Logics of Responsibility

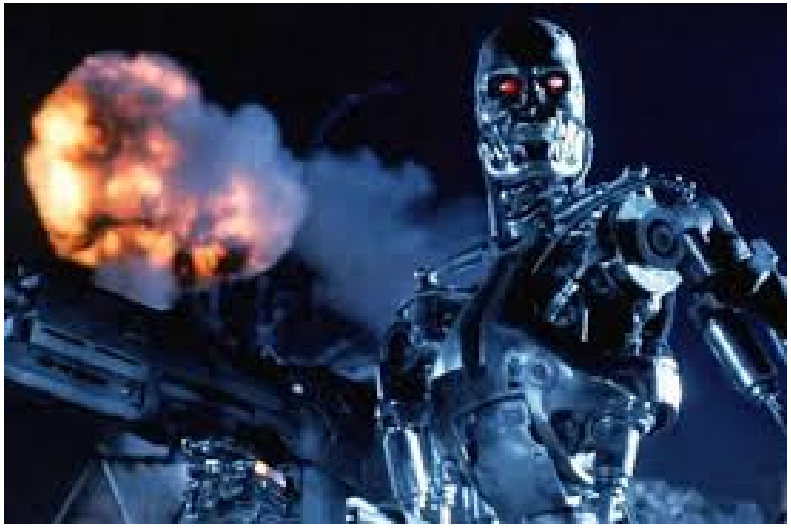
Seminario Inteligencia Artificial y Lógica

Aldo Iván Ramírez Abarca

May 6, 2025

Logics of Responsibility

How can we design and verify responsible AI?



Responsible AI

A topic of increasing importance nowadays

Responsible AI

A topic of increasing importance nowadays

- Transport industry (self-driving cars).

Responsible AI

A topic of increasing importance nowadays

- Transport industry (self-driving cars).
- Healthcare industry (prognosis and management of AI assistants).

Responsible AI

A topic of increasing importance nowadays

- Transport industry (self-driving cars).
- Healthcare industry (prognosis and management of AI assistants).
- Military industry (automated retaliation and reconnaissance technology).

Responsible AI

A topic of increasing importance nowadays

- Transport industry (self-driving cars).
- Healthcare industry (prognosis and management of AI assistants).
- Military industry (automated retaliation and reconnaissance technology).
- Market (algotrading).

- Elon Musk (CEO Tesla): “We need to be super careful with AI. Potentially more dangerous than nukes.”
- Bill Gates (CEO Microsoft): “If computers will run the algorithms that are in our heads, nobody knows what will happen.”
- Nick Bostrom (Professor at Oxford, Director of Future of Humanity Institute): “Machine intelligence is the last invention that humanity will ever need to make.”
- Stephen Hawking (World-class Physicist): “Machines with AI could spell the end of the human race.”


What should I do?

Proposal

Symbolic AI: Formal Methods

Proposal

Symbolic AI: Formal Methods

- Explicit models of an agent's knowledge and an agent's decision-making.

Proposal

Symbolic AI: Formal Methods

- Explicit models of an agent's knowledge and an agent's decision-making.
- Rule-based manipulation of symbols encoding such knowledge and decision-making.

Proposal

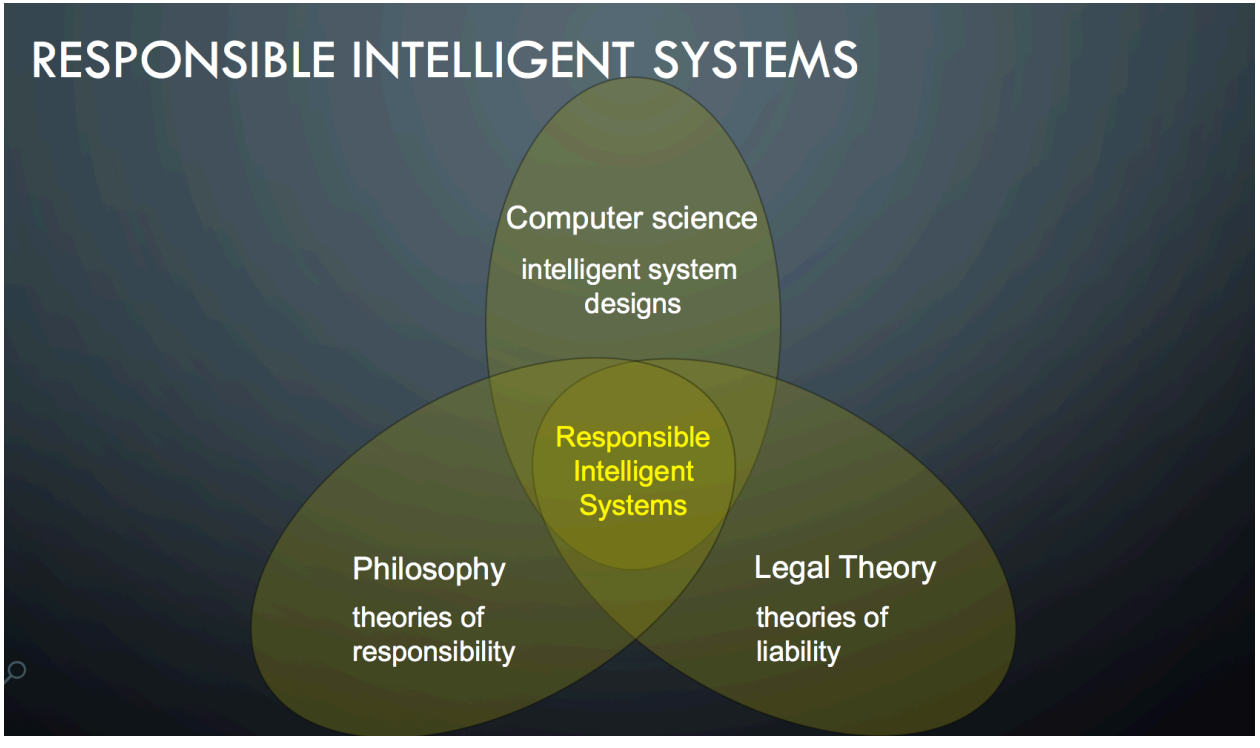
Symbolic AI: Formal Methods

- Explicit models of an agent's knowledge and an agent's decision-making.
- Rule-based manipulation of symbols encoding such knowledge and decision-making.

***Premise*: we can develop logic(s) that would allow computational checks of (ethical) specifications through Formal Verification.**

REINS

REINS (REsponsible Intelligent Systems, 2014–2022)



REINS

Two goals:

REINS

Two goals:

1. Develop logics to analyze responsibility and obligation (based on logics of action, deontic logic, epistemic logic, logics of intention).

REINS

Two goals:

1. Develop logics to analyze responsibility and obligation (based on logics of action, deontic logic, epistemic logic, logics of intention).
2. Implementation of the logics in **formal verification**.

REINS

Two goals:

1. Develop logics to analyze responsibility and obligation (based on logics of action, deontic logic, epistemic logic, logics of intention).
2. Implementation of the logics in **formal verification**.

These points roughly outline the agenda for today's talk.

Formal Theory of Responsibility

Responsibility: *a relation between the agents and the states of affairs of an environment, such that an agent is responsible for a state of affairs iff the agent's degree of involvement in the realization of that state of affairs warrants blame or praise (in light of a given normative system)*.

Components of responsibility:

Components of responsibility:

- Agents: the so-called bearers of responsibility, the authors of actions or the actors.

Components of responsibility:

- Agents: the so-called bearers of responsibility, the authors of actions or the actors.
- Actions: the processes by which agents bring about changes or effects in their environment.

Components of responsibility:

- Agents: the so-called bearers of responsibility, the authors of actions or the actors.
- Actions: the processes by which agents bring about changes or effects in their environment.
- Knowledge and beliefs.- mental attitudes concerning the information available in the environment. They constitute important *explanations* for agents' particular choices of action and therefore provide justifications for the situations in which agents cannot comply with their obligations.

Components of responsibility:

- Agents: the so-called bearers of responsibility, the authors of actions or the actors.
- Actions: the processes by which agents bring about changes or effects in their environment.
- Knowledge and beliefs.- mental attitudes concerning the information available in the environment. They constitute important *explanations* for agents' particular choices of action and therefore provide justifications for the situations in which agents cannot comply with their obligations.
- Intentions: agentive states that determine whether an action was done with the purpose of bringing about its effects.

Components of responsibility:

- Agents: the so-called bearers of responsibility, the authors of actions or the actors.
- Actions: the processes by which agents bring about changes or effects in their environment.
- Knowledge and beliefs.- mental attitudes concerning the information available in the environment. They constitute important *explanations* for agents' particular choices of action and therefore provide justifications for the situations in which agents cannot comply with their obligations.
- Intentions: agentive states that determine whether an action was done with the purpose of bringing about its effects.
- Obligations: the actions that agents should comply with, given by some normative system ruling over the interaction of agents in the environment. Such a normative system could be moral, judicial, legal, etc, and it is according to its tenets that agents can be either blamed or praised for the performance of some action.

Categories of Responsibility

1. ***Causal responsibility***: an agent is causally responsible for a state of affairs iff the agent is the material author of such a state of affairs. The component that this category involves is agency.
2. ***Informational responsibility***: an agent is informationally responsible for a state of affairs iff the agent is the material author and it behaved consciously while bringing about such a state of affairs. The components that this category involves is agency, knowledge, and belief.
3. ***Motivational responsibility***: an agent is motivationally responsible for a state of affairs iff the agent is the material author and it behaved consciously while bringing about such a state of affairs. The components that this category involves are agency and intentions.

A Logic of Agency

Stit Theory

A Logic of Agency

Stit Theory

- Language: given an agent α, $[\alpha]\varphi$ expresses that α brought about φ.
- Models: branching-time models

Basic Stit Theory

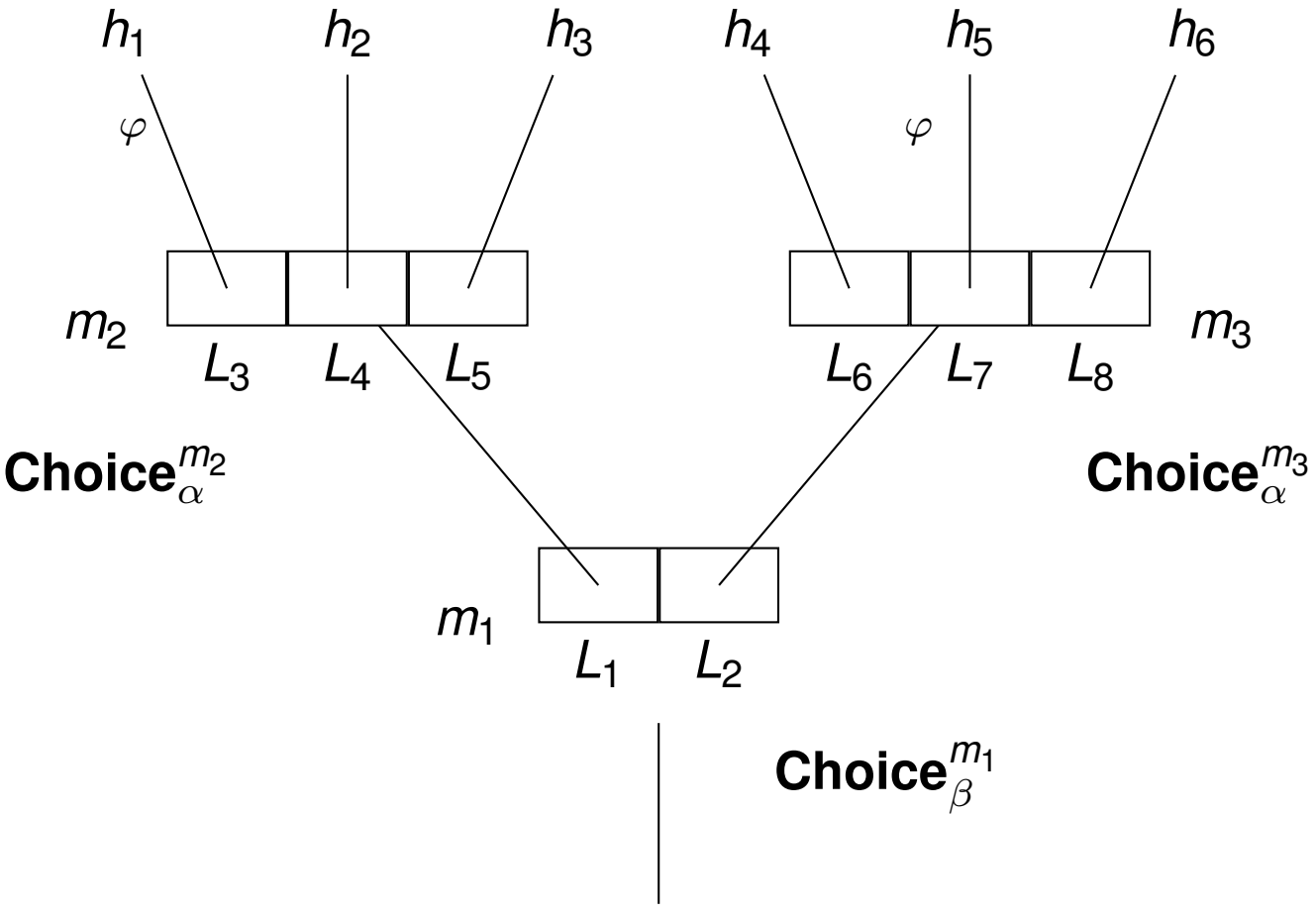
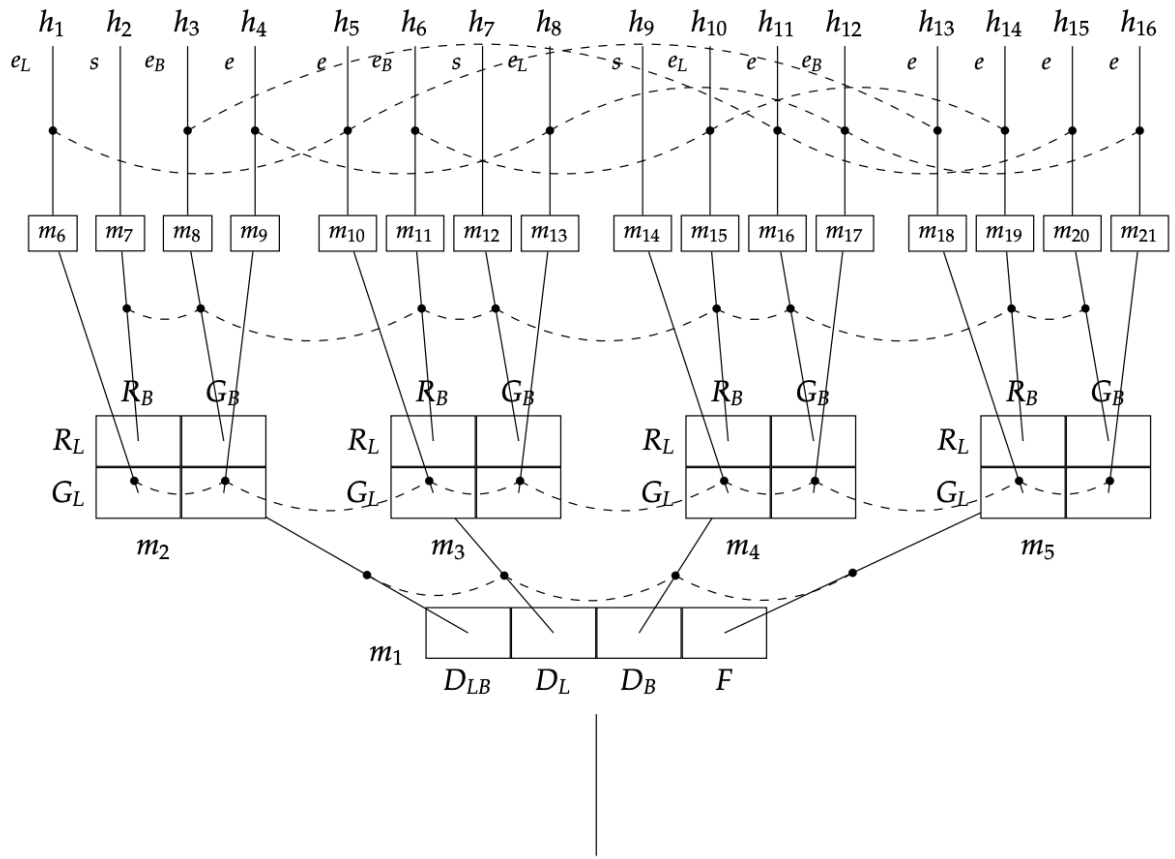


Figure: Stit diagram

Basic Stit Theory



Extensions of Stit Theory to Model Responsibility

- Knowledge: $K_\alpha\varphi$ - equivalence relations,
- Beliefs: $B_\alpha\varphi$ - probabilities.
- Intentions: $I_\alpha\varphi$ - topologies of present-directed intentions.
- Obligations: $\odot_\alpha\varphi$ - utilitarian value functions, dominance, and optimality of actions.

Basic Stit Theory

Heigh Ho...

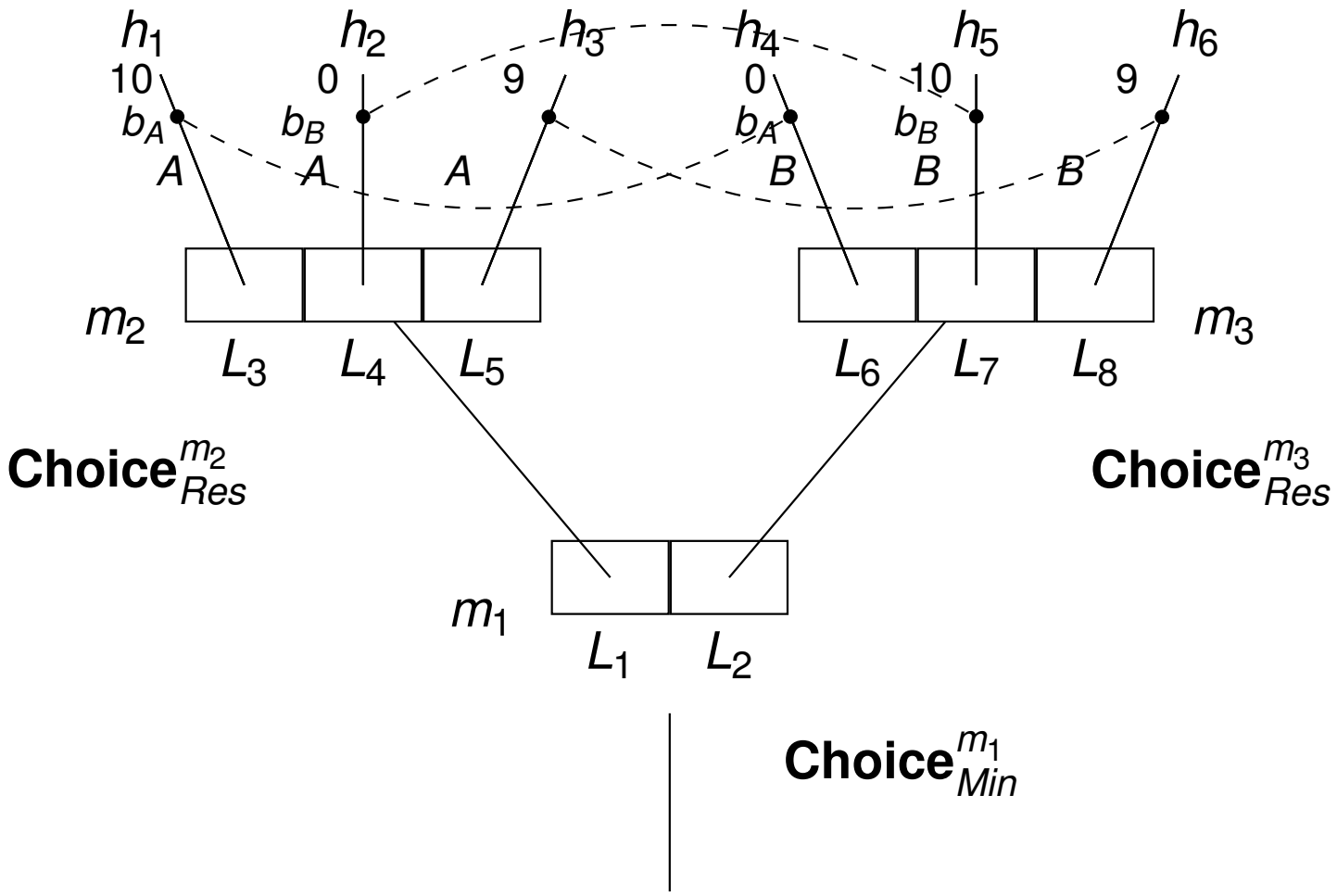


Figure: Miners paradox.

Categories of Responsibility

Category \ *Form*	Active (contributions)	Passive (omissions)
Causal	$[\alpha]\varphi \wedge \Diamond[\alpha]\neg\varphi$	$\varphi \wedge \Diamond[\alpha]\neg\varphi$
Informational	$K_\alpha[\alpha]\varphi \wedge \Diamond K_\alpha[\alpha]\neg\varphi$	$\varphi \wedge K_\alpha\neg[\alpha]\neg\varphi \wedge$ $\Diamond K_\alpha[\alpha]\neg\varphi$
Motivational	$K_\alpha[\alpha]\varphi \wedge I_\alpha[\alpha]\varphi \wedge$ $\Diamond K_\alpha[\alpha]\neg\varphi$	$\varphi \wedge K_\alpha\neg[\alpha]\neg\varphi \wedge$ $I_\alpha\neg[\alpha]\neg\varphi \wedge \Diamond K_\alpha[\alpha]\neg\varphi$

Table: Main sub-categories.

Intentional epistemic deontic STIT logic: semantics

Definition (Intentional epistemic act-utilitarian branching-time frames)

An intentional epistemic act-utilitarian branching-time frame (*ieaubt*-frame for short) is a tuple $\left\langle M, \sqsubset, Ags, \mathbf{Choice}, \{\sim_\alpha\}_{\alpha \in Ags}, \mathbf{Value}, \tau \right\rangle$ such that:

- M is a non-empty *finite* set of moments and $\sqsubset$ is a strict partial ordering on M satisfying 'no backward branching.' Each maximal $\sqsubset$-chain is called a history, which represents a way in which time might evolve. H denotes the set of all histories, and for each $m \in M$, $H_m := \{h \in H; m \in h\}$. Tuples $\langle m, h \rangle$ are called *indices* iff $m \in M$, $h \in H$, and $m \in h$.
- **Choice** is a function that maps each agent α and moment m to a partition $\mathbf{Choice}_\alpha^m$ of H_m, where the cells of such a partition represent α's available actions at m.

Intentional epistemic deontic STIT logic: semantics

Definition

- For $\alpha \in Ags$, $\sim_\alpha$ is the epistemic indistinguishability equivalence relation for agent α, which satisfies the following constraints: (OAC) *Own action condition*: For every index $\langle m_*, h_* \rangle$, if $\langle m_*, h_* \rangle \sim_\alpha \langle m, h \rangle$ for some $\langle m, h \rangle$, then $\langle m_*, h'_* \rangle \sim_\alpha \langle m, h \rangle$ for every $h'_* \in \mathbf{Choice}_\alpha^{m_*}(h_*)$. We refer to this constraint as the 'own action condition' because it implies that agents do not know more than what they perform. (Unif − H) *Uniformity of historical possibility*: For every index $\langle m_*, h_* \rangle$, if $\langle m_*, h_* \rangle \sim_\alpha \langle m, h \rangle$ for some $\langle m, h \rangle$, then for every $h'_* \in H_{m_*}$ there exists $h' \in H_m$ such that $\langle m_*, h'_* \rangle \sim_\alpha \langle m, h' \rangle$. Combined with (OAC), this constraint is meant to capture a notion of uniformity of strategies, where epistemically indistinguishable indices should have the same available actions for the agent to choose upon.
  For each index $\langle m, h \rangle$ and $\alpha \in Ags$, we define α's *ex ante information set* at $\langle m, h \rangle$ as $\pi_\alpha^{\Box}[\langle m, h \rangle] := \{\langle m', h' \rangle; \langle m, h \rangle \sim_\alpha \langle m', h'' \rangle \text{ for some } h'' \in H_{m'}\}$.
- **Value** is a deontic function that assigns to each history $h \in H$ a real number, representing the utility of h.

Intentional epistemic deontic STIT logic: semantics

Definition

- τ is a function that assigns to each $\alpha \in \mathit{Ags}$ and index $\langle m, h\rangle$ a topology $\tau_\alpha^{\langle m,h\rangle}$ on $\pi_\alpha^{\Box}[\langle m, h\rangle]$. This is the *topology of* α*'s intentionality at* $\langle m, h\rangle$, where any open set is interpreted as a p-d intention of α at $\langle m, h\rangle$. Additionally, τ must satisfy the following conditions:
  - (CI) *Consistency of intention*: for every non-empty $U, V \in \tau_\alpha^{\langle m,h\rangle}$, $U \cap V \neq \emptyset$. In other words, every non-empty $U \in \tau_\alpha^{\langle m,h\rangle}$ is $\tau_\alpha^{\langle m,h\rangle}$-dense.
  - (KI) *Knowledge of intention*: for $\alpha \in \mathit{Ags}$ and indices $\langle m, h\rangle$ and $\langle m', h'\rangle$, if $\pi_\alpha^{\Box}[\langle m, h\rangle] = \pi_\alpha^{\Box}\left[\langle m', h'\rangle\right]$, then $\tau_\alpha^{\langle m,h\rangle} = \tau_\alpha^{\langle m',h'\rangle}$. In other words, α has the same topology of p-d intentions at all indices lying within α's current *ex ante* information set.

Epistemic deontic STIT logic: semantics

Definition

A eaubt-frame is extended to a model $\mathcal{M} = \langle M, \sqsubset, \texttt{Choice}, \{\sim_\alpha\}_{\alpha \in Ags}, \texttt{Value}, \mathcal{V}\rangle$ by adding $\mathcal{V} : P \to 2^{T \times H}$.
For a situation $\langle m, h\rangle$,

$$
\begin{array}{lcl}
\langle m,h\rangle \models \Box\varphi & \Leftrightarrow & \forall h' \in H_m, \langle m,h'\rangle \models \varphi \\
\langle m,h\rangle \models [\alpha]\varphi & \Leftrightarrow & \forall h' \in \texttt{Choice}^m_\alpha(h), \\
 & & \langle m,h'\rangle \models \varphi \\
\langle m,h\rangle \models K_\alpha\varphi & \Leftrightarrow & \forall \langle m',h'\rangle \text{ such that} \\
 & & \langle m,h\rangle \sim_\alpha \langle m',h'\rangle, \\
 & & \langle m',h'\rangle \models \varphi \\
\langle m,h\rangle \models \odot[\alpha]\varphi & \Leftrightarrow & \forall L \in \texttt{Optimal}^m_\alpha, \\
 & & h' \in L \text{ implies that} \\
 & & \langle m,h'\rangle \models \varphi \\
\langle m,h\rangle \models \odot_S[\alpha]\varphi & \Leftrightarrow & \forall L \in \texttt{S}-\texttt{optimal}^m_\alpha, \\
 & & \forall m' \text{ such that} \\
 & & m \sim_\alpha m', \langle m',h'\rangle \models \varphi \\
 & & \text{for every } h' \in [L]^{m'}_\alpha \\
\mathcal{M}, \langle m,h\rangle \models I_\alpha\varphi & \Leftrightarrow & \exists U \in \tau^{\langle m,h\rangle}_\alpha \text{ s.t. } U \subseteq \|\varphi\|.
\end{array}
$$

Metalogic Results

Axiomatization

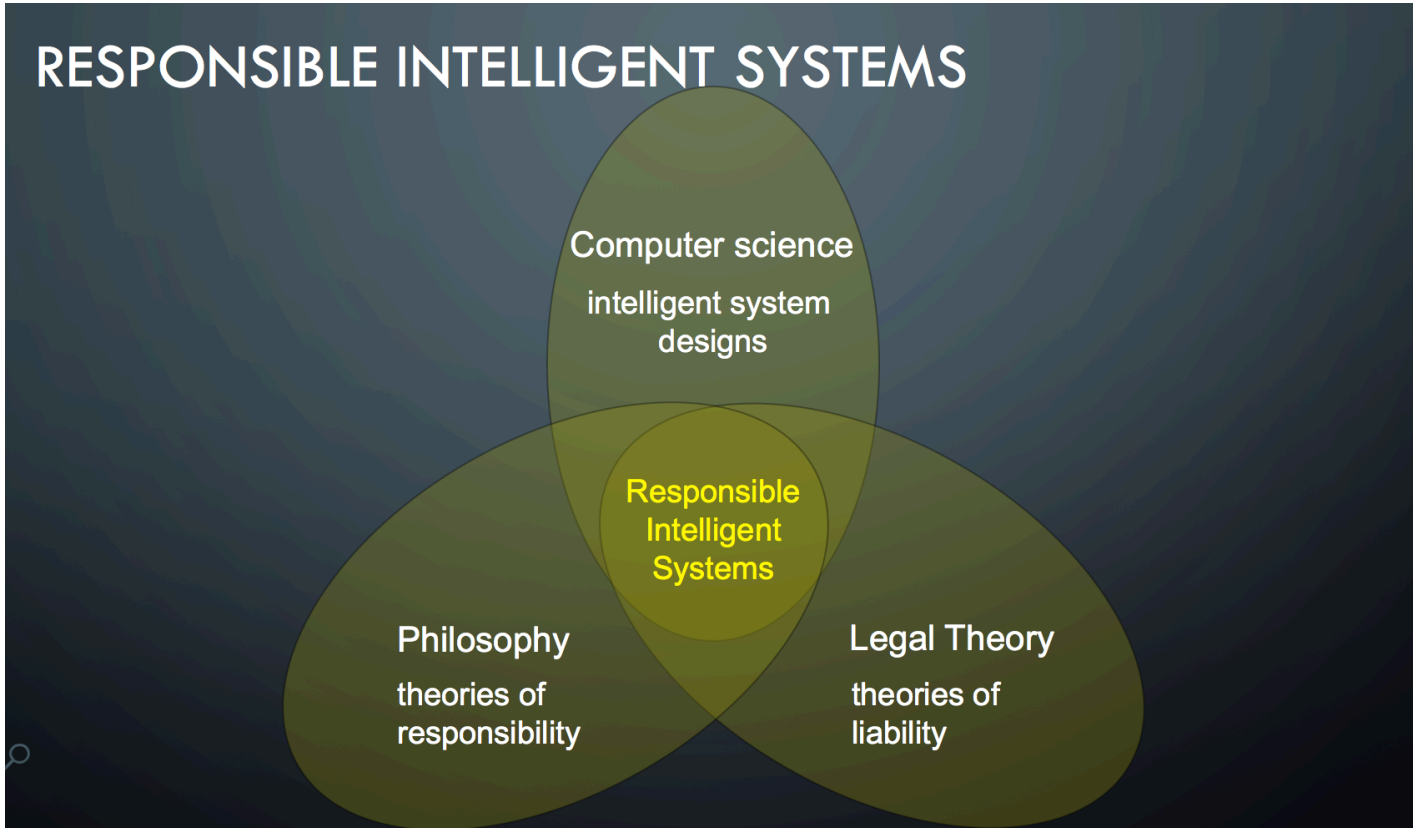
Metalogic Results

Axiomatization

- Soundness.
- Completeness.
- Decidability.

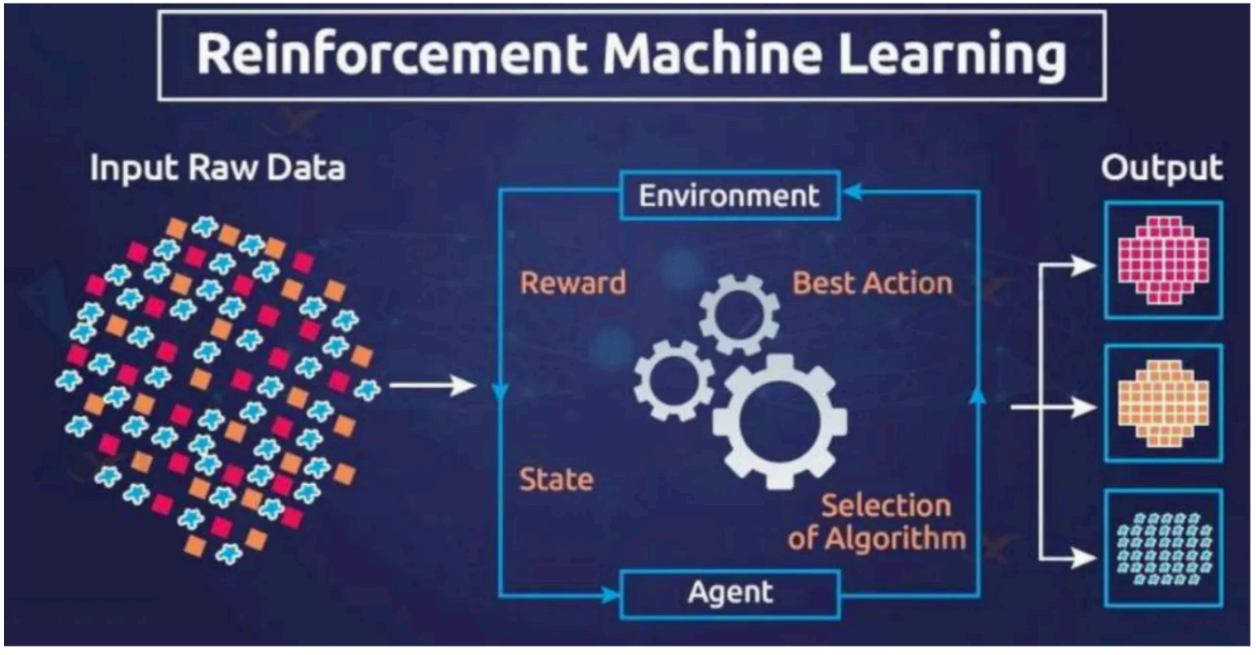
Oh, nice... but what about AI?!

Well...


RESPONSIBLE INTELLIGENT SYSTEMS
Computer science
intelligent system designs
Responsible Intelligent Systems
Philosophy
theories of responsibility
Legal Theory
theories of liability

Reinforcement Learning

Reinforcement Learning



Reinforcement Learning

Reinforcement Learning

Reinforcement Learning is a type of machine learning where an agent learns to make decisions by interacting with an environment. The agent gets rewards or penalties based on its actions and learns to maximize the total reward over time.

Reinforcement Learning: Models

The typical models for reinforcement learning (RL) are called **Markov Decision Processes** (MDPs):

- A set of environment and agent states (the **state space**), $\mathcal{S}$.
- A set of actions (the **action space**), $\mathcal{A}$, available to the agent.
- A transition probability function:

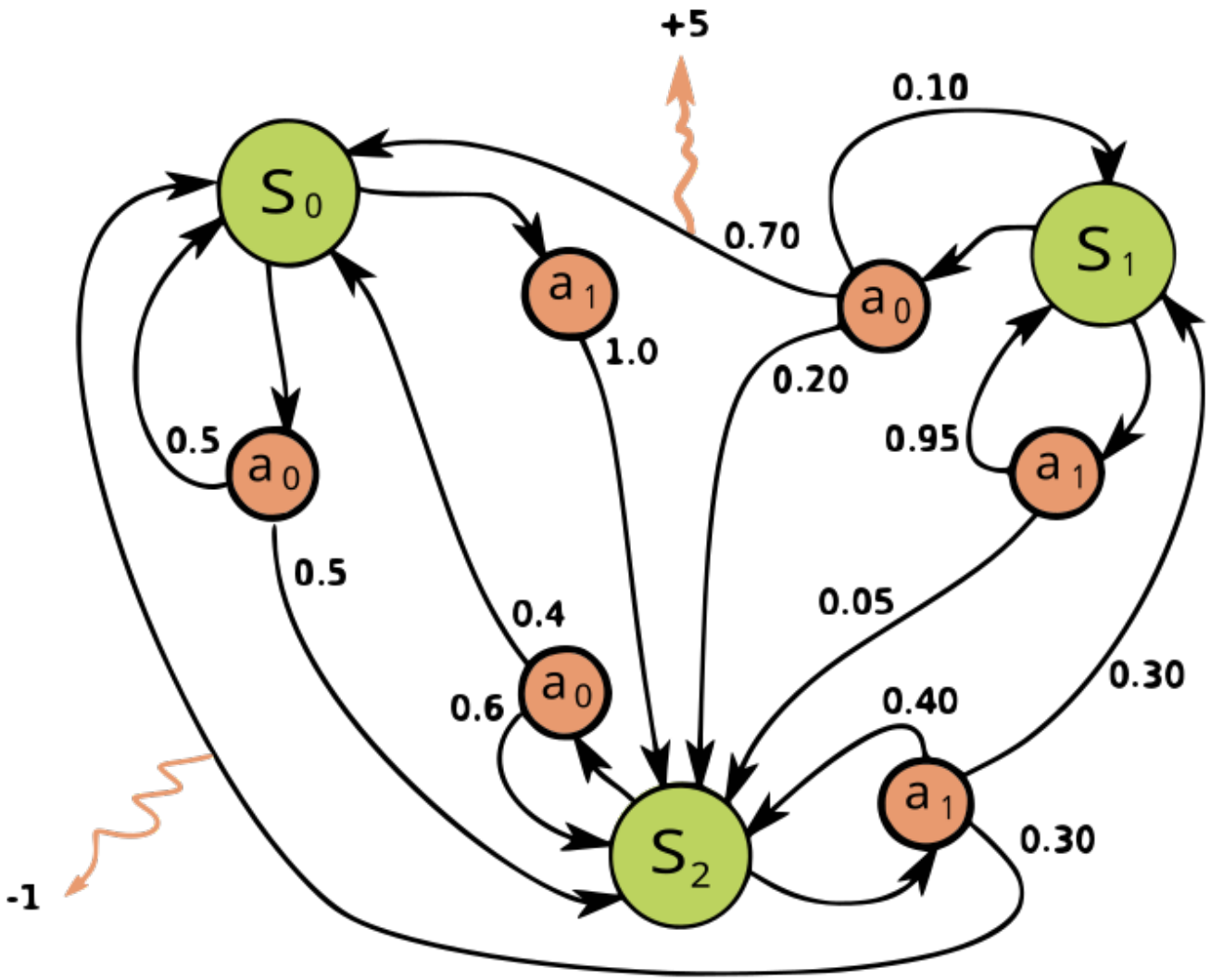
$$P_a(s, s') = \Pr(S_{t+1} = s' \mid S_t = s, A_t = a),$$

  which defines the probability of transitioning from state s to state s' under action a at time t.
- A reward function:

$$R_a(s, s'),$$

  which represents the immediate reward received after transitioning from state s to state s' under action a.

Reinforcement Learning: Models



Reinforcement Learning

The purpose of reinforcement learning is for the agent to learn an optimal (or near-optimal) policy π that maximizes the reward function or other user-provided reinforcement signals accumulated from immediate rewards.

Reinforcement Learning

A basic reinforcement learning agent interacts with its environment in discrete time steps. At each time step t:

- The agent receives the current state S_t and reward R_t.
- It chooses an action A_t from the set of available actions.
- The environment responds by transitioning to a new state S_{t+1}.
- The environment also returns the next reward R_{t+1} associated with the transition (S_t, A_t, S_{t+1}).

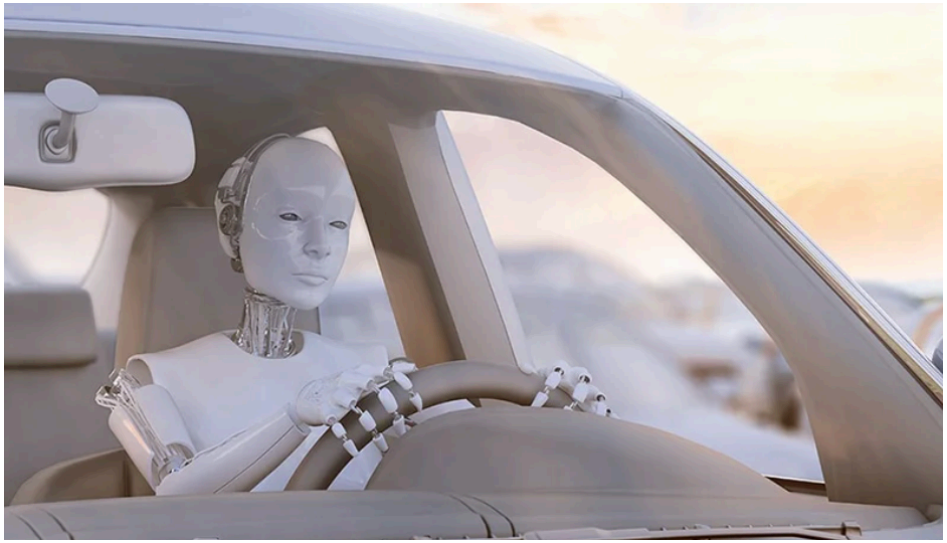
The agent's objective is to learn a policy:

$$\pi : \mathcal{S} \times \mathcal{A} \to [0, 1], \quad \pi(s, a) = \Pr(A_t = a \mid S_t = s),$$

which defines the probability of taking action a when in state s.

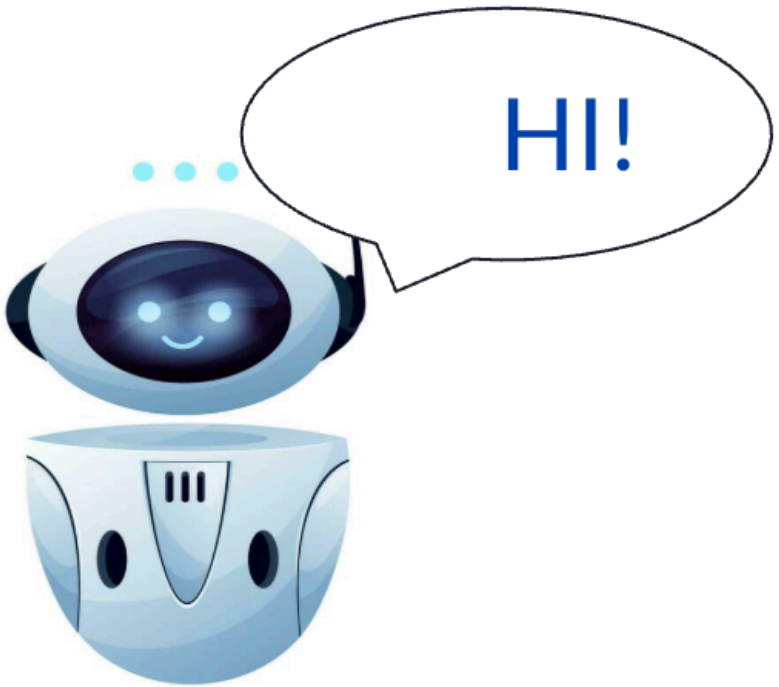
Reinforcement Learning Applications in AI

- **Robotics:** Deep reinforcement learning (DRL) aids in robot decision-making when operating in unpredictable environments. While simple tasks are straightforward, complex behaviors like driving or mimicking humans require learning from dynamic sensory input.



Reinforcement Learning Applications in AI

- **Natural Language Processing:** DRL is also used in chatbots, where it outperforms other methods in generating effective, context-aware responses.



Connection

There is a special connection between (act-utilitarian) **Stit Theory** and **Reinforcement Learning**!

Connection

Expected Act Utilitarian PCTL Stit Theory uses Probabilistic Computation Tree Logic (PCTL) to describe states of affairs in the world, and adds modalities to speak of action and obligation. Letting φ be a PCTL formula and α an agent, the syntax of this logic is defined by the following grammar,

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid [\alpha]\varphi \mid \odot_\alpha \varphi$$

Intuitively, a PCTL formula φ describes a state of affairs, such as $P_{\geq 0.9}\Diamond g$: the probability of eventually g occurring is at least 0.9.

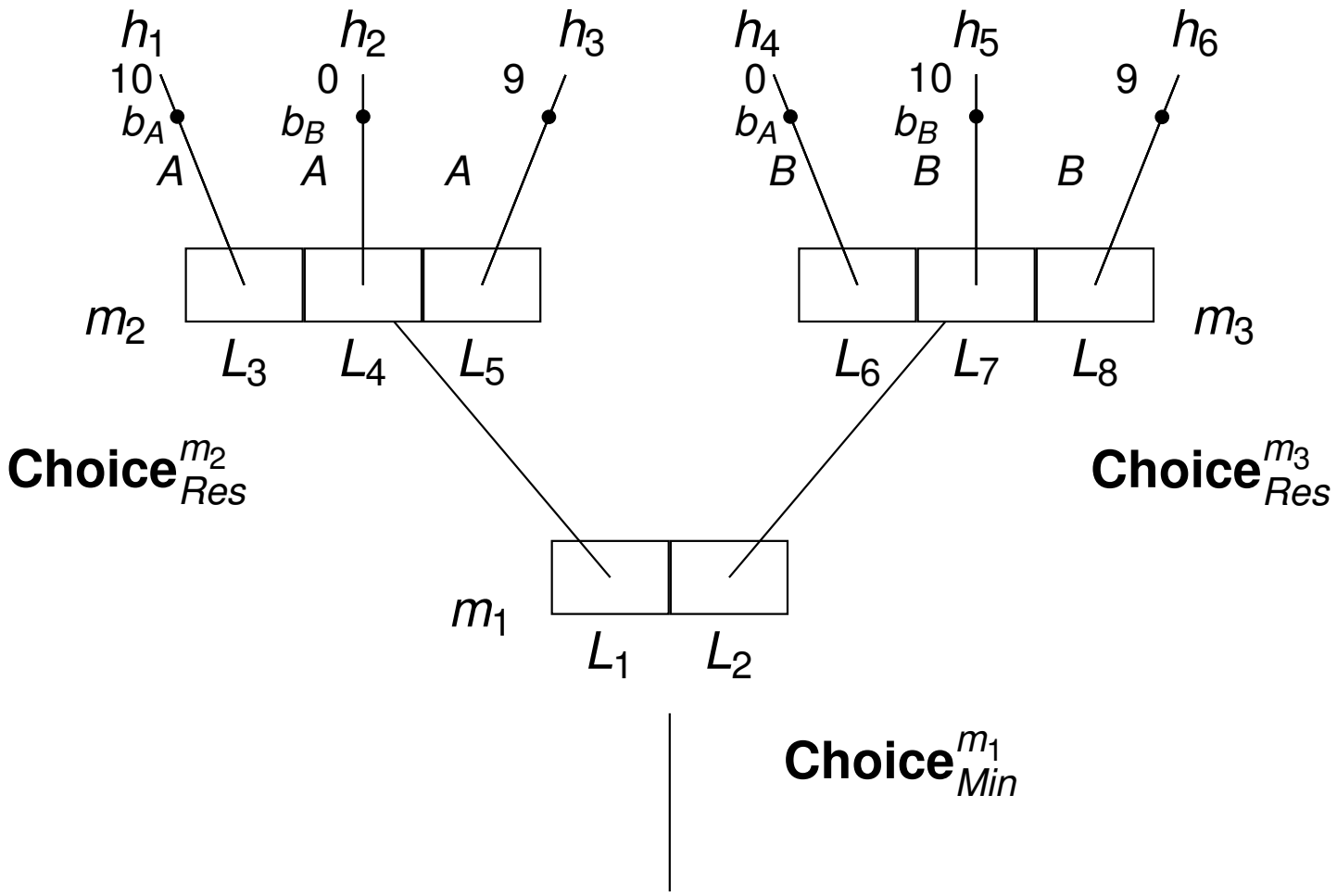


Figure: EAU-PCTL Stit Model

Connection

They're the models we had used! But with **probabilities**:

- $Pr_\alpha(m'|m)$: the probability of agent α moving from m to m', assuming that the agent takes some action K that leads to m' (formally, $K \subseteq H_m$ and $K \cap H_{m'} \neq \emptyset$).

Connection

The **quality** of an action, $Q(K)$, is defined as:

$$Q(K) = \sum_{m' \in M_K} Pr_a(m'|m) \max_{K' \in Choice^a_{m'}} Q(K')$$

where M_K is the set of moments that follow m by taking action K.

An agent's set of **optimal** actions, then, can be defined as the action(s) with the best quality at the moment:

$$\textit{E-Optimal}^m_\alpha := \left\{ K \in Choice^m_\alpha \,\middle|\, \nexists K' \in Choice^m_\alpha \text{ such that } Q(K) < Q(K') \right\}$$

Connection

Definition (Expected Ought)

With α an agent and A an obligation in a model $\mathcal{M}$,

$$\mathcal{M}, m/h \models \odot_\alpha \varphi \iff K \subseteq |\varphi|^{\mathcal{M}}_m \quad \text{for all } K \in \textit{E-Optimal}^m_\alpha$$

Connection

An EAU-PCTL model can be seen as the **roll-out** of an MDP, where moments are states, and probabilities are derived from **transition probabilities**.

(Shea-Blymyer and Abbas, 2022) prove a correspondence result between EAU-PCTL stit models and MDPs, such that the Q-function derived for the MDP is isomorphic to the aforementioned Q.

HOORAY! We can do Model Checking with respect to (strategic) obligations

Connection

(Shea-Blymyer and Abbas, 2024) prove the following points:

- ***At design time, we can specify and verify whether the optimal policy (with respect to Q) of the RL agent complies with an obligation.***

Connection

(Shea-Blymyer and Abbas, 2024) prove the following points:

- ***At design time, we can specify and verify whether the optimal policy (with respect to Q) of the RL agent complies with an obligation.***
- ***At design time, we can update any policy so that it complies with an obligation and maintains a high enough reward.***

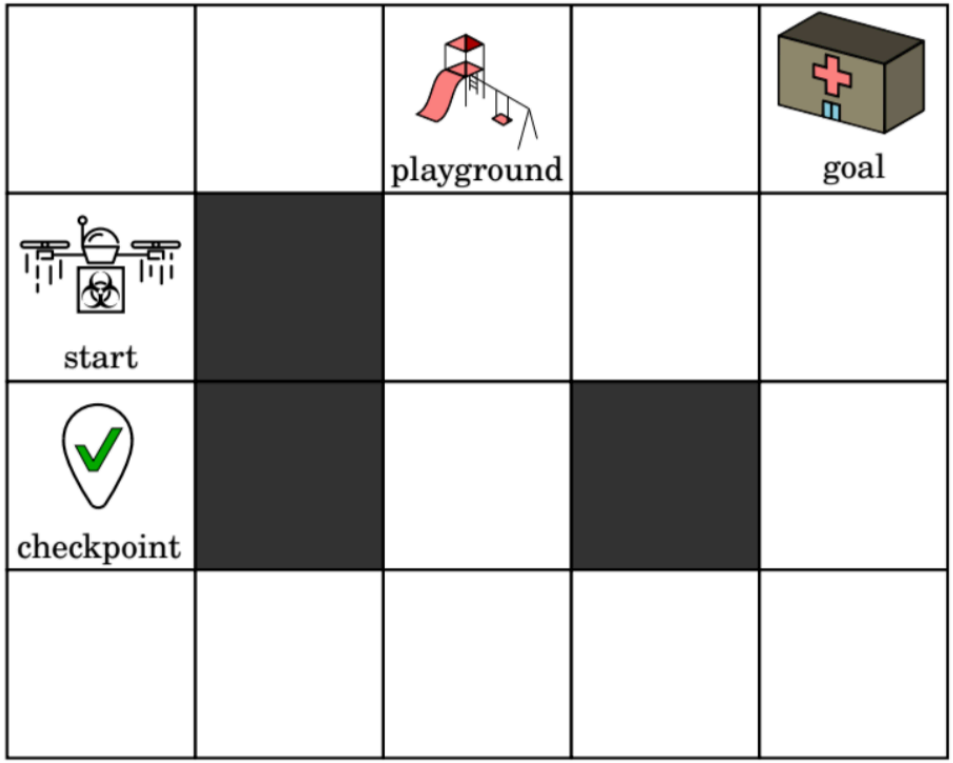
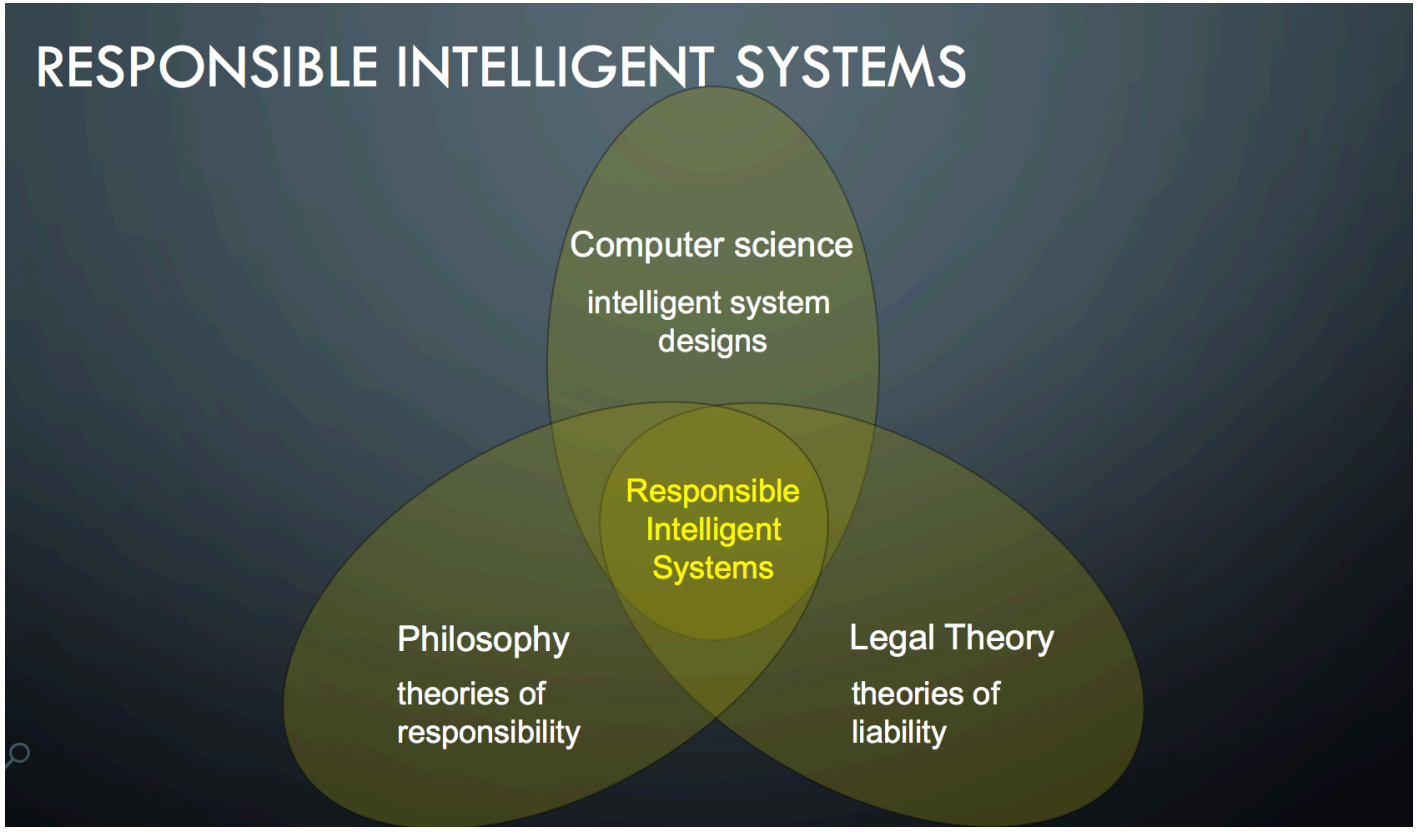


Figure 1: The windy-drone MDP. Darkened cells are inaccessible states. The goal state is an absorbing state. An agent in this MDP has 4 actions available to it in any state: up, down, left, and right. A chosen action has a 70% chance of success, and on a failure the agent "slips" in one of the unchosen directions. An action result that would move the agent into a wall, or other inaccessible state, leaves the agent in the state it acted from.

This is GREAT NEWS!



References

- Alur, R., Henzinger, T. A., Mang, F.Y., Qadeer, S., Rajamani, S. K., & Tasiran, S. (1998). Mocha: Modularity in model checking. In *International Conference on Computer Aided Verification* (pp. 521–525).
- Arkoudas, K., Bringsjord, S., & Bello, P. (2005). Toward ethical robots via mechanized deontic logic. In *AAAI Fall Symposium on Machine Ethics* (pp. 17–23).
- Broersen, J. (2014). Responsible intelligent systems. *KI-Künstliche Intelligenz*, 28(3), 209–214.
- Calegari, R., Ciatto, G., Denti, E., & Omicini, A. (2020). Logic-based technologies for intelligent systems: State of the art and perspectives. *Information*, 11(3), 167.
- Horty, J. F. (2001). *Agency and deontic logic*. Oxford University Press.

References

- Gunning, D. (2017). Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency* (DARPA), 2(2).
- Kwiatkowska, M., Norman, G., & Parker, D. (2002, April). Prism 4.0: verification of probabilistic real-time systems. In *Proceedings of 23rd international conference on computer aided verification (CAV 11)* (pp. 585–591), Lecture Notes in Computer Science, vol 6806. Springer, Berlin.
- Shea-Blymyer, Colin, and Houssam Abbas. Formal Ethical Obligations in Reinforcement Learning Agents: Verification and Policy Updates. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* Vol. 7. 2024.
- Shea-Blymyer, Colin, and Houssam Abbas. Generating Deontic Obligations From Utility-Maximizing Systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* 2022.
- Urban, C., & Miné, A. (2021). A review of formal methods applied to machine learning. *arXiv preprint arXiv:2104.02466.*