CIFAR CANADIAN INSTITUTE FOR ADVANCED RESEARCH | ICRA INSTITUT CANADIEN DE RECHERCHES AVANCÉES

Université de Montréal

Mila

From Conscious Processing to System 2 Deep Learning

Yoshua Bengio

UCSD Rockwood Memorial Lecture,
July 26th, 2021

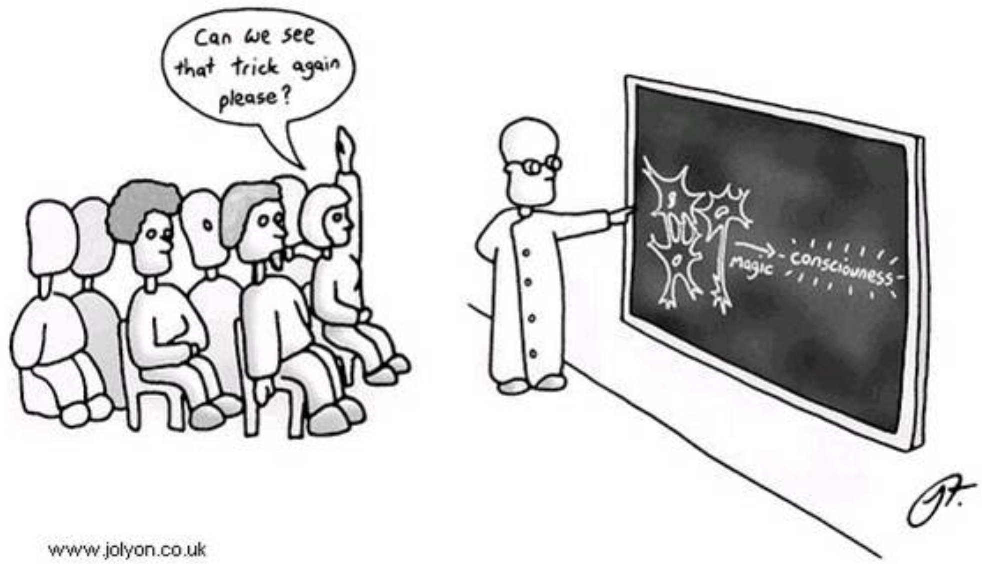
Missing from Current ML: Understanding & Generalization – Beyond the Training Distribution

- Learning theory only deals with generalization within the same distribution
- Models learn but do not generalize well (or have high sample complexity when adapting) to modified distributions, non-stationarities, etc.
- ***Humans do a lot better!!!***

Missing from Current ML: Understanding & Generalization – Beyond the Training Distribution

- If not iid, need alternative assumptions, otherwise no reason to expect generalization
  - Inductive biases inspired from brains
- How do distributions change?
- How can human-verbalizable knowledge be represented & re-used?

ML FOR CONSCIOUSNESS & CONSCIOUSNESS FOR ML



- Formalize and test **specific hypothesized functionalities of consciousness**
- Get the magic out of consciousness
- Understand evolutionary advantage of consciousness: computational and statistical (e.g. systematic generalization)
- Provide these advantages to learning agents

Mila

CONSCIOUS PROCESSING HELPS HUMANS DEAL WITH OOD SETTINGS

Faced with novel or rare situations, humans call upon conscious attention to combine on-the-fly the appropriate pieces of knowledge, to reason with them and imagine solutions.

→ we do not follow our habitual routines, we think hard to solve new problems.

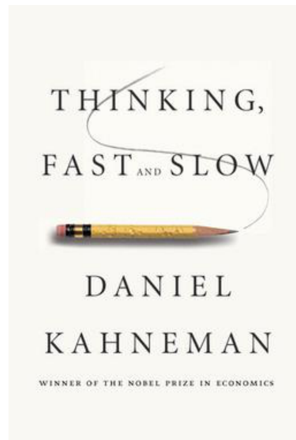
SYSTEM 1 VS. SYSTEM 2 COGNITION

2 systems (and categories of cognitive tasks):

Manipulates high-level / semantic concepts, which can be recombined combinatorially

System 1

- Intuitive, fast, **UNCONSCIOUS**, 1-step parallel, non-linguistic, habitual
- Implicit knowledge
- Current DL

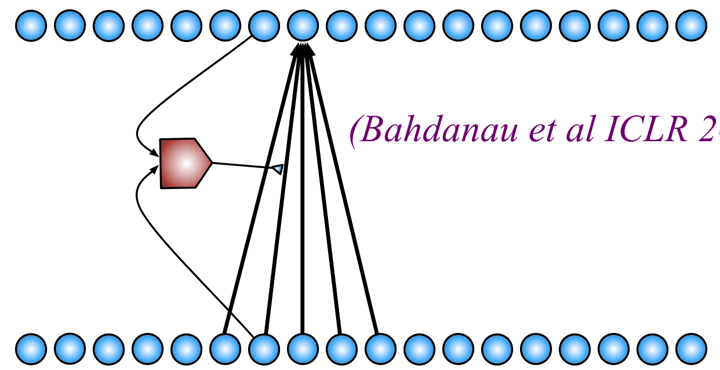


System 2

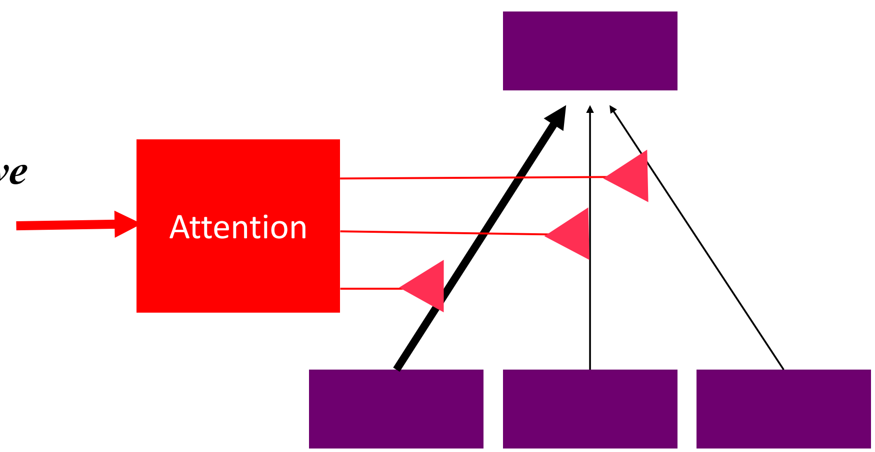
- Slow, logical, **sequential**, **CONSCIOUS**, linguistic, algorithmic, planning, reasoning
- Explicit knowledge
- DL 2.0

CORE INGREDIENT FOR CONSCIOUS REASONING: **ATTENTION**

- **Focus** on a one or a few elements at a time in order to reason / resolve coherent interpretation among these variables / modules
- **Content-based soft attention** is convenient (NLP SOTA), can backprop to *learn where to attend, what to think about*
- Attention is an **internal action**, needs a **learned attention policy***, may explain* ***subjective experience*** *(Graziano 2013), Attention Schema Theory*
- Operating on unordered SETS of (key, value) pairs
- Modules communicating through attention: RIMs, *Goyal et al arXiv:1909.10893*

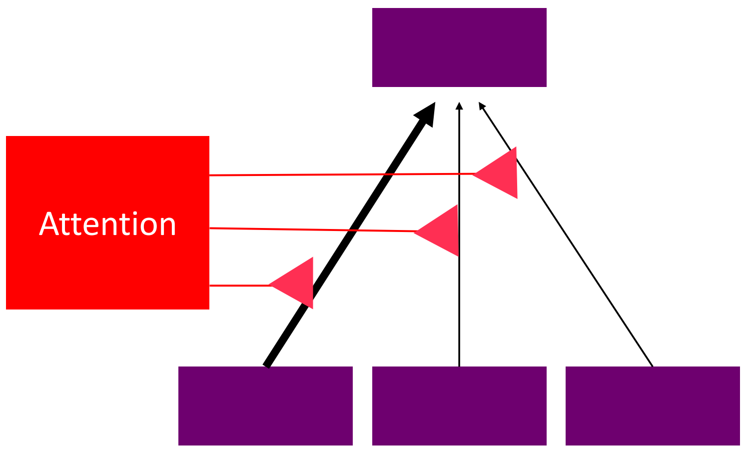


(Bahdanau et al ICLR 2015)



Mila

FROM ATTENTION TO **INDIRECTION**



- Attention = dynamic connection
- Receiver gets the selected value
- Value of what? From where?

  → Also send 'name' (or key) of sender
- Keep track of 'named' objects: indirection
- Manipulate sets of objects (transformers)

P.S. contrary to convnets doing object recognition, sequential tasks involving memory and attention typically involve a more difficult optimization problem, and fighting underfitting (including the issue of long-term dependencies)

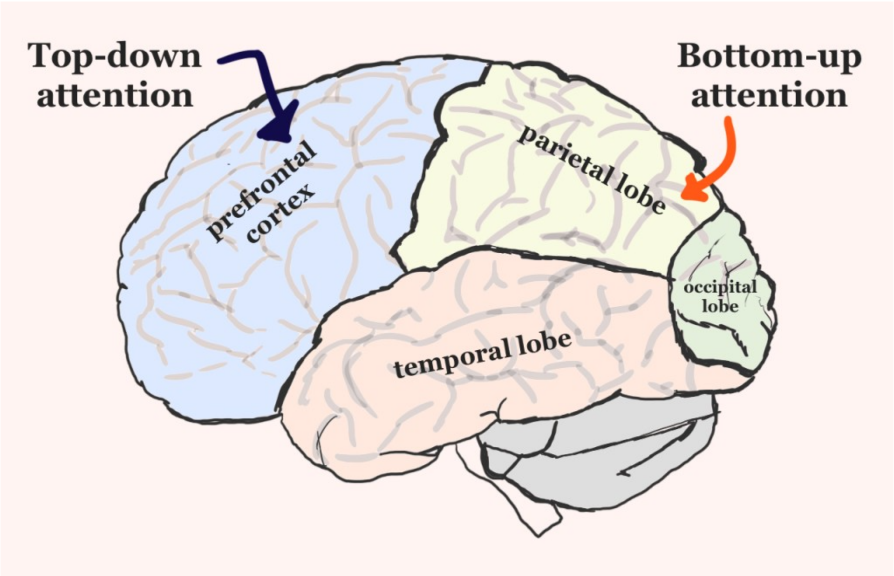
FROM ATTENTION TO **CONSCIOUSNESS**

C-word not taboo anymore in cognitive neuroscience

Global Workspace Theory

(Baars 1988++, Dehaene 2003++)

- Bottleneck of conscious processing
  - *WHY A BOTTLENECK?*
- Selected item is broadcast, stored in short-term memory, conditions perception and action
- System 2-like sequential processing, conscious reasoning & planning & imagination
- Can only run 1 simulation at a time, unlike a movie, only few abstract concepts involved at each step



THOUGHTS, CONSCIOUSNESS, LANGUAGE

- Consciousness: from humans reporting
- High-level representations ⟺ language
- High-level concepts: meaning anchored in low-level perception and action → **tie system 1 & 2**
- Grounded high-level concepts

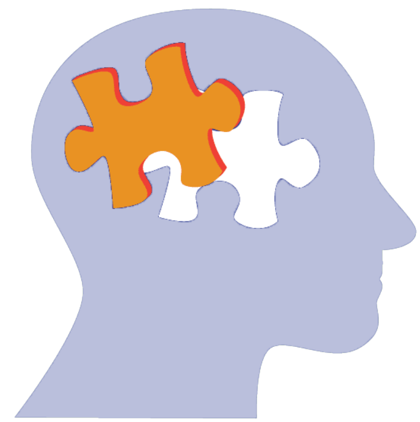
  → better natural language understanding

  → language = clues about high-level concepts

- **Grounded language learning**
  e.g. BabyAI: *(Chevalier-Boisvert and al ICLR 2019)*

FROM REASONING TO OOD GENERALIZATION?

- **Current industrial-strength ML (including in NLP) suffers from robustness issues due to poor performance OOD**
- Humans use higher-level cognition (system 2) for out-of-distribution generalization
- Why and how does it help?
- How is that related with agency? causality?
- How do we incorporate these principles in deep learning to obtain both system 1 and system 2 deep learning?



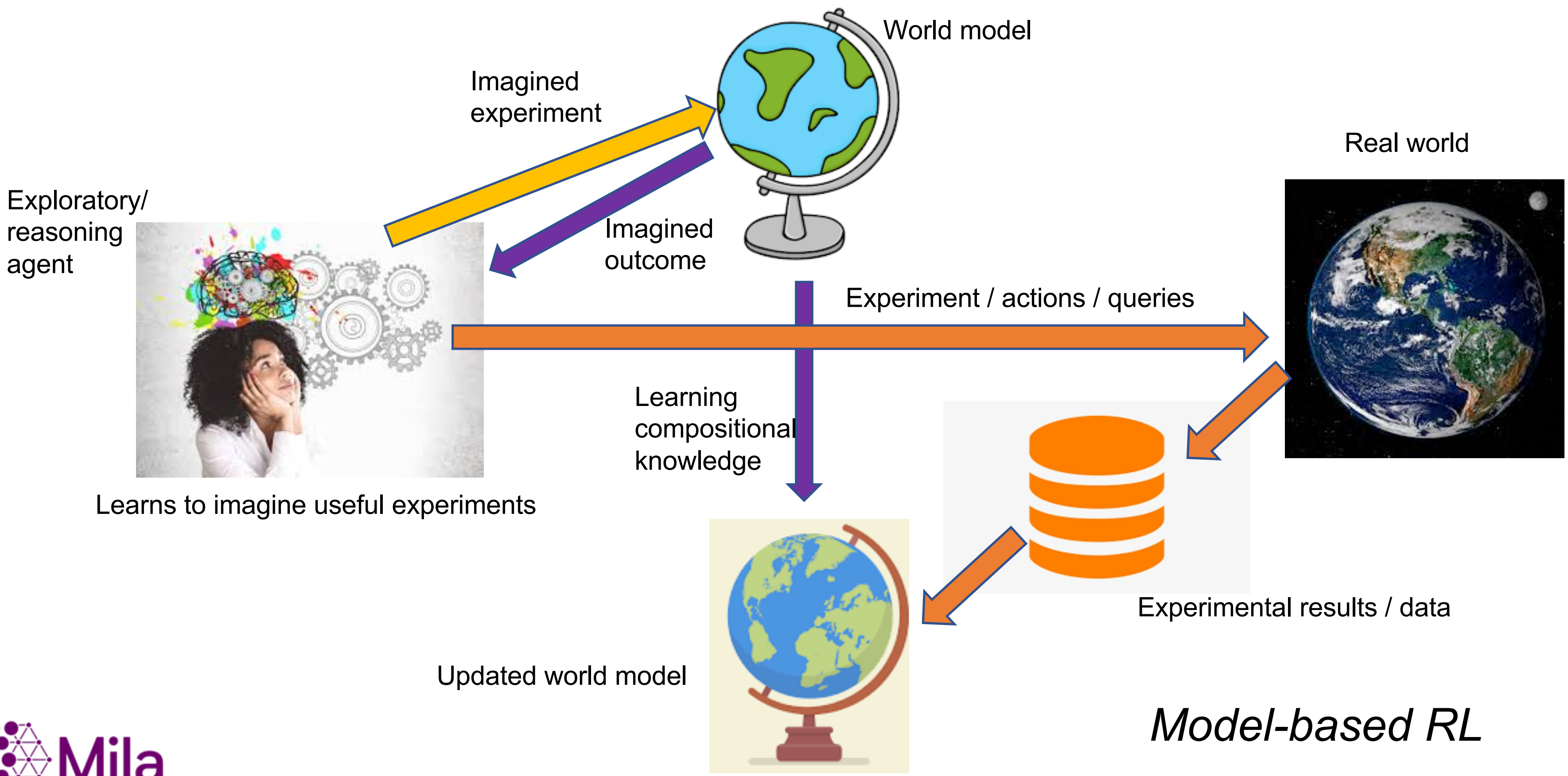
CAUSAL UNDERSTANDING → PREDICT EFFECT OF INTERVENTIONS → OOD GENERALIZATION

- **Causal understanding = decomposing knowledge into pieces (causal mechanisms) = building an abstract model of how the world works**

- Losing the IID hypothesis, we need other hypotheses OOD

- Causal understanding rests on the notion of INTERVENTION and the assumption that **causal mechanisms are stationary**
- Intervention = action which breaks the default flow of causality
- Good causal model: requires a **world model** of the effect of actions
- Good causal model: can infer what intervention explains a change in distribution and can predict the effect of these actions by combining , even if they never happened in the past

World Model, External Policy & Internal Policy



World Model, External Policy & Internal Policy

Why do we need all these pieces?

- Dangerous world: Try actions in your head (world model) first
- Compositional knowledge: World model's knowledge decomposed into its independent mechanisms (not easy to do that with fast policy)
- Need to act quickly: Searching through all possible plans and evaluating them with world model is too expensive → train a fast-acting external policy
- Expensive actions: Training ext. policy through direct experimentation = waste (need to iterate), better to train the external policy by interrogating the model
- Internal vs external policy: Avoid danger, internal exploration to train external policy & plan external actions, internal policy = thinking

Mila

HUMAN INSPIRATION FOR INDUCTIVE BIASES: IMPLICIT VS VERBALIZABLE KNOWLEDGE

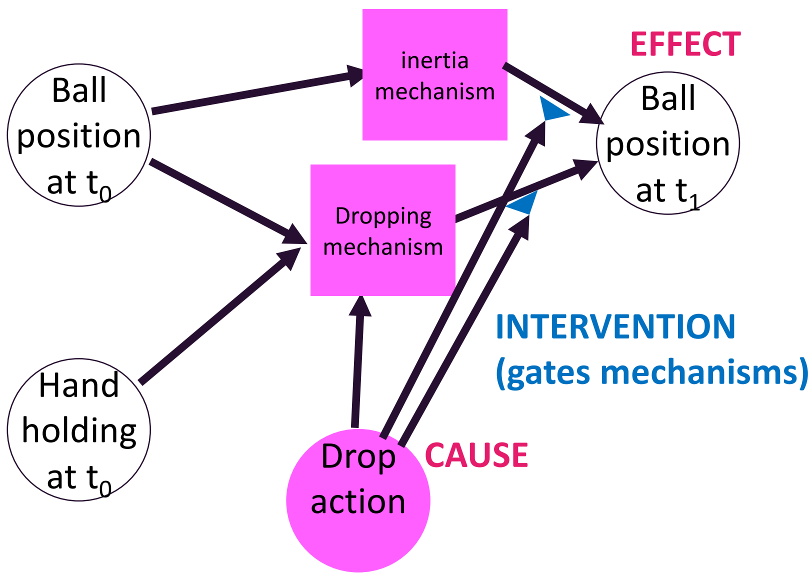
- Most knowledge in our brain is implicit and **not verbalizable** (hence the explainability challenge, even for humans)
- Some of our knowledge is verbalizable and we can reason and plan explicitly with it, using system 2
- The concepts manipulated in this way are those we can name with language, allow us to reason OOD

➔ clarify these assumptions as priors to be able to embed them in ML architectures and training frameworks which bridge abstract perception, abstract reasoning and abstract action.

SPARSE DEPENDENCIES BETWEEN ABSTRACT VARIABLES

Also consistent with Baar’s Global Workspace Theory (1997) of conscious processing.

Linguistic example: *"if I drop the ball, it will fall on the ground"*

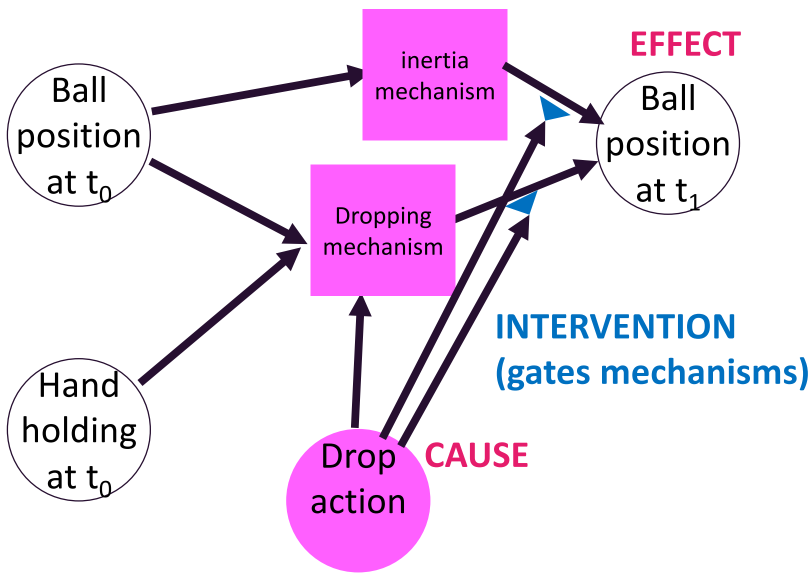


An abstract outcome can be predicted accurately from very few conditioning abstract variables

ABSTRACT VARIABLES PLAY A CAUSAL ROLE

COUNTERFACTUAL

Linguistic example: *"if I had dropped the ball, it would have fallen on the ground"*

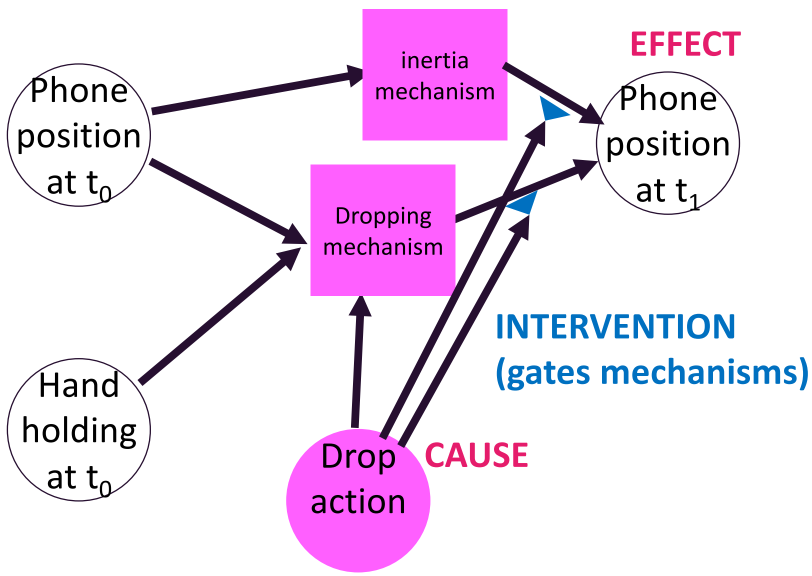


Variables play the role of cause, effect, agent, action, intervention

REUSABLE CAUSAL MECHANISMS

COUNTERFACTUAL

Linguistic example: *"if I had dropped the phone, it would have fallen on the ground"*

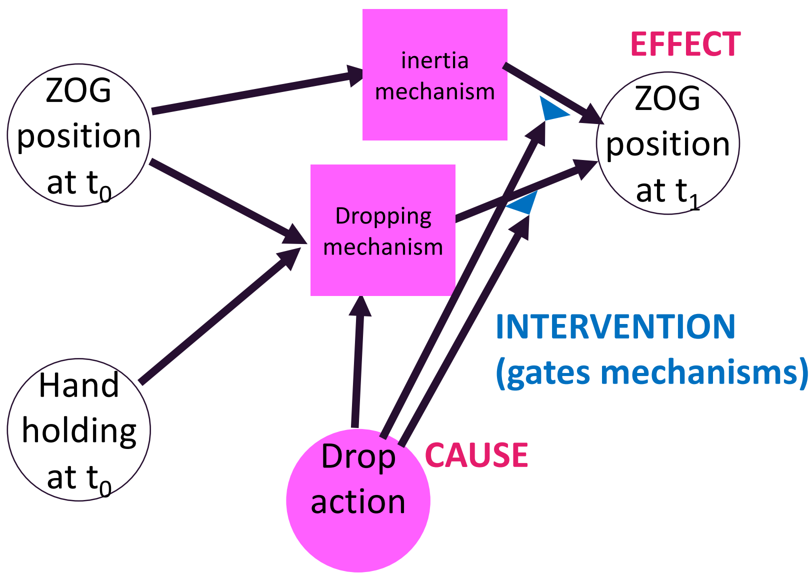


The same mechanism can be reused on many instance tuples

SYSTEMATIC GENERALIZATION

COUNTERFACTUAL

Linguistic example: *"if I had dropped the ZOG, it would have fallen on the ground"*

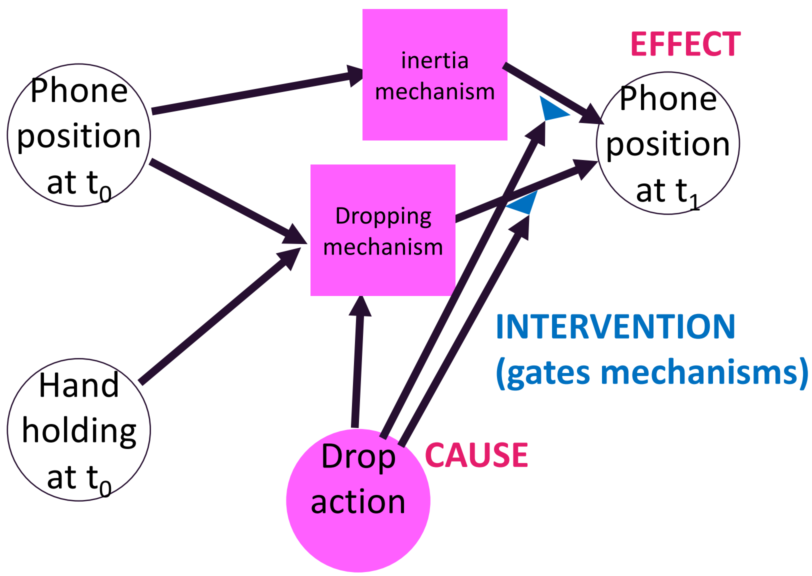


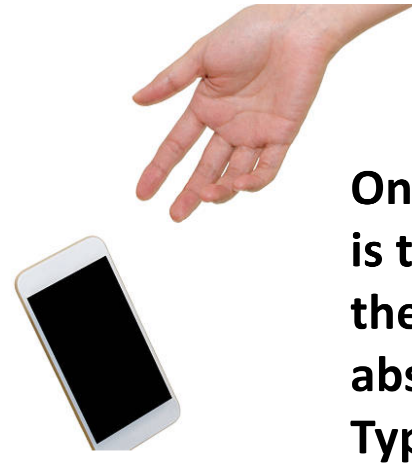
We make inferences assuming that the same mechanism can be reused on novel instances (if the object has the right affordances / type)

SPARSE LOCALIZED INTERVENTIONS

PLANNING

Linguistic example: *"if I decided to drop the phone, it would fall on the ground"*





Only one abstract entity is typically affected by the abstract action = abstract intervention. Typically only one attribute of that entity is directly affected.

INDEPENDENT MECHANISMS

Scholkopf et al 2012

Updating a verbalizable fact about the world generally does not affect any other piece of knowledge.

Consider how we try to factorize code into reusable but independent pieces:

Ideally, knowledge is factorized into independent 'pieces of code', i.e., which cannot be better compressed by merging them.

Better having a separate piece of code for dropping and for watching.

DISCRETE, SYMBOLIC, ABSTRACT CONCEPTS

- Language allows communication of simplified, DISCRETE, messages among humans
- Thoughts manipulate such discrete entities
- Evidence that hippocampus represents discrete concepts
- **The bottleneck of discretization in the communication between brain modules may further facilitate systematic generalization, making different brain modules hot-swappable for one another** (e.g. replace a noun by another in a sentence)

DIRECT MAPPING BETWEEN ABSTRACT VARIABLES AND ABSTRACT ACTIONS

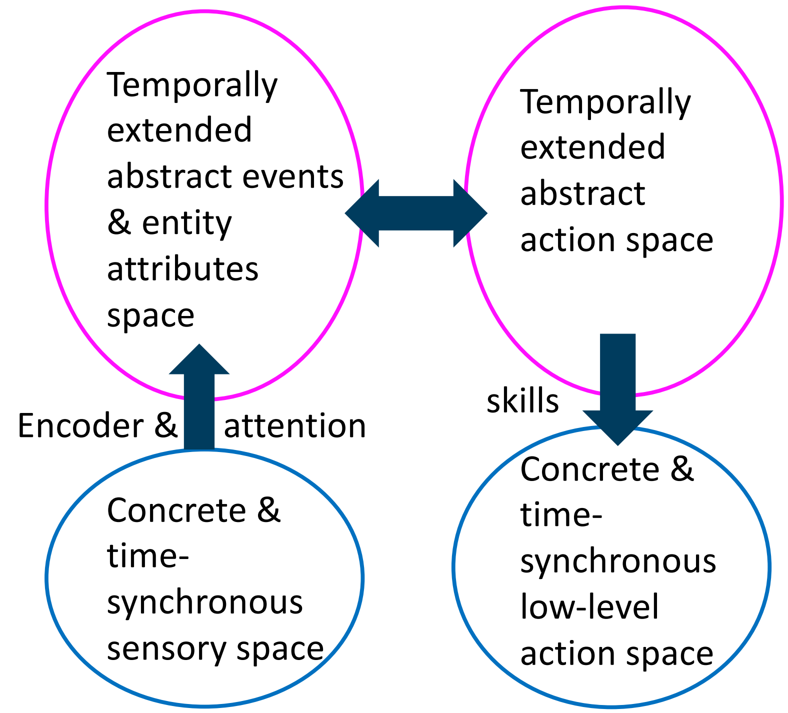
"Dropping" ⟷ *"the phone"*

Follows up on (E. Bengio et al, 2017; V. Thomas et al, 2017; more recently see Kim et al ICML 2019)

For each instantiated abstract action, there is generally one abstract entity, and one abstract attribute of that entity, which that abstract action intends to change (although there may be changes in intermediate elements and downstream effects as well).

However, the same entity (object) can be affected or controlled in many different ways, different abstract actions (verbs) by many different agents (subjects).

The same action type (verb) can of course be applied on many different entities (objects).



WHAT **CAUSES** CHANGES IN DISTRIBUTION?

Underlying physics: actions are localized in space and time.

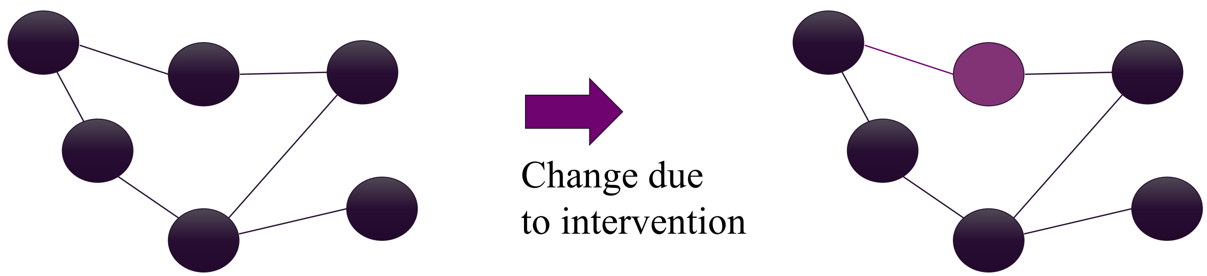
Hypothesis to replace iid assumption:

changes = consequence of an intervention on few causes or mechanisms

Extends the hypothesis of (informationally) Independent Mechanisms *(Scholkopf et al 2012)*

ICLR 2020: A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms, Bengio, Deleu, Rahaman, Ke, Lachapelle, Bilaniuk, Goyal, Pal

➔ **local inference or adaptation in the right model**

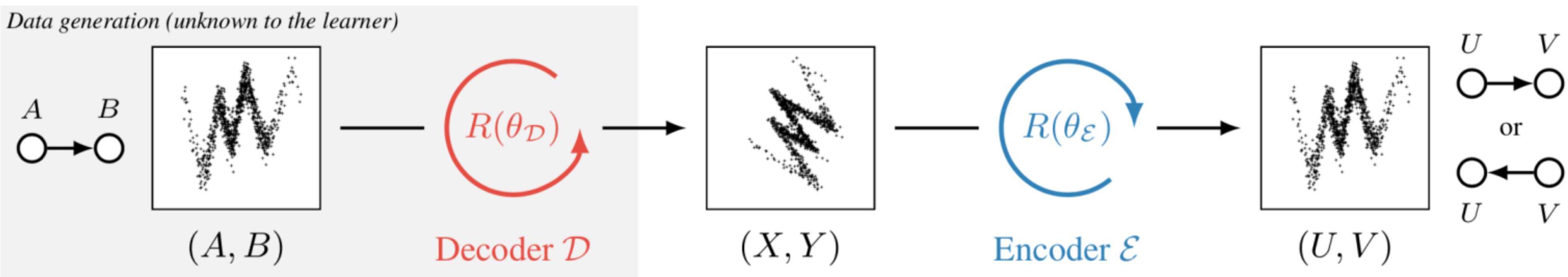


Ke et al 2019, 2020; Brouillard et al NeurIPS 2020

Mila

DISENTANGLING THE CAUSES

- Realistic settings: causal variables are not directly observed.
- Need to learn an encoder which maps raw data to causal space.
- Consider both the encoder parameters and the causal graph structural parameters as meta-parameters trained together wrt proposed meta-transfer objective.



- Simplest possible scenario: linear mixing (rotating decoder) and unmixing (rotating decoder)

DISCOVERING LARGER CAUSAL GRAPHS

Learning Neural Causal Models from Unknown Interventions

Ke, Bilaniuk, Goyal, Bauer, Scholkopf, Larochelle, Pal & Bengio 2019 arXiv:1910.01075

See also ***Brouillard et al NeurIPS 2020***

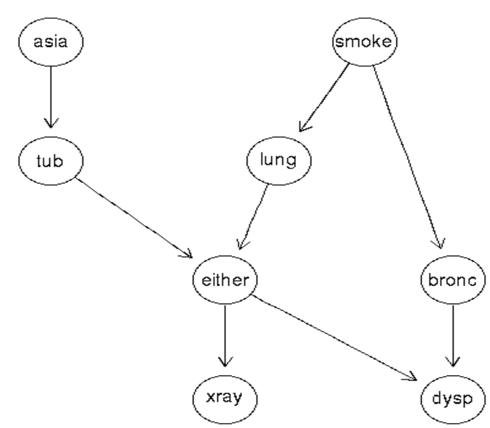
- Learning small causal graphs, avoid exponential explosion of # of graphs by parametrizing factorized distribution over graphs

- With enough observations of changes in distribution: perfect recovery of the causal graph without knowing the intervention; converges faster on sparser graphs

- Inference over the intervention:
  faster causal discovery

Asia graph, CE on ground truth edges, comparison against other causal induction methods

Our method	(Eaton & Murphy, 2007a)	(Peters et al., 2016)	(Zheng et al., 2018)
0.0	0.0	10.7	3.1

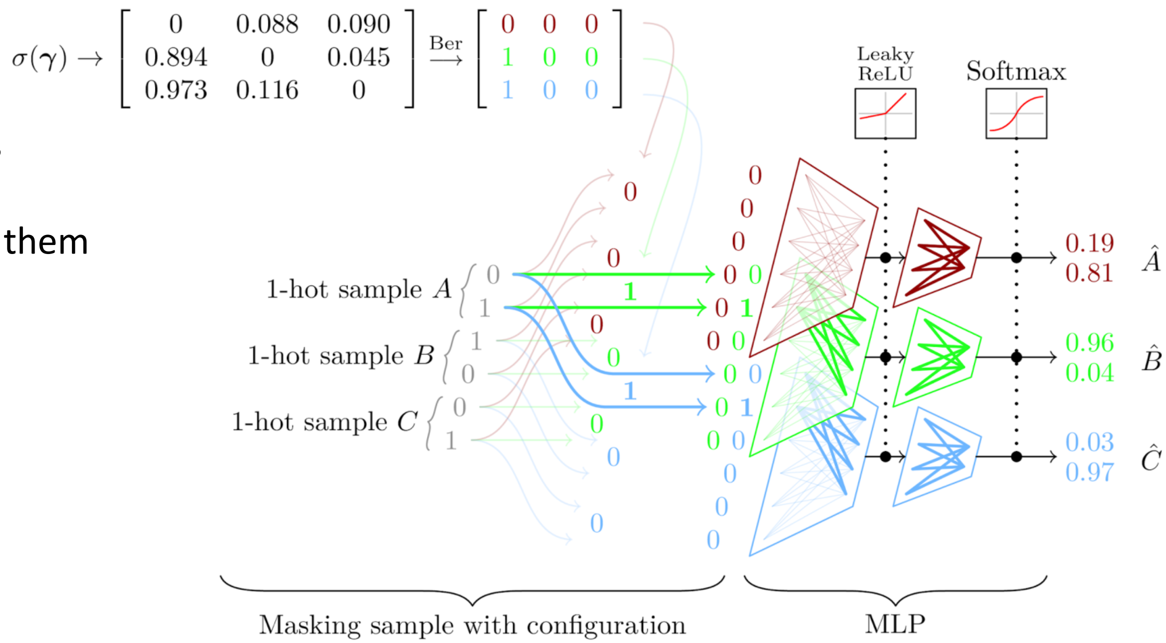


MODEL ARCHITECTURE

Use N neural networks to represent causal graph with N variables

Each neural network models:

- Who are the direct causal parents
  - *Structural parameters*
- What is the relationship between them
  - *Functional parameters*



RIMS: MODULARIZE COMPUTATION AND OPERATE ON SETS OF NAMED AND TYPED OBJECTS

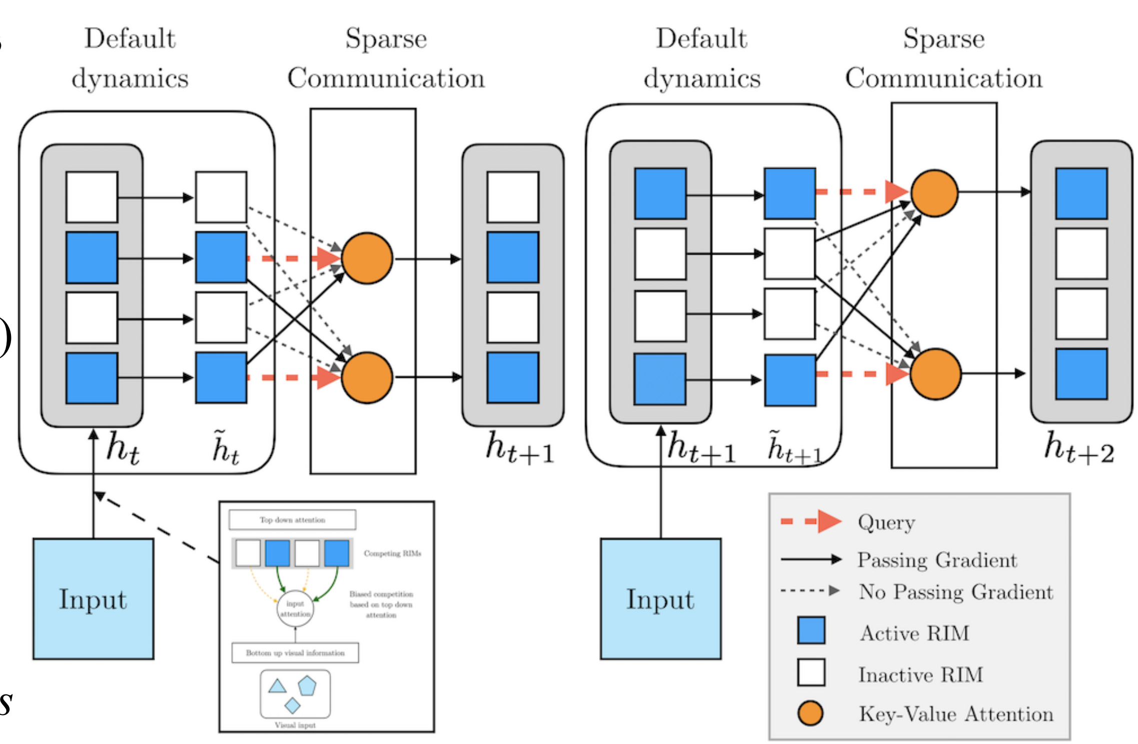
Goyal et al 2019, arXiv:1909.10893, ICLR 2021

Recurrent Independent Mechanisms

Multiple recurrent sparsely interacting modules, each with their own dynamics, with object (key/value pairs) input/outputs selected by multi-head attention

Results: better ood generalization

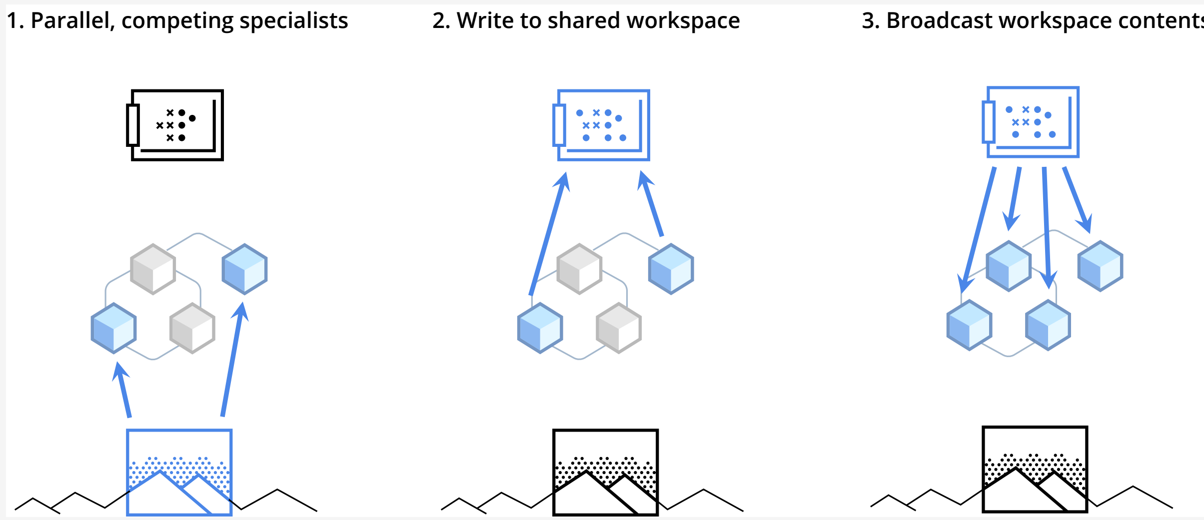
Ongoing work: hierarchy, top-down broadcasting, spatial layout of modules



Builds on rich recent litterature on object-centric representations (mostly for images)

Modules + Global Workspace

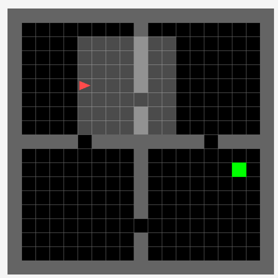
GWT: Baars 1997,…,
Dehaene et al 2017

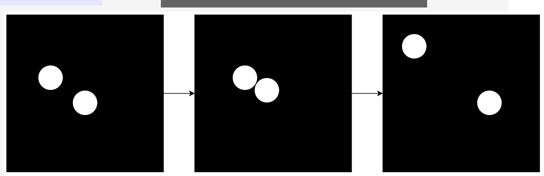


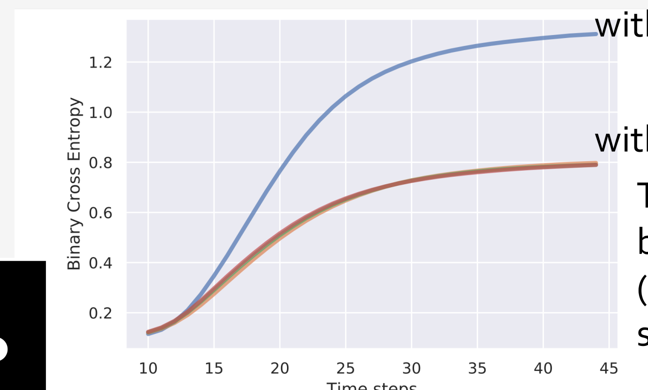
Adding to RIMS a shared global workspace similar to the GWT greatly improves OOD behavior

Table 2: **FourRoom Navigation Task:** Success Rate of the proposed method vs. the baselines on the FourRoom navigation environment illustrated on the right, with the agent in red, its field of visibility greyed out, and the object to get in green.

RIMs	RMC	LSTM	Ours
0.72 ± 0.02	0.67 ± 0.05	0.62 ± 0.02	0.96 ± 0.02







without GW

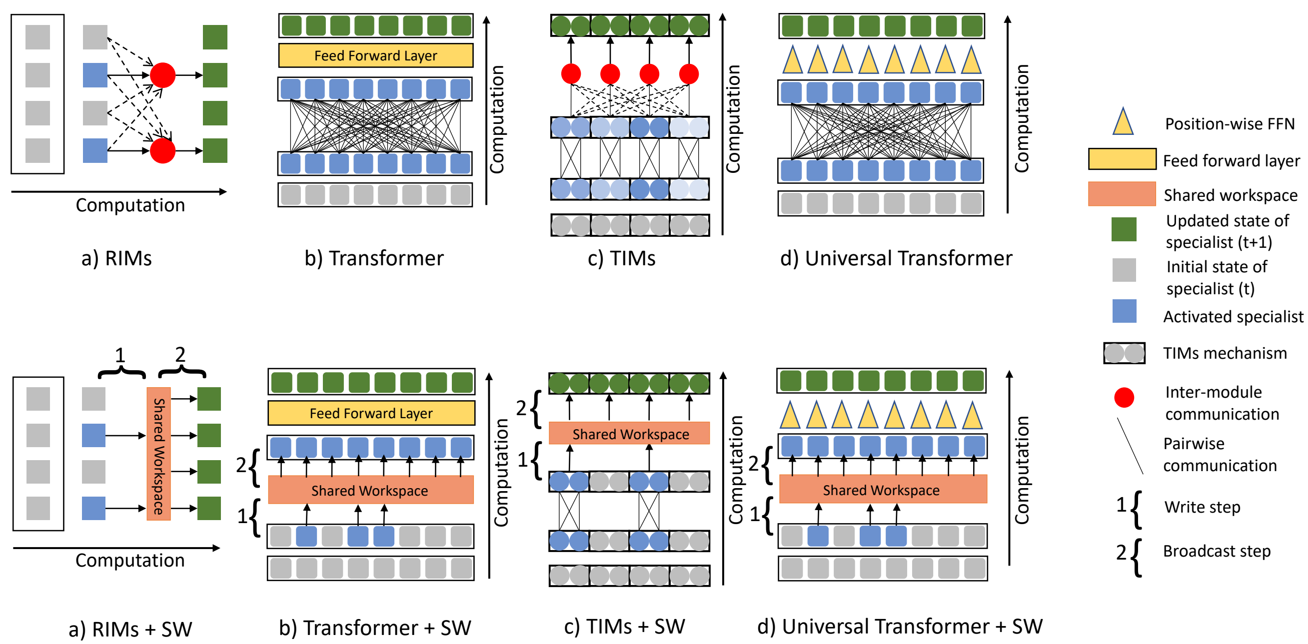
with GW

Tracking bouncing balls (Goyal et al 2021, submitted)

GLOBAL WORKSPACE ARCHITECTURE

Create global coherence through a communication bottleneck replacing full pairwise communication.

Activated specialists are denoted by a blue shade and the intensity depends on the degree of activation.



2-step process (1 and 2 in figures), bottom half:

1) specialists compete for write access to workspace, a subset of is activated (in blue).
2) shared content broadcast to all the specialists.

SCHEMAS AND SLOTS

Separate values (slots) from rules (schemas)

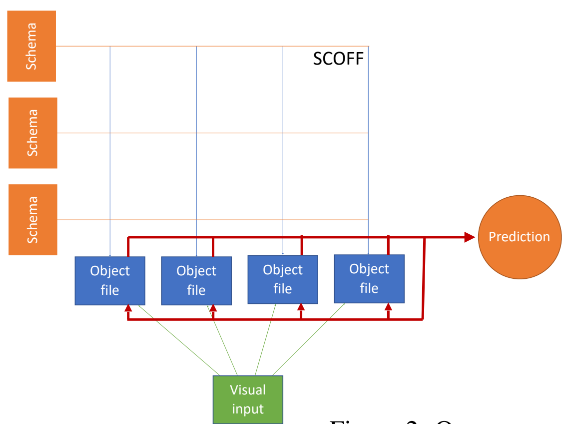


Figure 2: Our SCOFF model. Schemata are sets of parameters that specify the dynamics of objects. Object files are active modules that maintain the time-varying state of an object, seek information from the input, and select schemata for updating.

Object Files	Schema 1 Pacman	Schema 2 Normal Ghost	Schema 3 Scared Ghost
Top Frame			
A	✓		
B		✓	
C		✓	
D		✓	
E		✓	
Bottom Frame			
A	✓		
B			✓
C			✓
D			✓
E			✓

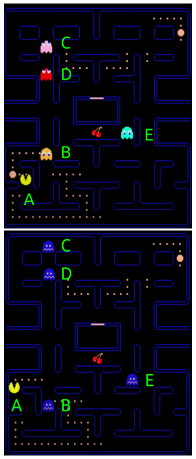


Figure 1: As a motivating example, we show two successive frames of the game PacMan and show how procedural and declarative knowledge must be dynamically factorized. The "B" ghost has a persistent object file (with its location and velocity), yet its procedure mostly depends on whether it is in its *scared* or *normal* routine.

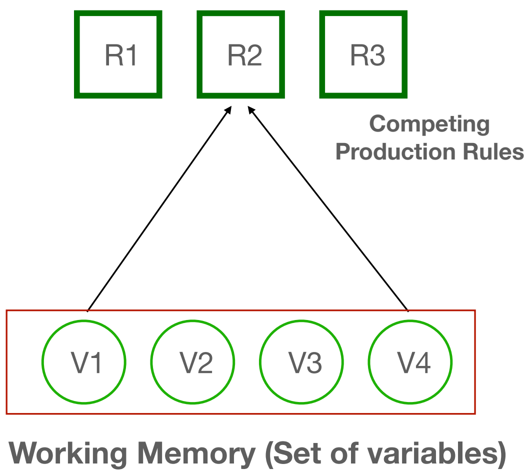
Object Files and Schemata: factorizing declarative and procedural knowledge in dynamical systems
Lamb, Goyal, Blundell, Mozer, Beaudoin, Levine & Bengio, ICLR 2021

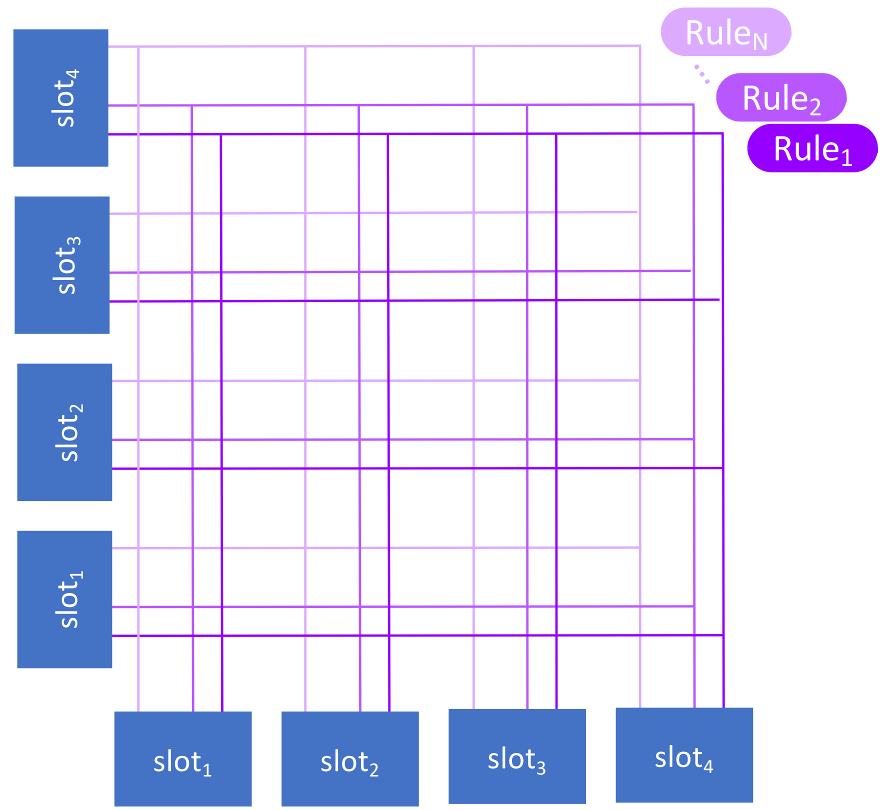
NEURAL PRODUCTION SYSTEMS

Mechanisms (rules) only take 1, 2 or 3 arguments and modify one of their arguments.

Sequentially trigger only one mechanism at a time which best fits with a subset of variables in working memory

Each rule takes 2 or more arguments, evaluate more or less greedy selection procedures

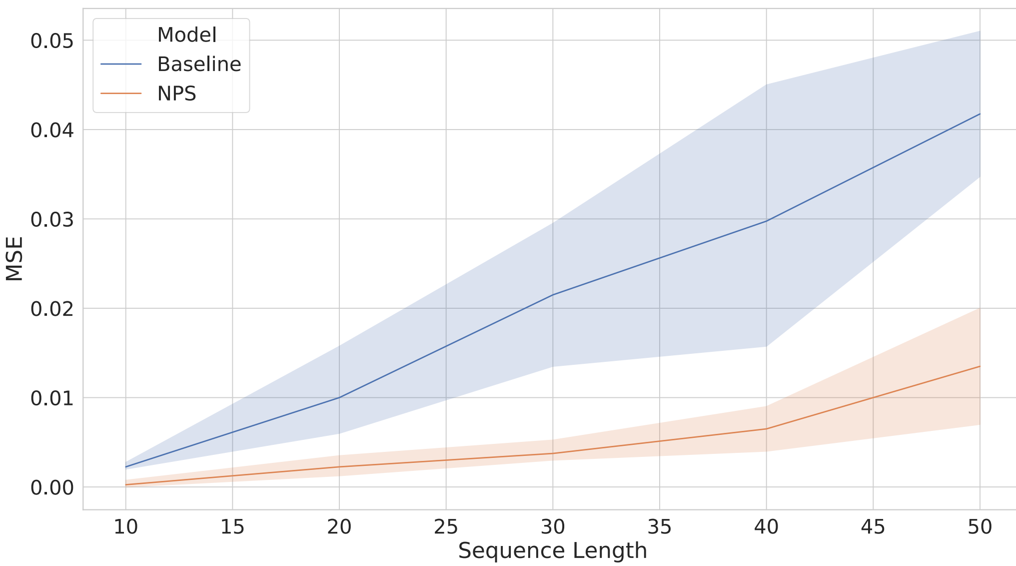




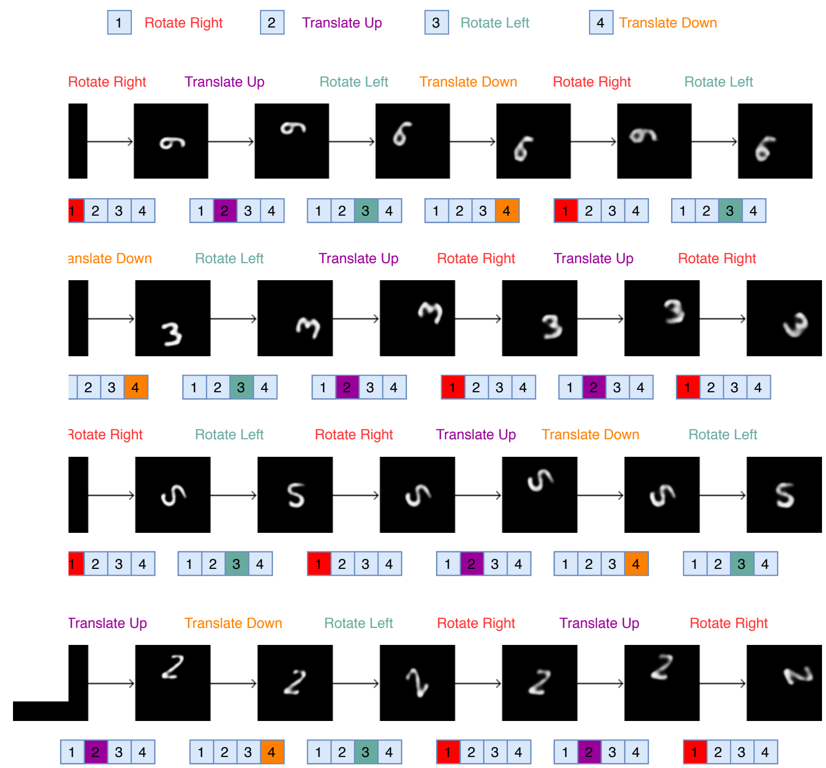
Goyal et al 2021, submitted

NPS TOY EXPERIMENTS

Learn to parse and compute Reverse Polish Notation sequences. Baseline = GRU RNN.



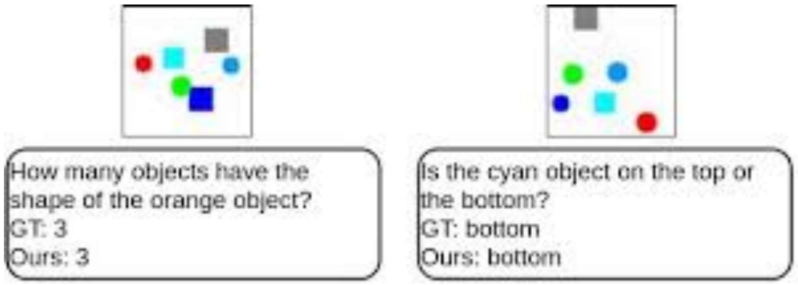
Learn to discover, disentangle and apply geometric transformations to MNIST digits



NPS disentangles the three underlying operations (**+**, **×**, **-**)

Each rule converges to one of the underlying operations

DISCRETE-VALUED NEURAL COMMUNICATION



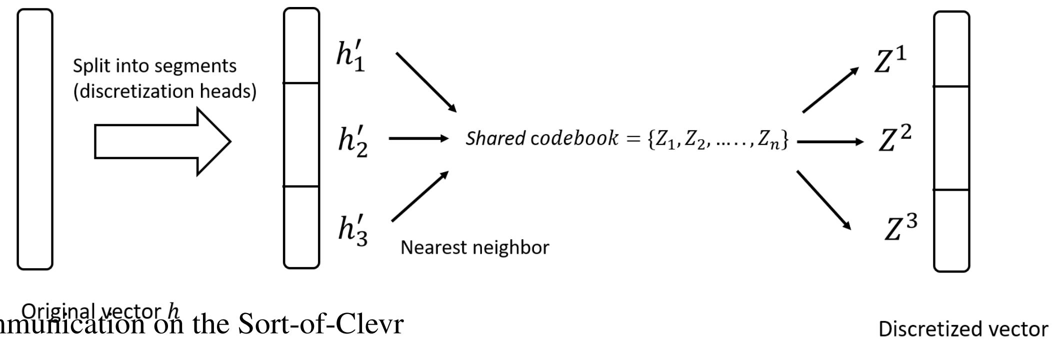
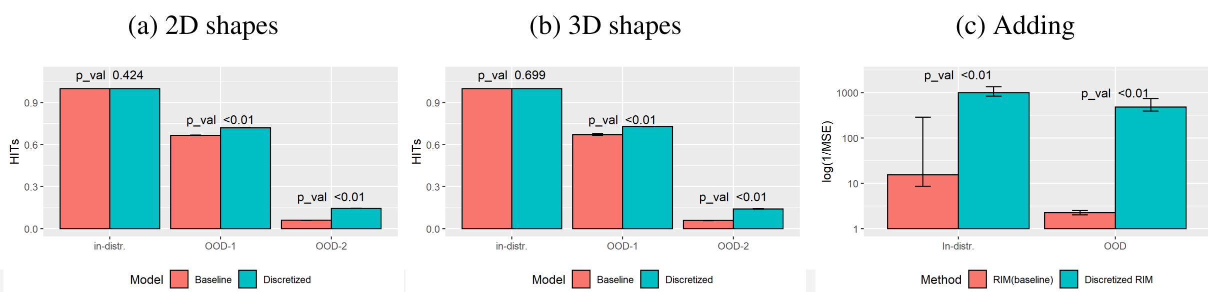


Table 2: Performance of transformer models with discretized communication on the Sort-of-Clevr visual reasoning task.

Method	Ternary Accuracy	Binary Accuracy	Unary Accuracy
Transformer baseline	57.25 ± 1.30	76.00 ± 1.41	97.75 ± 0.83
Discretized transformer (G=16)	61.33 ± 2.62	84.00 ± 2.94	98.00 ± 0.89
Discretized transformer (G=8)	62.67 ± 1.70	88.00 ± 0.82	98.75 ± 0.43
Discretized transformer (G=1)	58.50 ± 4.72	80.50 ± 7.53	98.50 ± 0.50

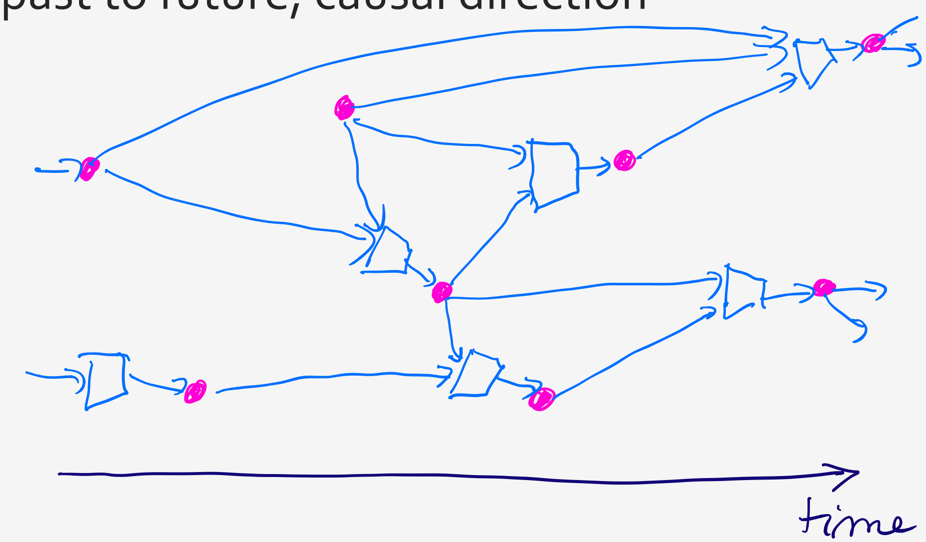


- Modular architectures (transformers, RIMs, GNNs)
- Quantize value vector in attention mechanism
- Each attention head uses a different code, but from same codebook
- Better OOD generalization

(Liu et al, submitted, 2021)

Causal reasoning over events factor graph

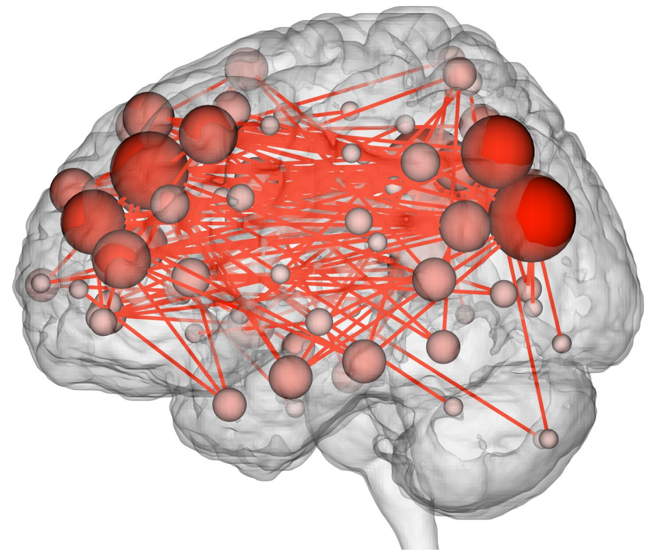
- Node of graph = event at particular time, involving a small set of variables
  - Content of episodic memory
- Factor = causal mechanism
  - Generic knowledge about a few high-level variables, cortical module
- Directed edges: from past to future, causal direction



Mila

LEARNING TO REASON & PLAN

- Reasoning, long-range credit assignment and planning are inference, inherently computationally expensive
- Brains do not use exhaustive search but instead **generate** good candidates
- Conscious processing seems involved in evaluating them for global coherence across the brain's modules
- Attention mechanisms are part of the reasoning policy, converting declarative knowledge into selective computations for inference and decision-making



CONTRAST WITH **THE SYMBOLIC AI PROGRAM**

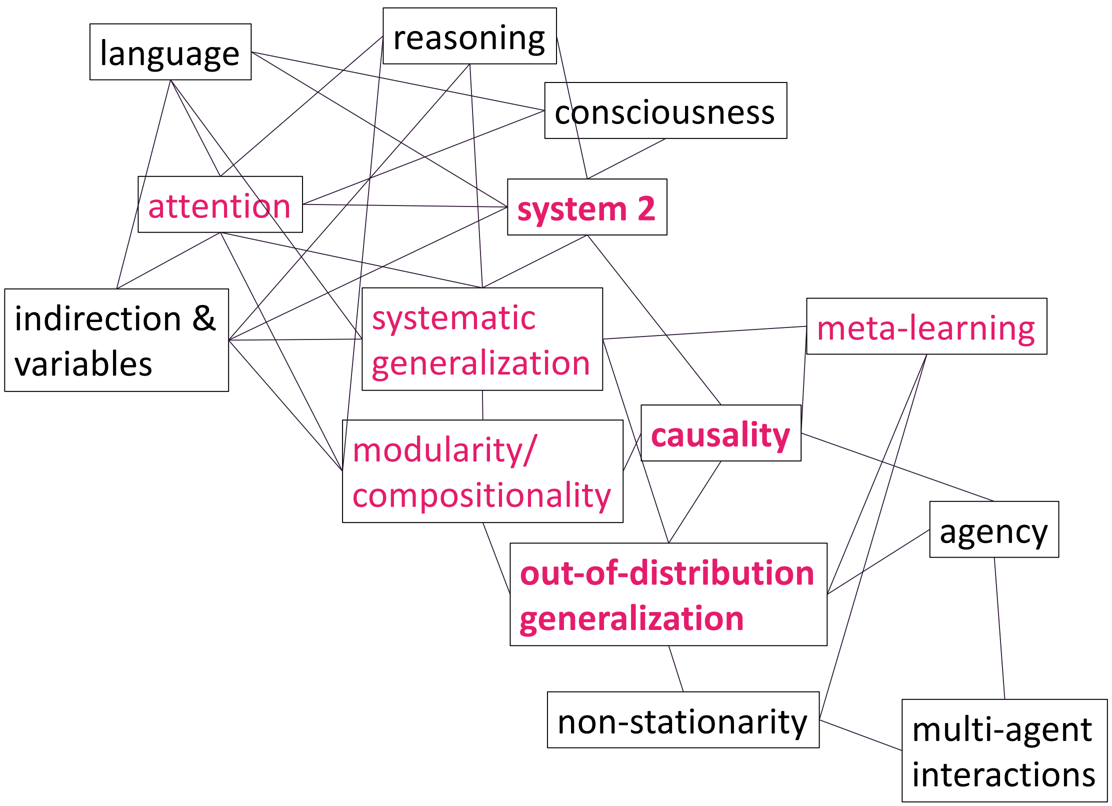
Avoid pitfalls of classical AI rule-based symbol-manipulation

- Need efficient large-scale learning
- Need semantic grounding in system 1 (implicit knowledge)
- Need distributed representations for generalization
- Need efficient = trained search (also system 1)
- Need uncertainty handling

But want

- Systematic generalization
- Factorizing knowledge in small exchangeable pieces
- Manipulating variables, instances, references & indirection

CONSCIOUSNESS PRIORS



- Sparse factor graph in space of high-level semantic variables
- Semantic variables are causal: agents, intentions, controllable objects
- Many of these variables are discrete
- Simple mapping between high-level semantic variables / thoughts and words / sentences
- Shared 'rules' across instance tuples (as arguments), w/ variables & indirection
- Distributional changes due to localized causal interventions (in semantic space)
- Meaning (e.g. grounded by an encoder) is stable & robust wrt changes in distribution
- Credit assignment is only over short causal chains

SOME OPEN QUESTIONS WHICH COULD USE BRAIN INSPIRATION

1. How to jointly learn the encoder, the inference machinery, the mechanisms and how they form an explanatory graph?
2. How to handle ambiguous abstract variables (given sensors) and manage the resulting inference?
3. How to jointly learn the tied abstract variable space and abstract action space?
4. How to learn an inference & attention policy which selects what event / object / attribute to attend?
   - How to combine system 1 habitual inference (VAE-like?) with system 2 iterative inference (MCMC?)?
5. What heuristics to exploit short-term and long-term memory to rapidly select relevant entities, events, agents, objects and causal mechanisms for inference and credit assignment?
6. How to efficiently search / plan in the space of abstract actions anchored on abstract events?
   - How to generate interesting relevant hypothetical explanatory graphs & plans?
7. How to efficiently perform credit assignment across long time spans through the causal graph?

THANK YOU!