Computational Responsibility for Trustworthy Citizen-Centric AI

Vahid Yazdanpanah, Jennifer Williams, Sebastian Stein

Citizen-Centric Artificial Intelligence Systems

Background: To develop and effectively deploy trustworthy citizen-centric AI, we need computational tools to reason about and determine if, and to what extent, AI agents, human users, or human-AI collectives are

1. able to deliver an outcome/task and take the role to do so (**prospective responsibility**)
2. to be seen accountable, blameworthy, and liable for potential failures (**retrospective responsibility**)

Dimensions of Responsibility:

- **Normativity:** responsibility for a given desirable/undesirable outcome
- **Strategic aspects:** verifying the ability of agents to influence the occurrence of outcomes
- **Epistemic dynamics:** what information is available to agents
- **Temporality:** one may be responsible now but not a week ago

Methods:

- Game Theory
- Temporal Logics and Game Structures
- Formal Verification

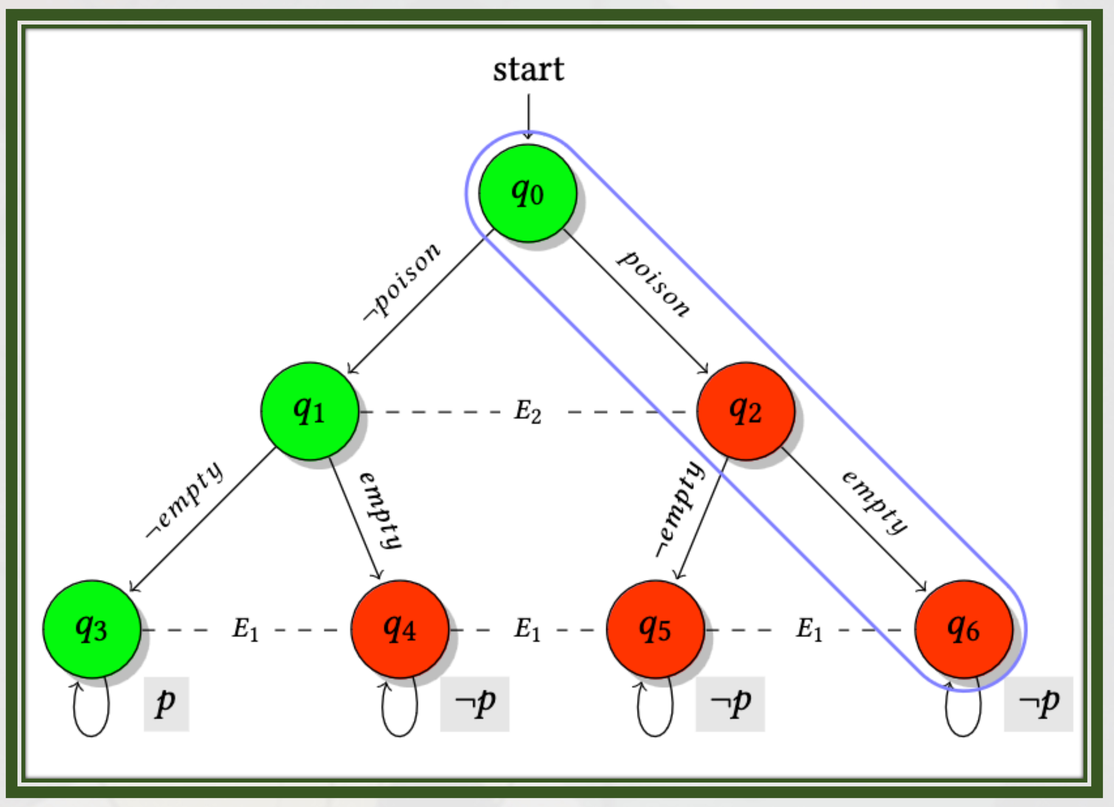
Concurrent Epistemic Game Structures (CEGS):

A CEGS is a tuple M = $\langle \Sigma, Q, Act, \sim 1, \ldots, \sim n, d, o \rangle$ where:

- $\Sigma = \{a1, \ldots, an\}$ is the set of agents;
- Q is the set of states;
- Act is the set of actions;
- $\sim a \subseteq Q \times Q$ is an epistemic (equivalence) indistinguishability relation, $q \sim a\ q'$ indicates that states q and q′ are indistinguishable to a;
- $d : \Sigma \times Q\ 7 \rightarrow P(Act)$ specifies the sets of actions available to agents at each state, same actions are available in indistinguishable states;
- o is a deterministic transition function.

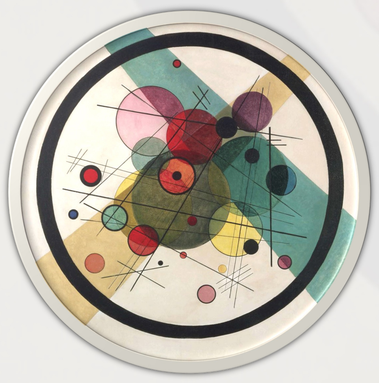
Results:

- Computational techniques for **quantifying responsibility** and sharing blames
- Formalisations for **determining and distinguishing responsibility**, blame, accountability, and liability in human-AI systems



A Passenger and Two Enemies (McLaughlin, 1925)

In q0, E1 may poison the water. In q1 and q2, E2 may empty the canteen. As a result, P is alive in q3 (represented by proposition p) and dead in q4, q5, and q6 (represented by ¬p). The path outlined in blue denotes the history.



With more **autonomy** comes more, and different forms of **responsibility**

Limitations: Our formal responsibility models requires an **exhaustive specification** of the context.
Future work: Hybrid models that use data-driven methods **to learn** about the context and formal methods **to reason** about responsibilities.

- Stein and Yazdanpanah. Citizen-centric multiagent systems (2023)
- Williams, Yazdanpanah and Stein. Safe audio AI services in smart buildings (2022)
- Yazdanpanah, Gerding, Stein, et al. Different forms of responsibility in multiagent systems: Sociotechnical characteristics and requirements (2021)
- Yazdanpanah, Gerding, Stein, et al. Reasoning about responsibility in autonomous systems: challenges and opportunities (2022)

UK Research and Innovation
University of Southampton

Citizen-Centric AI Systems (CCAIS)
www.ccais.ac.uk