AI in a vat:
Limits of efficient world modelling for agent sandboxing and interpretability

Fernando E. Rosas, PhD

Department of Informatics, University of Sussex

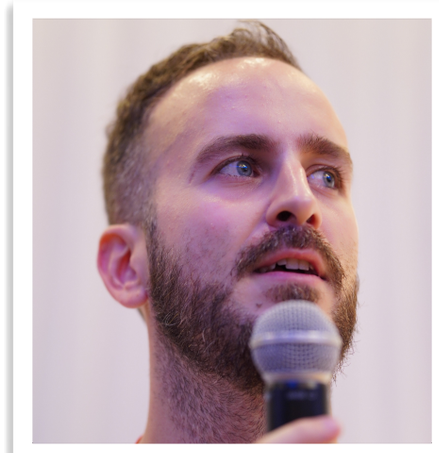
Department of Brain Sciences and Centre for Complexity Science, Imperial College London

Centre for Eudaimonia and Human Flourishing, University of Oxford

Work developed in synergistic collaboration with :

Alexander Boyd

(University of Sussex, BITS)



Manuel Baltieri

(Araya)

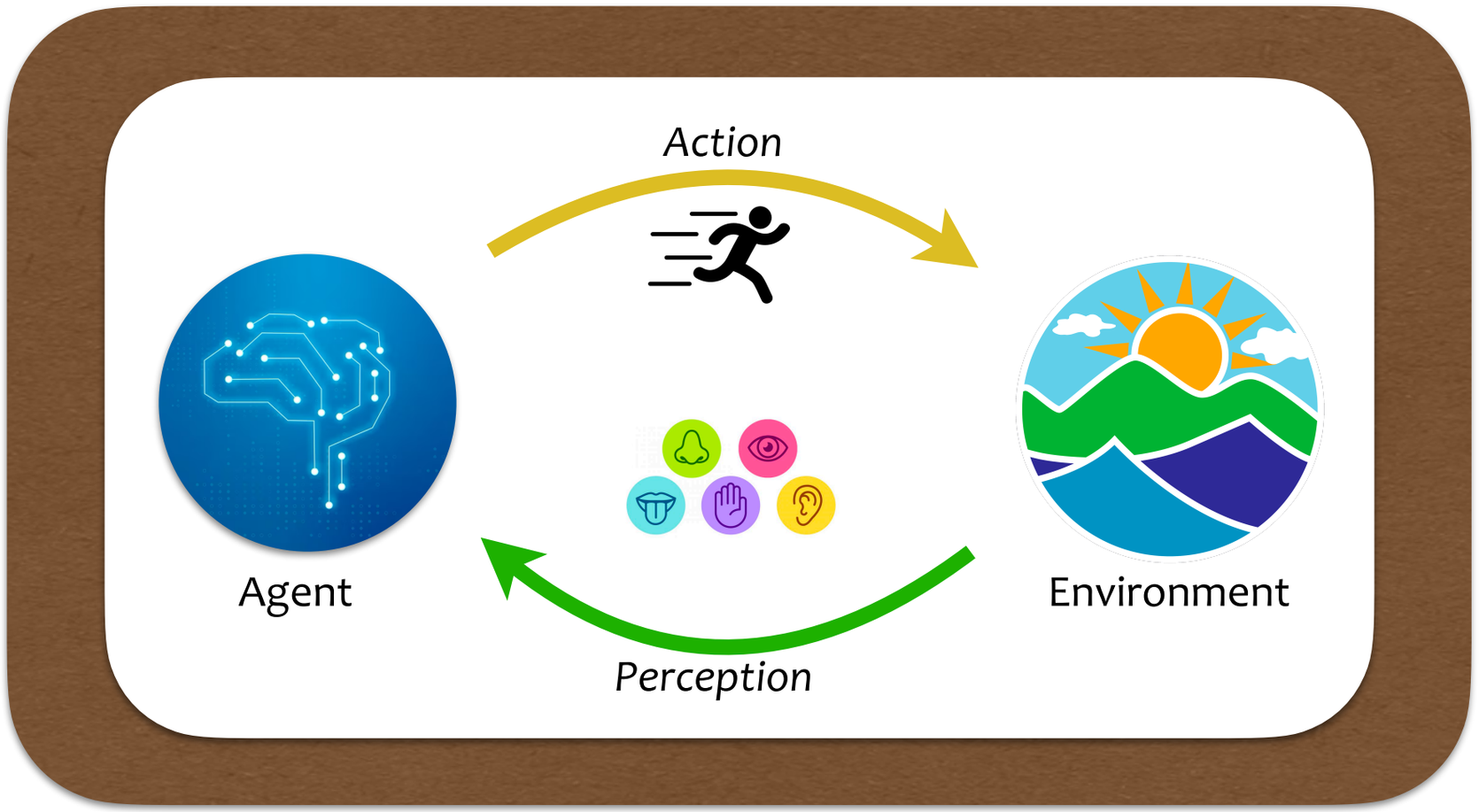
This work has been supported funded by the ARIA's *Safeguraded AI* programme and by the PIBBSS affiliateship program.

Contents

AI sandboxing

Scenario:

- A given AI system needs testing before deployment
- There is a model of the world available for sandboxing
- The model allows to estimate probabilities



AI sandboxing

Scenario:

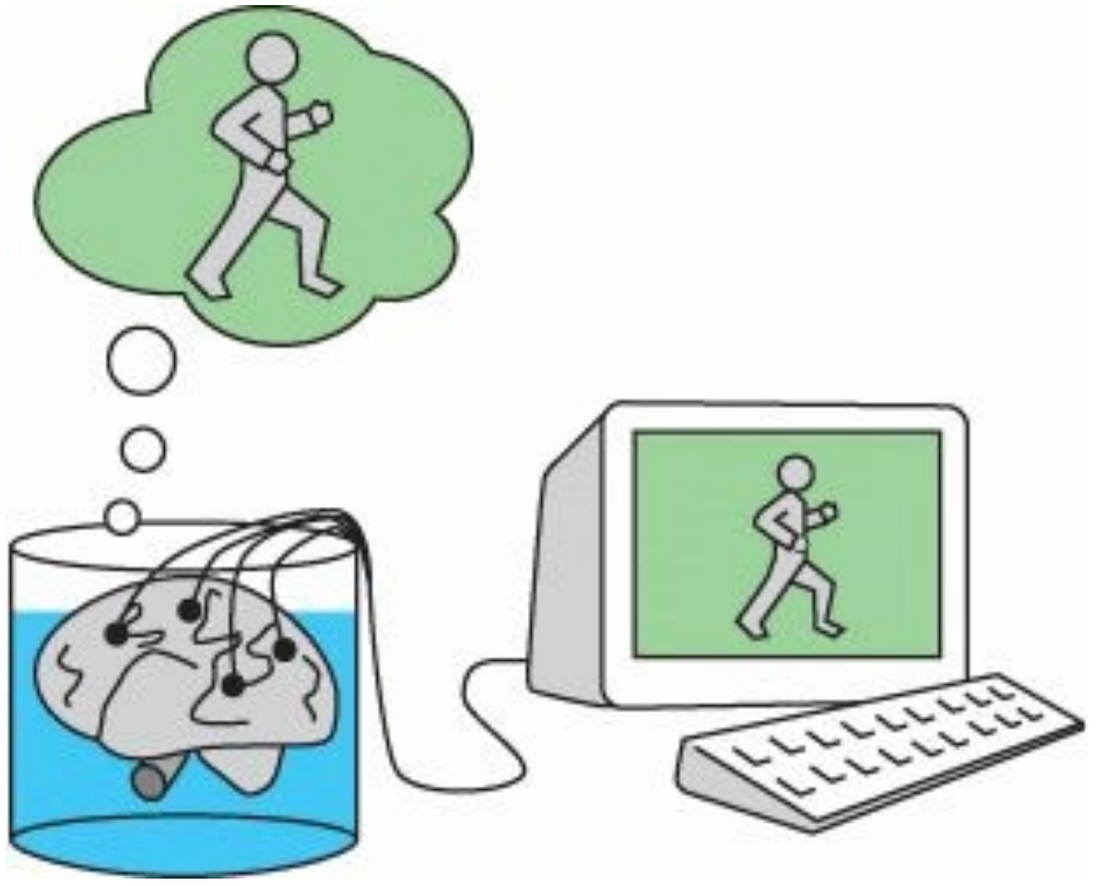
- A given AI system needs testing before deployment
- There is a model of the world available for sandboxing
- The model allows to estimate probabilities

Reasons to do better:

- The model may be too detailed (e.g. a physical model of the whole planet)
- Agent may not “see” some resolution (e.g. classical agents cannot see quantum)
- The agent may only interact with a corner of the world (e.g. a bacteria)

Thought experiment: Brain in a Vat

How can one distinguish between reality and a simulation?



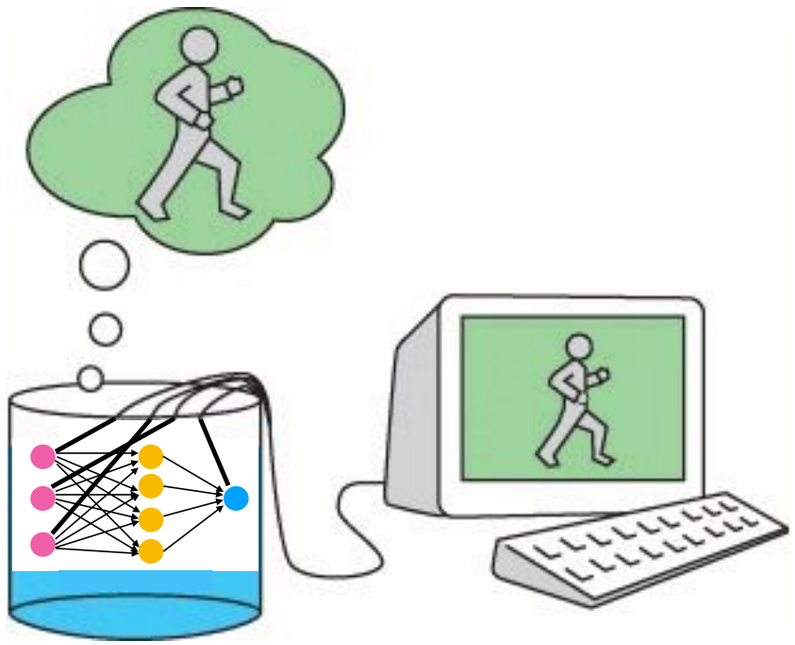
- Plato's allegory of the cave, *The Republic*, (~375 BC)
- Zhuangzi's dream of being a butterfly (circa 476-221 BC)
- Decartes' evil demon, *Meditations* (1641)
- Hilary Putnam, *Brains in a vat* (1981)
- Etc etc…

Thought experiment: Brain in a Vat

How can one distinguish between reality and a simulation?

Key insight ("AI in a vat" approach):

—> Forget about the world, and focus on how inputs turn into outputs (the interface)

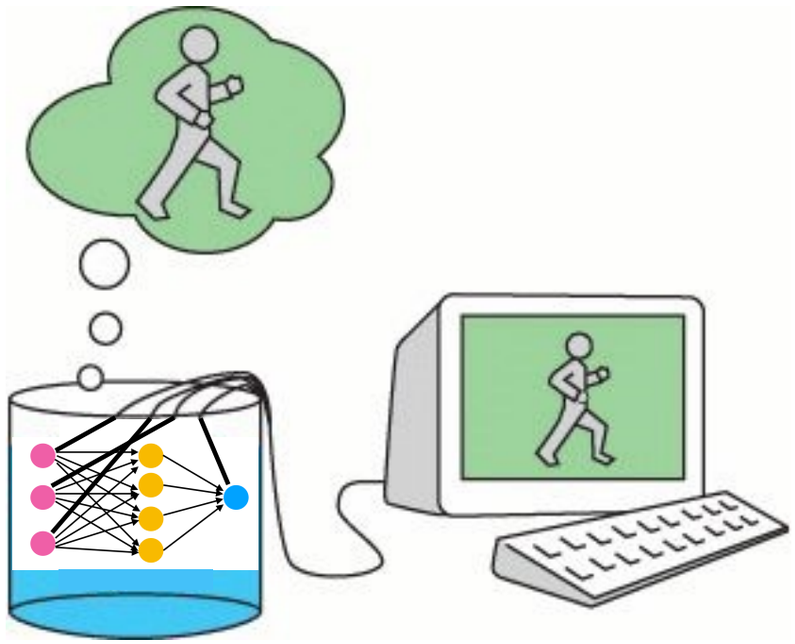


Thought experiment: Brain in a Vat

How can one distinguish between reality and a simulation?

Key insight ("AI in a vat" approach):

—> Forget about the world, and focus on how inputs turn into outputs (the interface)



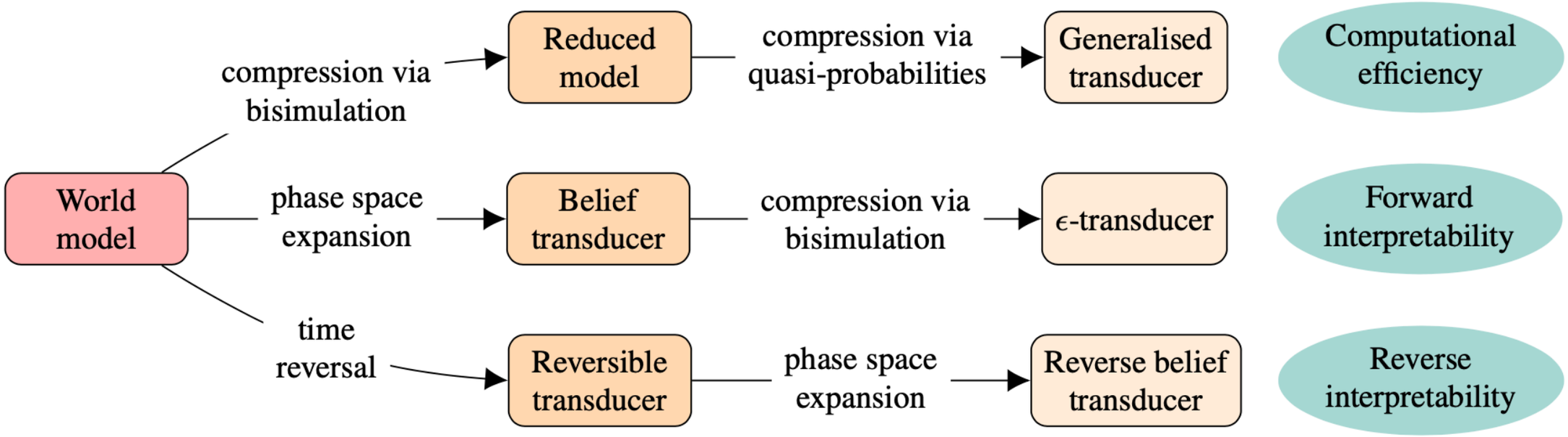
How can we best simulate an interface? (*computational goal*)

Can we characterise what can be learned through it? (*epistemic goal*)

Summary of results: possible world models

World models can be optimised for two purposes:

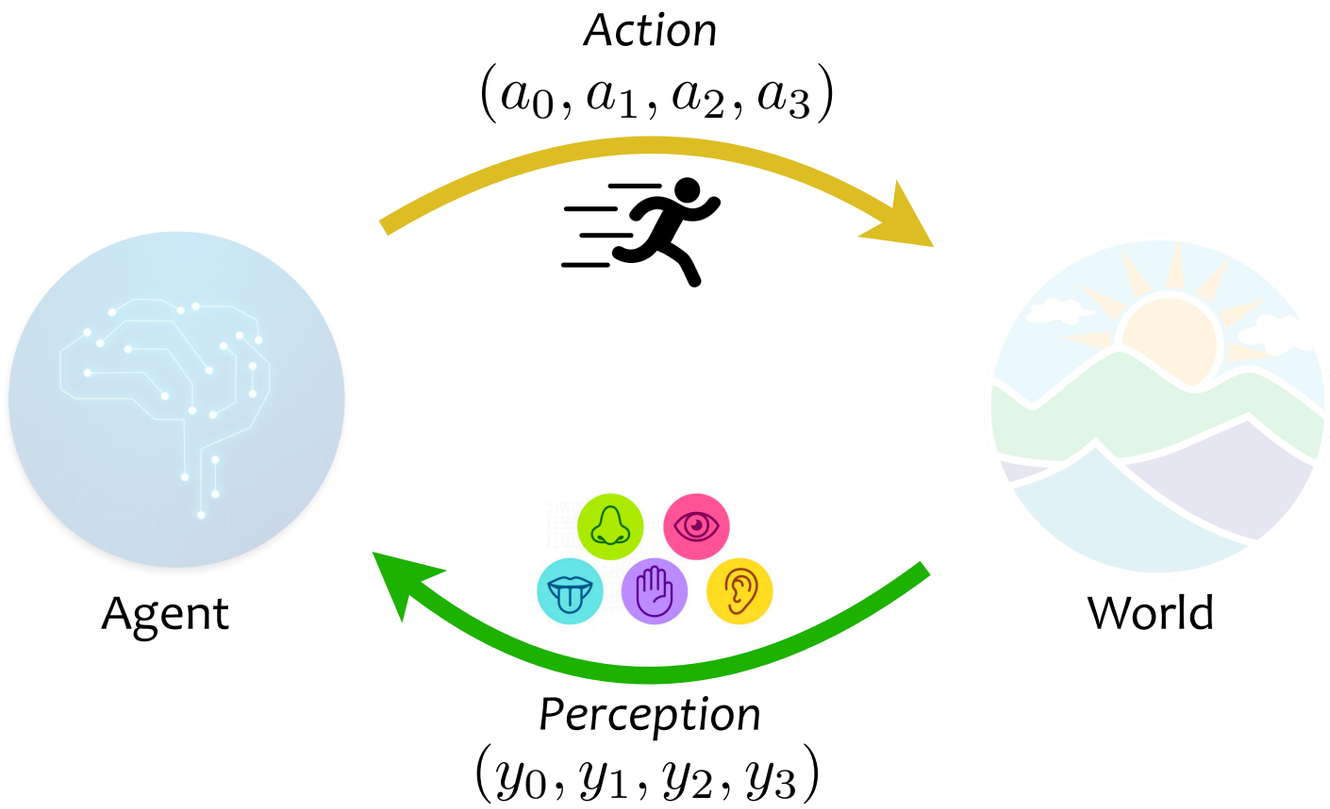
- Computational efficiency
- Interpretability: forward and backward



Contents

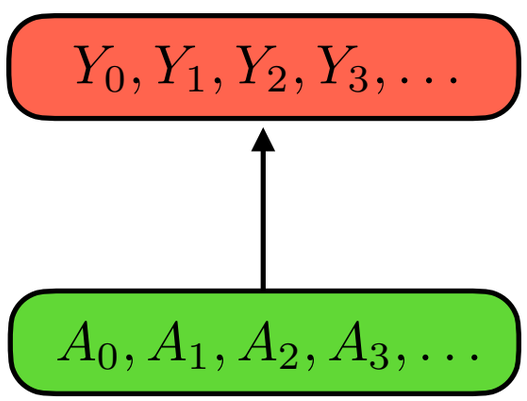
Formal foundations

The **interface of an agent** $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ is a collection of conditional distributions $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:})$ turning sequences of actions into outcomes, satisfying $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t})$.



Formal foundations

The **interface of an agent** $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ is a collection of conditional distributions $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:})$ turning sequences of actions into outcomes, satisfying $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t})$.

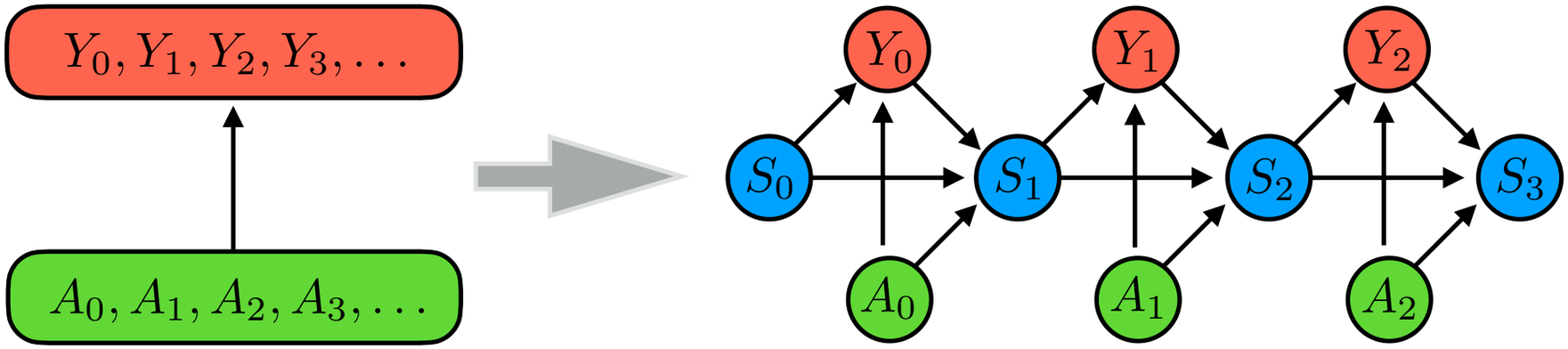


Formal foundations

The **interface of an agent** $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ is a collection of conditional distributions $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:})$ turning sequences of actions into outcomes, satisfying $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t})$.

A **transducer** for a given interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ is another collection of distributions $p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:})$ corresponding to an auxiliary process satisfying:

$$p(\boldsymbol{y}_{:t}\boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = p(s_0)\prod_{\tau=0}^{t}\kappa_\tau(y_\tau, s_{\tau+1}|a_\tau, s_t)$$



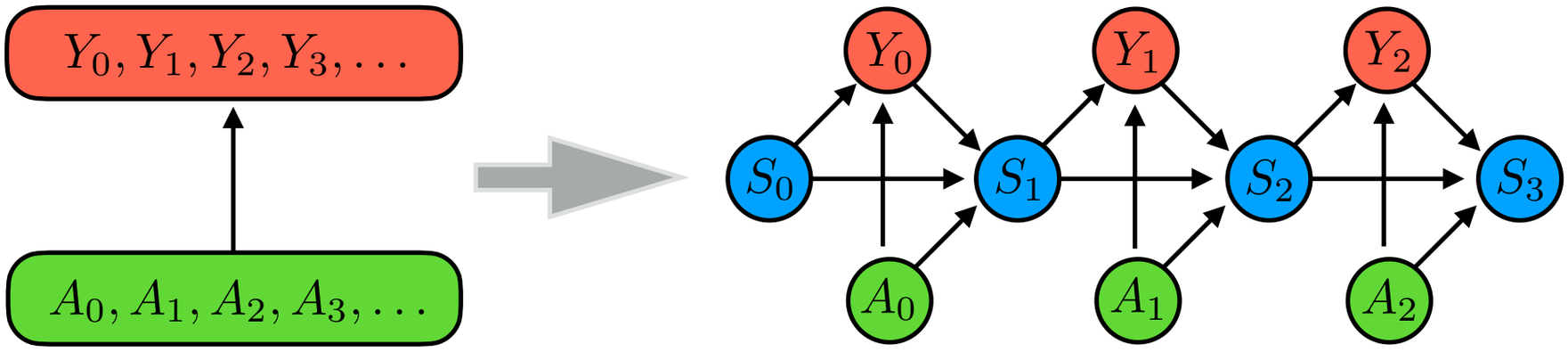
$$h_t := (y_t, a_t)$$

Formal foundations

The **interface of an agent** $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ is a collection of conditional distributions $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:})$ turning sequences of actions into outcomes, satisfying $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t})$.

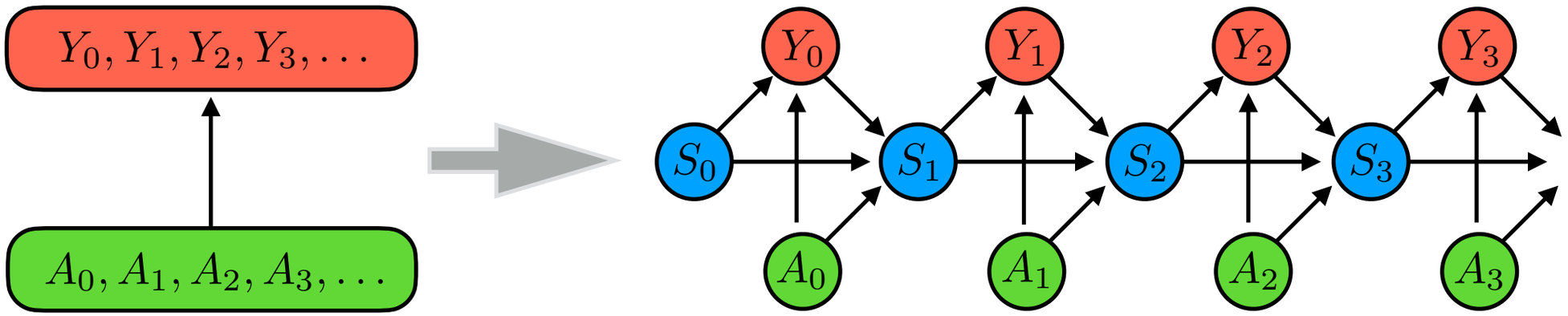
A **transducer** for a given interface $\mathcal{I}(\boldsymbol{Y}|\boldsymbol{A})$ is another collection of distributions $p(\boldsymbol{s}_{:t}|\boldsymbol{h}_{:})$ corresponding to an auxiliary process satisfying:

$$p(\boldsymbol{y}_{:t}\boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = p(s_0)\prod_{\tau=0}^{t}\kappa_\tau(y_\tau, s_{\tau+1}|a_\tau, s_t)$$



There are no assumptions about how smart the agent is, its architecture, or policy!

Properties of transducers

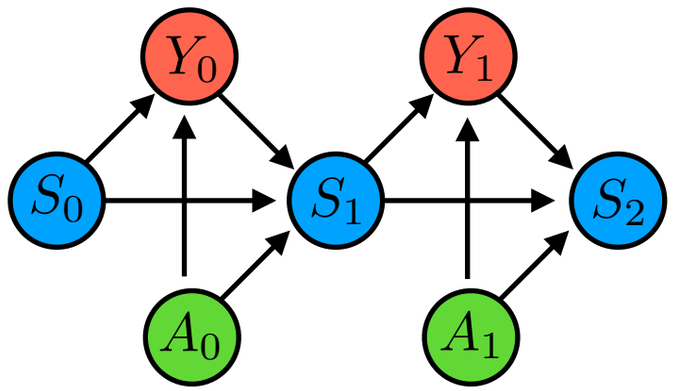


Transducers capture the memory of interfaces:

- Trivial transducers $S_t = 0$ <—> Memoryless interfaces $p(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:}) = \prod_{\tau=0}^{t} p(y_\tau|a_\tau)$
- Fully observability $S_t = Y_t$ <—> Markov interfaces $p(y_{t+1}|\boldsymbol{y}_{:t}, \boldsymbol{a}_{:}) = p(y_{t+1}|y_t, a_t)$

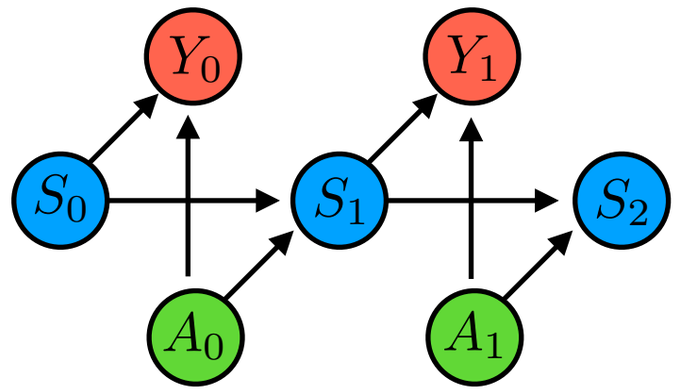
Different types of transducers

Full transducer



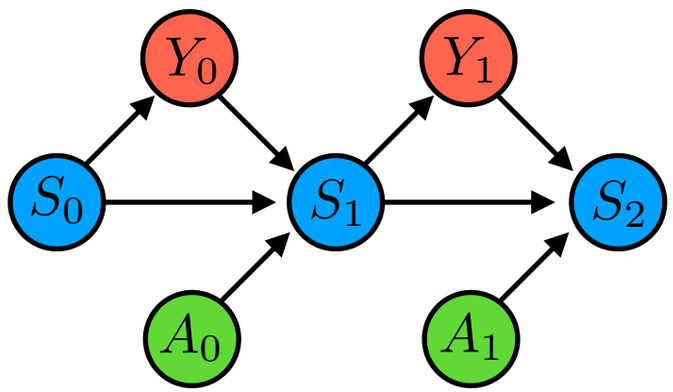
$\kappa_\tau(y, \tilde{s}|a, s)$

POMDP — action first



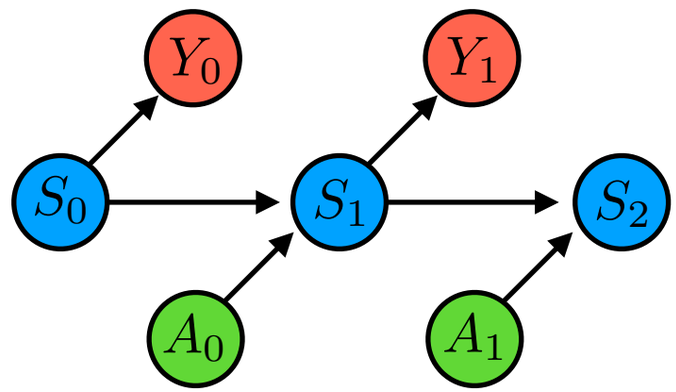
$\kappa_\tau(y, \tilde{s}|a, s) = \mu_\tau(y|a, s)\nu_\tau(\tilde{s}|a, s)$

Belief transducer



$\kappa_\tau(y, \tilde{s}|a, s) = \mu_\tau(y|s)\nu_\tau(\tilde{s}|y, a, s)$

POMDP — observation first

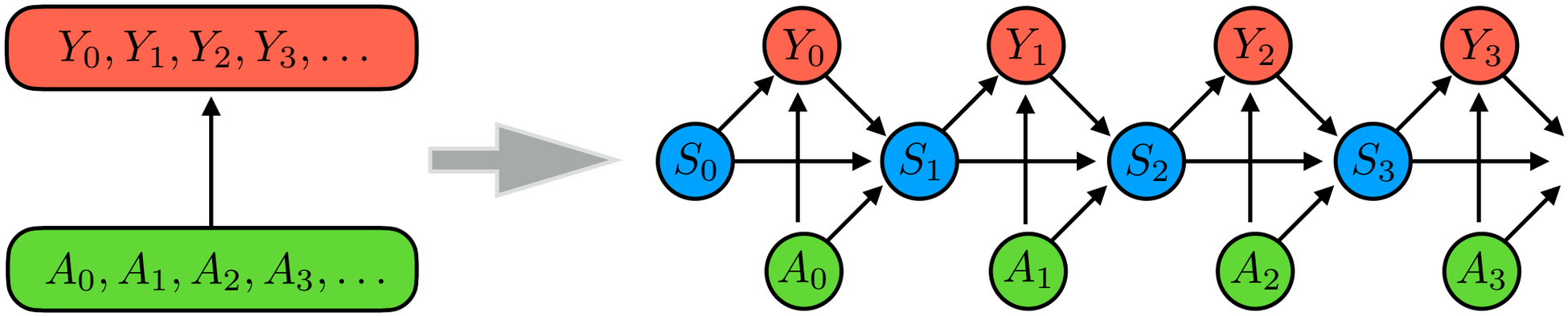


$\kappa_\tau(y, \tilde{s}|a, s) = \mu_\tau(y|s)\nu_\tau(\tilde{s}|a, s)$

Contents

Reducing transducers

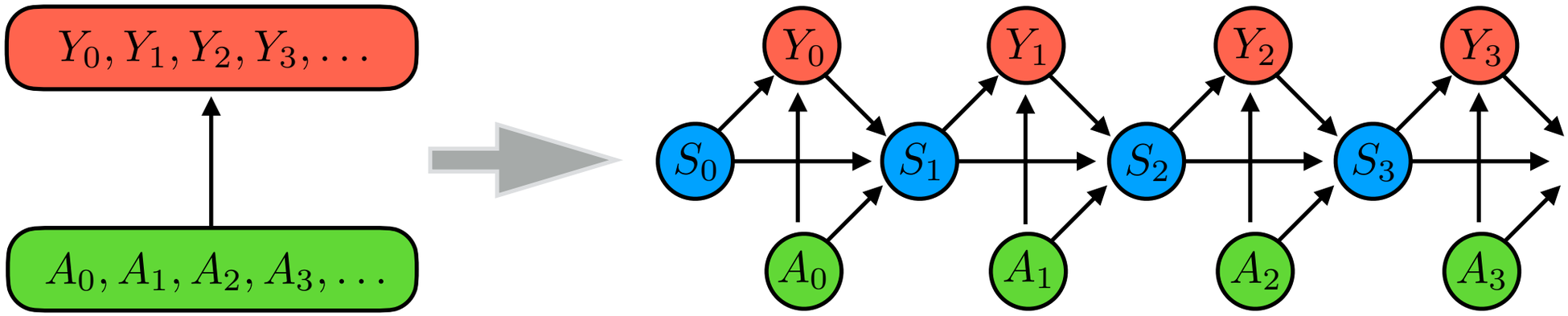
Can all interfaces be simulated via a transducer?



$$p(\boldsymbol{y}_{:t}\boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = p(s_0)\prod_{\tau=0}^{t}\kappa_\tau(y_\tau, s_{\tau+1}|a_\tau, s_t)$$

Reducing transducers

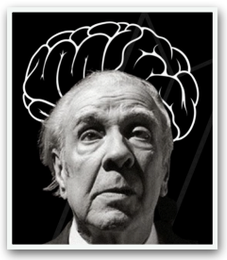
Can all interfaces be simulated via a transducer?



$$p(\boldsymbol{y}_{:t}\boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = p(s_0)\prod_{\tau=0}^{t}\kappa_\tau(y_\tau, s_{\tau+1}|a_\tau, s_t)$$

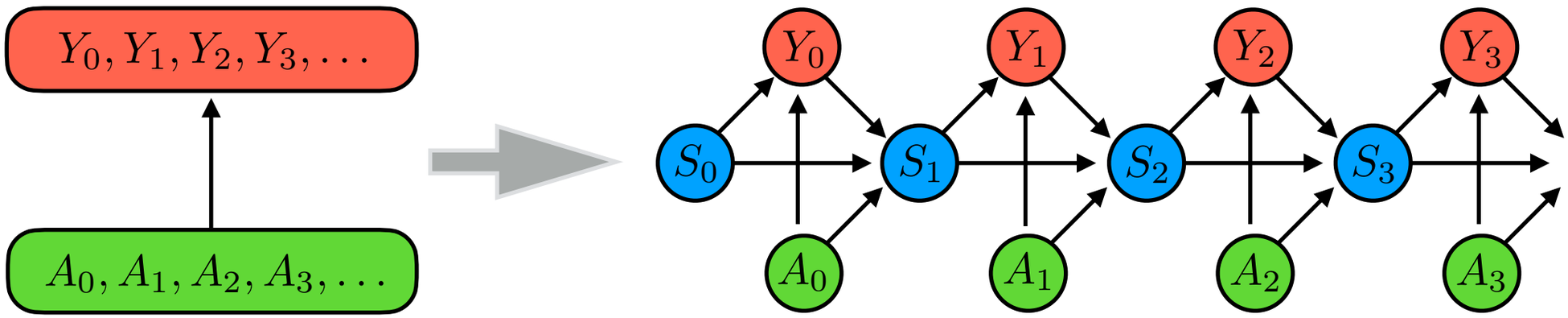
Yes! Consider the transducer "Funes" with world model $S_t = \boldsymbol{H}_{:t-1}$.

$$S_0 = 0, \quad S_1 = (A_0, Y_0), \quad S_2 = (\boldsymbol{A}_{:1}, \boldsymbol{Y}_{:1}), \ldots$$



Reducing transducers

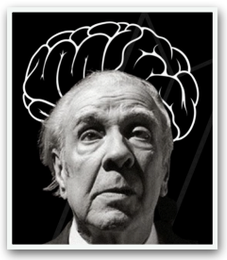
Can all interfaces be simulated via a transducer?



$$p(\boldsymbol{y}_{:t}\boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = p(s_0)\prod_{\tau=0}^{t}\kappa_\tau(y_\tau, s_{\tau+1}|a_\tau, s_t)$$

Yes! Consider the transducer "Funes" with world model $S_t = \boldsymbol{H}_{:t-1}$.

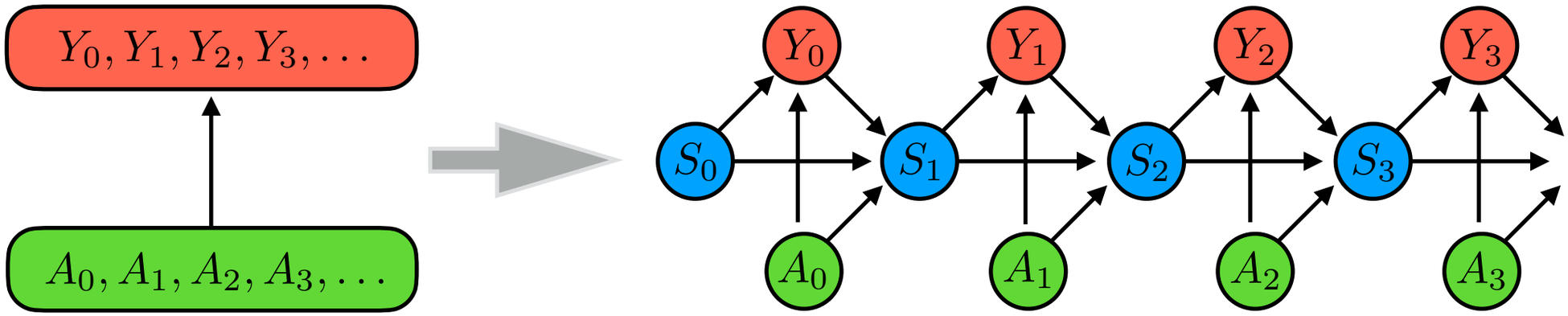
$$S_0 = 0, \quad S_1 = (A_0, Y_0), \quad S_2 = (\boldsymbol{A}_{:1}, \boldsymbol{Y}_{:1}), \ldots$$



—> It works, but requires exponential amounts of memory

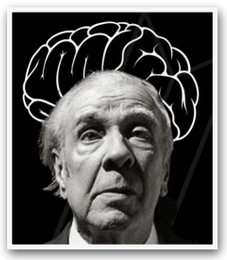
Reducing transducers

Can all interfaces be simulated via a transducer?



$$p(\boldsymbol{y}_{:t}\boldsymbol{s}_{:t+1}|\boldsymbol{a}_{:}) = p(s_0)\prod_{\tau=0}^{t}\kappa_\tau(y_\tau, s_{\tau+1}|a_\tau, s_t)$$

Yes! Consider the transducer "Funes" with world model $S_t = \boldsymbol{H}_{:t-1}$.



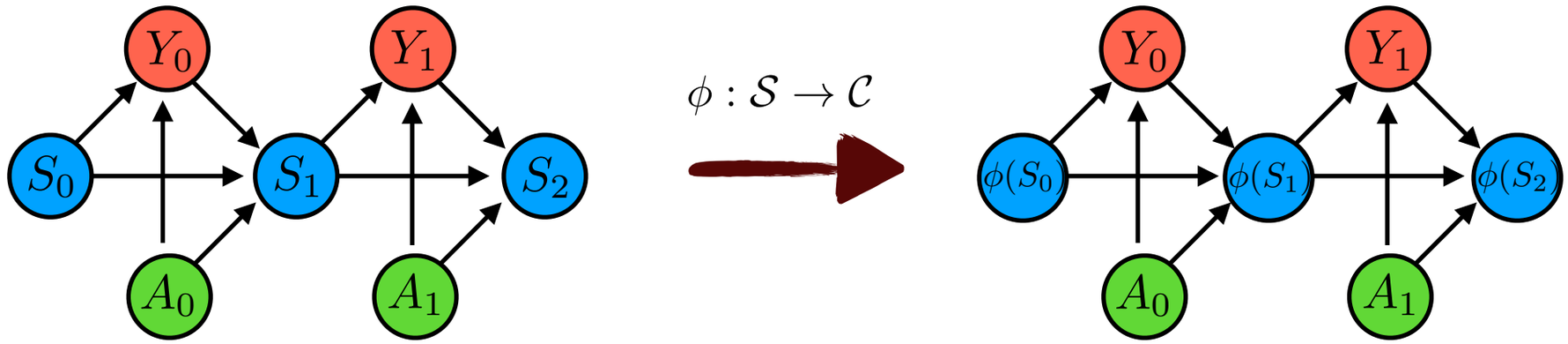
$$S_0 = 0, \quad S_1 = (A_0, Y_0), \quad S_2 = (\boldsymbol{A}_{:1}, \boldsymbol{Y}_{:1}), \ldots$$

—> It works, but requires exponential amounts of memory

How can one "**reduce**" a transducer?

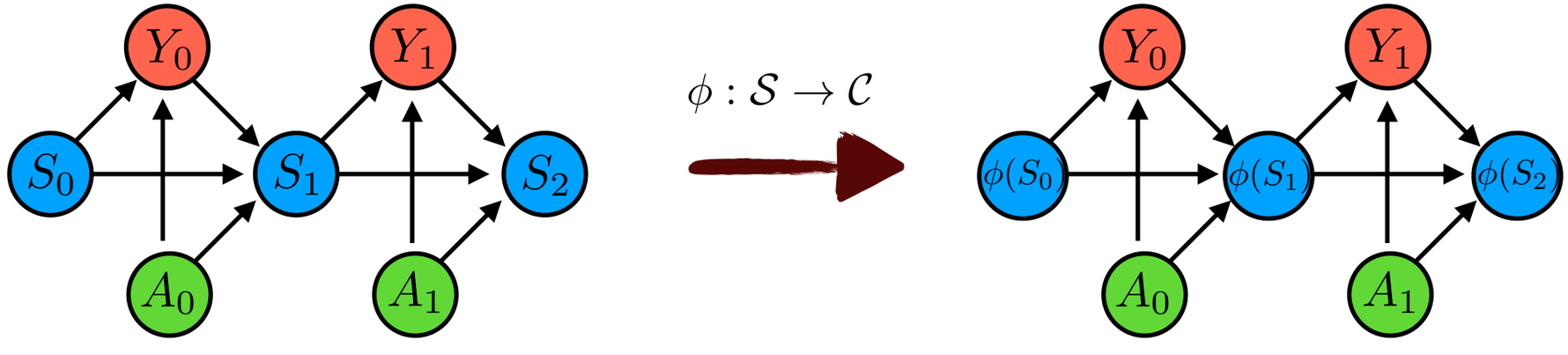
Reducing transducers: tools

World reductions: coarse-grains that keep the interface stays the same (equivalent to **bisimulation**).

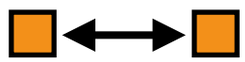


Reducing transducers: tools

World reductions: coarse-grains that keep the interface stays the same (equivalent to **bisimulation**).

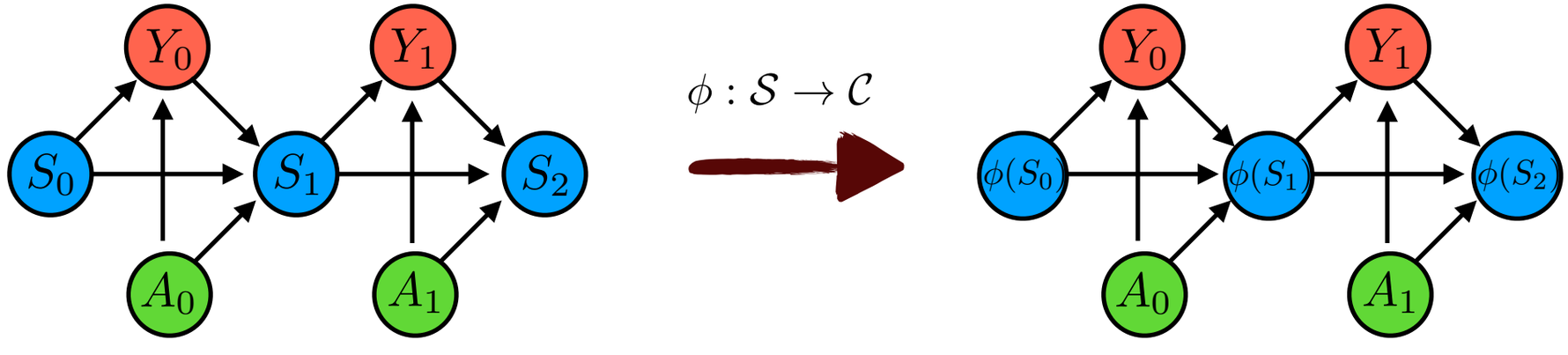


Isomorphic worlds: it is possible to reduce one into each other.



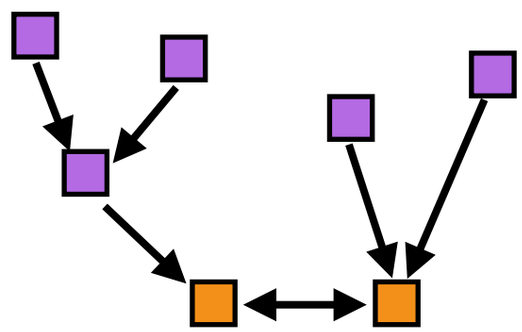
Reducing transducers: tools

World reductions: coarse-grains that keep the interface stays the same (equivalent to **bisimulation**).



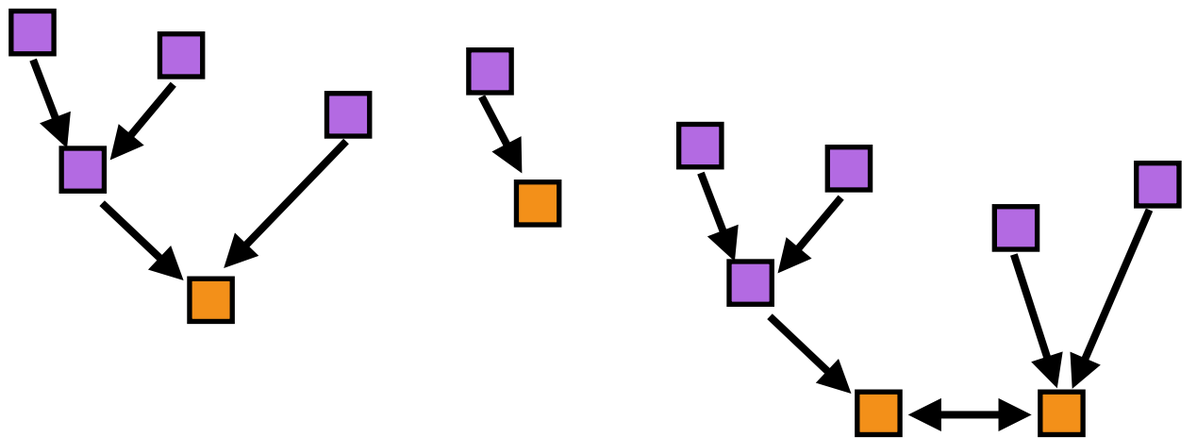
Isomorphic worlds: it is possible to reduce one into each other.

Minimal worlds: all reductions are isomorphic to itself.



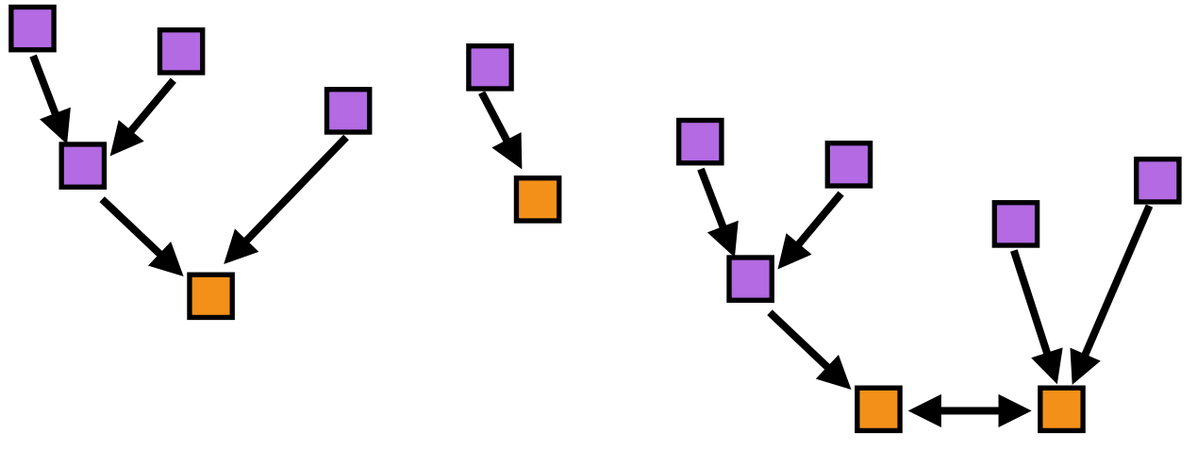
Limitations of bisimulation

Problem: bisimulation usually doesn't have a global minima



Limitations of bisimulation

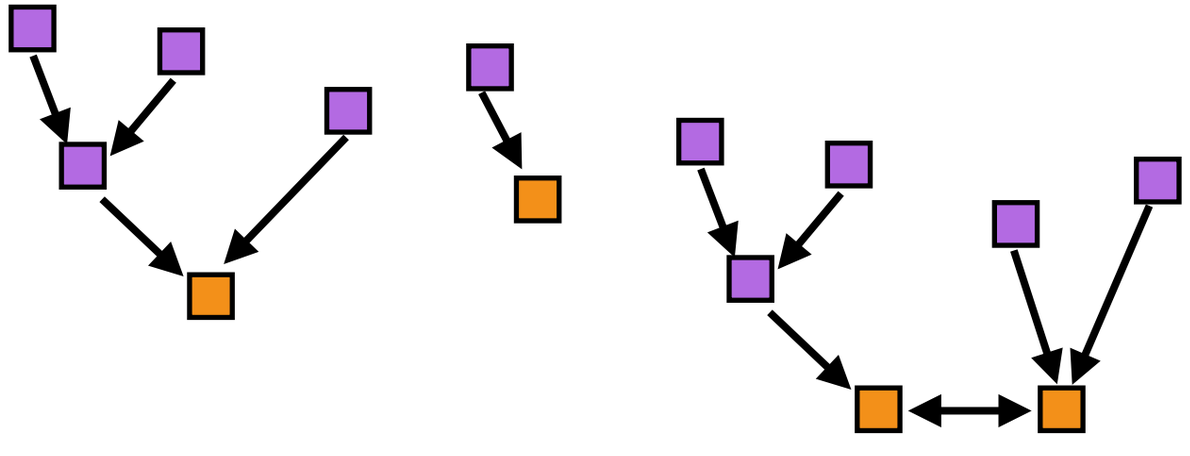
Problem: bisimulation usually doesn't have a global minima



Reason: bisimulation can replace pairwisely identical states, but cannot compress redundancy between groups of three or more states…

Limitations of bisimulation

Problem: bisimulation usually doesn't have a global minima



Reason: bisimulation can replace pairwisely identical states, but cannot compress redundancy between groups of three or more states…

Solution: accept what you get, or embrace negative probabilities!

Generalised transducers

A quasi-stochastic vector sums 1 but can have negative numbers.

$$p = (0.6, -0.1, 0.5)$$

Generalised transducers

A quasi-stochastic vector sums 1 but can have negative numbers.

Generalised transducers have quasi-stochastic states and transitions:

$$\Pr(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t}) = \boldsymbol{u}^\intercal \cdot \left(\prod_{i=0}^{t} A_i^{(y_i|a_i)} \right) \cdot \boldsymbol{v}$$

Generalised transducers

A quasi-stochastic vector sums 1 but can have negative numbers.

Generalised transducers have quasi-stochastic states and transitions:

$$\Pr(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t}) = \boldsymbol{u}^\intercal \cdot \left(\prod_{i=0}^{t} A_i^{(y_i|a_i)} \right) \cdot \boldsymbol{v}$$

Theorem: generalised transducers can always be reduced up to a unique minimum!

Generalised transducers

A quasi-stochastic vector sums 1 but can have negative numbers.

Generalised transducers have quasi-stochastic states and transitions:

$$\Pr(\boldsymbol{y}_{:t}|\boldsymbol{a}_{:t}) = \boldsymbol{u}^{\intercal} \cdot \left(\prod_{i=0}^{t} A_i^{(y_i|a_i)} \right) \cdot \boldsymbol{v}$$

Theorem: generalised transducers can always be reduced up to a unique minimum!

Limitation: substantial lack of interpretability, inability to sample, etc

Contents
1. Problem setting
2. Fundamental ideas
3. Minimal world models
4. Interpretable world models
5. Ideas to take home

Interpretable models: what can agents learn?

What makes a world model learnable?

1. *Predictive*: world state have no future information.
2. *Observable*: world state is a (deterministic) function of the history, so $S_t = f(\boldsymbol{H}_{:t-1})$
3. *Unifilar*: world state can be deterministically updated as $S_{t+1} = \hat{f}(S_t, Y_t, A_t)$

Interpretable models: what can agents learn?

What makes a world model learnable?

1. *Predictive*: world state have no future information.
2. *Observable*: world state is a (deterministic) function of the history, so $S_t = f(\boldsymbol{H}_{:t-1})$
3. *Unifilar*: world state can be deterministically updated as $S_{t+1} = \hat{f}(S_t, Y_t, A_t)$

How to create observable world models?

—-> By switching the phase space into distributions, known as “beliefs”.

Interpretable models: what can agents learn?

What makes a world model learnable?

1. *Predictive*: world state have no future information.
2. *Observable*: world state is a (deterministic) function of the history, so $S_t = f(\boldsymbol{H}_{:t-1})$
3. *Unifilar*: world state can be deterministically updated as $S_{t+1} = \hat{f}(S_t, Y_t, A_t)$

How to create observable world models?

—-> By switching the phase space into distributions, known as "beliefs".

In fact, Bayesian beliefs are unifilar and yield a transducer that generate the same interface!

$$b_{t+1}(s_{t+1}) = \frac{1}{Z} \sum_{s_t} p(y_t, s_{t+1}|a_t, s_t) b_t(s_t)$$

Minimal predictive models

Let's build equivalence classes of histories where $\boldsymbol{h}_{:t-1} \sim_\epsilon \boldsymbol{h}'_{:t-1}$ if and only if

$$p(\boldsymbol{y}_{t:t+T}|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:t+T}) = p(\boldsymbol{y}_{t:t+T}|\boldsymbol{h}'_{:t-1}, \boldsymbol{a}_{t:t+T}), \quad \forall \boldsymbol{y}_{t:t+L}, \boldsymbol{a}_{t:t+L}, L \in \mathbb{N}.$$

These are the states of the "e-transducer".

Minimal predictive models

Let's build equivalence classes of histories where $\boldsymbol{h}_{:t-1} \sim_\epsilon \boldsymbol{h}'_{:t-1}$ if and only if

$$p(\boldsymbol{y}_{t:t+T}|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:t+T}) = p(\boldsymbol{y}_{t:t+T}|\boldsymbol{h}'_{:t-1}, \boldsymbol{a}_{t:t+T}), \quad \forall \boldsymbol{y}_{t:t+L}, \boldsymbol{a}_{t:t+L}, L \in \mathbb{N}.$$

These are the states of the "e-transducer".

Findings:

1. It is the minimal bisimulation of the Funes transducer
2. It is a transducer.
3. It is the minimal bisimulation of ***any*** predictive world model that generates the interface.
4. Thus, it is the minimal predictive world model that generates the interface.

Minimal predictive models

Let's build equivalence classes of histories where $\boldsymbol{h}_{:t-1} \sim_\epsilon \boldsymbol{h}'_{:t-1}$ if and only if

$$p(\boldsymbol{y}_{t:t+T}|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:t+T}) = p(\boldsymbol{y}_{t:t+T}|\boldsymbol{h}'_{:t-1}, \boldsymbol{a}_{t:t+T}), \quad \forall \boldsymbol{y}_{t:t+L}, \boldsymbol{a}_{t:t+L}, L \in \mathbb{N}.$$

These are the states of the "e-transducer".

Findings:

1. It is the minimal bisimulation of the Funes transducer
2. It is a transducer.
3. It is the minimal bisimulation of ***any*** predictive world model that generates the interface.
4. Thus, it is the minimal predictive world model that generates the interface.

Conclusion 1: agents with same interface but beliefs in different world models still share the e-transducer.

Minimal predictive models

Let's build equivalence classes of histories where $\boldsymbol{h}_{:t-1} \sim_\epsilon \boldsymbol{h}'_{:t-1}$ if and only if

$$p(\boldsymbol{y}_{t:t+T}|\boldsymbol{h}_{:t-1}, \boldsymbol{a}_{t:t+T}) = p(\boldsymbol{y}_{t:t+T}|\boldsymbol{h}'_{:t-1}, \boldsymbol{a}_{t:t+T}), \quad \forall \boldsymbol{y}_{t:t+L}, \boldsymbol{a}_{t:t+L}, L \in \mathbb{N}.$$

These are the states of the "e-transducer".

Findings:

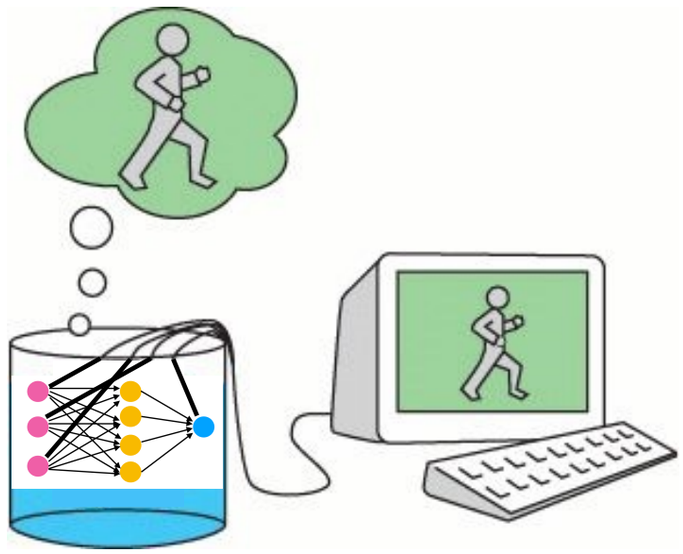
1. It is the minimal bisimulation of the Funes transducer
2. It is a transducer.
3. It is the minimal bisimulation of ***any*** predictive world model that generates the interface.
4. Thus, it is the minimal predictive world model that generates the interface.

Conclusion 1: agents with same interface but beliefs in different world models still share the e-transducer.

Conclusion 2: the e-transducer contains all the predictive information about the world that is learnable from the interface.

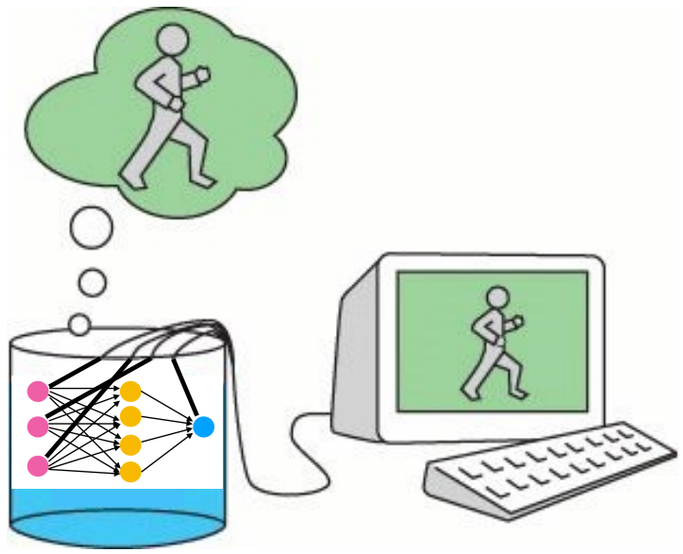
Contents

What have we learned?



- The class of world models that can be used to simulate a given agent's interface is *surprisingly rich*.

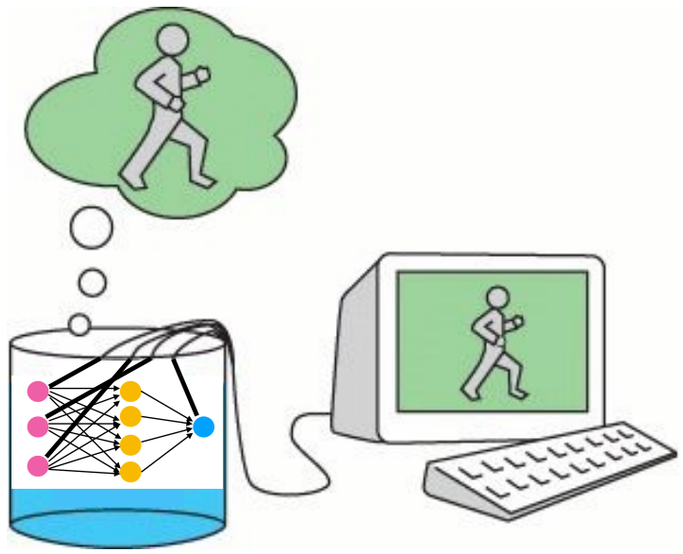
What have we learned?



- The class of world models that can be used to simulate a given agent's interface is *surprisingly rich*.

- There exists a fundamental trade-off between computational efficiency and interpretability of world models.

What have we learned?

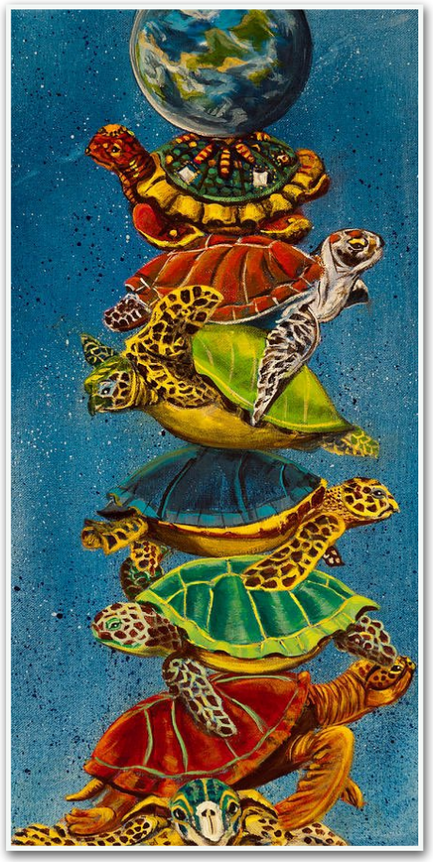


- The class of world models that can be used to simulate a given agent's interface is *surprisingly rich*.

- There exists a fundamental trade-off between computational efficiency and interpretability of world models.

- Interpretability has various flavours — we explored forward based on e-transducer and backwards based on reversible transducers.

arXiv



LessWrong

Thank you!

Contact information: Fernando E. Rosas

`f.rosas@sussex.ac.uk`