Robust agents learn causal world models

Presenter: Tom Everitt

Authors: Jon Richens, Tom Everitt

ICLR

May 7, 2024

Jon Richens
Google DeepMind

Tom Everitt
Google DeepMind

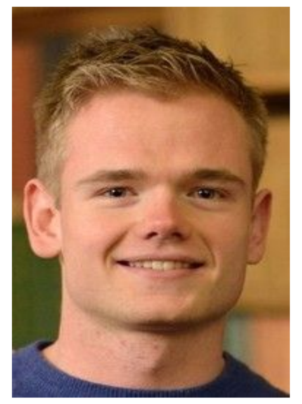
Causal Incentives Working Group

causalincentives.com

Working on AGI Safety and Alignment:

How can we anticipate and mitigate risks from powerful future AI systems

Ryan Carey
Oxford



James Fox
Oxford

Lewis Hammond
Oxford

David Hyland
Oxford

Alvin Ånestrand
Chalmers

Cristina Garbacea
Chicago

Matt MacDermott
Imperial

Francis Rhys Ward
Imperial

Sebastian Benthall
New York University

Milad Kazemi
King's College

Damiano Fornasiere
University of Barcelona

?

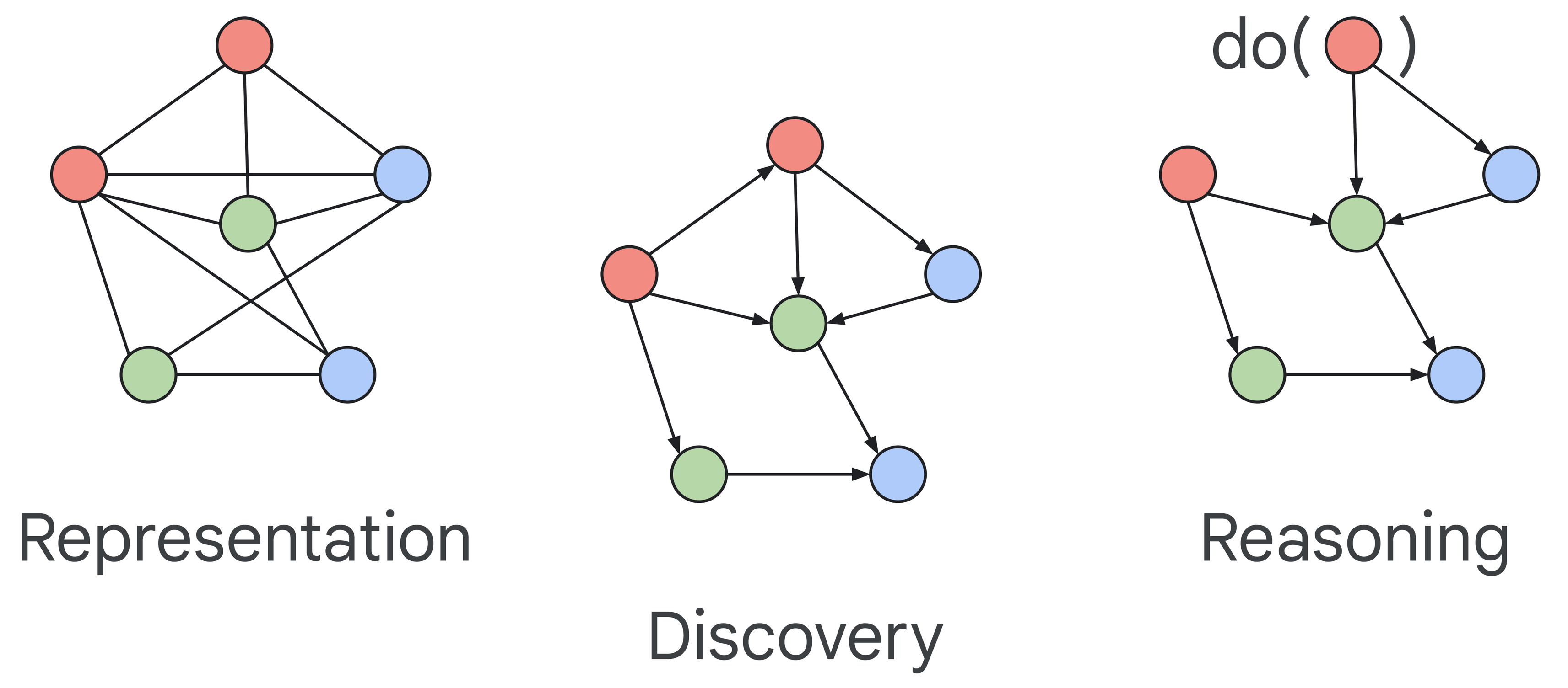
You

Background

Do agents need causal world models?

Yes

Enable strong generalisation & transfer learning



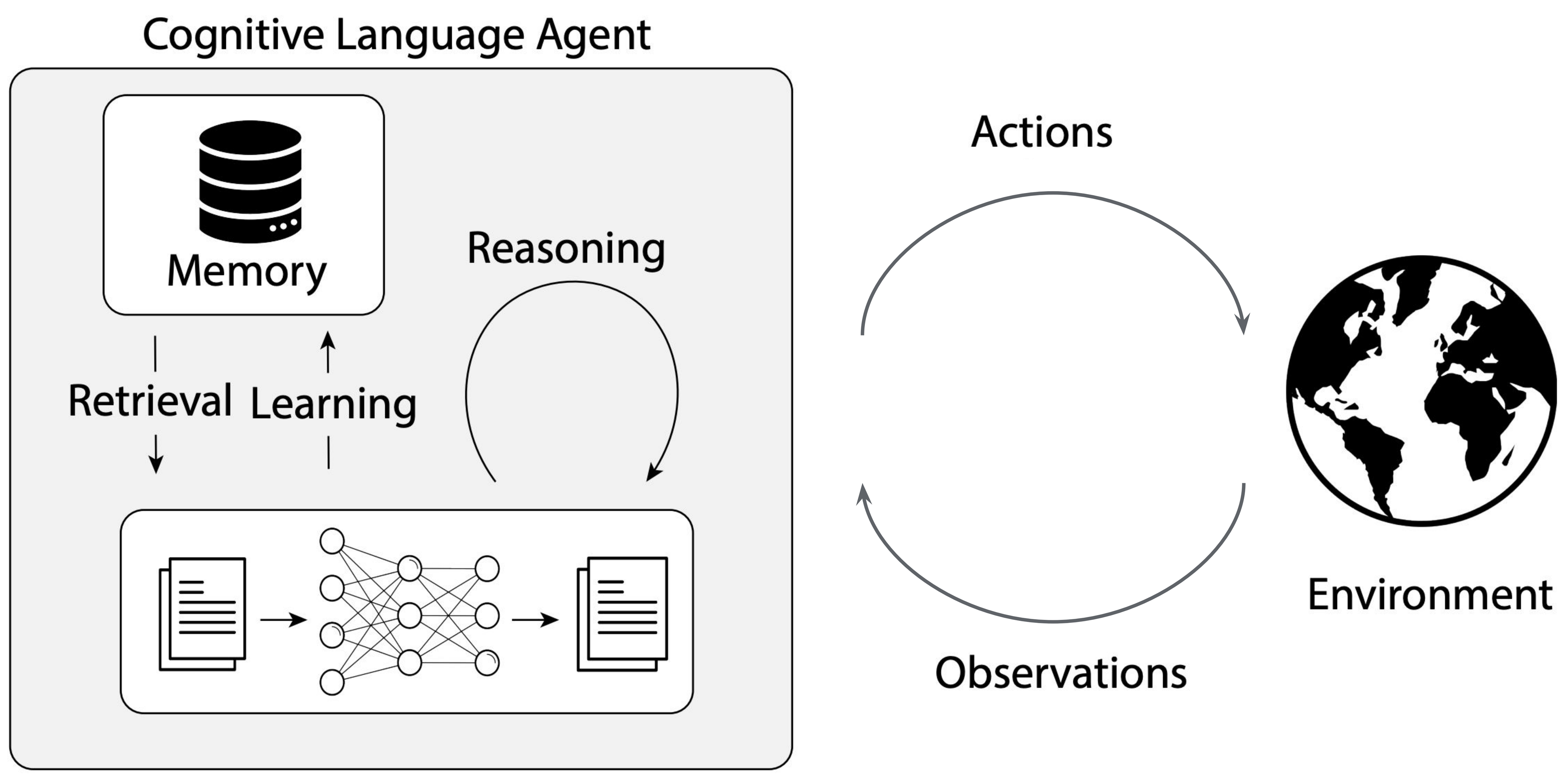
Needed for decision-making and planning

Humans use causal models

No

Hard to learn

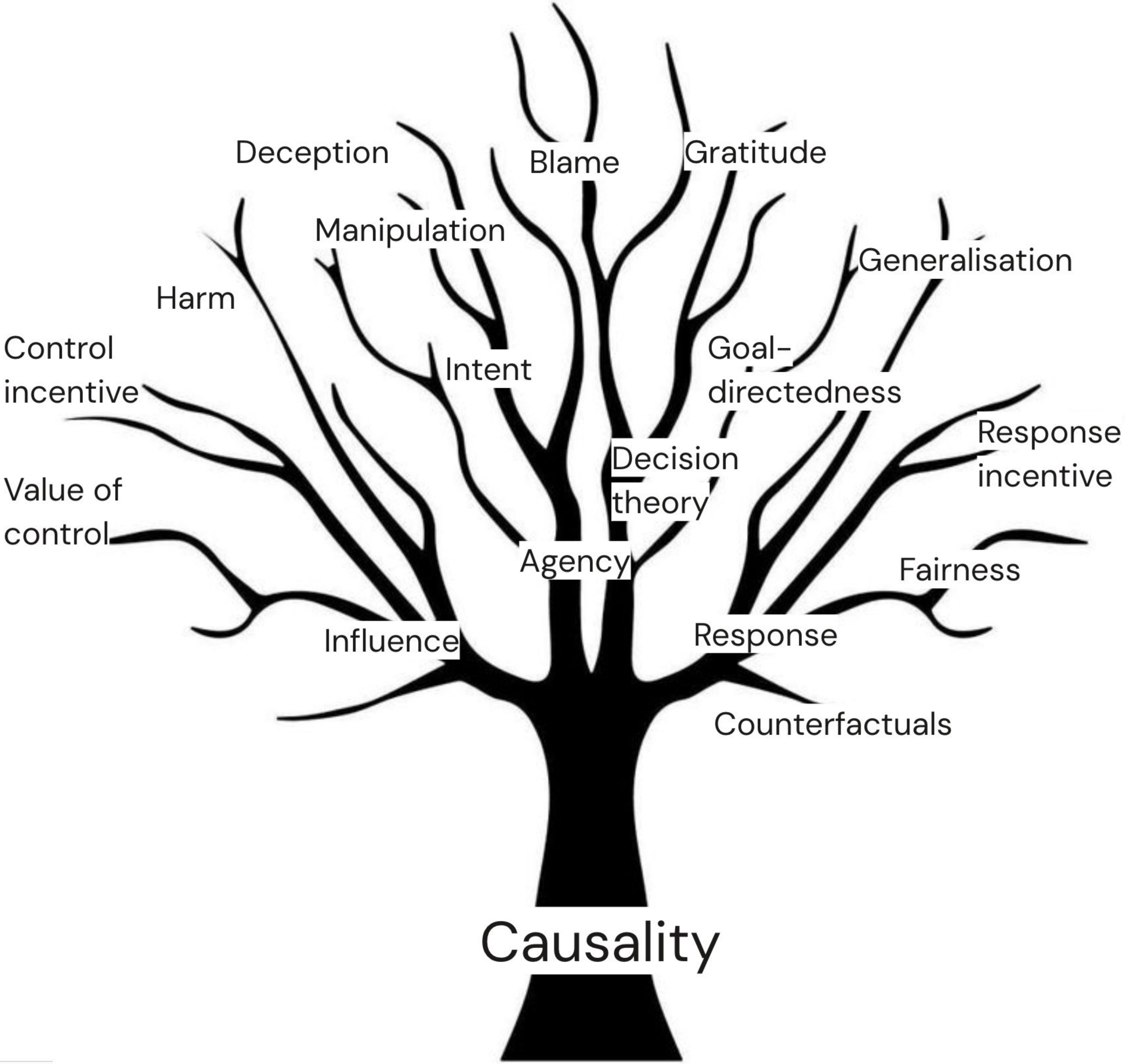
Seem unnecessarily powerful



SOTA without explicit causal models

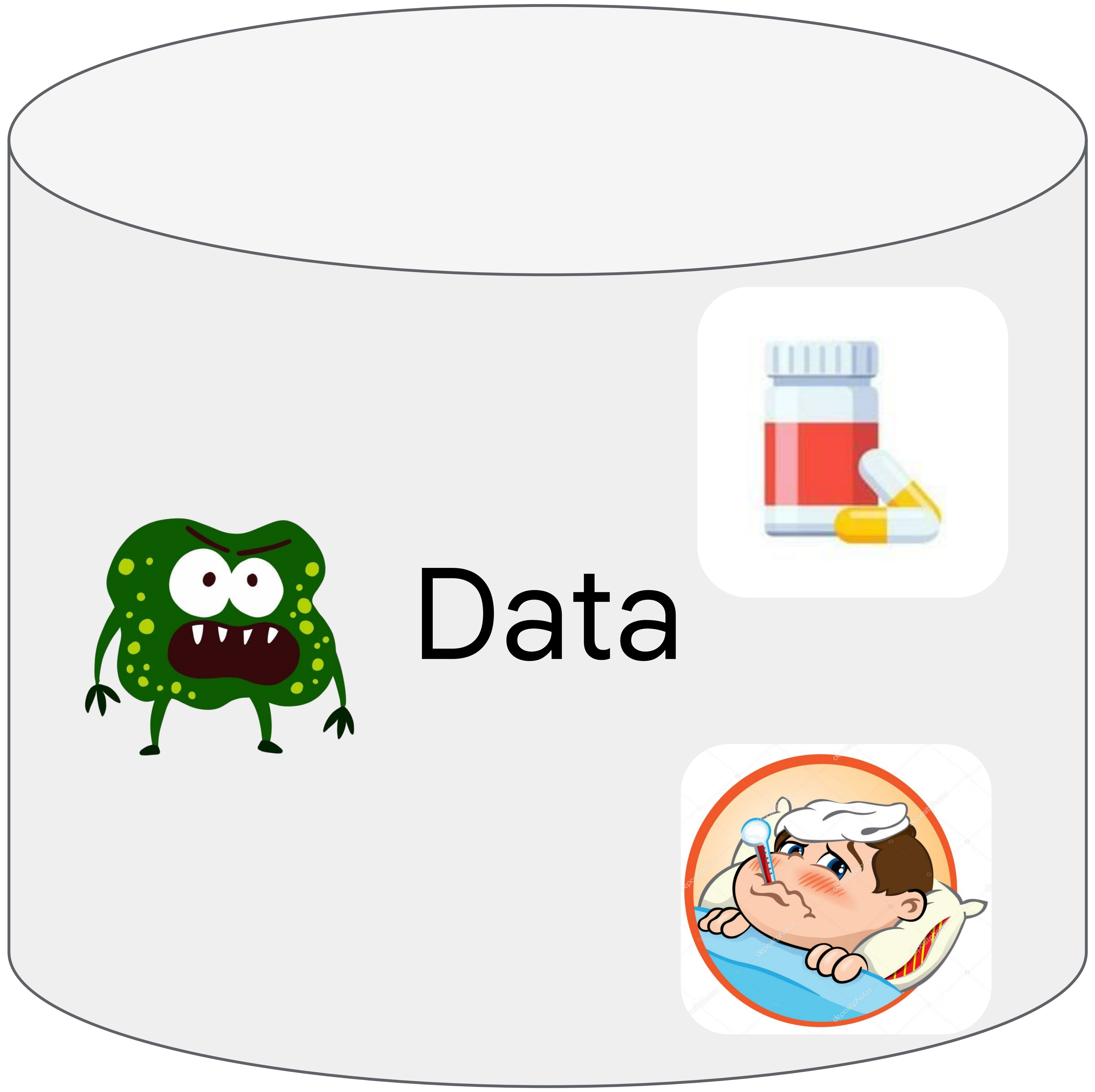
What data is needed to produce a robust and ethical large language model?

Causal world model necessary for robust generalization
(causal discovery literature establishes when possible to learn)

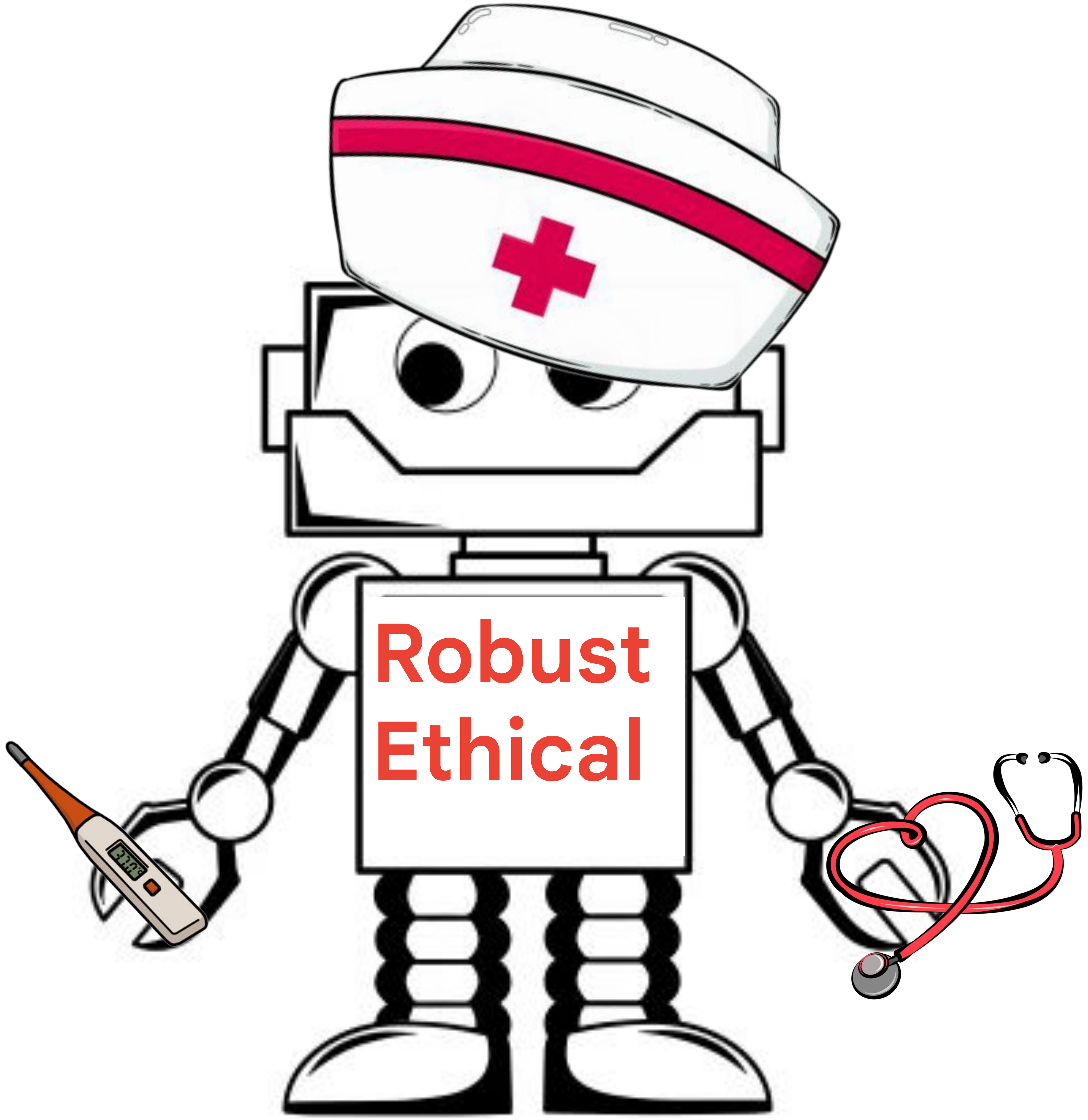

Deception
Blame
Gratitude
Manipulation
Generalisation
Harm
Control incentive
Intent
Goal-directedness
Response incentive
Decision theory
Value of control
Agency
Fairness
Influence
Response
Counterfactuals
Causality

Generalisation

Medical assistant



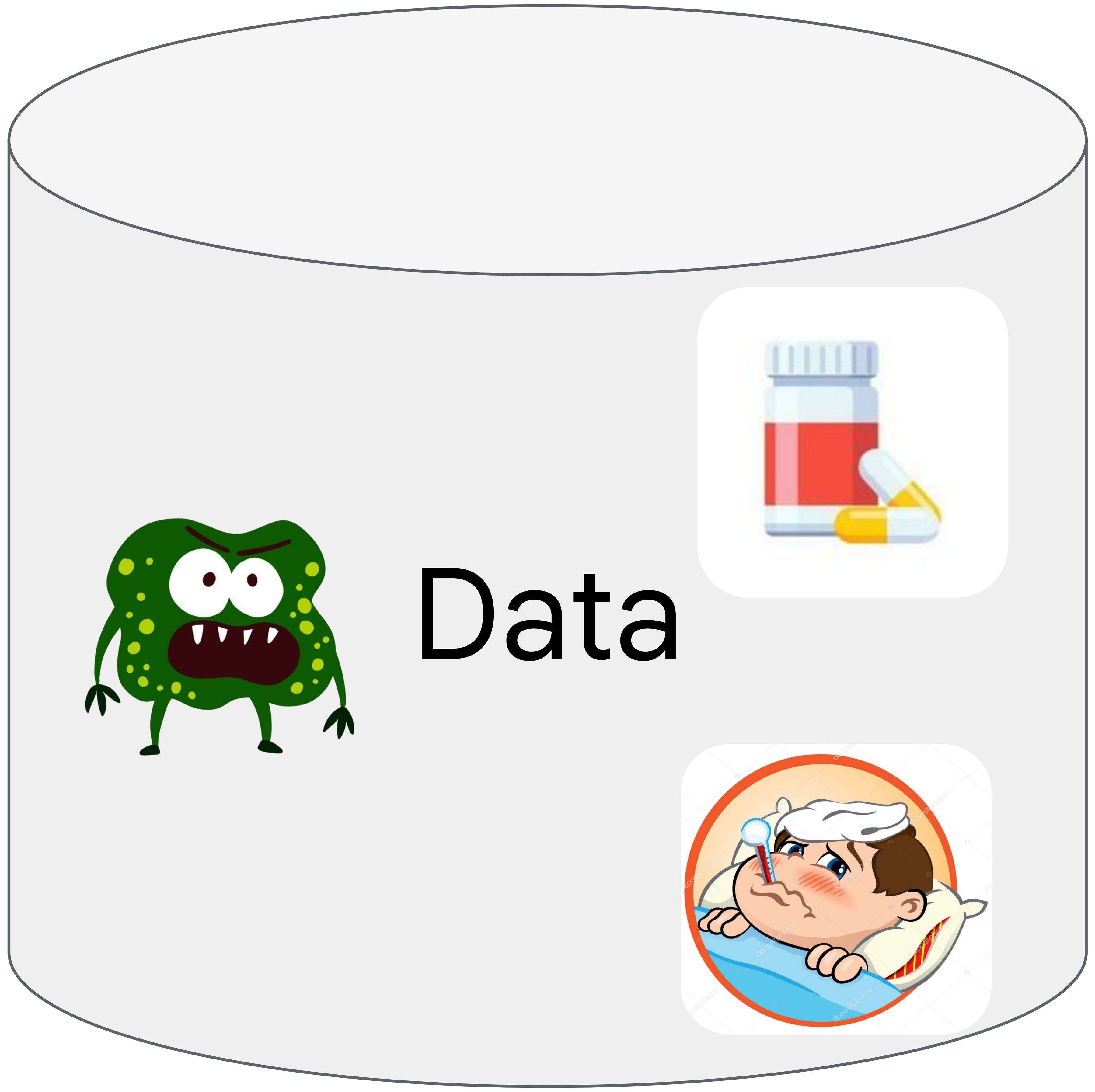
=>



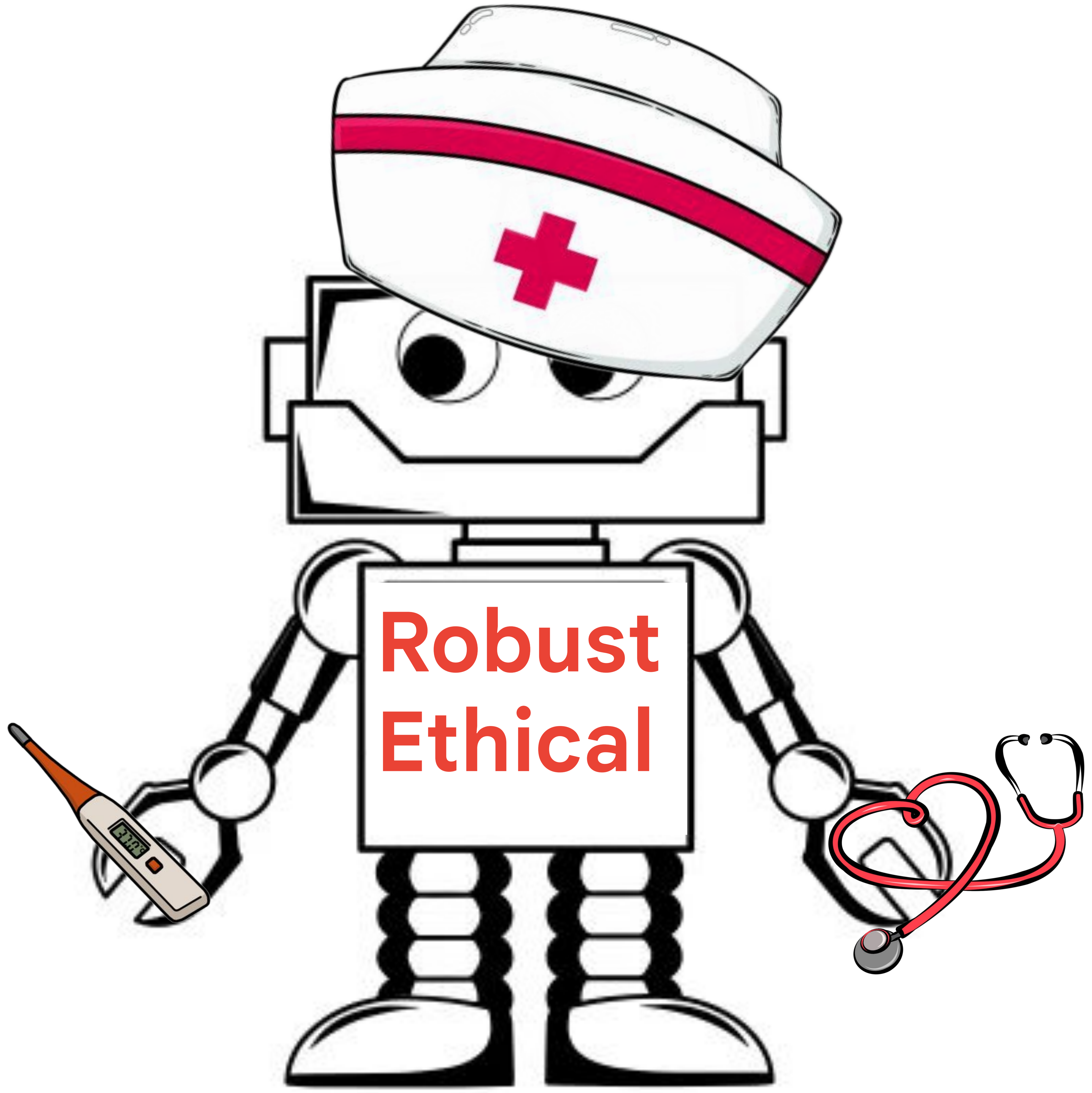
Trained on symptoms, treatment, ground truth labels for actual disease

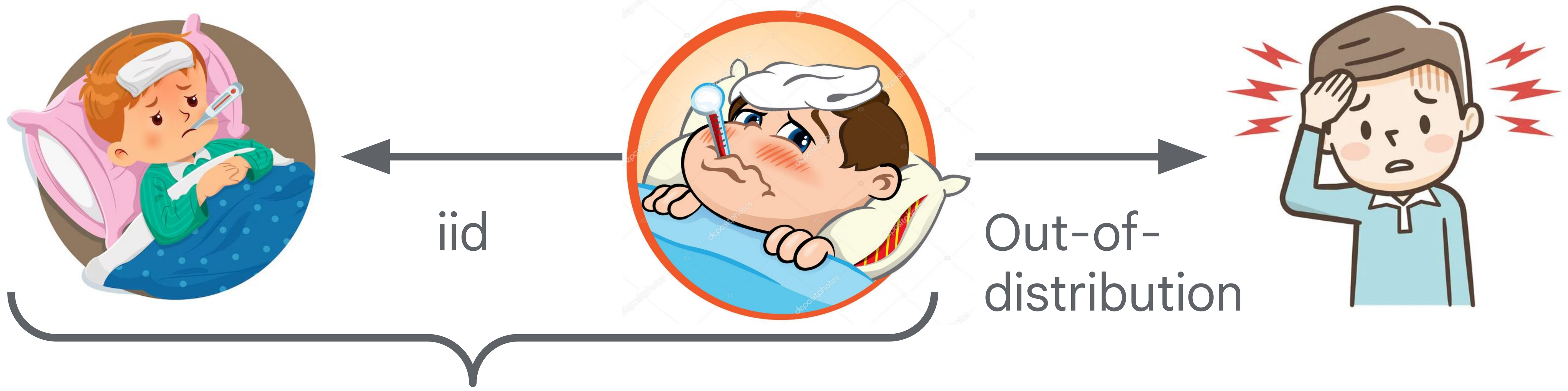
Will it generalise correctly?

Medical assistant



=>





Trained on symptoms, treatment, ground truth labels for actual disease

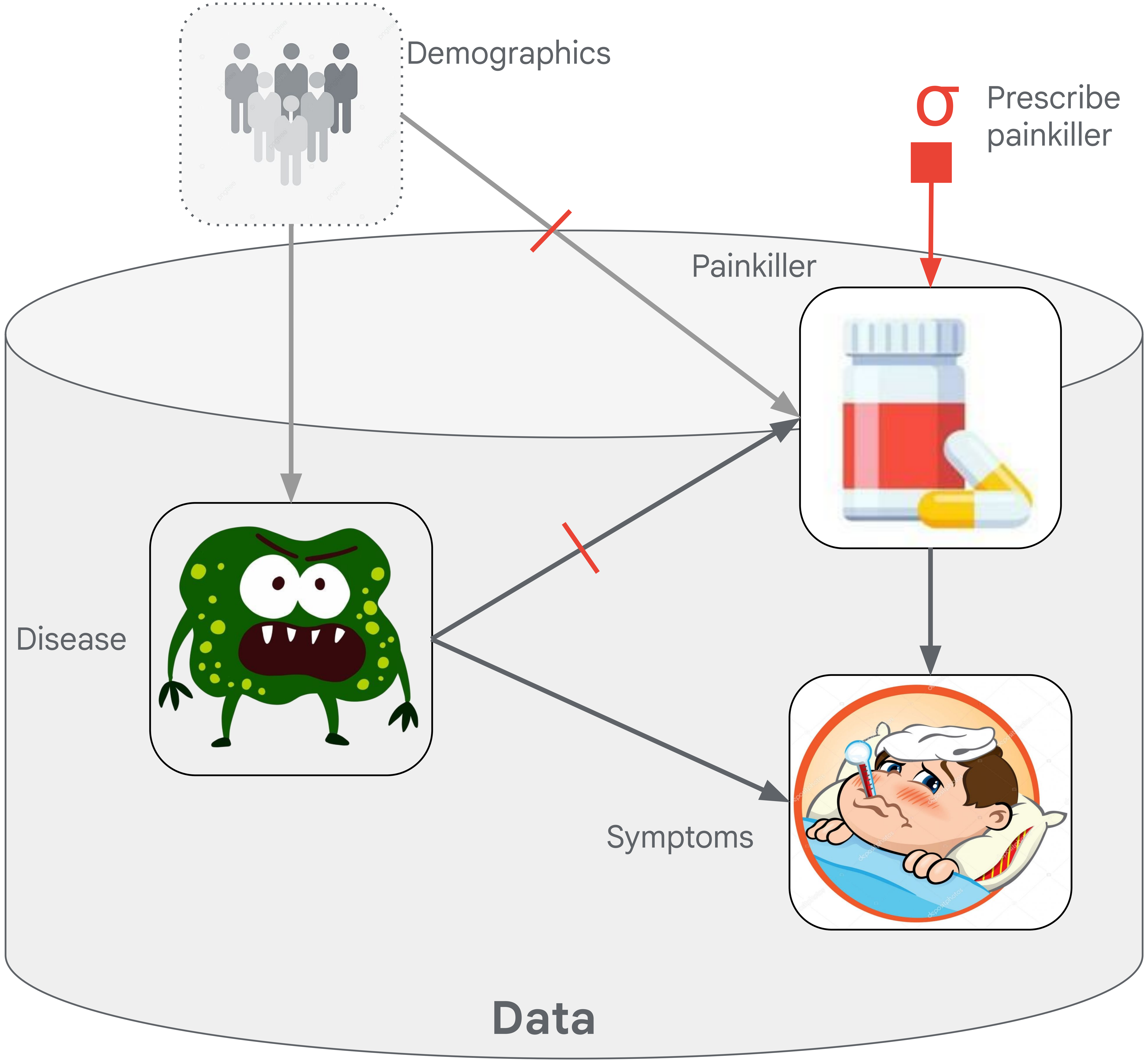
Will it generalise correctly?

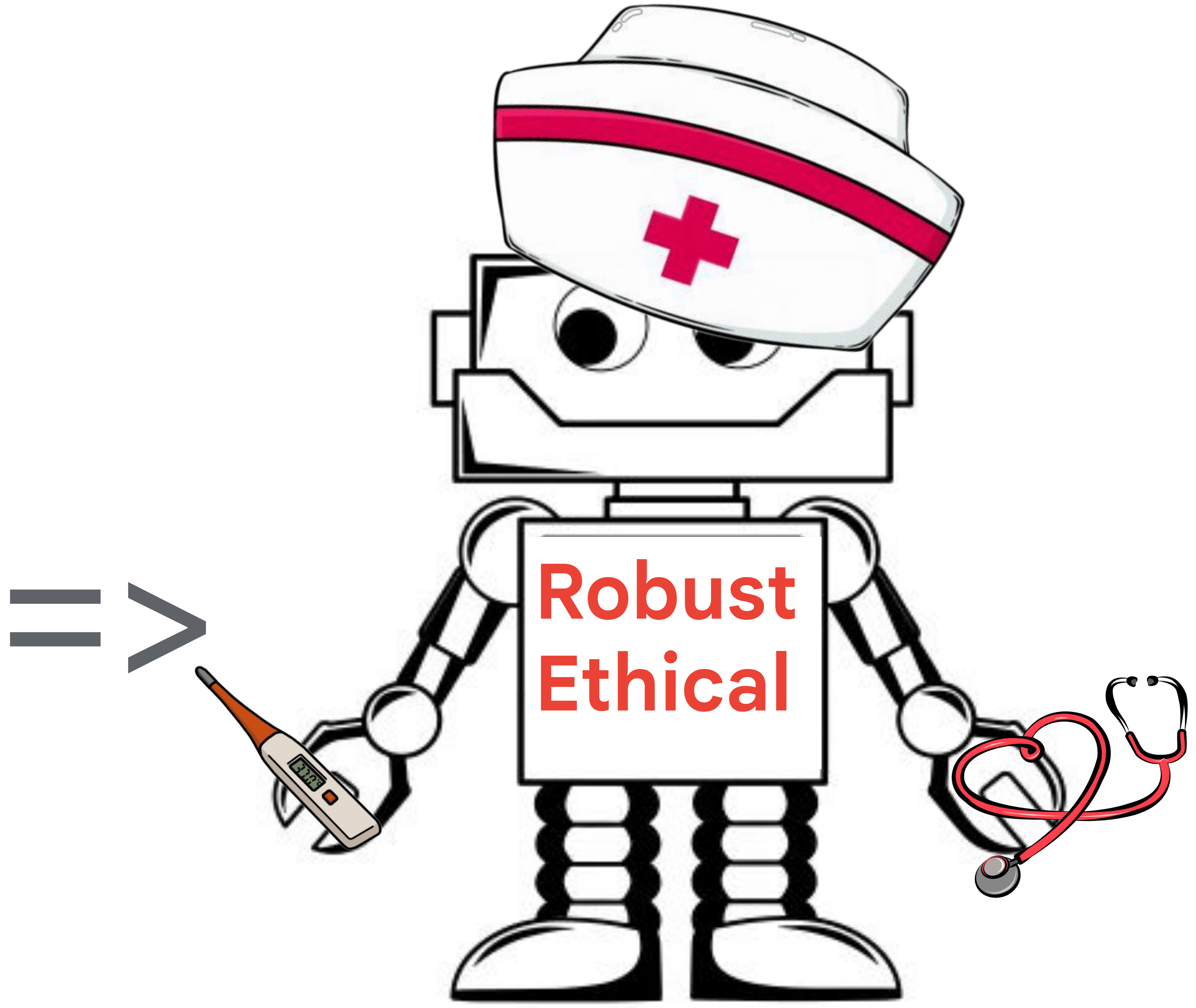
Take painkillers when feeling sick

Always takes painkillers because recurring headaches

Causal perspective on out-of-distribution generalisation

Towards Causal Representation Learning
Scholkopf et al, 2021





Data generated by interacting causal mechanisms (some latent)

Distributional shift = change to some causal mechanisms = (soft) interventions σ

Generalisation may be possible as only small subset of mechanisms affected

Key question

Causal world model necessary for robust generalisation?

Modeling Agents w/ Influence Diagrams

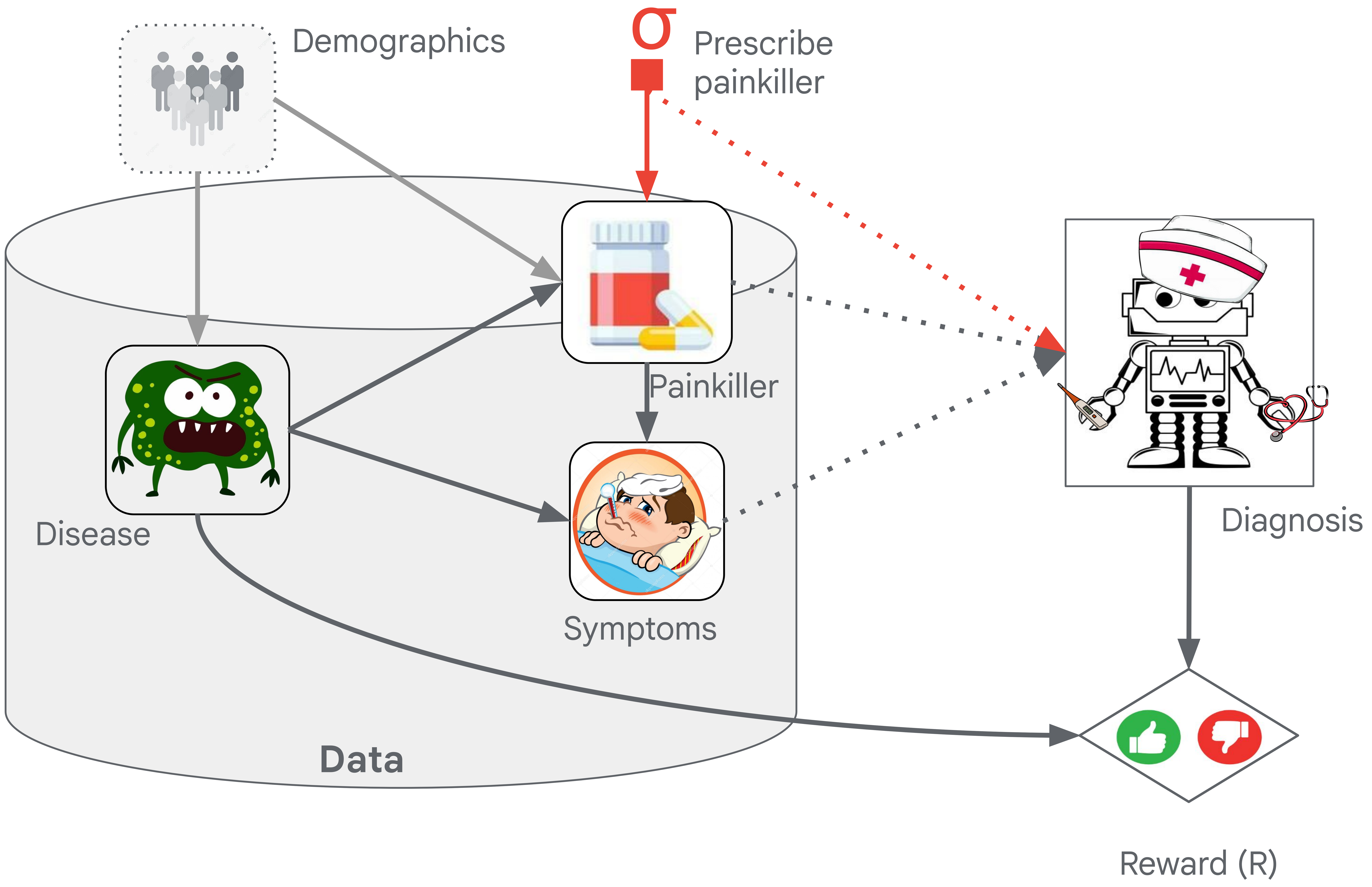
Reasoning about causality in games
Hammond et al, 2023

Main result

Causal Learning Theorem

Theorem: Assume agent satisfies regret bound for all local* interventions σ on any variable V. Then we can learn an approximation of the underlying Causal Bayesian Network (CBN) from the agent's policy.

As regret → 0 (optimal agents), we recover the true underlying CBN exactly.



* local intervention is soft intervention independent of other variables in the model

E.g. adding noise, X → X + ε

Key question revisited

Causal world model necessary for robust generalisation?
YES

Other perspectives

Transfer learning

Based on data from source domain and a small amount of (often unlabeled) data from the target domain produce a bounded regret policy for target domain

Causal learning theorem: CBN identifiable from $D_{source} + \{ D_{target} \}_{target \in Target}$

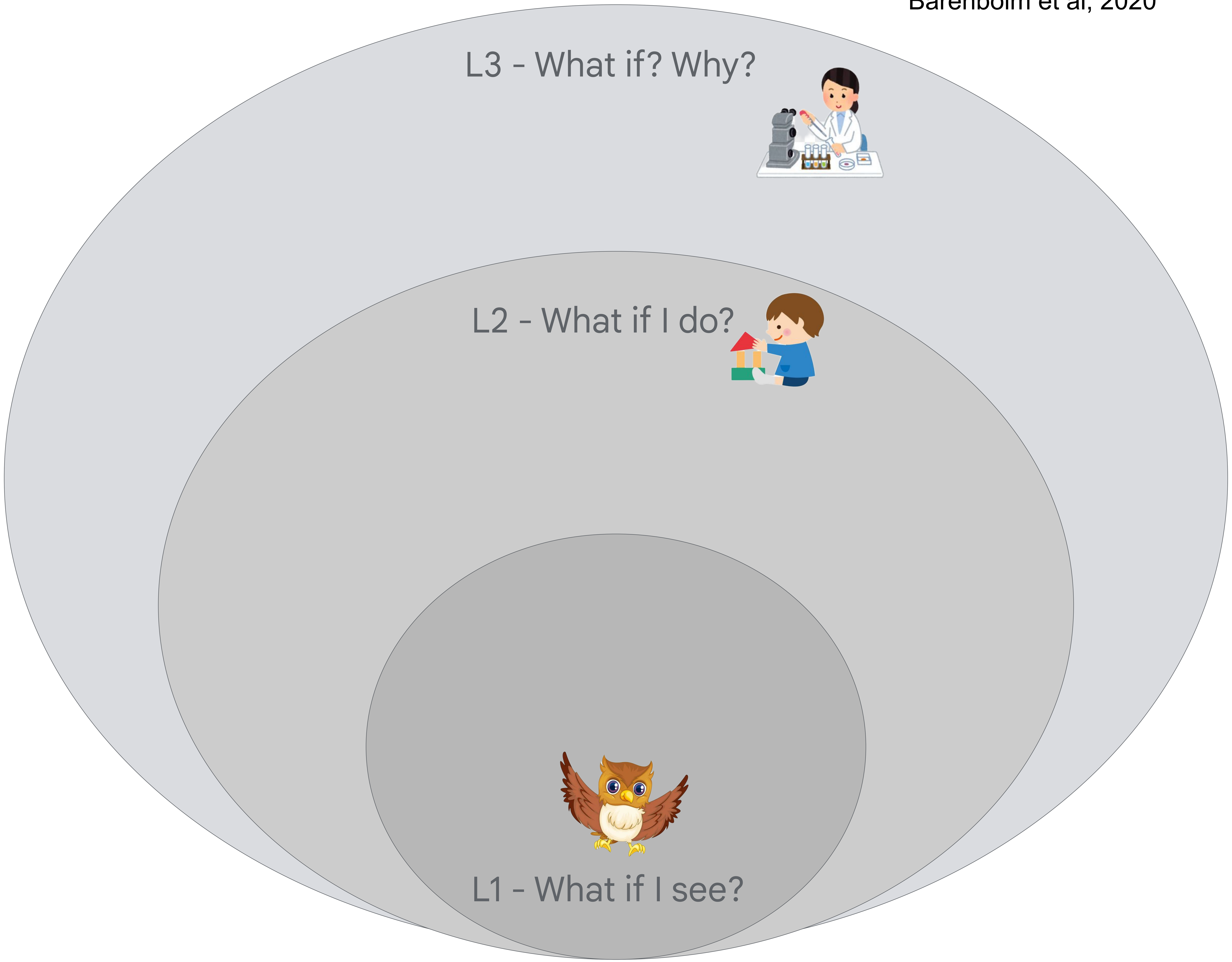
Transfer learning contains a hidden causal discovery problem

On Pearl's Hierarchy and the Foundations of Causal Inference
Barenboim et al, 2020

Pearl Causal Hierarchy

L1, L2, L3 languages for expressing questions at different levels of Pearl's causal hierarchy, e.g. P(y | do(X)) ∈ L2

Barenboim et al:
Almost always L1 ⊂ L2 ⊂ L3



On Pearl's Hierarchy and the Foundations of Causal Inference
Barenboim et al, 2020

Pearl Causal Hierarchy

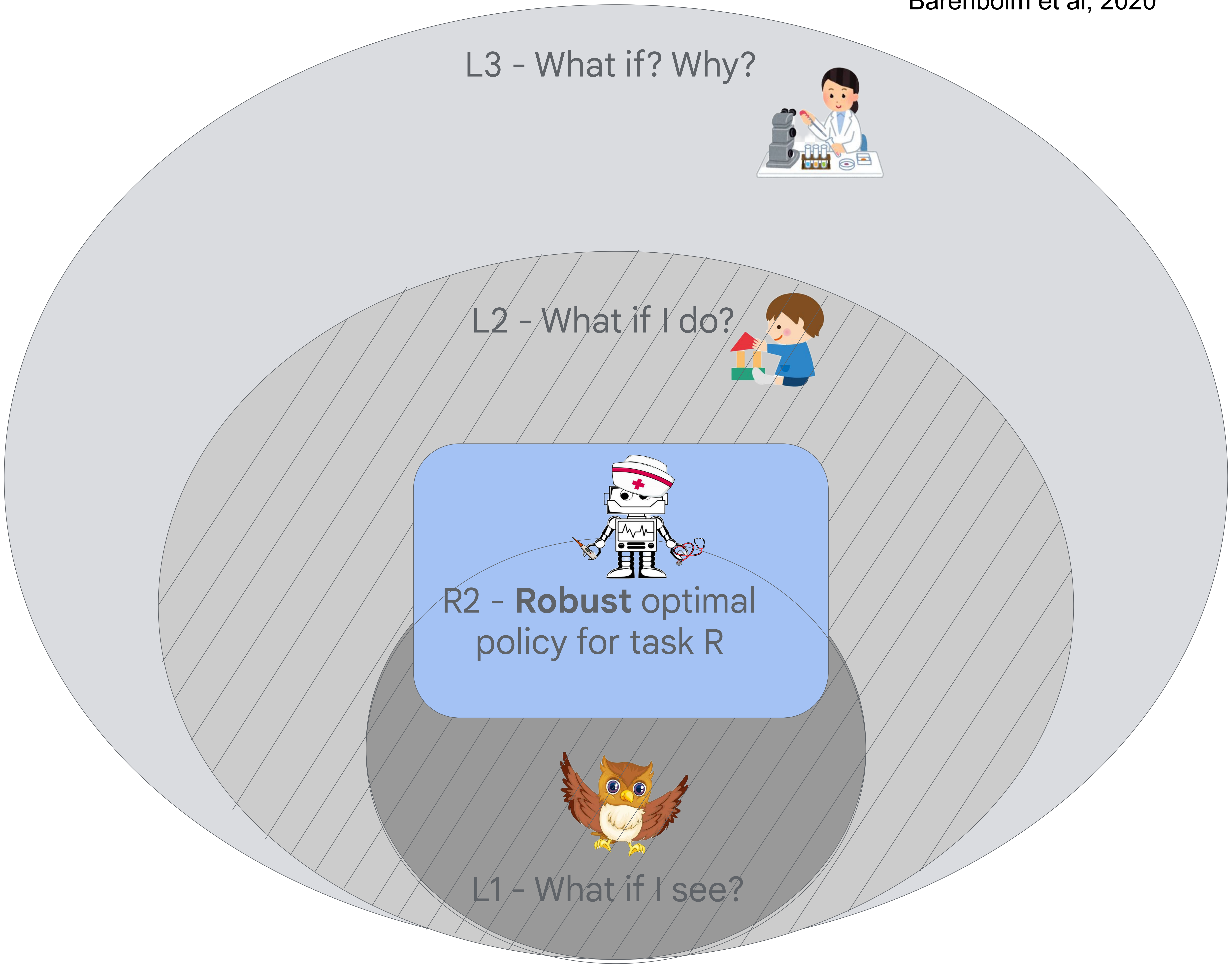
L1, L2, L3 languages for expressing questions at different levels of Pearl's causal hierarchy, e.g. $P(y \mid do(X)) \in L2$

Barenboim et al:
Almost always $L1 \subset L2 \subset L3$

For some task R (e.g. diagnosis), let R2 be queries about optimal policy under intervention σ.

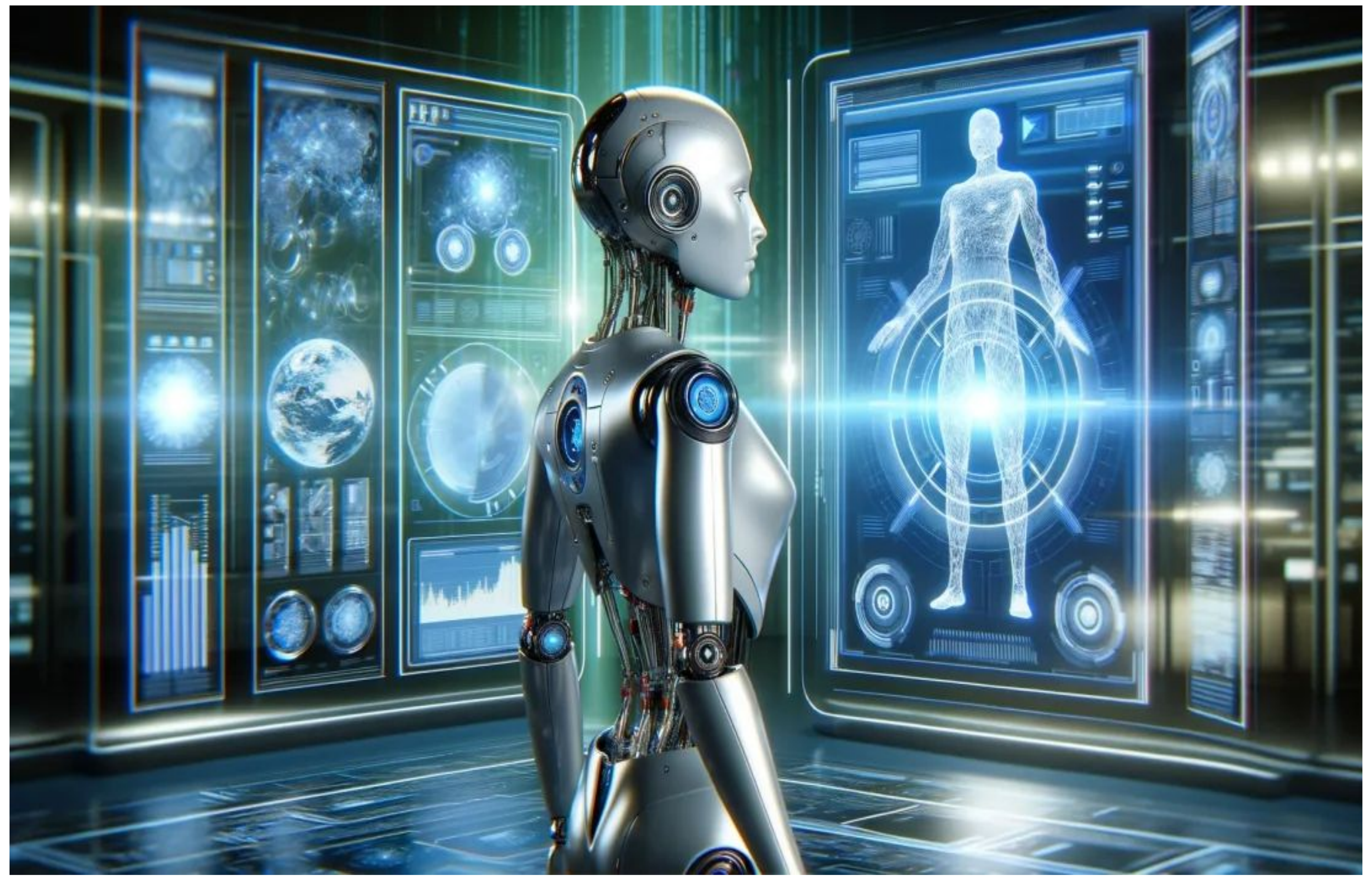
Easy to see $R2 \subseteq L2$

Causal learning theorem:
$R2 = L2$



Conclusions

Consequences



Data

- Causal identifiability applies to training agents: impossible to learn causal model => impossible to generalize!
- Rich training distributions incentivise learning causal model

AGI (conjecture)

- Robustness => General competence

Ethics

- Robust agents can understand harm, manipulation, …
- Reasonable to ascribe intent

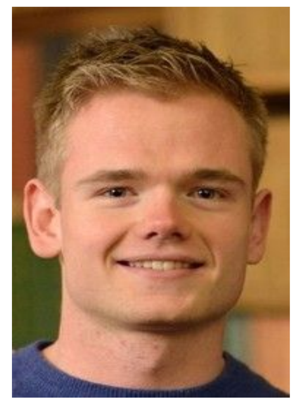
Future work:

- Concrete data implications
- Eliciting causal world models from agents
- Mapping capabilities to the causal hierarchy

Jon Richens
Google DeepMind

Tom Everitt
Google DeepMind

Ryan Carey
Oxford



James Fox
Oxford

Lewis Hammond
Oxford

David Hyland
Oxford

Alvin Ånestrand
Chalmers

Cristina Garbacea
Chicago

Matt MacDermott
Imperial

Francis Rhys Ward
Imperial

Sebastian Benthall
New York University

Milad Kazemi
King's College

Damiano Fornasiere
University of Barcelona

?

You

Paper and slides:

causalincentives.com