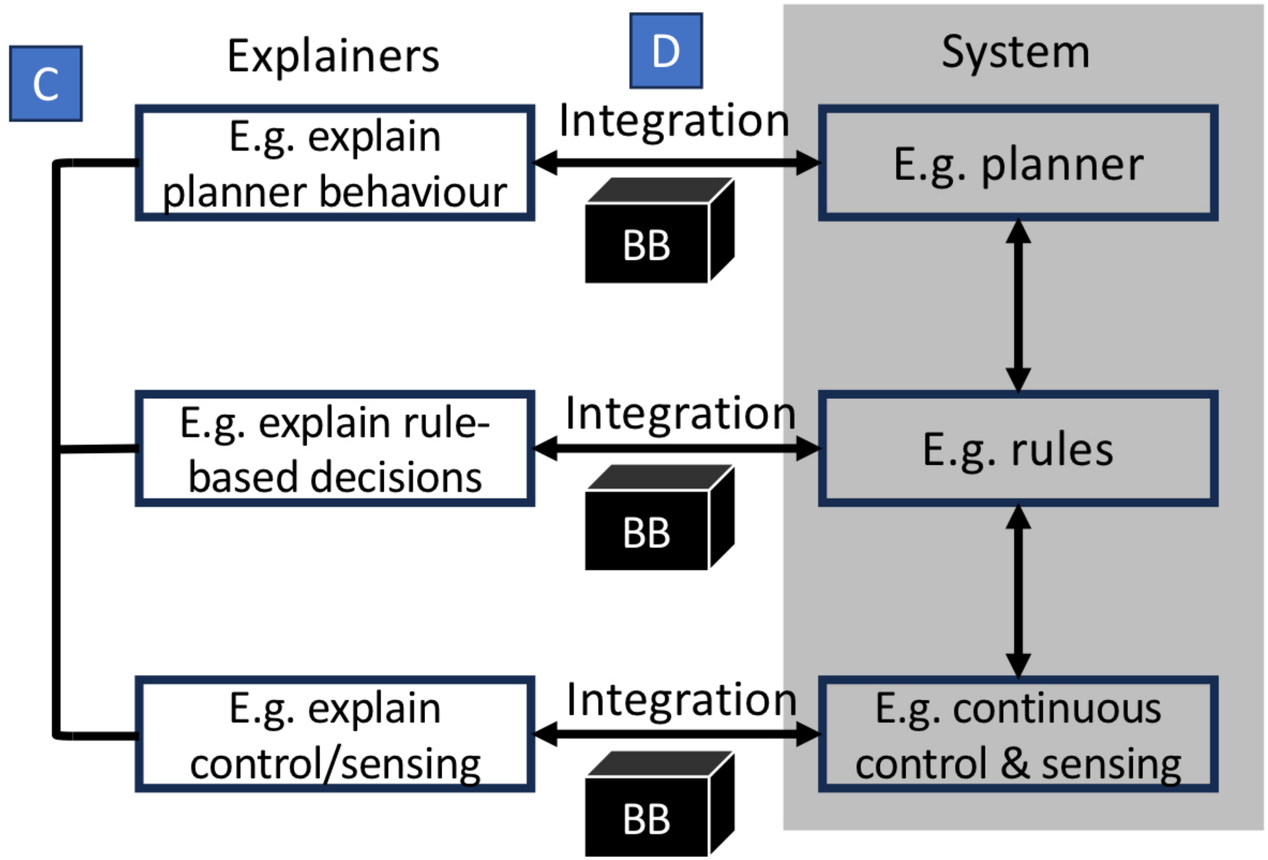


This paper

- Calls for extending XAI to multi-component systems
- Proposes an **architecture** for explainable multi-component systems, including considering:
  - the forms of **questions** and of **answers;**
  - number of **design decisions**.
- Raises a number of **integration-related issues**
- Poses **research questions**

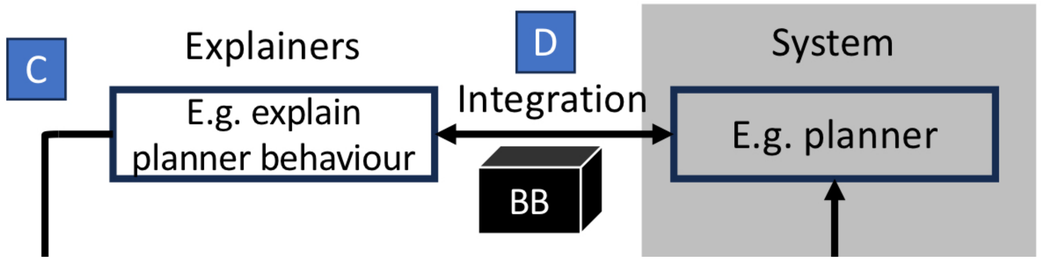
Architecture



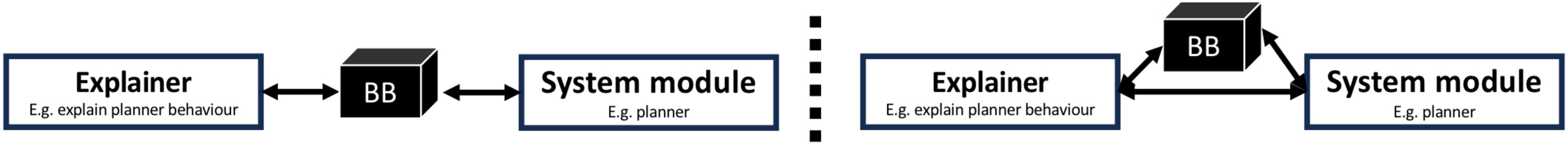


Integration Issues

- How to determine target explainer for a given question?
  1. Static indexing
  2. Tag actions
  3. Send to all



- Indirect or direct black box integration?



Research Challenges (broad)

- Broader context of use: e.g. development process, situations in which system tends to fail
- Endeavour of research: test beds, benchmarks

Research Challenges (specific)

1. How manage tagging of actions with the responsible component? Additional: for “Why X and not Y” need to identify how Y might have occurred
2. How do explainer agents interact with black boxes?
3. What is captured in the black boxes? How?
4. How ensure that explanations can be verified to be authentic and honest?
5. Would the interface need to share user model information with the explainers?

Also more XAI-oriented questions (e.g. specifying properties of the desired answer, expressing confidence/scope, more question types)

Summary

- Calls for extending XAI to multi-component systems
- Proposes an **architecture** for explainable multi-component systems (including questions, answers, design decisions)
- Poses **research questions** (including integration issues)