Causality and Responsibility in Multi-agent Systems

Mehdi Dastani
Utrecht University

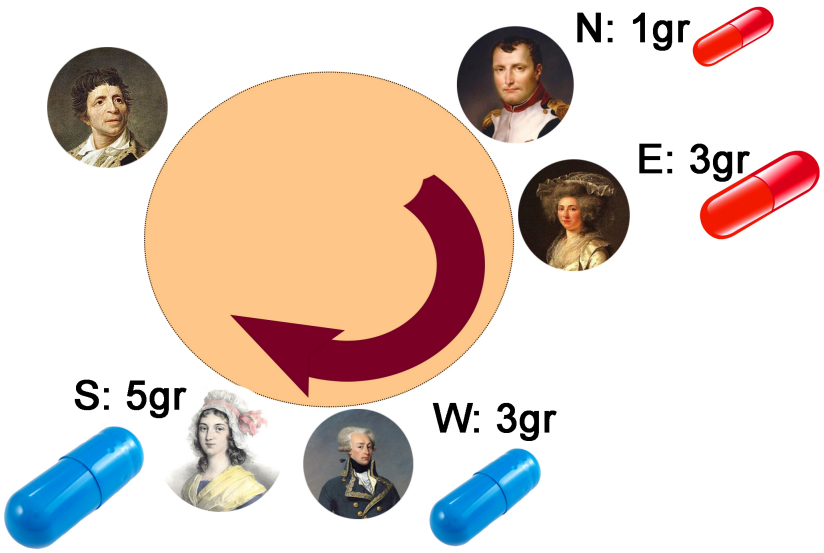
March 26, 2025

Responsibility in Artificial Intelligence

Safety of AI systems is a well-recognised and important concern. If an unsafe outcome occurs, it is important to determine which actions by which agents have caused it, and whether it could have been prevented.

Reasoning about Responsibility

The Traveller and Two Enemies: (Proximate Cause, McLaughlin, 1925) A traveller T needs water to survive a trip across the desert. T has two enemies E_1 and E_2. The night before T's departure, E_1 adds poison to the water in T's canteen. Later, but before T departs, E_2 empties the (poisoned) water from the canteen. T dies of thirst in the desert.



Blue and red cancel each other.
Four effective grams of each kills.

Flash crash in financial markets.

Views on Responsibility

- **Harry Frankfort, 1969**: An individual is responsible for what he has done only if she could have done otherwise.

- **Braham & van Hees, 2012**: An agent is responsible for an outcome iff
  1. the agent is autonomous, intentional, and capable of distinguishing right and wrong, and good and bad,
  2. there exists a causal relation between the action of the agent and the outcome in question, and
  3. the agent has had a reasonable opportunity to have done otherwise (the principle of *avoidance potential* or *alternative possibilities*).

- **Halpern, 2016**: A specific set of events causes an outcome iff
  1. the set of events and the outcome have taken place,
  2. the outcome depends counterfactually on the set of events, and
  3. the set of events is minimal in the sense that no subset of those events could be the cause.

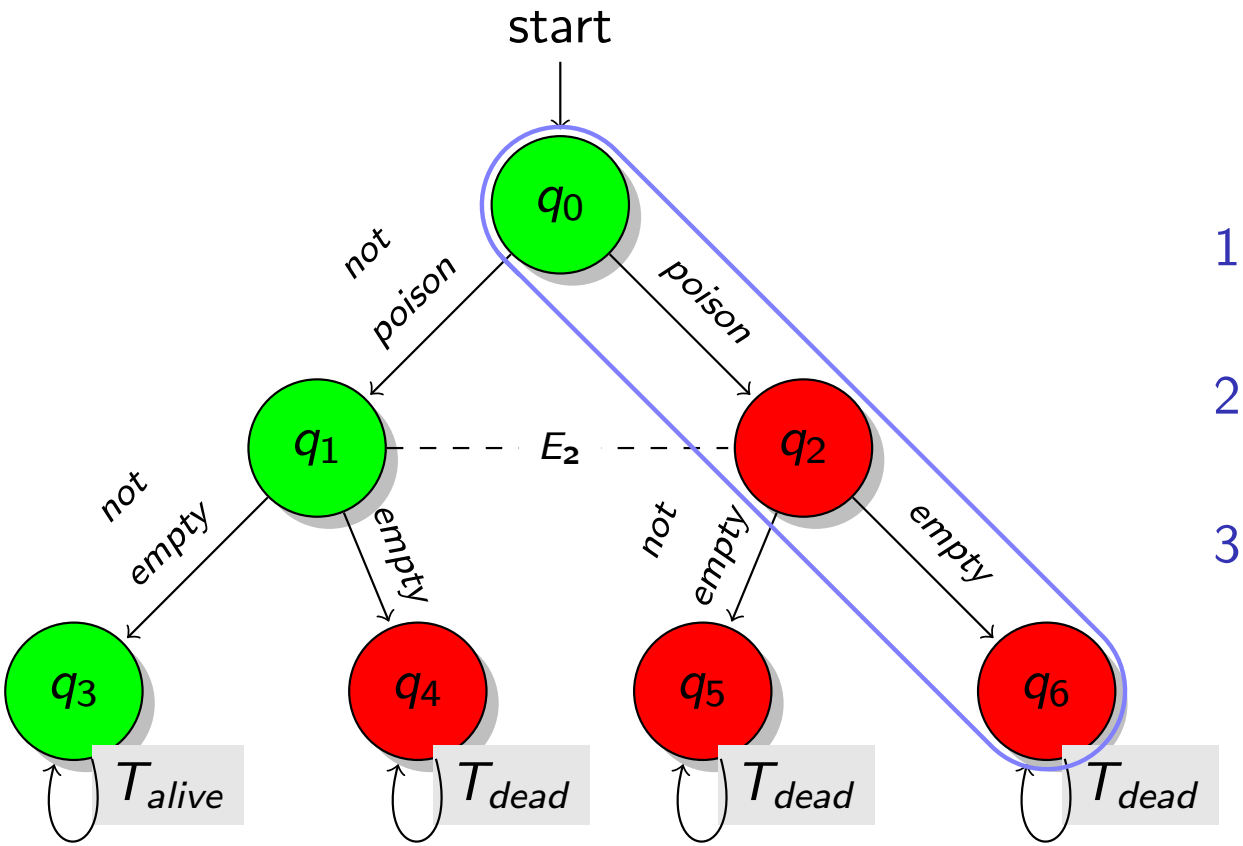
Responsibility in Multiagent Setting

A group A is responsible for a state of affairs $\mathcal{S}$ iff $\mathcal{S}$ occurs and A was able (had a strategy) to preclude it given their knowledge, i.e., iff the following three conditions hold:

1. $\mathcal{S}$-relevant history: a sequence of states h ending in $\mathcal{S}$;
2. A's ability to preclude $\mathcal{S}$: A had the potential to avoid $\mathcal{S}$ in some state on h;
3. Minimality of A: there is no subgroup $B \subset A$ that was able to preclude $\mathcal{S}$ in any state on h.

Note: Ability to preclude captures both the **strategic ability** of a group and their **epistemic uncertainty**.

Responsibility in Multiagent Setting



1. $\mathcal{S}$-relevant history
2. A's ability to preclude $\mathcal{S}$
3. Minimality of A

Given the history mentioned above, enclosed by blue, we have that **neither E_1 nor E_2 are individually responsible**, but they are **collectively responsible** for the dead of the traveller.

Responsibility in $ATL_{i,r}$

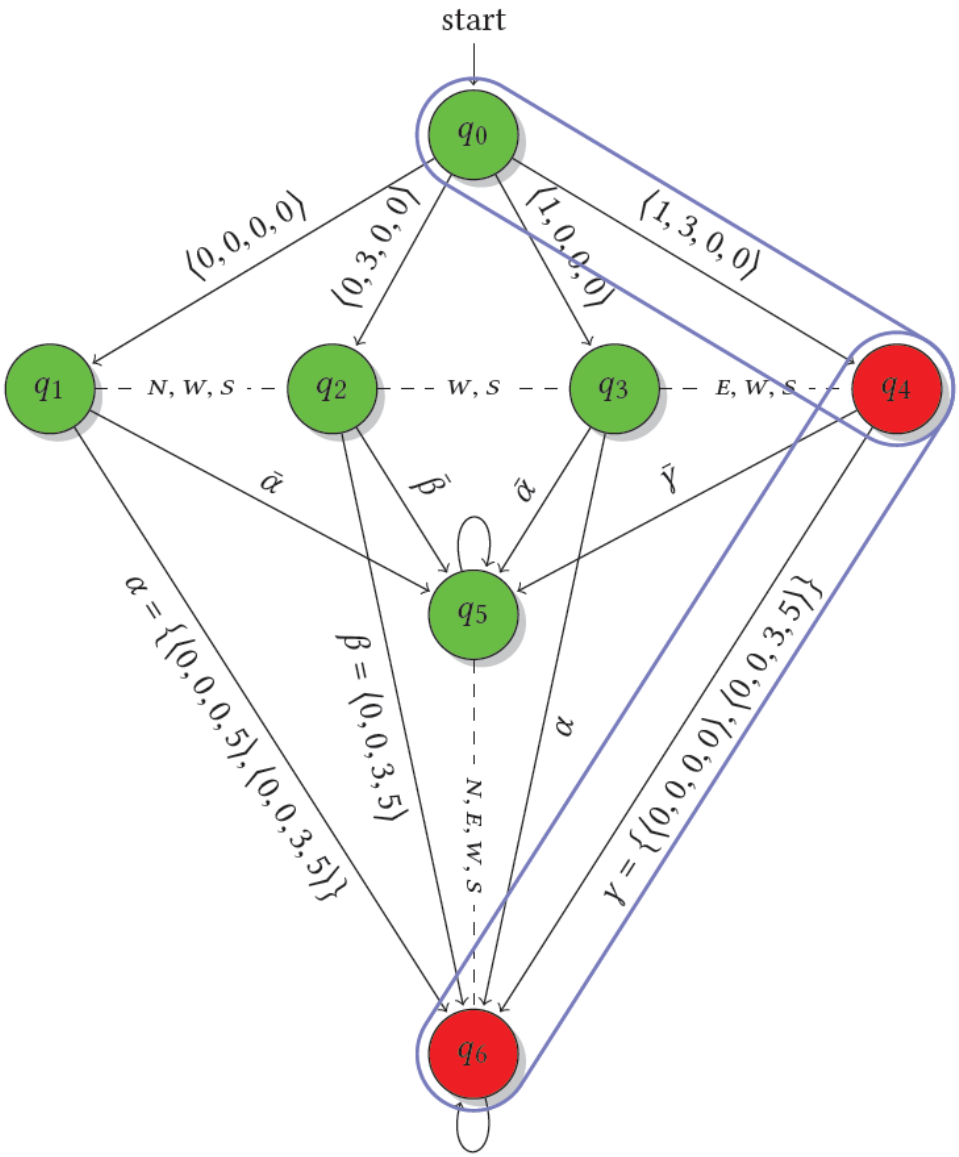
Γ is responsible for φ given the q-history $\lambda[q_i, k]$ iff:

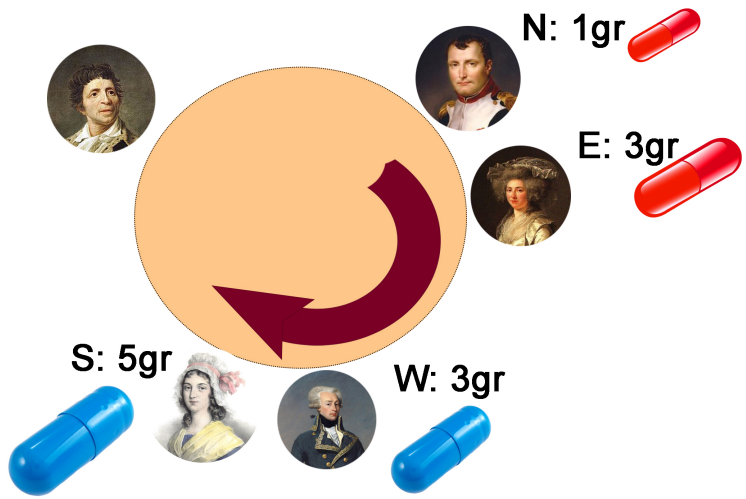
1. $\mathcal{M}, q \models \varphi$ (*relevance of the history*) and
2. $\exists q' \in \lambda[q_i, k]$, s.t. $\mathcal{M}, q' \models \langle\langle \Gamma \rangle\rangle \Box \neg \varphi$, and
3. $\forall \Gamma' \subset \Gamma$, $\mathcal{M}, q' \not\models \langle\langle \Gamma' \rangle\rangle \Box \neg \varphi$.

Theorem (Complexity)

The problem of checking whether a group Γ is responsible for φ given $\lambda[q_i, k]$ is Δ_2^P-complete (w.r.t. the size of $\mathcal{M}$ and φ, and the length k).

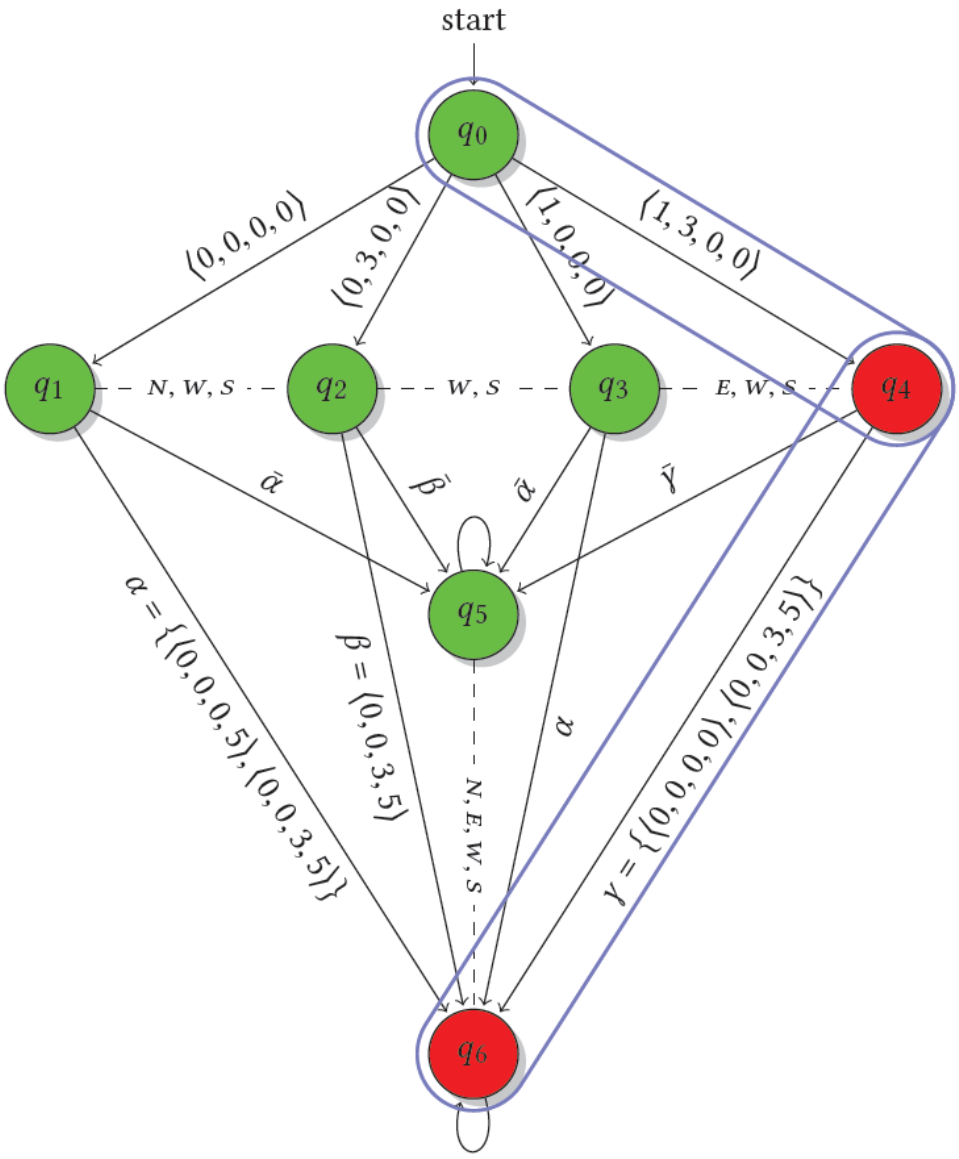
Responsibility in Multiagent Setting

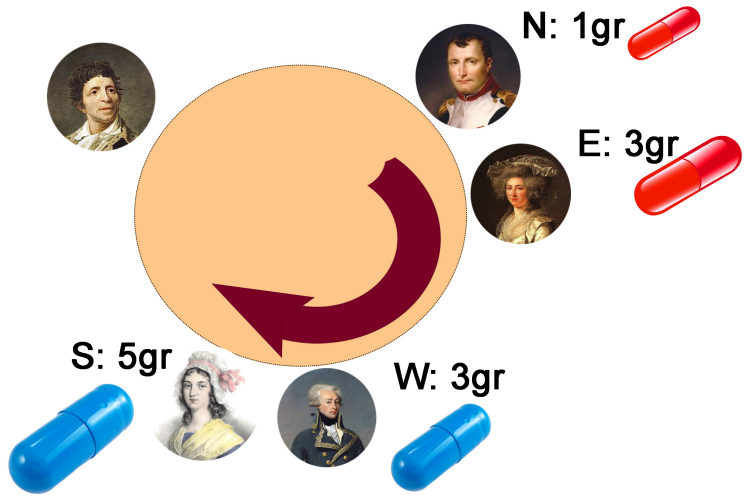




$\langle$ **N** , **E** , **W** , **S** $\rangle$

Responsibility in Multiagent Setting

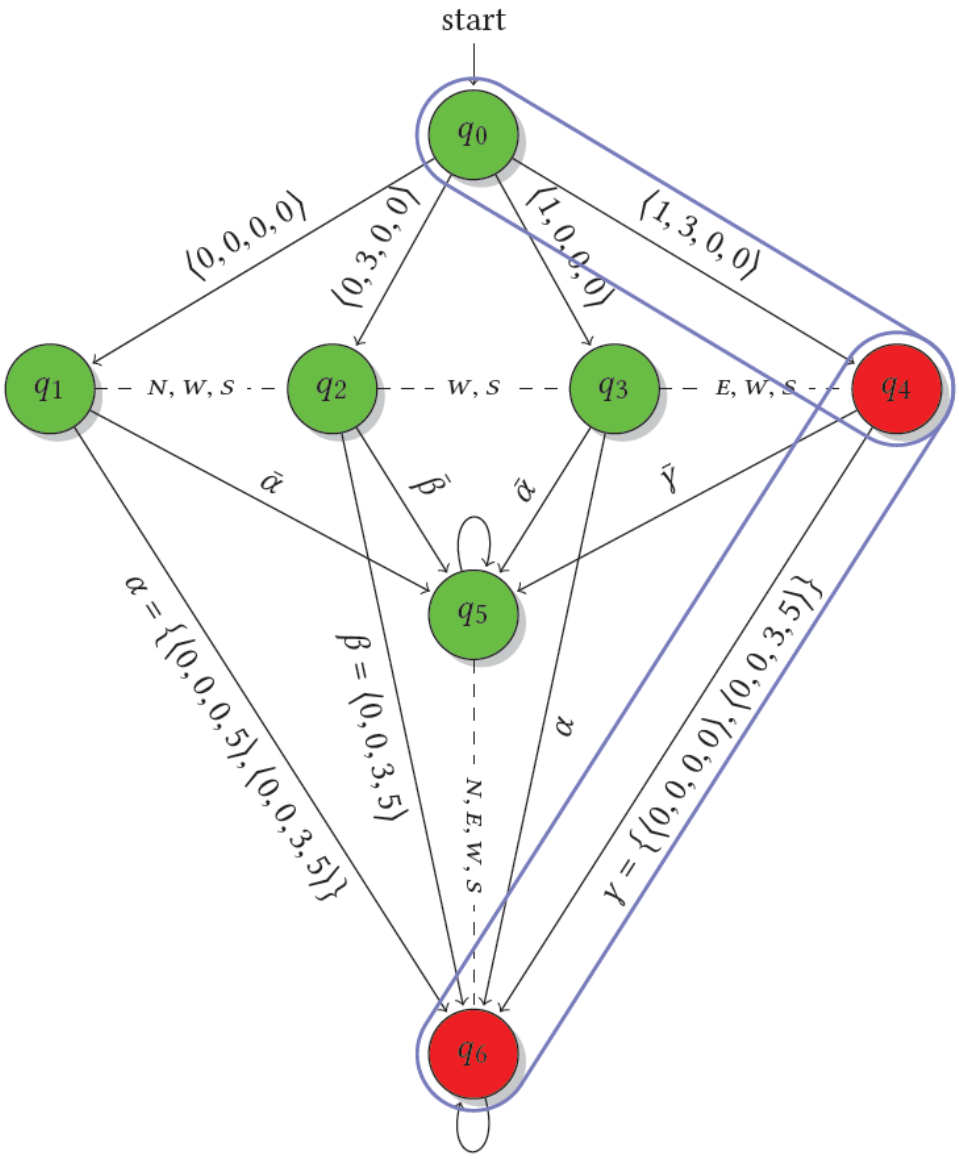


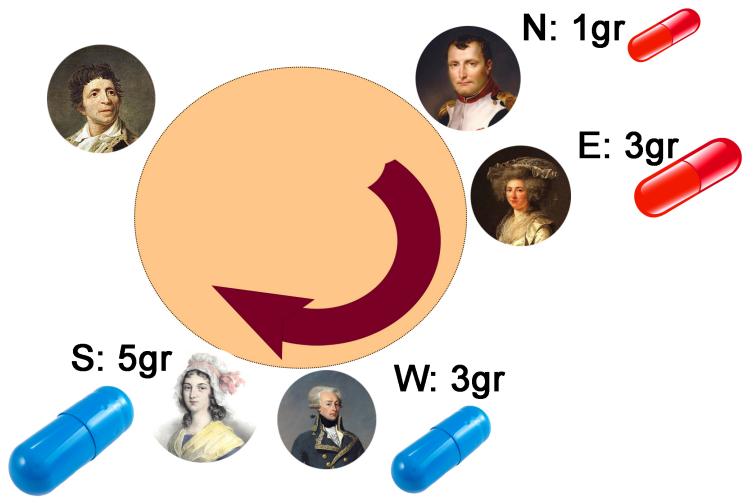


$\langle$ **N** , **E** , **W** , **S** $\rangle$

- *WS*: drop, not

Responsibility in Multiagent Setting

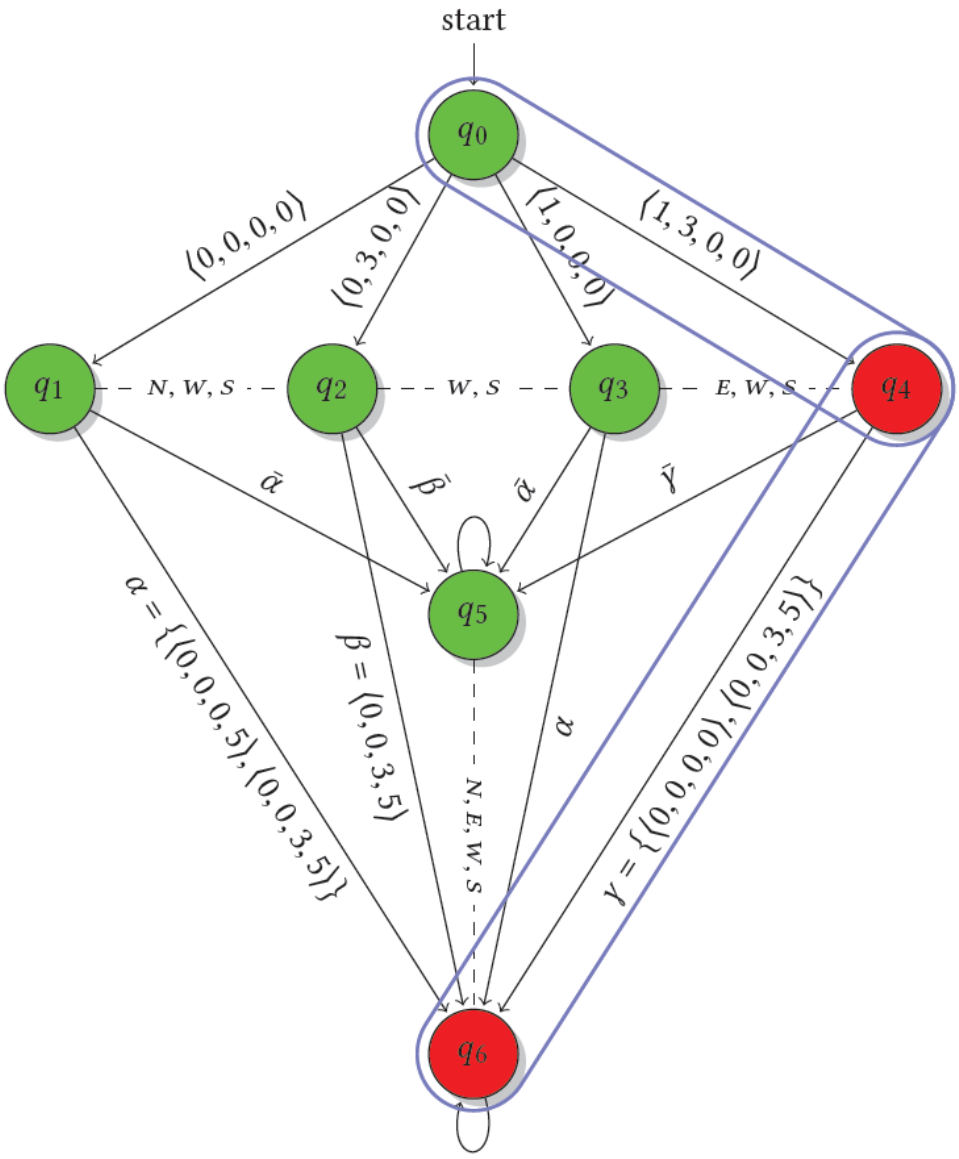


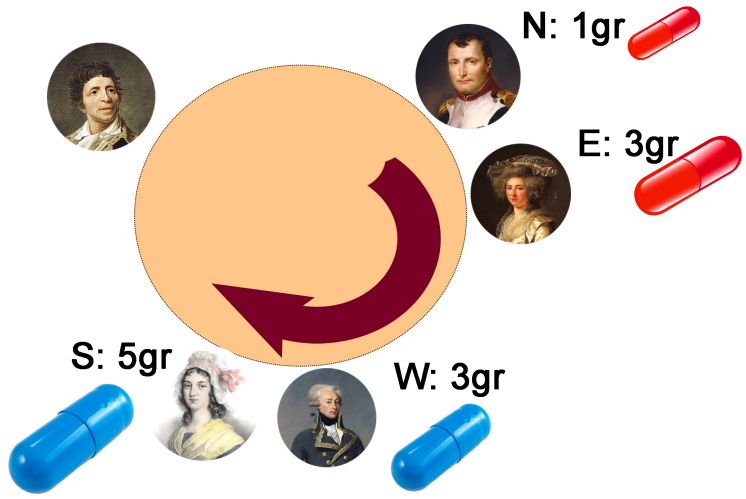


$\langle$ **N** , **E** , **W** , **S** $\rangle$

- *WS*: drop, not
- *NS*:

Responsibility in Multiagent Setting

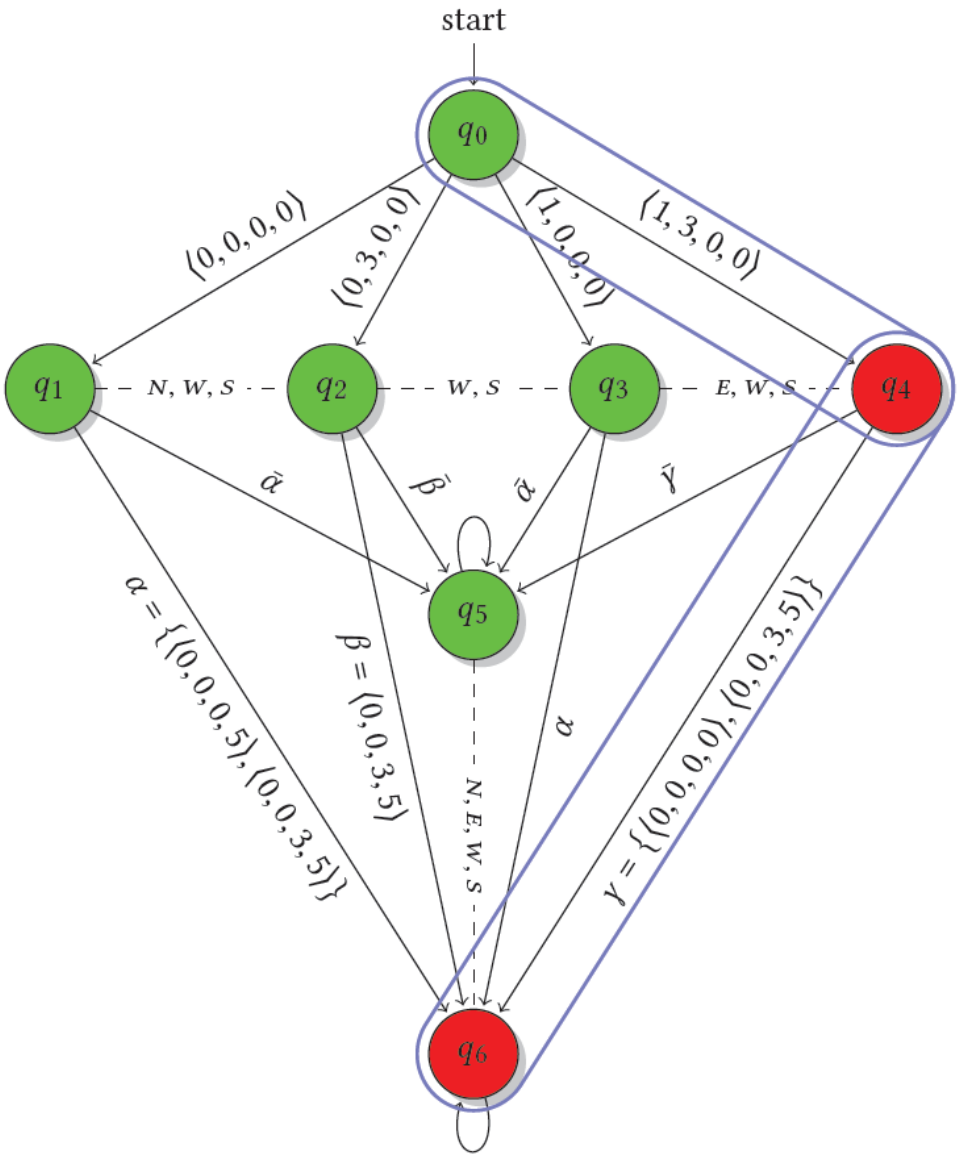


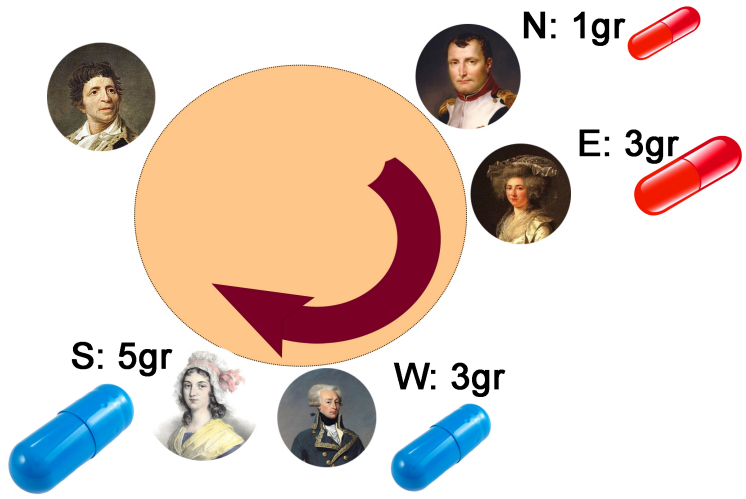


$\langle$ **N** , **E** , **W** , **S** $\rangle$

- *WS*: drop, not
- *NS*: not, not

Responsibility in Multiagent Setting

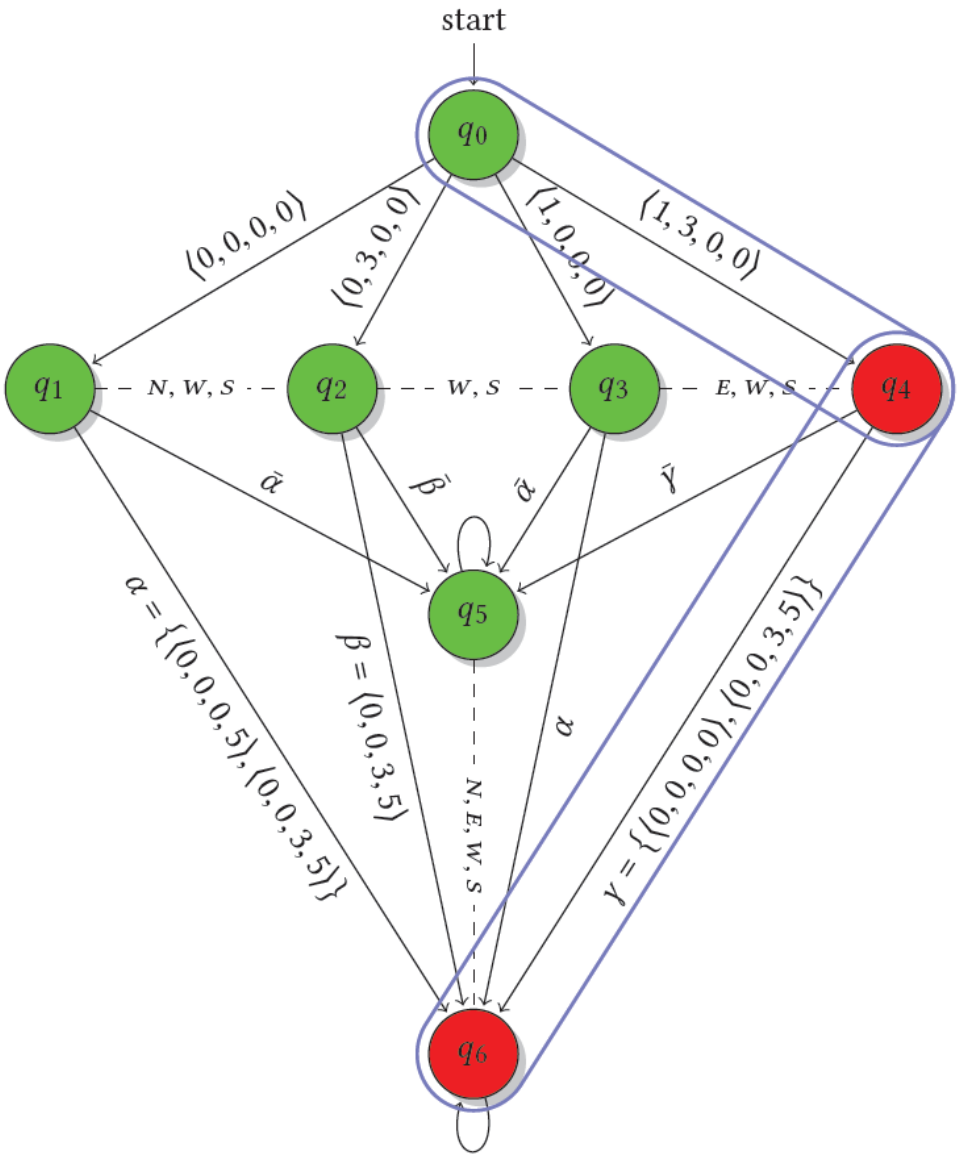


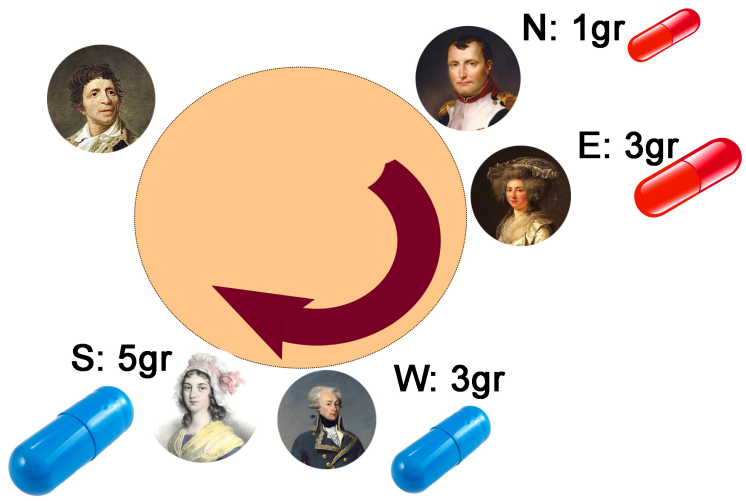


$\langle$ **N** , **E** , **W** , **S** $\rangle$

- *WS*: drop, not
- *NS*: not, not
- *ES*: not, not

Responsibility in Multiagent Setting

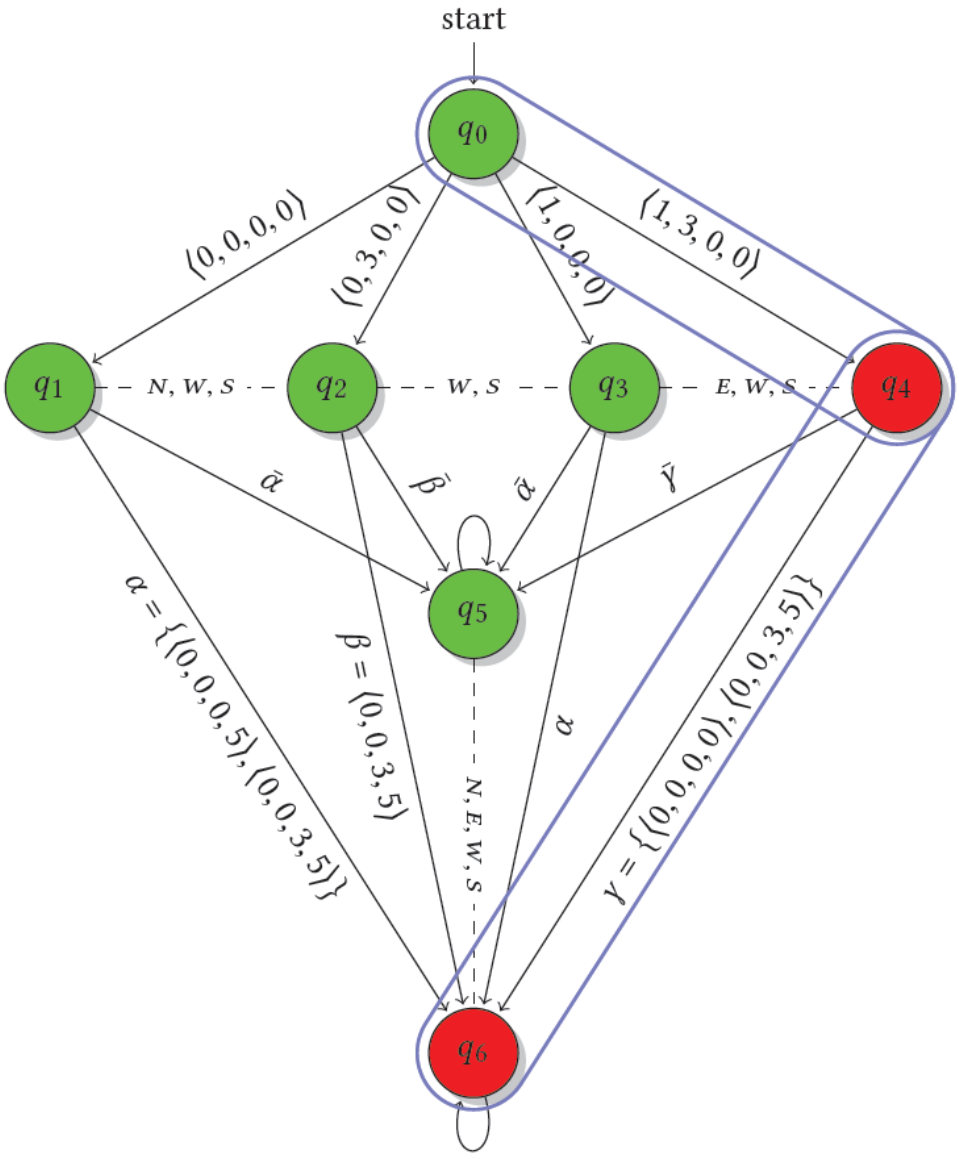


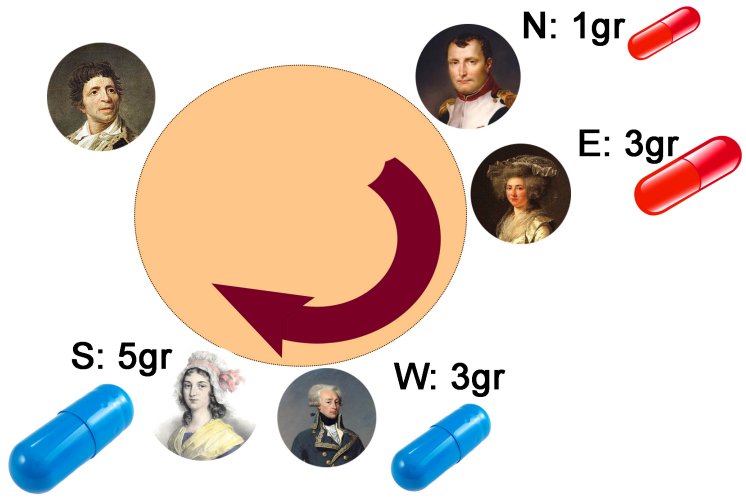


$\langle$ **N** , **E** , **W** , **S** $\rangle$

- *WS*: drop, not
- *NS*: not, not
- *ES*: not, not
- *NEW*:

Responsibility in Multiagent Setting





$\langle$ **N** , **E** , **W** , **S** $\rangle$

- *WS*: drop, not
- *NS*: not, not
- *ES*: not, not
- *NEW*: not, drop, not

Group Responsibility for Risky Behaviour

Motivation

Most of existing approaches assume that the responsibility for an (unsafe or undesirable) outcome can be assigned to a group of agents if **the outcome actually holds**.

However, unsafe outcomes are not necessarily the states of affairs where a bad event **has actually happened**, but also the states of affairs where **the probability of bad events is unacceptably high**. So, it is just as important to consider 'near misses' and 'risky' situations, e.g.

- Leaving very young children alone at home
- Drinking while driving

Legal Doctrines

Endangerment Offence: The chance that harm will occur is unacceptably high.

An agent is held responsible because

- the (group of) agent(s) has created a risky situation where the probability of (remote) harm is unacceptably high, and
- this (group of) agent(s) could act differently to prevent such risky situation.

Note also that in such scenarios being able to **prevent a risk of a (remote) harm** does not usually mean being able to **completely prevent the harm**.

Example

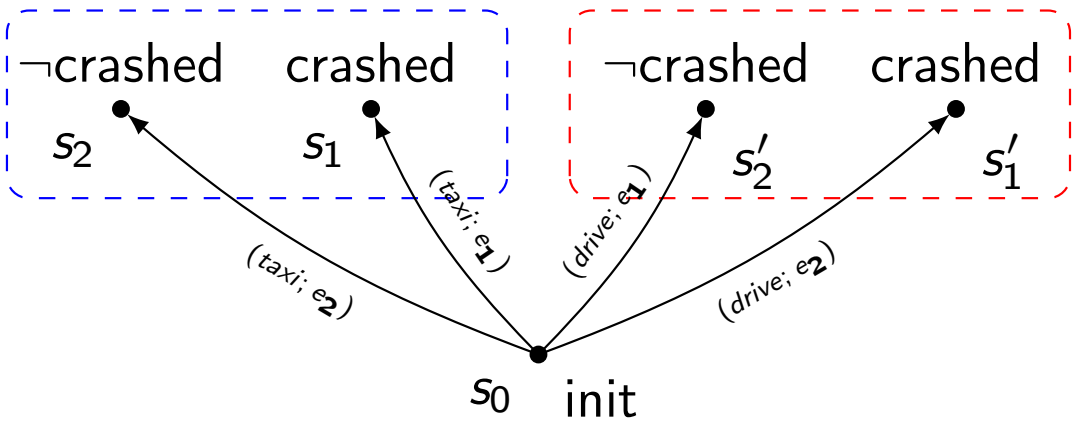
After a party Alice is facing a choice to take a taxi and make a safe trip home, leaving her car at the bar, or to break the law and choose unsafe (above acceptable risk level) trip driving her car home.

- Alice has created a risky situation where the probability of (remote) harm is unacceptably high
- She could act differently to prevent such risky situation
- Note that in both cases an accident may still happen

Example: Drink-driving

After a party Alice is facing a choice to take a taxi and make a safe trip home, leaving her car at the bar, or to break the law and choose unsafe (above acceptable risk level) trip driving her car home.

$$P(s_1)(\text{crashed}) = P(s_2)(\text{crashed}) = .01 \qquad P(s_1)(\neg\text{crashed}) = P(s_2)(\neg\text{crashed}) = .99$$
$$P(s_1')(\text{crashed}) = P(s_2')(\text{crashed}) = .2 \qquad P(s_1')(\neg\text{crashed}) = P(s_2')(\neg\text{crashed}) = .8$$



Language GRR: Syntax

GRR combines coalition ability operator from (Pauly 2002) with a probabilistic operator from (Heifetz and Mongin 2001) that allow reasoning about probabilities and their changes.

Definition (Language)

The language of GRR is defined by the following grammar

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid L_\alpha\varphi \mid [G]\varphi \mid \boxminus\varphi,$$

where $p \in Prop \cup \{init\}$, G ranges over $2^{\mathbb{AG}}$ and α is any rational in $[0, 1]$.

- $L_\alpha\varphi$: "probability of φ is at least α"
  - $M_\alpha\varphi \equiv L_{1-\alpha}\neg\varphi$: "probability of φ is at most α"
  - $I_\alpha\varphi \equiv L_\alpha\varphi \wedge M_\alpha\varphi$: "probability of φ is equal (identical) to α"
- $[G]\varphi$: "group G can enforce φ to be true"
- $\boxminus\varphi$: "φ was true at the previous step"

Language GRR: Semantics

Given a model $\mathcal{M}$ and a state $s \in S$ we define $\models$ relation in the following way:

- $\mathcal{M}, s \models p$ iff $s \in V(p)$;
- $\mathcal{M}, s \models \neg\varphi$ iff $\mathcal{M}, s \not\models \varphi$;
- $\mathcal{M}, s \models \varphi \wedge \psi$ iff $\mathcal{M}, s \models \varphi$ and $\mathcal{M}, s \models \psi$;
- $\mathcal{M}, s \models [G]\varphi$ iff there is a strategy str_G for G, such that for all $s' \in o(s, str_G)$ it holds that $\mathcal{M}, s' \models \varphi$;
- $\mathcal{M}, s \models L_\alpha\varphi$ iff $P(s)([\varphi]^{\mathcal{M}}) \geq \alpha$, where $[\varphi]^{\mathcal{M}} = \{s \in S \mid \mathcal{M}, s \models \varphi\}$;
- $\mathcal{M}, s \models \boxminus\varphi$ iff $\forall s' \in \mathrm{Past}(s) : M, s' \models \varphi$.

Responsibility in GRR

Responsibility for taking risk can be expressed as a formula of GRR:

$$Resp_G(\varphi, \alpha) \equiv_{def} \neg M_\alpha \varphi \wedge \Diamond\!\!\!-[G] M_\alpha \varphi \wedge \bigwedge_{H \subset G} \Diamond\!\!\!-\neg [H] M_\alpha \varphi$$

The correspondence between $Resp_G(\varphi, \alpha)$ and the responsibility conditions:

1. $\mathcal{M}, s \vDash \neg M_\alpha \varphi$ iff $P(s)([\varphi]^{\mathcal{M}}) > \alpha$
2. $\mathcal{M}, s \vDash \Diamond\!\!\!-[G] M_\alpha \varphi$ iff in the unique $s^- \in \text{Past}(s)$ there is a strategy str_G for G, such that for all $s' \in o(s^-, str_G)$ it holds that $P(s')([\varphi]^{\mathcal{M}}) \leq \alpha$
3. $\mathcal{M}, s \vDash \bigwedge_{H \subset G} \Diamond\!\!\!-\neg [H] M_\alpha \varphi$ iff no proper subset of G satisfies $[H] M_\alpha \varphi$ in the unique $s^- \in \text{Past}(s)$

Results

Theorem (Completeness)

Logic GRR *is complete wrt* $\mathcal{M}^{\mathsf{GRR}}$, *i.e.* $\vDash \varphi$ *iff* $\vdash_{\mathsf{GRR}} \varphi$.

Theorem (Decidability)

The satisfiability problem for GRR *is decidable.*

Proposition (Model checking)

The model checking problem for GRR *is decidable in polynomial time.*

Some Properties

Proposition

$\not\models Resp_G(\varphi, \alpha) \rightarrow Resp_G(\varphi, \alpha')$, *where* $\alpha' > \alpha$ *or* $\alpha' < \alpha$.

Proposition

$\models Resp_G(\varphi, \alpha) \wedge Resp_D(\varphi, \beta) \rightarrow \neg Resp_{G \cup D}(\varphi, \min(\alpha, \beta))$
where $D \cap G = \emptyset$.

$\models Resp_G(\varphi, \alpha) \wedge Resp_D(\varphi, \beta) \rightarrow \neg Resp_{G \cup D}(\varphi, \max(\alpha, \beta))$
where $D \cap G = \emptyset$.

Proposition

$\not\models Resp_G(\varphi, \alpha) \wedge Resp_G(\varphi \rightarrow \psi, \alpha) \rightarrow Resp_G(\psi, \alpha)$.

Proposition

$\not\models Resp_G(\varphi, \alpha) \wedge Resp_G(\psi, \alpha) \rightarrow Resp_G(\varphi \wedge \psi, \alpha)$.

$\not\models Resp_G(\varphi \wedge \psi, \alpha) \rightarrow (Resp_G(\varphi, \alpha) \wedge Resp_G(\psi, \alpha))$.

Causal Responsibility

Causality in Multiagent Setting

- **Causality** plays an important role in Artificial Intelligence.
- **Actual causality** concerns causal relations between concrete events.
- **Causality in multiagent settings** concerns causal dependence among agent decisions and environmental events.
- **Combining structural causal models and CGS** relates causality and strategic reasoning, allowing us to reason about responsibility.

Structural Causal Models

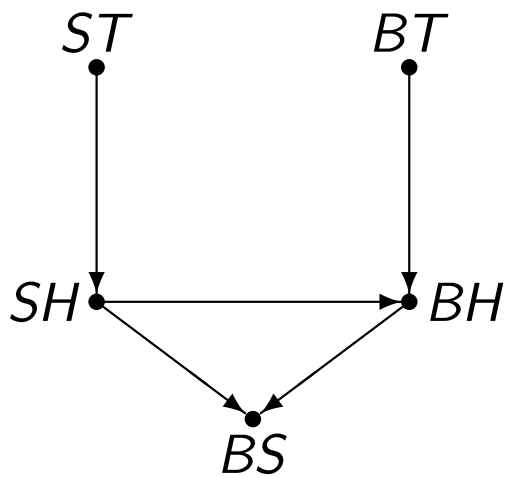
A **signature** $S = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, where

- $\mathcal{U}$ is a finite set of exogenous variables,
- $\mathcal{V}$ is a finite set of endogenous variables,
- $\mathcal{R}(Y)$ is the range of variable $Y \in \mathcal{U} \cup \mathcal{V}$.

A **causal model** $M = (\mathcal{S}, \mathcal{F})$, where

- $\mathcal{F} = \{\mathcal{F}_X \mid \mathcal{F}_X : \mathcal{U} \cup \mathcal{V} \longrightarrow \mathcal{R}(X)\ \}$

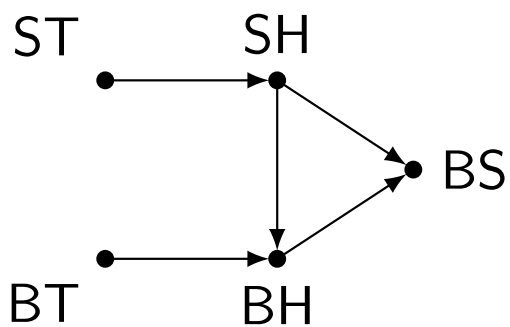
A **causal setting** $(M, \vec{u})$ consists of the causal model M in context $\vec{u}$, i.e., with values for all exogenous variables $\mathcal{U}$.



Example: Rock-throwing

Suzy and Billy both pick up rocks and throw them at a bottle (ST=1, BT=1). Suzy's rock gets there first (SH=1), shattering the bottle (BS=1). Billy's rock does not hit the bottle (BH=0). Billy's rock would've hit the bottle (BH=1) had it not been preempted by Suzy's throw (SH=0). Structural equations can be defined as follows:

- SH:=ST;
- BH:=(BT$\wedge\neg$SH);
- BS:=(SH$\vee$BH).



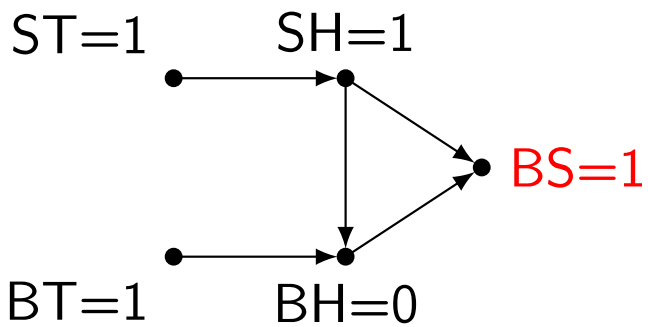
Reasoning in Causal Model

$(M, \vec{u}) \models Y = y$ iff $\mathcal{F}_Y(\vec{Z} = \vec{z}) = y$ for $Z = \mathcal{U} \cup \mathcal{V} \setminus \{Y\}$

"$Y = y$ holds in causal setting $(M, \vec{u})$".

- SH:=ST;
- BH:=(BT$\wedge\neg$SH);
- BS:=(SH$\vee$BH).

- $(\mathcal{M}, \vec{u}) \models ST = 1 \wedge BT = 1 \wedge$ BS=1



Interventions $[X \leftarrow x]$

- $(M, \vec{u}) \models Y = y \quad \wedge \quad [X \leftarrow x]Y = y'$

  "In the causal setting $(M, \vec{u})$, $Y = y$ holds, but if variable X were set to value x, then $Y = y'$ would have happened"

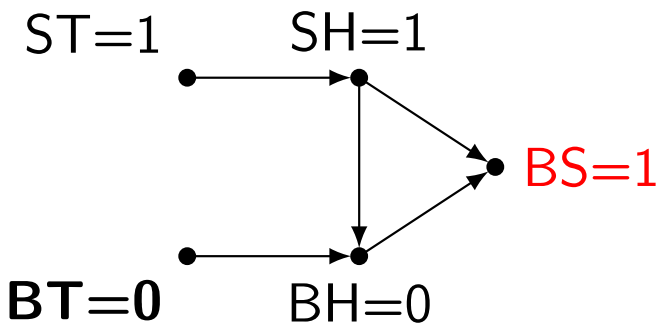
- Intervention $X \leftarrow x$ results in an updated model $\mathcal{M}^{X \leftarrow x}$, in which $\mathcal{F}_X$ is replaced with $\mathcal{F}^{X \leftarrow x}$, s.t. $\mathcal{F}^{X \leftarrow x}$ returns x on any input for X.

- Thus, $(M, \vec{u}) \models Y = y \ \wedge \ [X \leftarrow x]Y = y'$ is equivalent to

$$(M, \vec{u}) \models Y = y \qquad \textit{and} \qquad (\mathcal{M}^{X \leftarrow x}, \vec{u}) \models Y = y'$$

Example: Rock-throwing

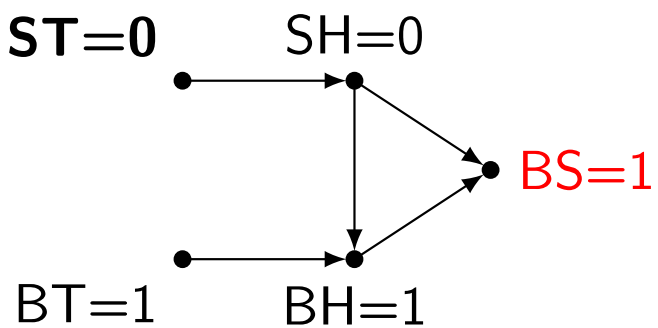
- SH:=ST;
- BH:=(BT$\wedge\neg$SH);
- BS:=(SH$\vee$BH).



- $\mathcal{M}, \vec{u} \vDash ST = 1 \wedge BT = 1 \wedge BS = 1$
- $\mathcal{M}, \vec{u} \vDash [\mathbf{BT} \leftarrow \mathbf{0}]ST = 1 \wedge$ BS=1

Example: Rock-throwing

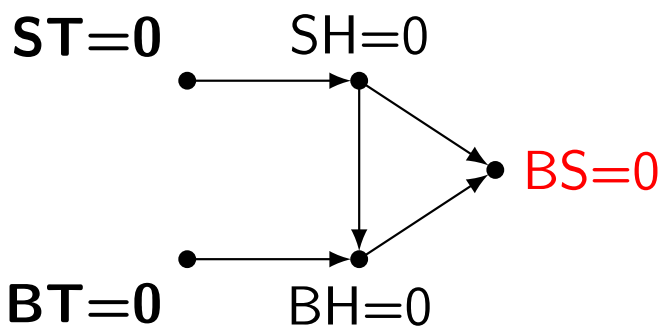
- SH:=ST;
- BH:=(BT$\wedge\neg$SH);
- BS:=(SH$\vee$BH).



- $\mathcal{M}, \vec{u} \models ST = 1 \wedge BT = 1 \wedge BS = 1$
- $\mathcal{M}, \vec{u} \models [BT \leftarrow 0]ST = 1 \wedge$ BS=1
- $\mathcal{M}, \vec{u} \models [\mathbf{ST} \leftarrow \mathbf{0}]BT = 1 \wedge BS = 1$

Example: Rock-throwing

- SH:=ST;
- BH:=(BT$\wedge\neg$SH);
- BS:=(SH$\vee$BH).



- $\mathcal{M}, \vec{u} \models ST = 1 \wedge BT = 1 \wedge BS = 1$
- $\mathcal{M}, \vec{u} \models [BT \leftarrow 0]ST = 1 \wedge$ BS=1
- $\mathcal{M}, \vec{u} \models [ST \leftarrow 0]BT = 1 \wedge BS = 1$
- $\mathcal{M}, \vec{u} \models [\mathbf{ST} \leftarrow \mathbf{0}, \mathbf{BT} \leftarrow \mathbf{0}]BS = 0$

HP^m Cause

Definition

We say that $\vec{X} = \vec{x}$ is an actual cause of φ in $(\mathcal{M}, \vec{u})$ if the following three conditions hold:

- **AC1**. $(\mathcal{M}, \vec{u}) \models (\vec{X} = \vec{x})$ and $(\mathcal{M}, \vec{u}) \models \varphi$
- **AC2**. There is a set $\vec{W}$ of variables in $\mathcal{V}$, such that if $(\mathcal{M}, \vec{u}) \models \vec{W} = \vec{w^*}$, then

$$(\mathcal{M}, \vec{u}) \models [\vec{X} \leftarrow \vec{x'}, \vec{W} \leftarrow \vec{w^*}]\neg\varphi$$

- **AC3**. $\vec{X}$ is minimal: no proper subset of $\vec{X}$ satisfies **AC2**.

Concurrent Game Structures

Concurrent Game Structure $GS = \langle N, G, D, \delta, \Pi, \pi \rangle$, where

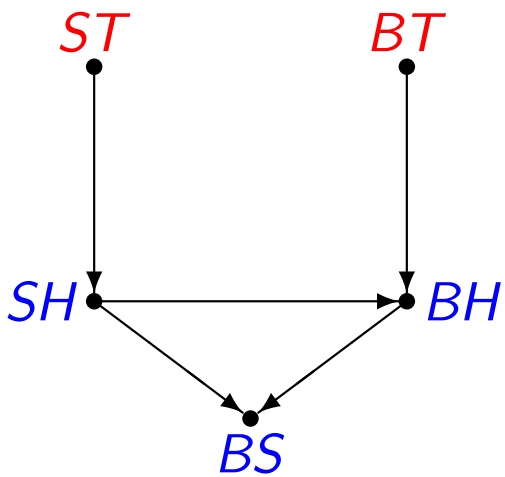
- N is the number of players
- Q is the set of states
- D, a function that for every state $q \in Q$ defines a set of move vectors available at that state
- δ is the transition function that maps a state and move vector to a new state
- Π is a set of propositions
- π is a function that assigns a set of true propositions to every state q

Overview

- Construct a **Causal CGS** from a recursive **SCM**
- $\mathcal{V} = V_a \cup V_e$: Endogenous variables are split into agent (V_a) and environment (V_e) variables.
- Agent actions are values of agent variables (e.g., $X = x$ the action of agent X is x).
- Counterfactual actions are interventions on the model.
- We order variables in the causal model to model the order that agents take actions.

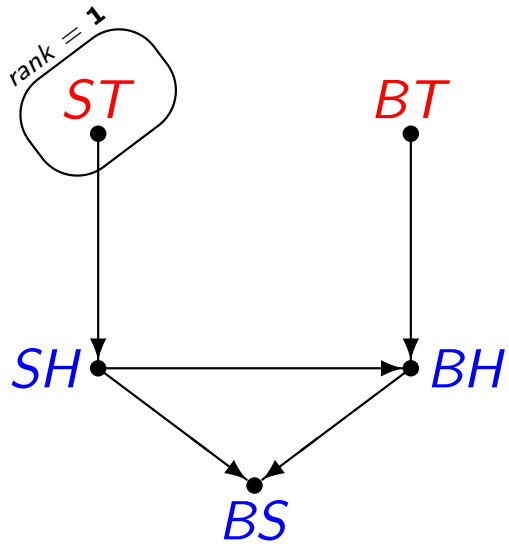
Rank

In the rock throwing example, ST and BT are agent variables, SH, BH and BS are environment variables.



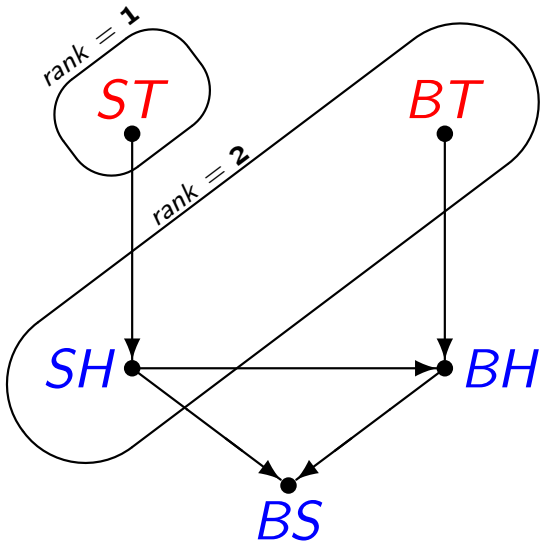
Rank

In the rock throwing example, ST and BT are agent variables, SH, BH and BS are environment variables.



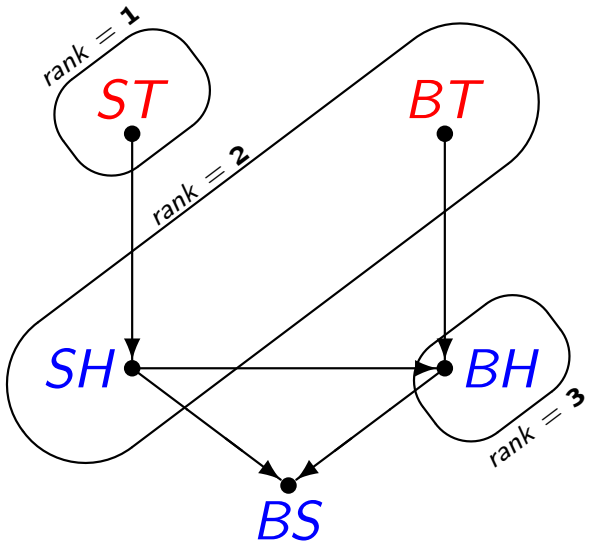
Rank

In the rock throwing example, ST and BT are agent variables, SH, BH and BS are environment variables.



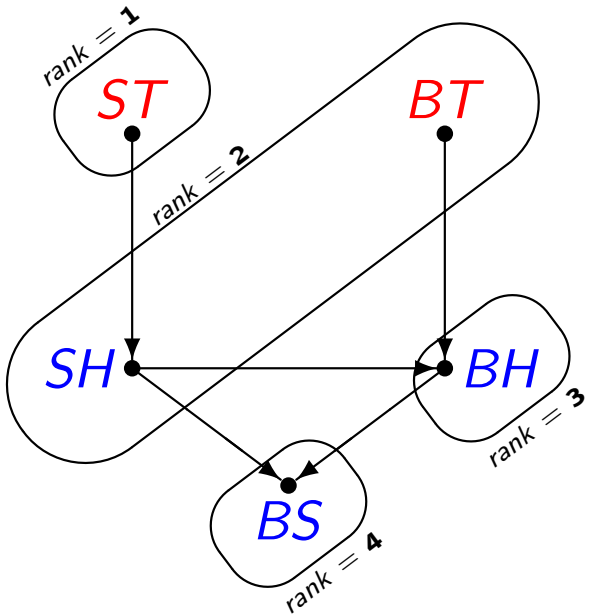
Rank

In the rock throwing example, ST and BT are agent variables, SH, BH and BS are environment variables.



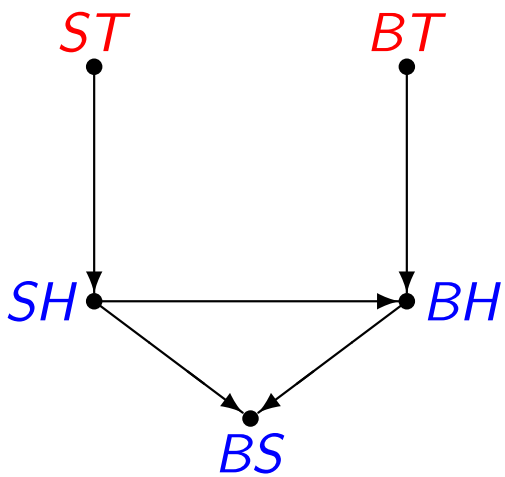
Rank

In the rock throwing example, ST and BT are agent variables, SH, BH and BS are environment variables.



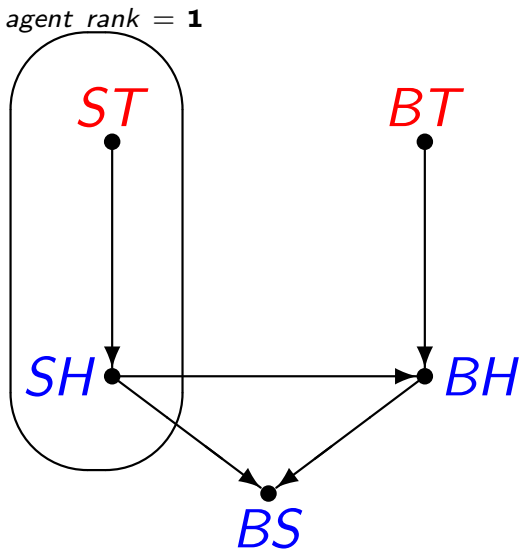
Agent Rank

Remember that *ST* had rank 1, *BT* and *SH* had rank 2, *BH* had rank 3 and *BS* had rank 4



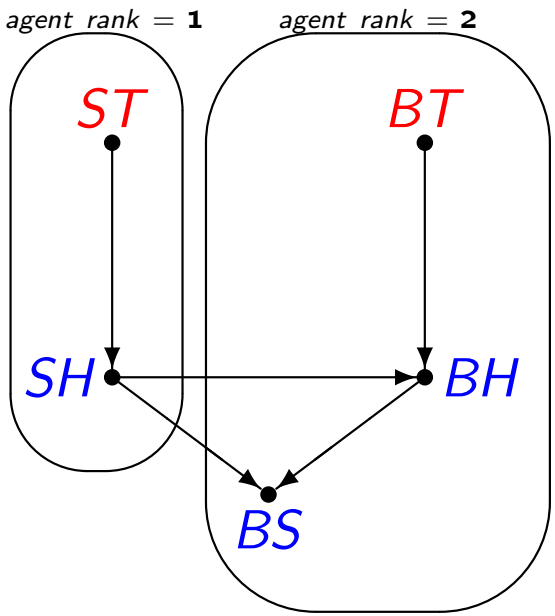
Agent Rank

Remember that *ST* had rank 1, *BT* and *SH* had rank 2, *BH* had rank 3 and *BS* had rank 4



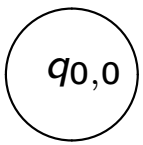
Agent Rank

Remember that *ST* had rank 1, *BT* and *SH* had rank 2, *BH* had rank 3 and *BS* had rank 4



Causal CGS (1)

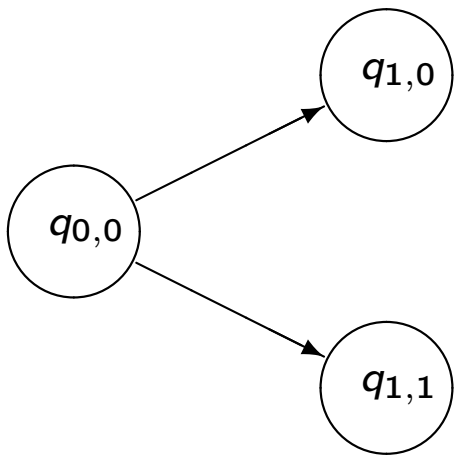
Agents get to take actions in a sequential order based on their agent rank



Causal CGS (1)

Agents get to take actions in a sequential order based on their agent rank

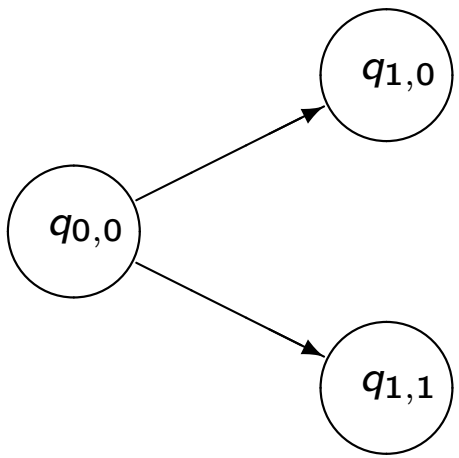
- ST is the only agent variable with agent rank 1



Causal CGS (1)

Agents get to take actions in a sequential order based on their agent rank

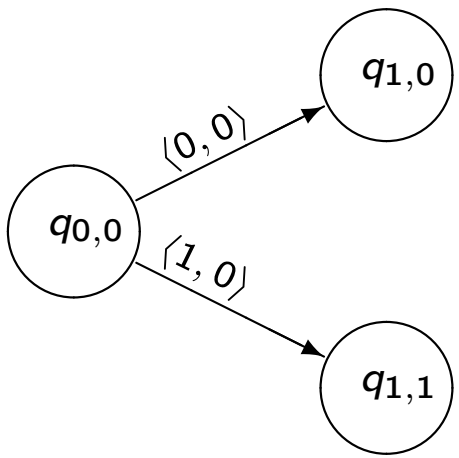
- ST is the only agent variable with agent rank 1
- $\mathcal{R}(ST) = \{0, 1\}$



Causal CGS (1)

Agents get to take actions in a sequential order based on their agent rank

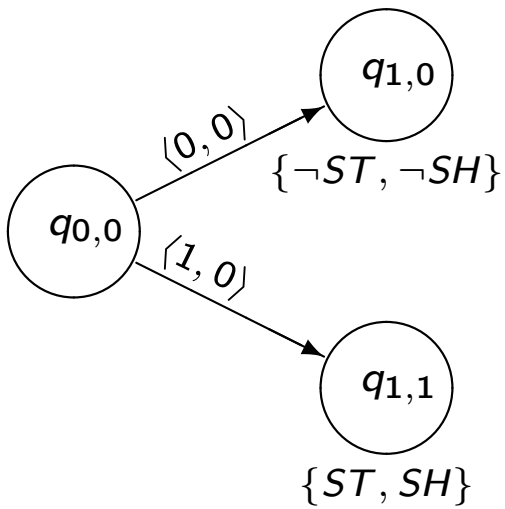
- ST is the only agent variable with agent rank 1
- $\mathcal{R}(ST) = \{0, 1\}$
- The possible moves will be $\langle 0, 0 \rangle$ and $\langle 1, 0 \rangle$



Causal CGS (1)

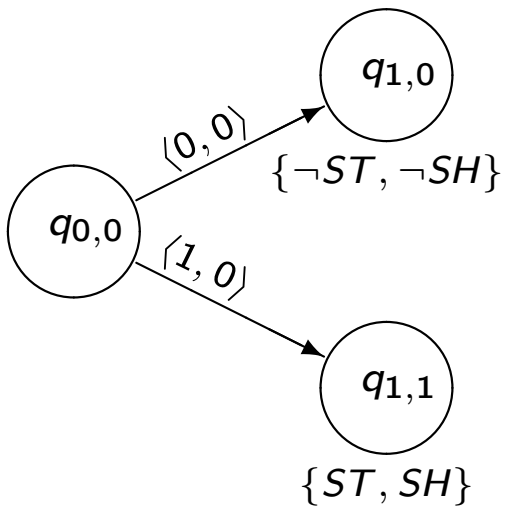
Agents get to take actions in a sequential order based on their agent rank

- ST is the only agent variable with agent rank 1
- $\mathcal{R}(ST) = \{0, 1\}$
- The possible moves will be $\langle 0, 0 \rangle$ and $\langle 1, 0 \rangle$



Causal CGS (2)

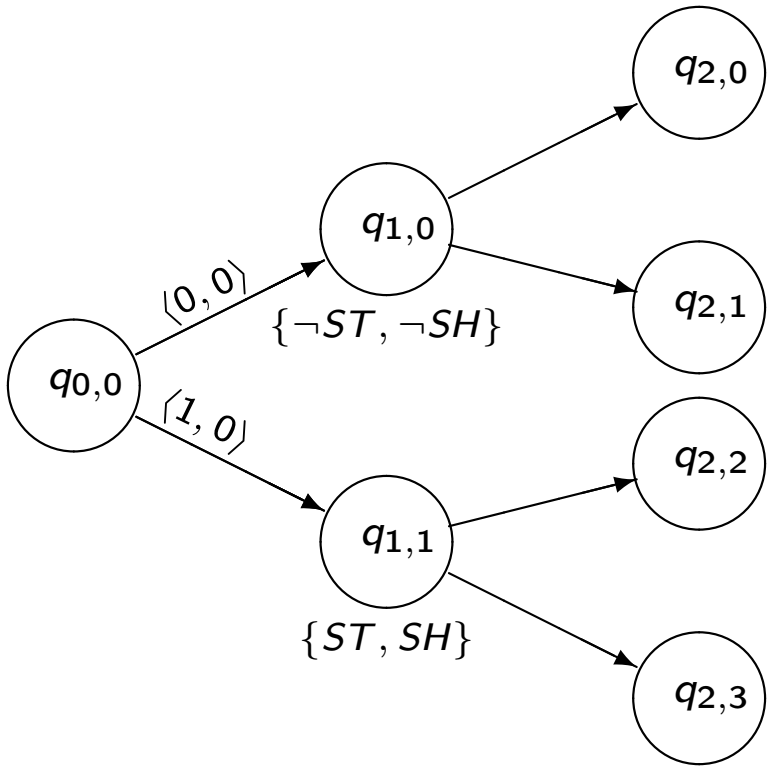
Agents get to take actions in a sequential order based on their agent rank



Causal CGS (2)

Agents get to take actions in a sequential order based on their agent rank

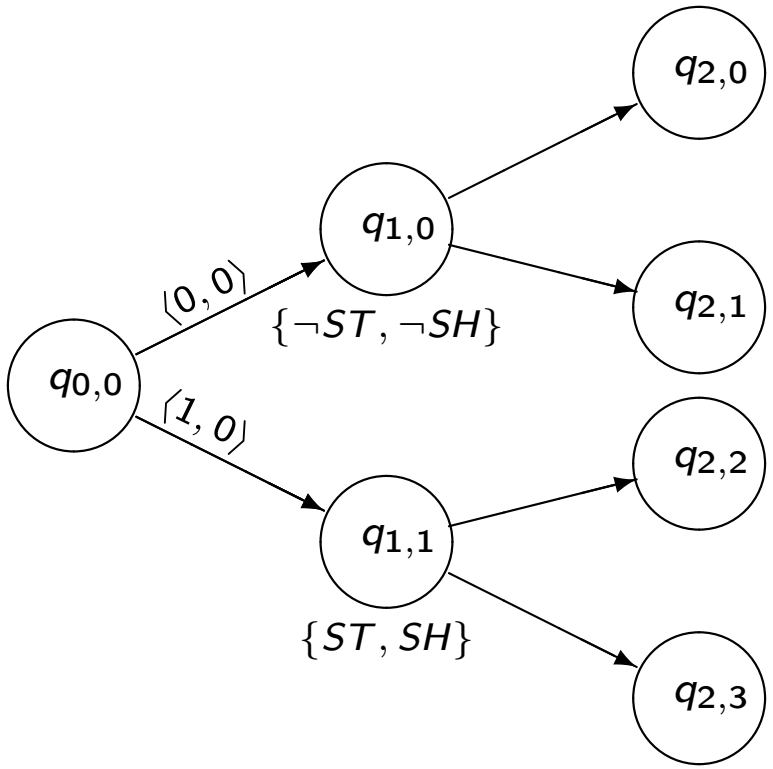
- BT is the only agent variable with agent rank 2



Causal CGS (2)

Agents get to take actions in a sequential order based on their agent rank

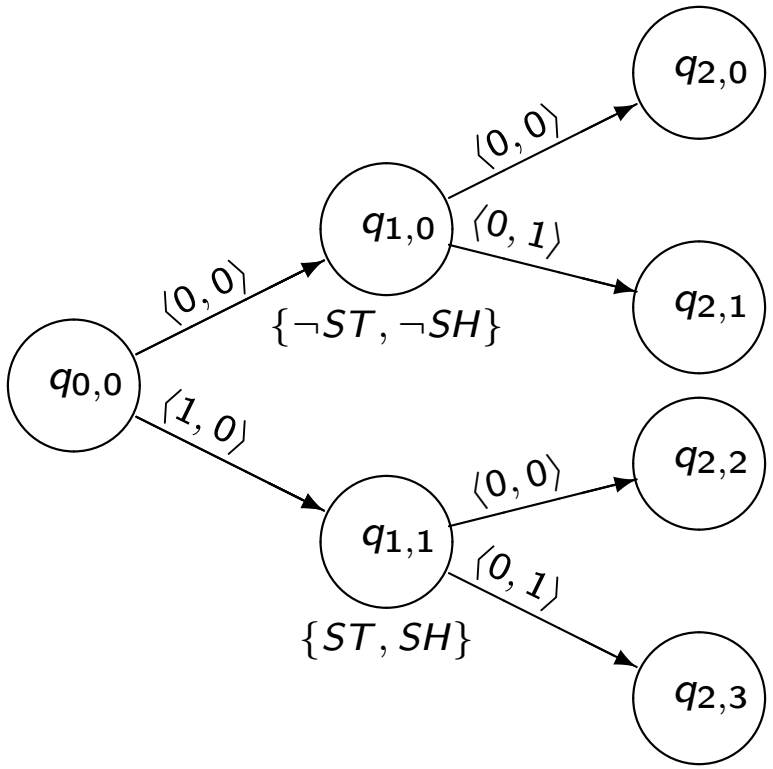
- BT is the only agent variable with agent rank 2
- $\mathcal{R}(BT) = \{0, 1\}$



Causal CGS (2)

Agents get to take actions in a sequential order based on their agent rank

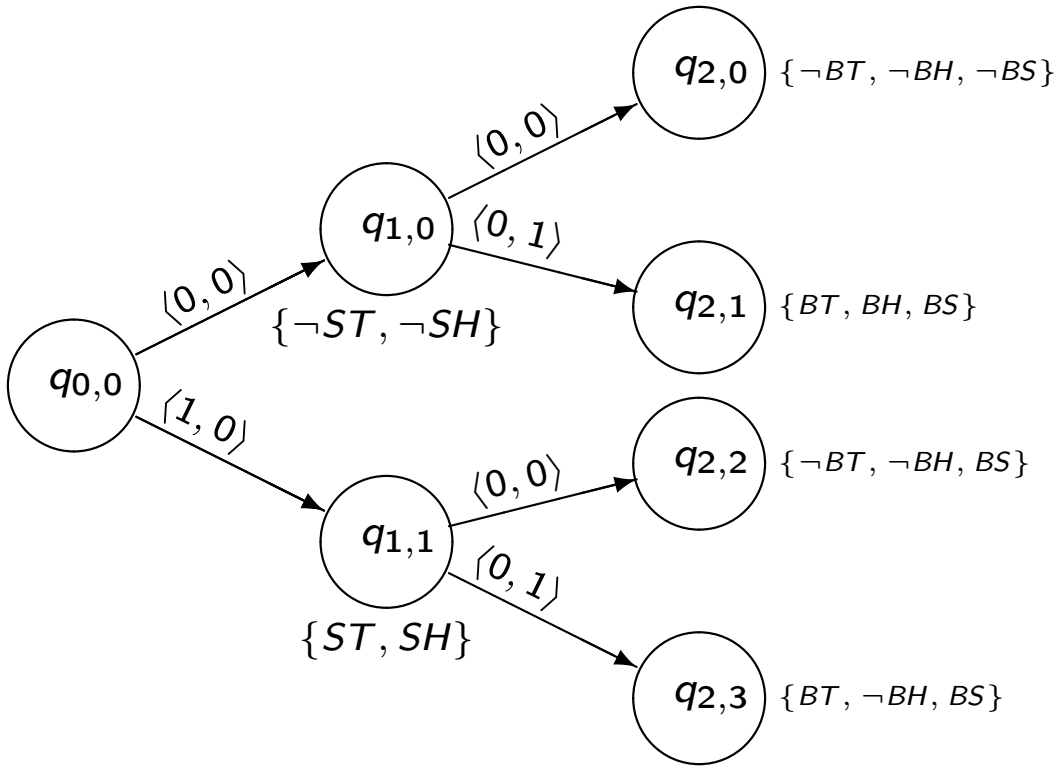
- BT is the only agent variable with agent rank 2
- $\mathcal{R}(BT) = \{0, 1\}$
- The possible moves will be $\langle 0, 0\rangle$ and $\langle 0, 1\rangle$



Causal CGS (2)

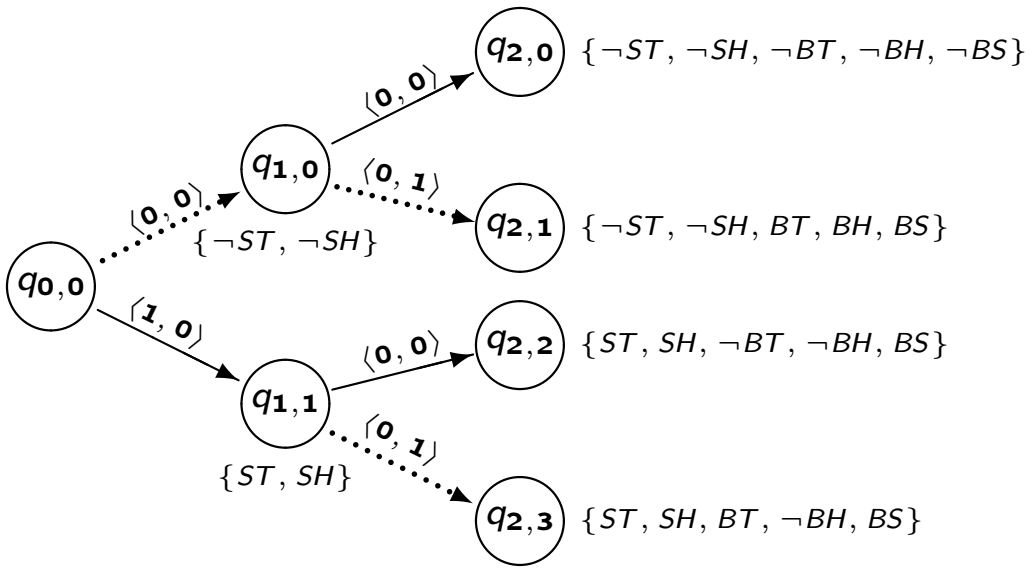
Agents get to take actions in a sequential order based on their agent rank

- BT is the only agent variable with agent rank 2
- $\mathcal{R}(BT) = \{0, 1\}$
- The possible moves will be $\langle 0, 0\rangle$ and $\langle 0, 1\rangle$



Causal Strategy Profile

We define the causal strategy profile to be those actions in the causal CGS that an agent would take if they would follow the original causal setting.

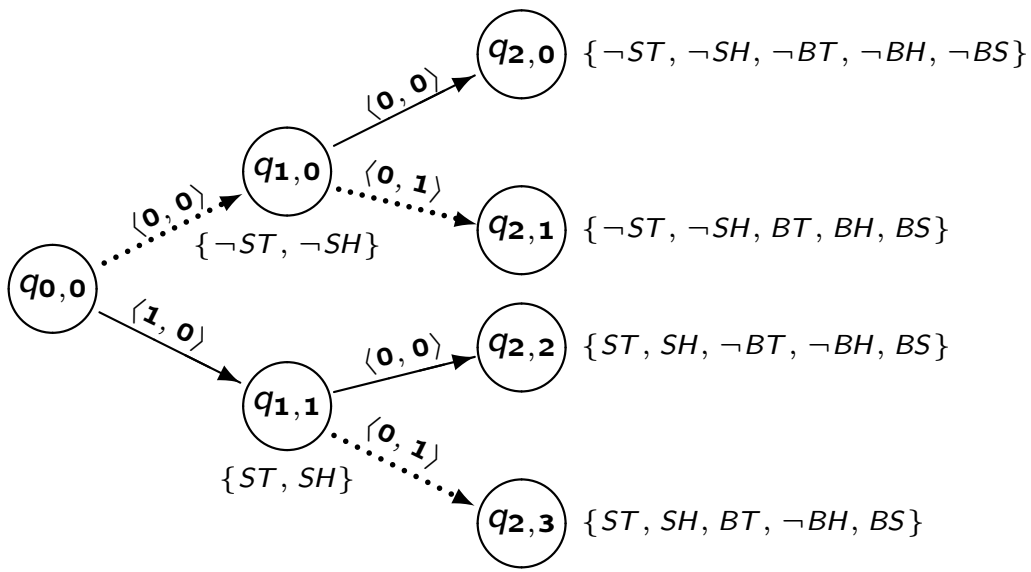


Relation Cause and Strategy

Theorem

Correspondence With the causal strategy profile we can show that a set of agent variables $\mathbf{X}$ *is a but-for cause of a causal formula* ϕ *if and only if the coalition of the agents controlling* $\mathbf{X}$ *has a strategy that makes* $\neg\phi$ *true in a leaf state, provided the agents not in the coalition follow the causal strategy profile.*

- In the causal setting for the rock throwing example with ST and $\neg BT$, ST is a but-for cause of the bottle shattering.
- If Suzy follows the strategy to always take action 0 and Billy follows the causal strategy profile, they will reach leaf-state $q_{2,0}$ where indeed $\neg BS$ holds



Conclusion

- Other conditions on responsibility: Epistemic, Norms, Motivation
- Other models of responsibility for risk taking: for not have taken the alternative to a lower risk, or to minimise risk
- Dynamic causality and responsibility
- Causality in multiagent organisations