Active Inference in String Diagrams

Sean Tull

QUANTINUUM

Association for Mathematical Consciousness Science

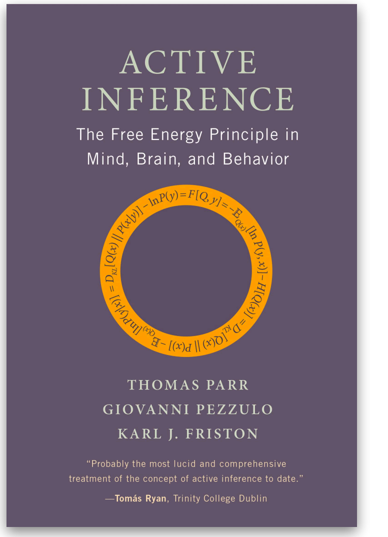
Paris Mathematical Models of Cognition and Consciousness Seminar

Sorbonne Université, 29th May 2024

TOPOS INSTITUTE

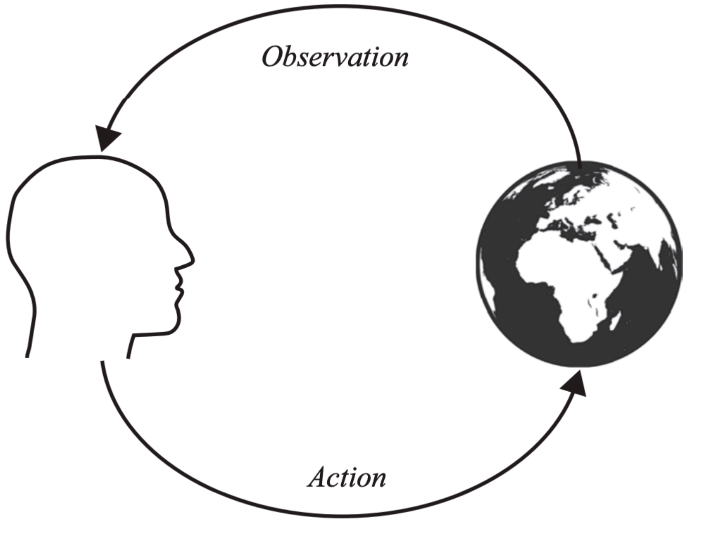
FQxI

Active Inference



Model of cognition applicable from single neuron to whole organism.

Agent comes with a **generative model**:



used to explain observations (**perception**) and choose **actions.**

Free Energy Principle: Achieved by approximate **Bayesian inference** through minimising **Free Energy.**

Image: T. Parr, G. Pezzulo, K. J. Friston. *Active inference: the free energy principle in mind, brain, and behavior*. 2022.

Formalising Active Inference

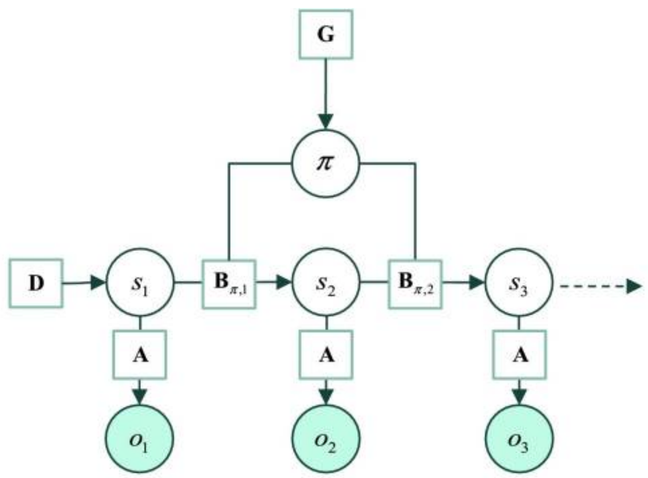
Further **formalisation** of active inference would help to:

- Clarify the 'core' of the theory
- Generalise the framework
- Make accessible to those with formal backgrounds and in AI

Most importantly, a clear conceptualisation should make active inference **simpler**.

A Diagrammatic Approach?

Generative models are highly **compositional** and naturally described in **diagrams**.
There have been calls to formalise active inference **graphically.**




The graphical brain: Belief propagation and active inference

Karl J. Friston,[1,*] Thomas Parr,[1] and Bert de Vries[2,3]

There is a well-established graphical formalism for processes and composition:

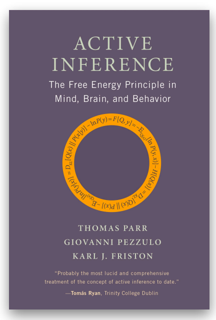
Category theory and the language of **string diagrams.**

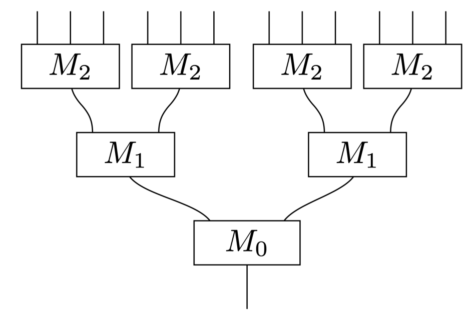
Several recent categorical treatments of **probability theory** and **causal models**.
We use (*Causal Models in String Diagrams,* Robin Lorenz, ST 2023).

This Work

Active Inference in String Diagrams: A Categorical Account of Predictive Processing and Free Energy

Sean Tull[1,2], Johannes Kleiner[2,3,4], and Toby St Clere Smithe[5,6]





Formalise active inference categorically via string diagrams.

Part of FQXi project on categorical approaches to **consciousness.**

Also related: **categorical cybernetics.**

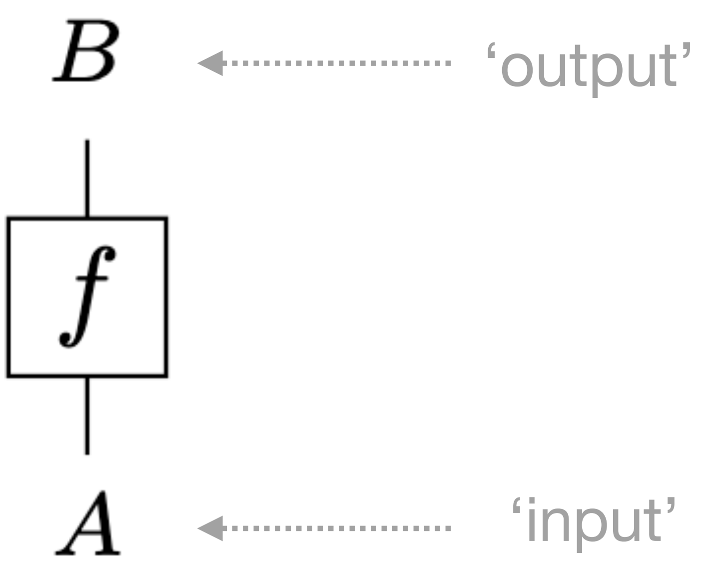
FQXi Project: *Categorical Theories of Consciousness: Bridging Neuroscience and Fundamental Physics.* Johannes Kleiner, ST, Quanlong Wang, Bob Coecke

Category Theory

and String Diagrams

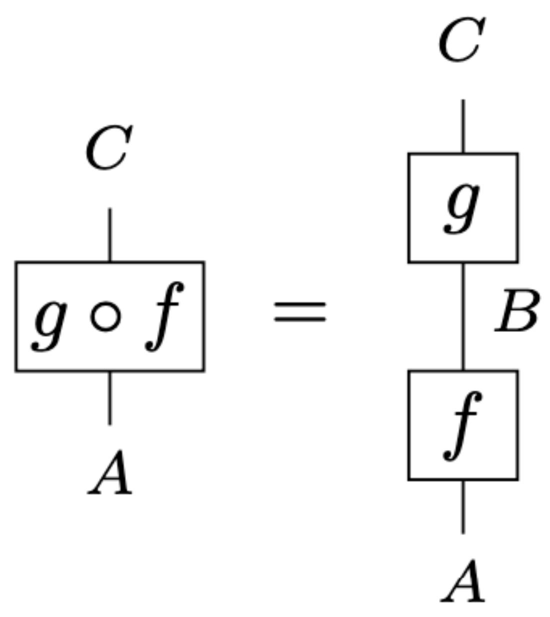
Categories

A **symmetric monoidal category** $\mathbf{C}$ consists of a collection of **objects** $A, B, C \ldots$ and **morphisms** or **processes** written $f \colon A \to B$ and depicted:

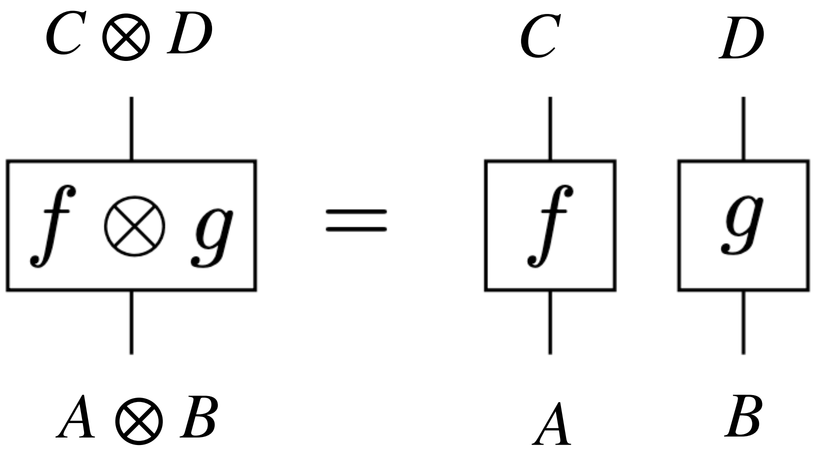


Categories

We can compose 'in sequence':

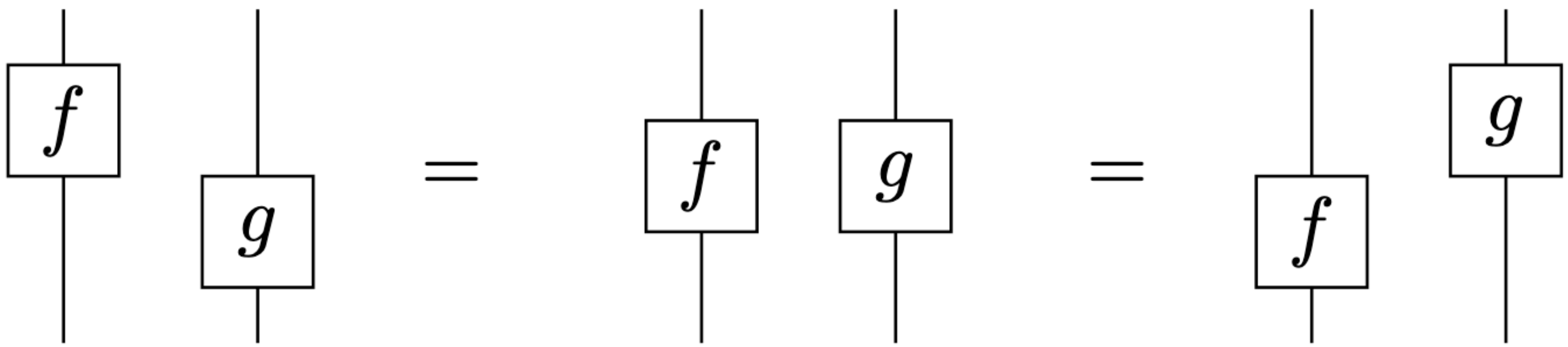


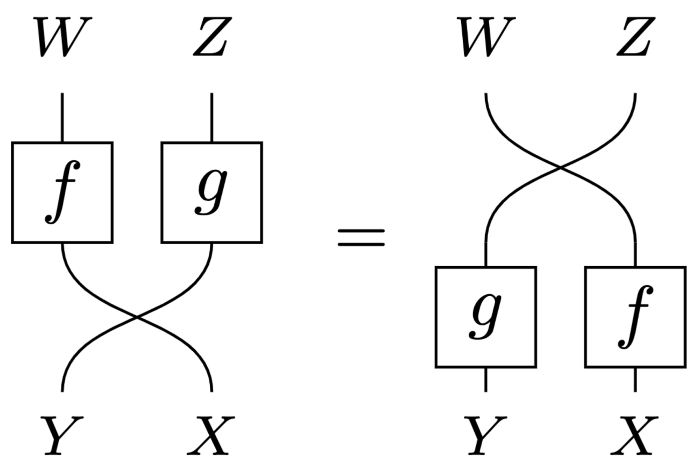
and 'in parallel' using the **tensor** of objects $A, B \mapsto A \otimes B$ and morphisms:



String Diagrams

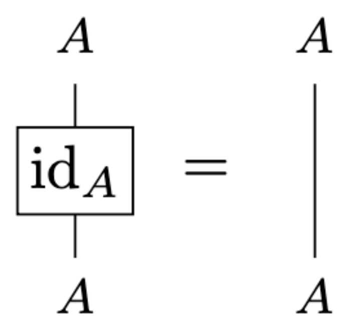
Categories satisfies various equations that come ‘for free’ in the diagrams:

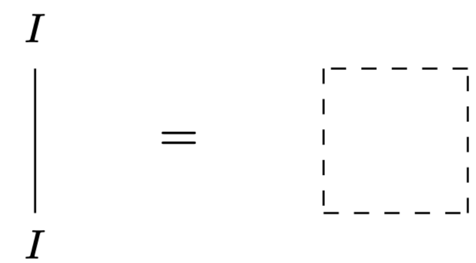




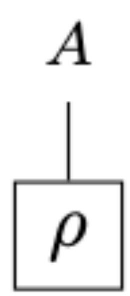
Categories

Every object has an **identity** morphism drawn as a blank wire, and there is a **unit object** I drawn as 'empty space':



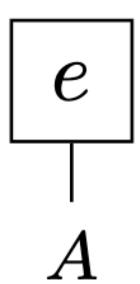


This lets us have morphisms with 'no' input or output:



State

$\rho \colon I \to A$



Effect

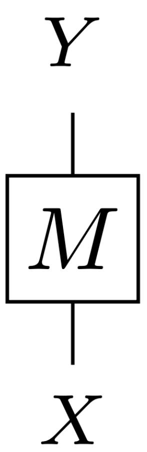
$e \colon A \to I$

Scalar

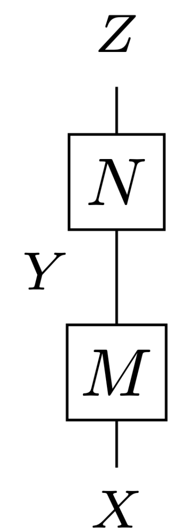
$r \colon I \to I$

Example: $\mathbf{Mat}_{\mathbb{R}^+}$

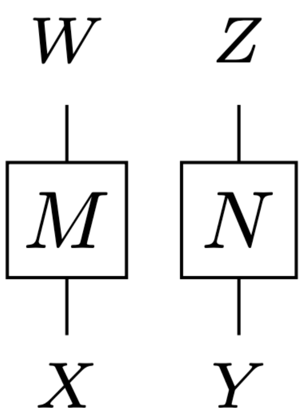
In the category $\mathbf{Mat}_{\mathbb{R}^+}$ objects are finite sets $X, Y\ldots$ and morphisms are positive matrices, with $X \otimes Y = X \times Y$.



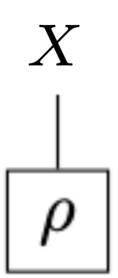
$$(x, y) \mapsto M(y \mid x) \in \mathbb{R}^+$$



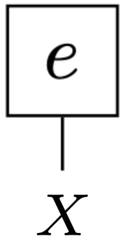
$$(x, z) \mapsto \sum_{y \in Y} N(z \mid y) M(y \mid x)$$



$$((x, y), (w, z)) \mapsto M(w \mid x) N(z \mid y)$$



$$x \mapsto \rho(x)$$



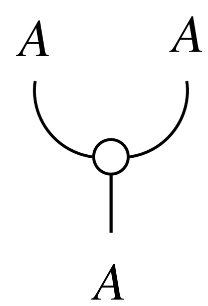
$$x \mapsto e(x)$$

$$r \in \mathbb{R}^+$$

Copying and Discarding

In a **copy-discard** (cd-)**category** each object comes with distinguished morphisms:

$(a, b, c) \mapsto \delta_{a,b,c}$

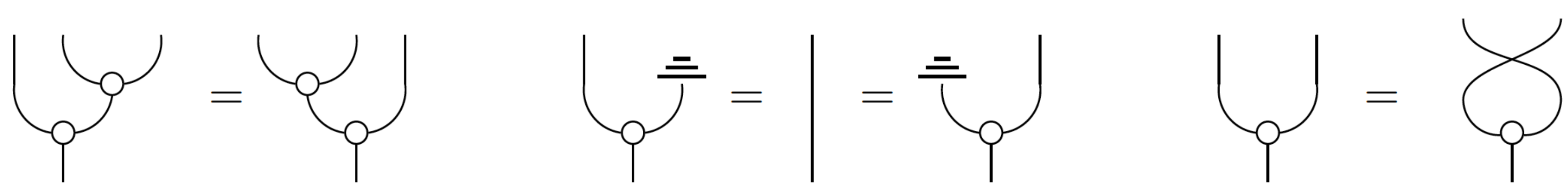


copy

discard

$a \mapsto 1$

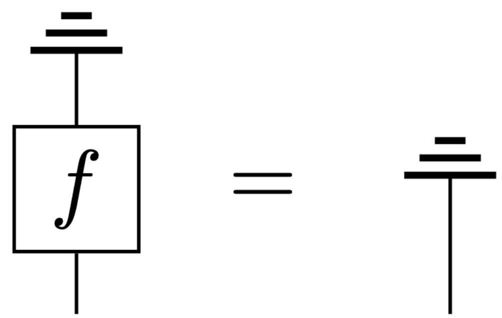
satisfying:



Major area of research in treating **probability theory** via cd-categories.

Categorical Probability

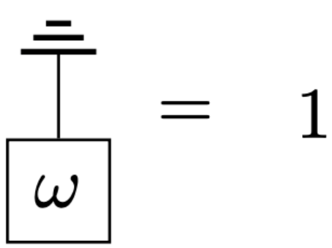
A **channel** is a morphism which preserves discarding:



$$\sum_y f(y \mid x) = 1$$

Probability channel
(Stochastic matrix).

A state ω is **normalised** when:

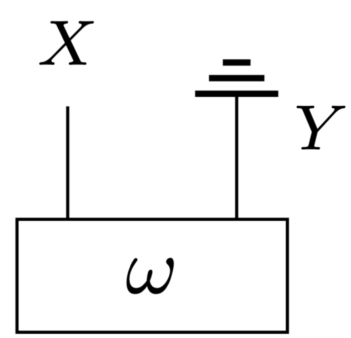


$$x \mapsto \omega(x)$$

Probability distribution

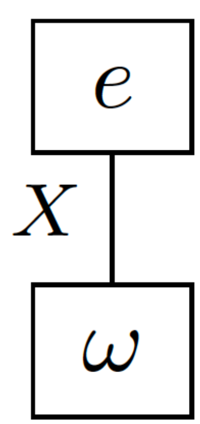
Categorical Probability

We can **marginalise** processes:



$$x \mapsto \sum_{y \in Y} \omega(x, y)$$

Composing a state and effect gives the **expectation** value:



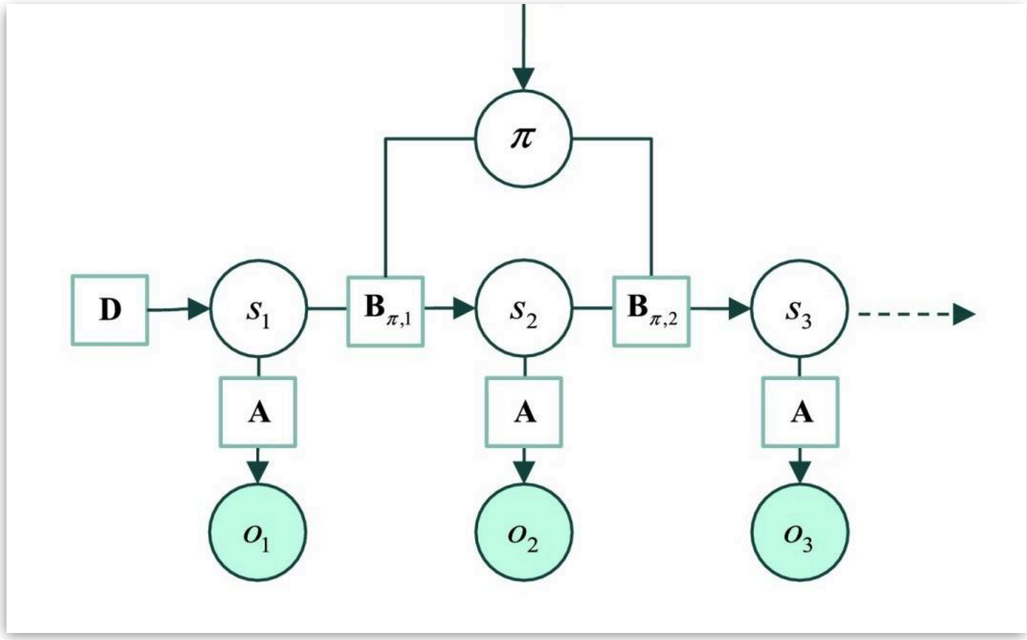
$$\mathop{\mathbb{E}}_{x \sim \omega}[e]$$

Generative Models

Generative Models

An agent uses a generative model relating actions, observations and world states.

Usually a (**causal**) **Bayesian network**: a DAG *G* with probability channels $P(X_i \mid \mathrm{Pa}(X_i))$.



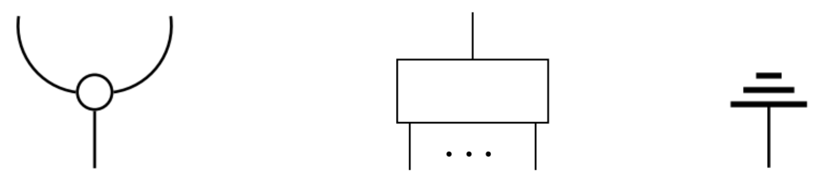
$$P(V) = \prod_i P(X_i \mid \mathrm{Pa}(X_i))$$

But active inference literature is independently converging on string diagrams!

Image: K. J. Friston, T. Parr, and B. de Vries. *The graphical brain: belief propagation and active inference*". 2017.

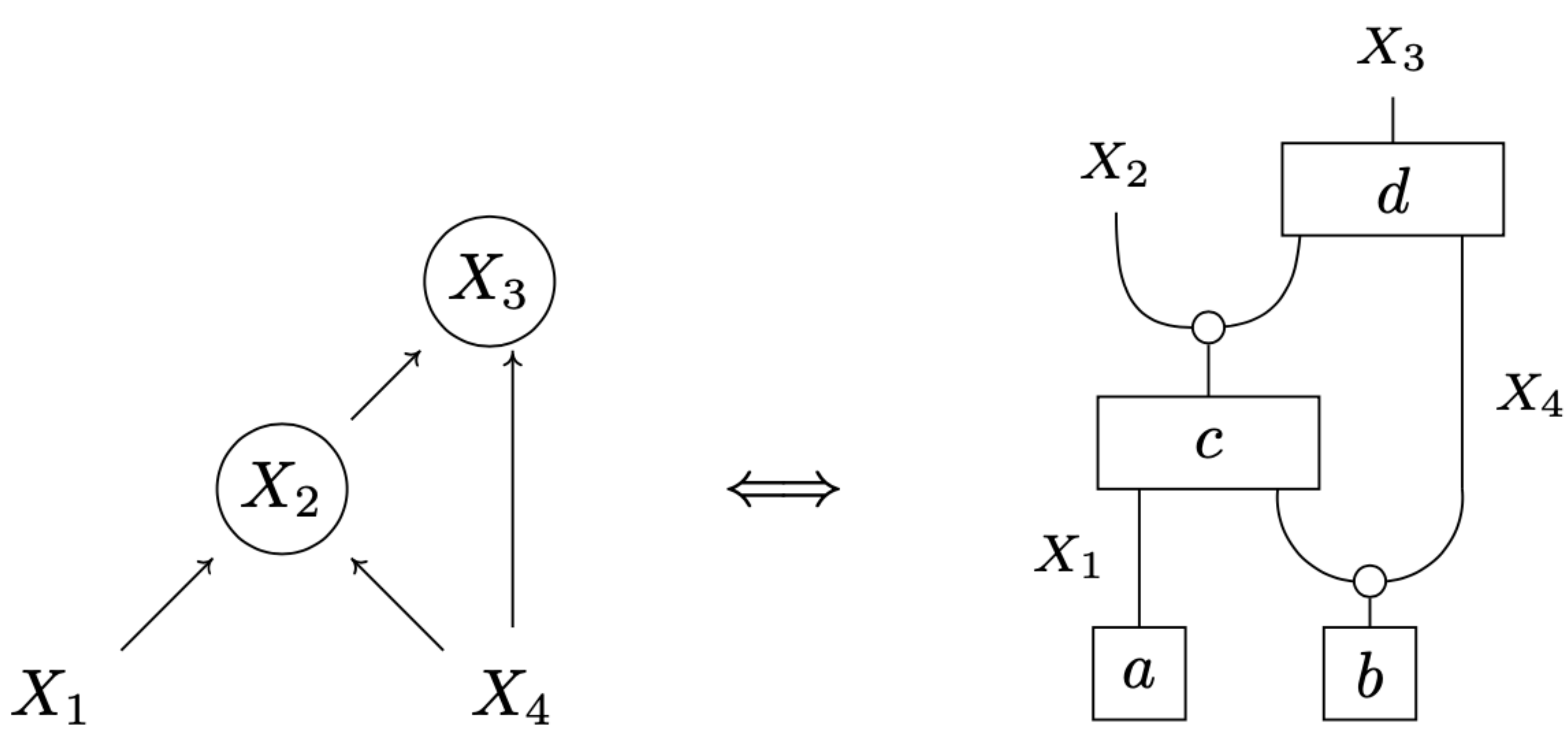
DAGs as Diagrams

A **network diagram** is a string diagram built from copy, single output boxes and discarding such that each wire appears as an output or input to any box at most once.



Proposition

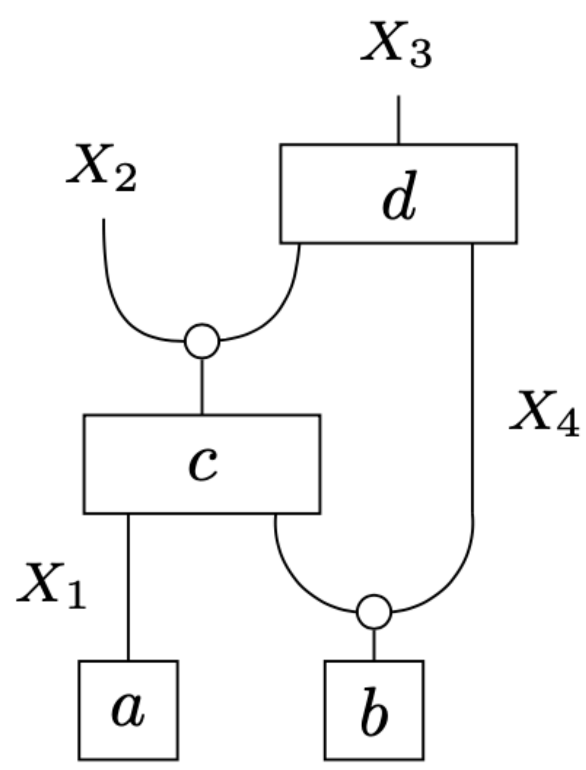
A DAG G with chosen outputs O is equivalent to a network diagram with outputs O and no inputs.



(B. Fong 2013), (B. Jacobs, A. Kissinger, F. Zanasi 2018).

Generative Models

A **generative model** $\mathbb{M}$ in a cd-category $\mathbf{C}$ is a network diagram with no inputs, along with a **representation** as objects and channels in $\mathbf{C}$.



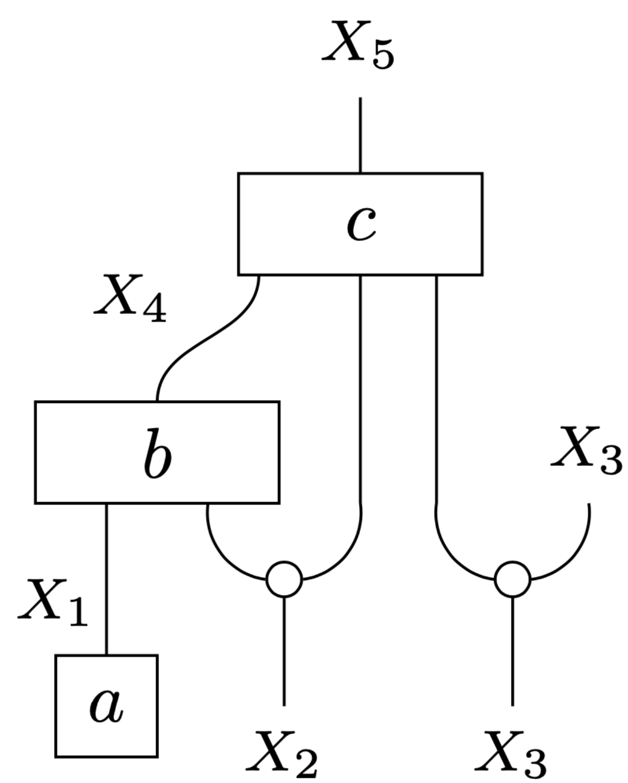
Outputs to the diagram are **observed** variables, the rest are **hidden**.

Example

In $\mathbf{Mat}_{\mathbb{R}^+}$: a causal Bayesian Network.

Open Generative Models

An **open generative model** $\mathbb{M}$ in a cd-category $\mathbf{C}$ is a network diagram, along with a **representation** as objects and channels in $\mathbf{C}$.



Equivalent to an *open causal model* in $\mathbf{C}$.

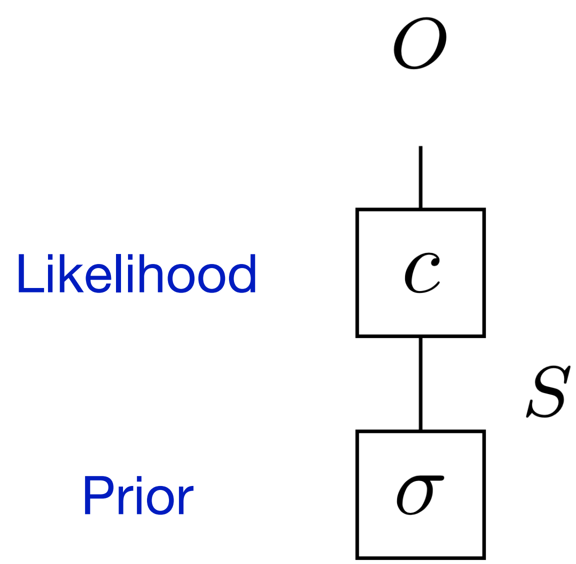
Example

In $\mathbf{Mat}_{\mathbb{R}^+}$: a causal Bayesian Network, with optional inputs.

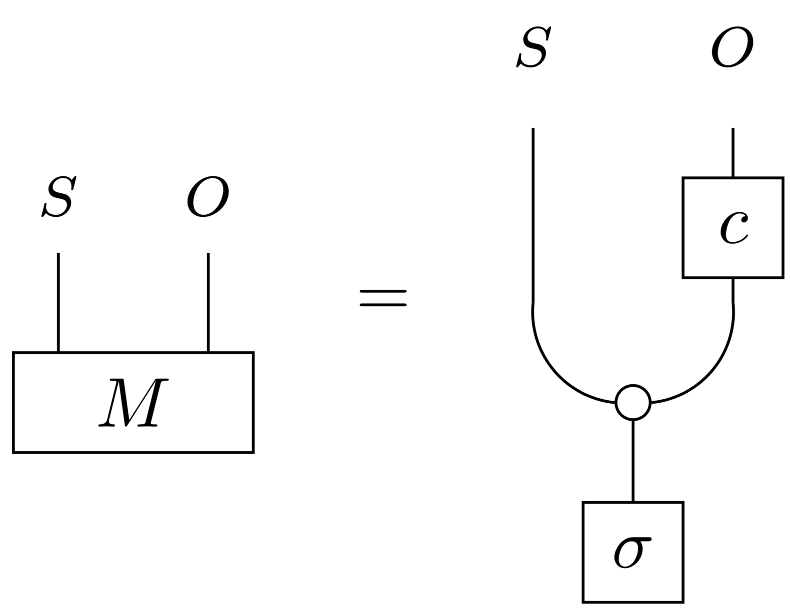
R. Lorenz, ST. *Causal models in string diagrams.* 2023.

A Simple Generative Model

A generative model $\mathbb{M}$ of how hidden **states** S lead to **observations** O :

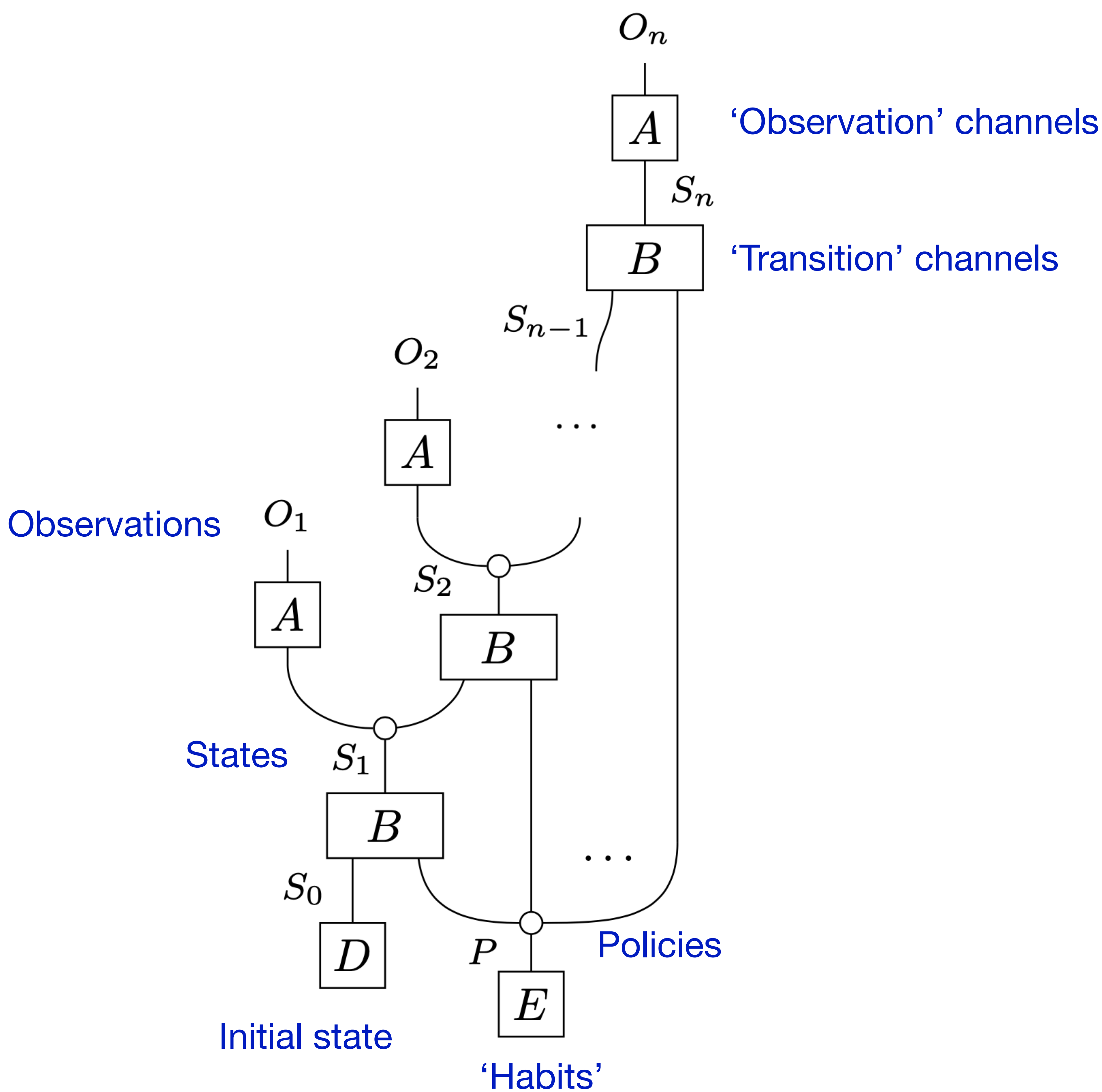


Induces a **total distribution** over S, O :



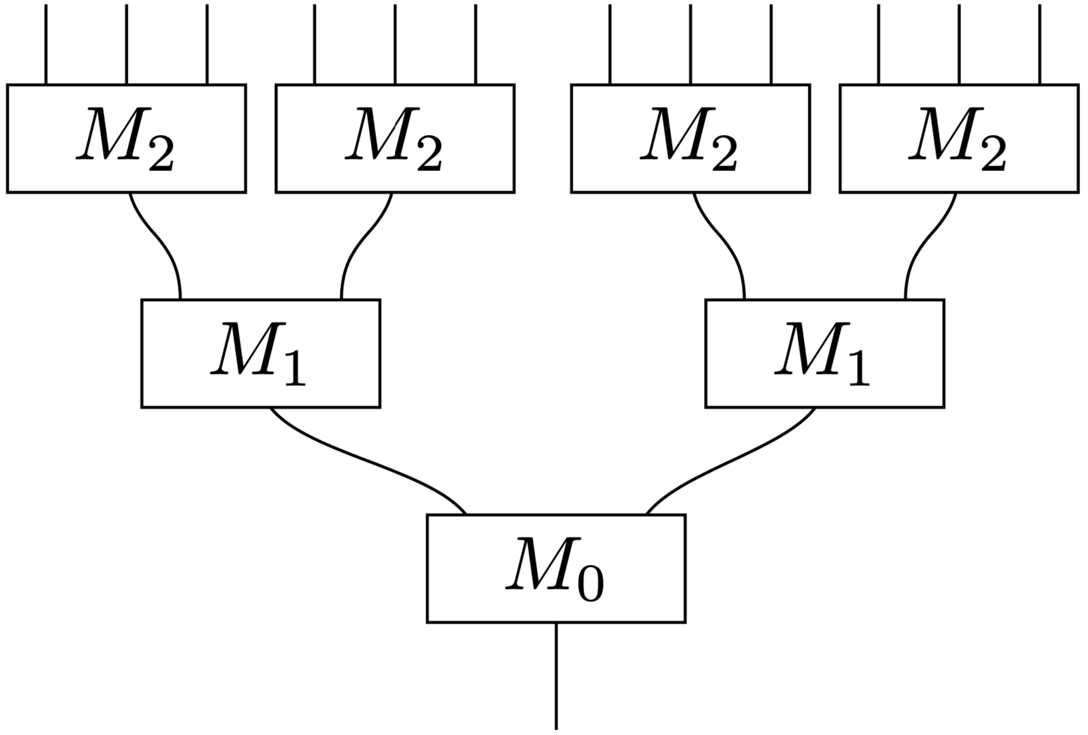
$$M(s, o) = c(o \mid s)\sigma(s)$$

Discrete Time Models



Hierarchical Models

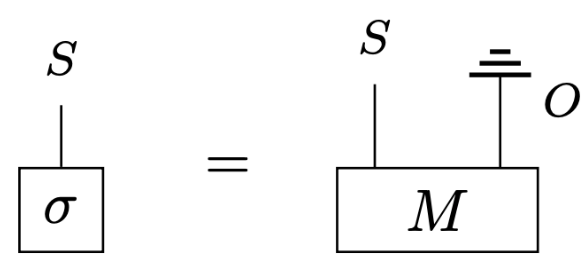
A typical model in active inference is given by composing open models in a **hierarchy**:



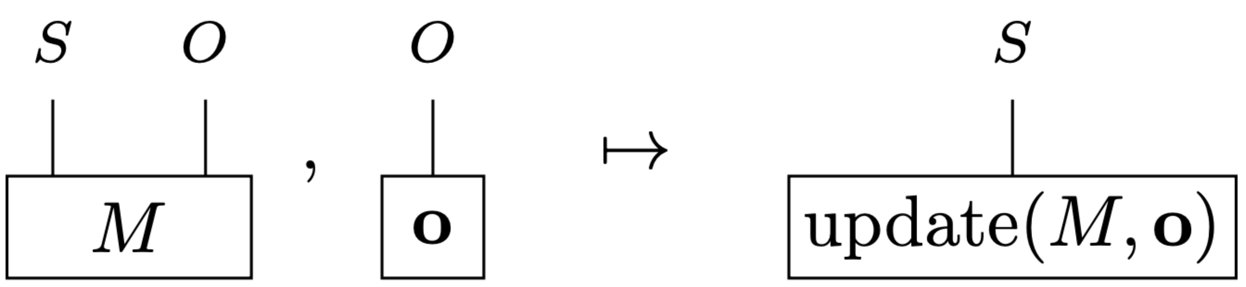
Updating Models

Updating

Suppose an agent has model $\mathbb{M}$ with prior beliefs about the hidden state:



Given a soft observation (distribution) they wish to **update** these beliefs:



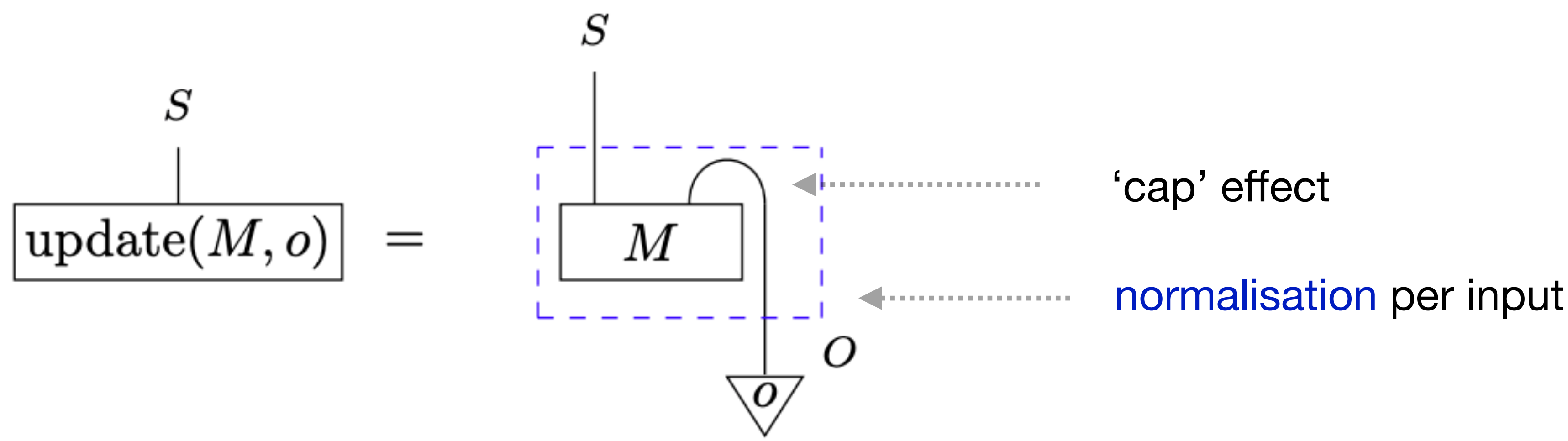
Examples

Perception = updating state S, given observation O

Planning = updating policies P, given future preferences F

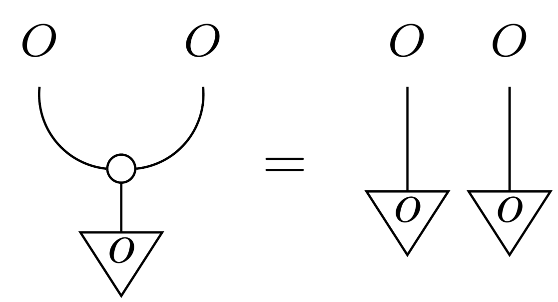
Sharp Updating

When an observation is sharp we ideally update by **Bayesian conditioning**:



$$s \mapsto \begin{cases} \frac{M(s,o)}{M(o)} & M(o) > 0 \\ 0 & \text{otherwise} \end{cases}$$

A distribution o is **sharp** when:

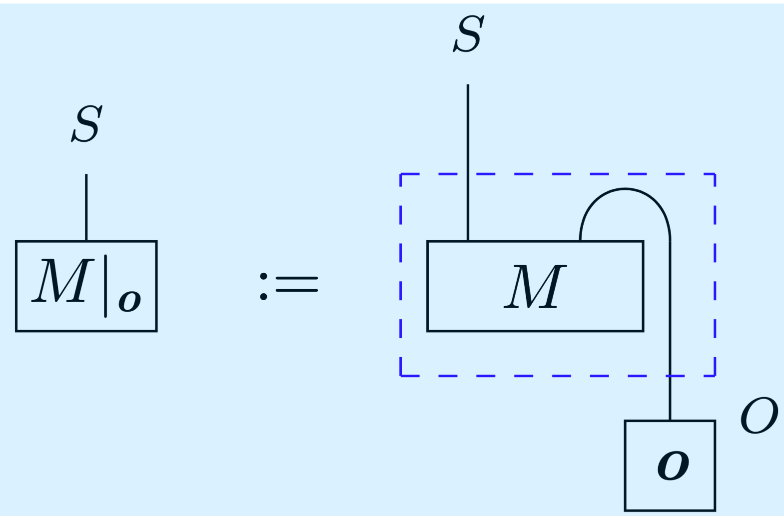


point distribution δ_o at $o \in O$

Soft Updating

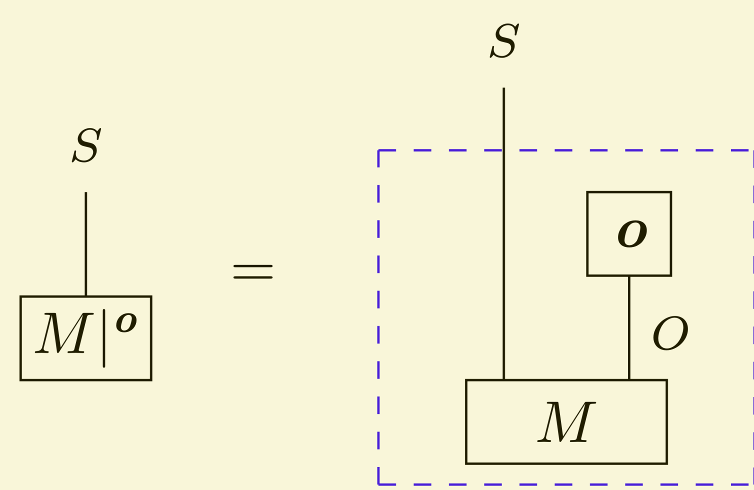
For soft observations there are two ways to update, which coincide for sharp $o \in O$:

Jeffrey's update:



$$M|_{\boldsymbol{o}}(s) = \sum_{o} \frac{M(s,o)\,\boldsymbol{o}(o)}{\sum_{s'} M(s',o)}$$

Pearl's update:



$$M|^{\boldsymbol{o}}(s) = \frac{\sum_{o} M(s,o)\,\boldsymbol{o}(o)}{\sum_{s',o'} M(s',o')\,\boldsymbol{o}(o')}$$

Hard to compute in practice!

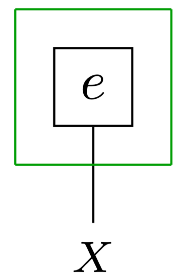
B. Jacobs. *The Mathematics of Changing one's Mind, via Jeffrey's or via Pearl's update rule*. 2019.

See also: E. Di Lavore, M. Román. *Evidential Decision Theory via Partial Markov Categories.* 2023.

Free Energy

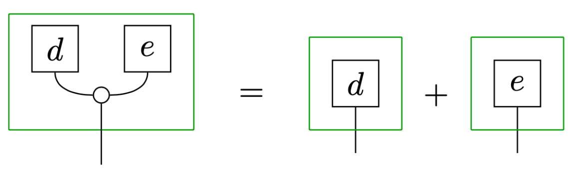
Log Boxes

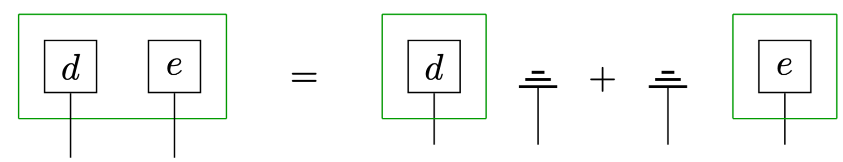
For any $e\colon X \to \mathbb{R}^+$ we depict the 'surprise' as:



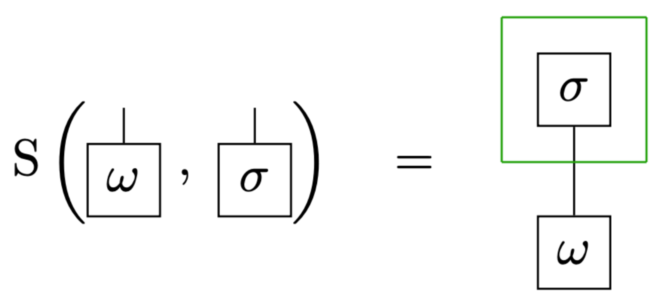
$$:: x \mapsto -\log e(x) \in (-\infty, \infty]$$

Properties of log give us graphical rules like:





The **surprise** of distribution σ relative to distribution ω is:



$$= \; - \mathop{\mathbb{E}}_{x \sim \omega} \log \sigma(x)$$

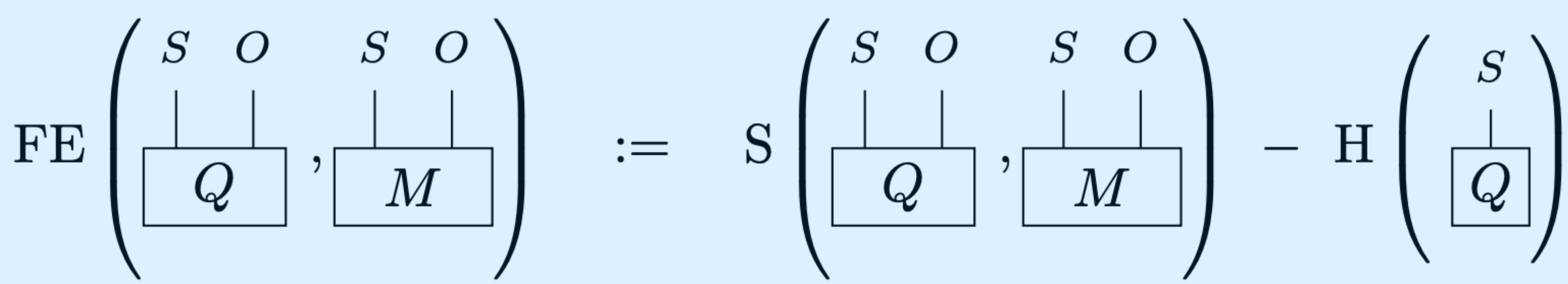
The **entropy** of ω is $H(\omega) = S(\omega, \omega)$.

The **KL-divergence** is $D(\omega, \sigma) = S(\omega, \sigma) - H(\omega)$.

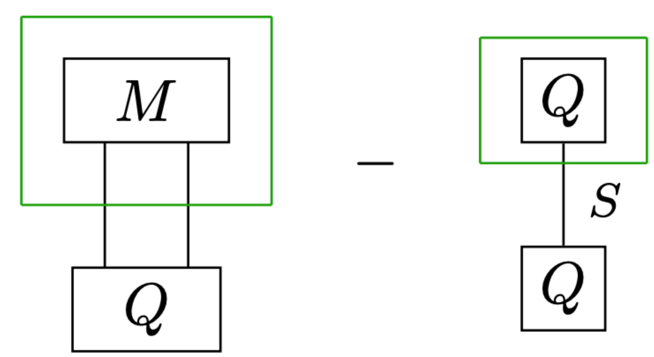
Free Energy

Let M be a distribution (induced by a generative model) over S, O.

The **Free Energy** of distribution Q over S, O is:



=

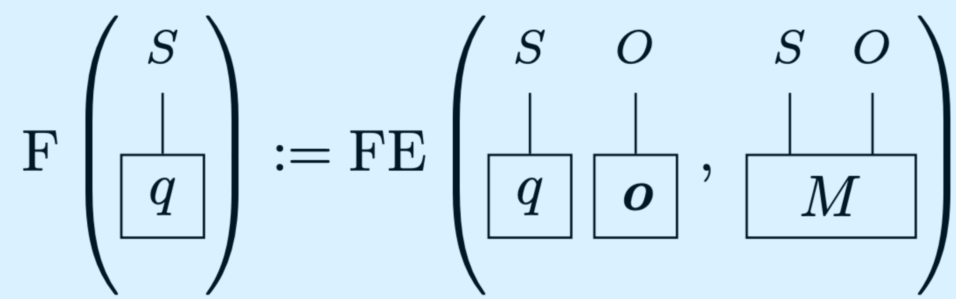


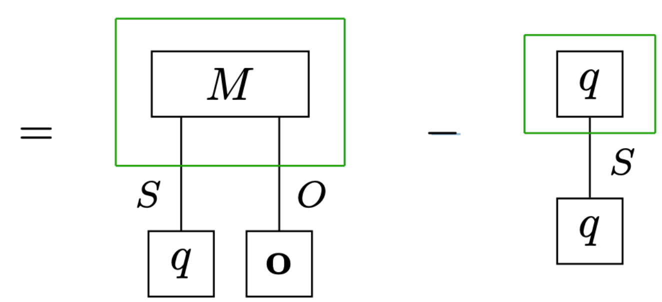
$$= \underset{(s,o)\sim Q}{\mathbb{E}} \left[-\log M(s,o) + \log Q(s) \right] \in \mathbb{R}^+$$

Variational Free Energy

Let $\mathbf{o}$ be a (soft) observation over O.

The **Variational Free Energy (VFE)** of 'beliefs' distribution q over S is:





$$\geq D(q, M|_{\mathbf{o}}) + S(\mathbf{o}, M)$$

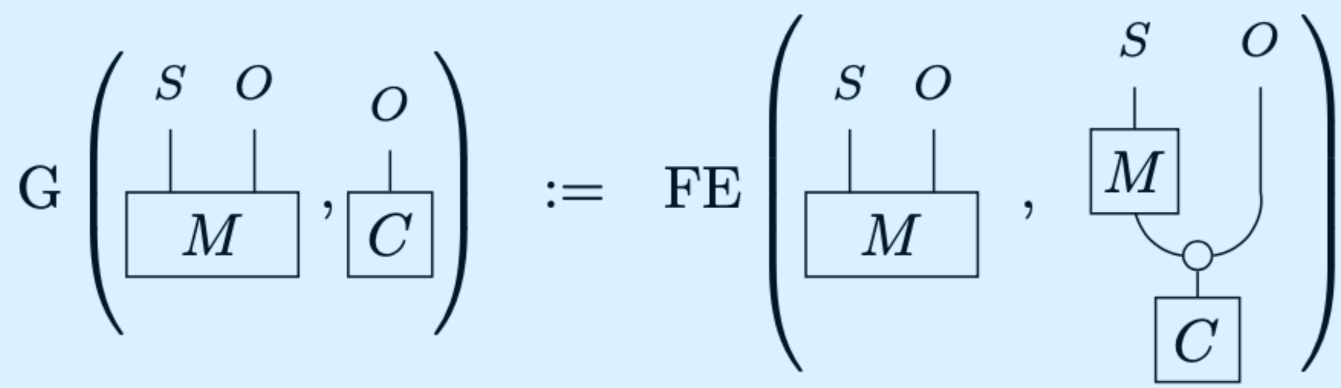
Minimising VFE $\implies q \approx M|_{\mathbf{o}}$

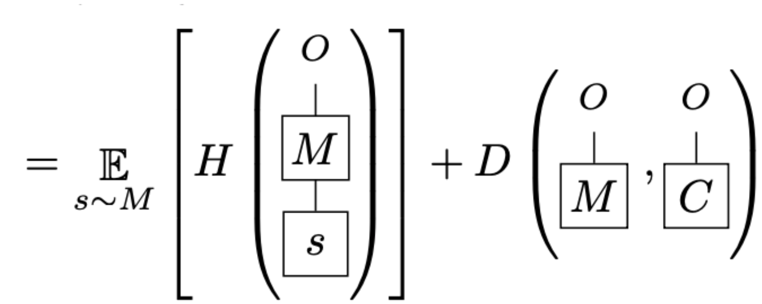
We call the minimal q the **VFE update** with respect to $\mathbf{o}$.

This gives a **third notion of updating** for soft observations.

Expected Free Energy

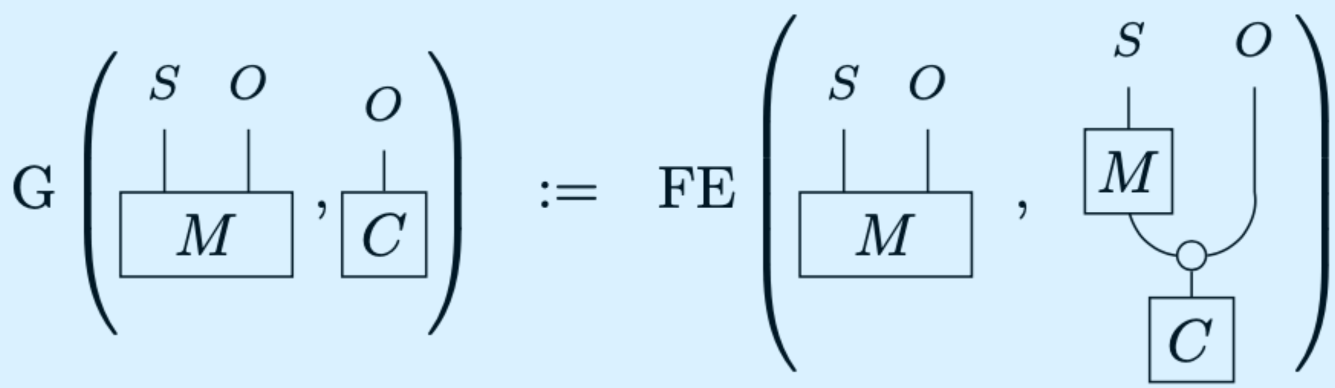
The **Expected Free Energy (EFE)** of ‘preferences’ distribution C over O is:

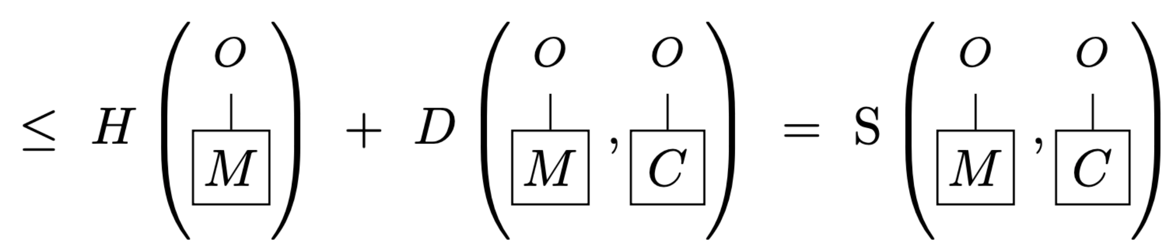




Expected Free Energy

The **Expected Free Energy (EFE)** of ‘preferences’ distribution C over O is:

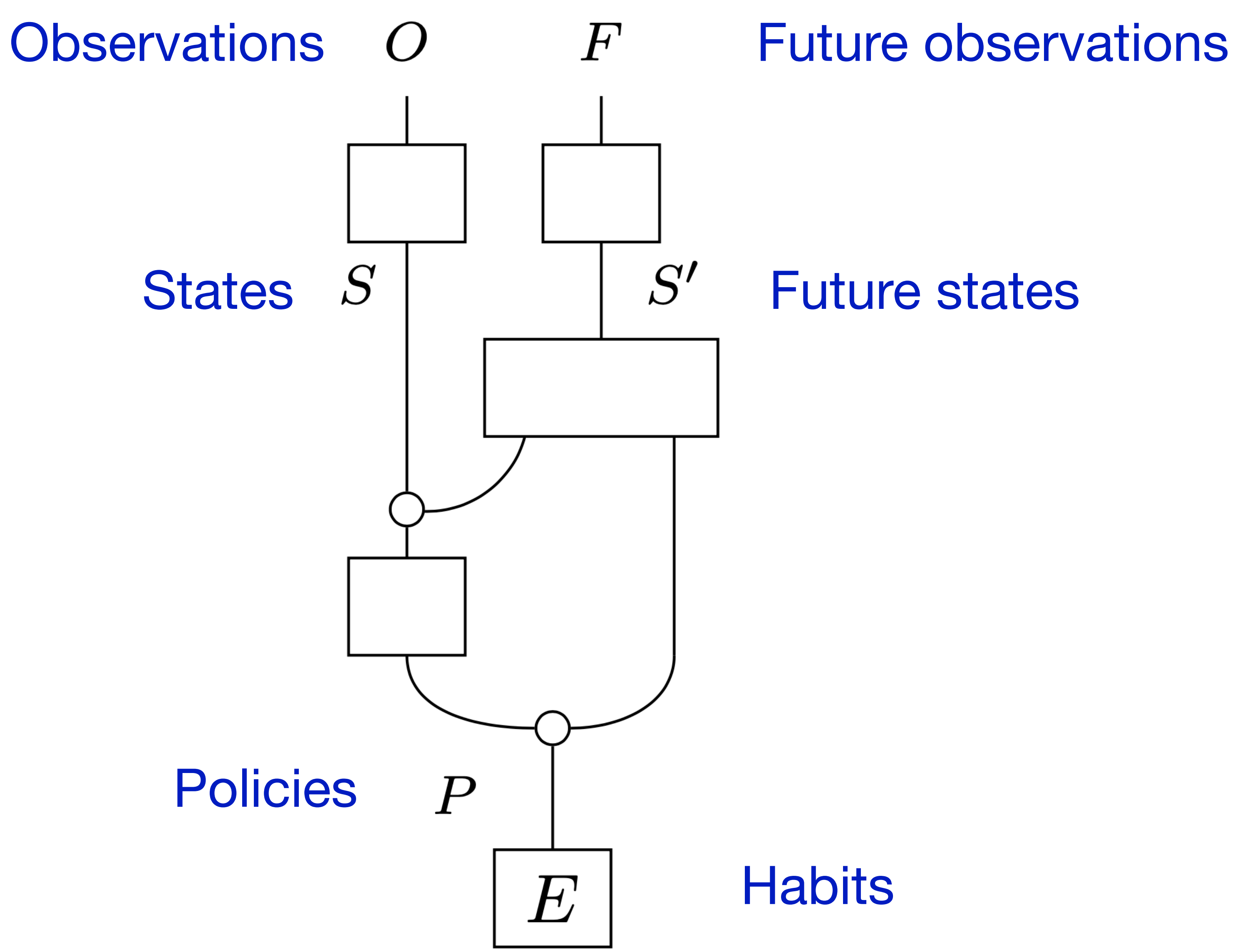




Active Inference

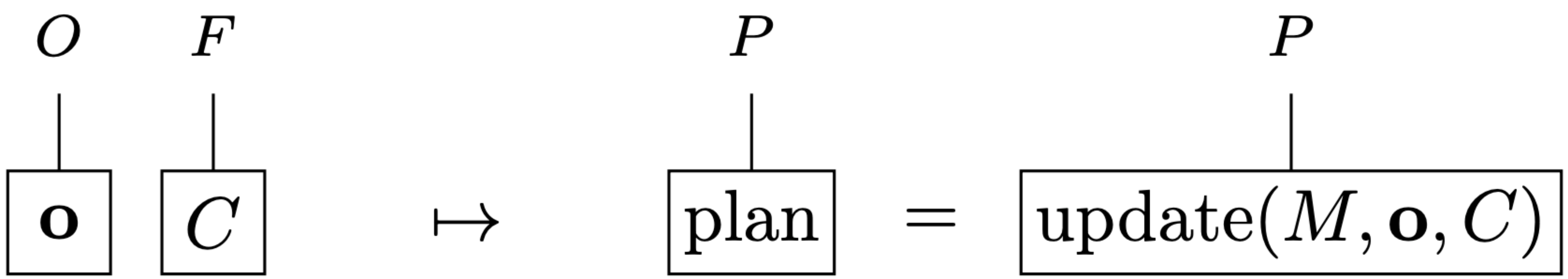
Active Inference

We consider a model $\mathbb{M}$ of the form:



Active Inference

Given an **observation** $\mathbf{o}$ and future **preferences** C the agent **plans** actions via approximate updating:



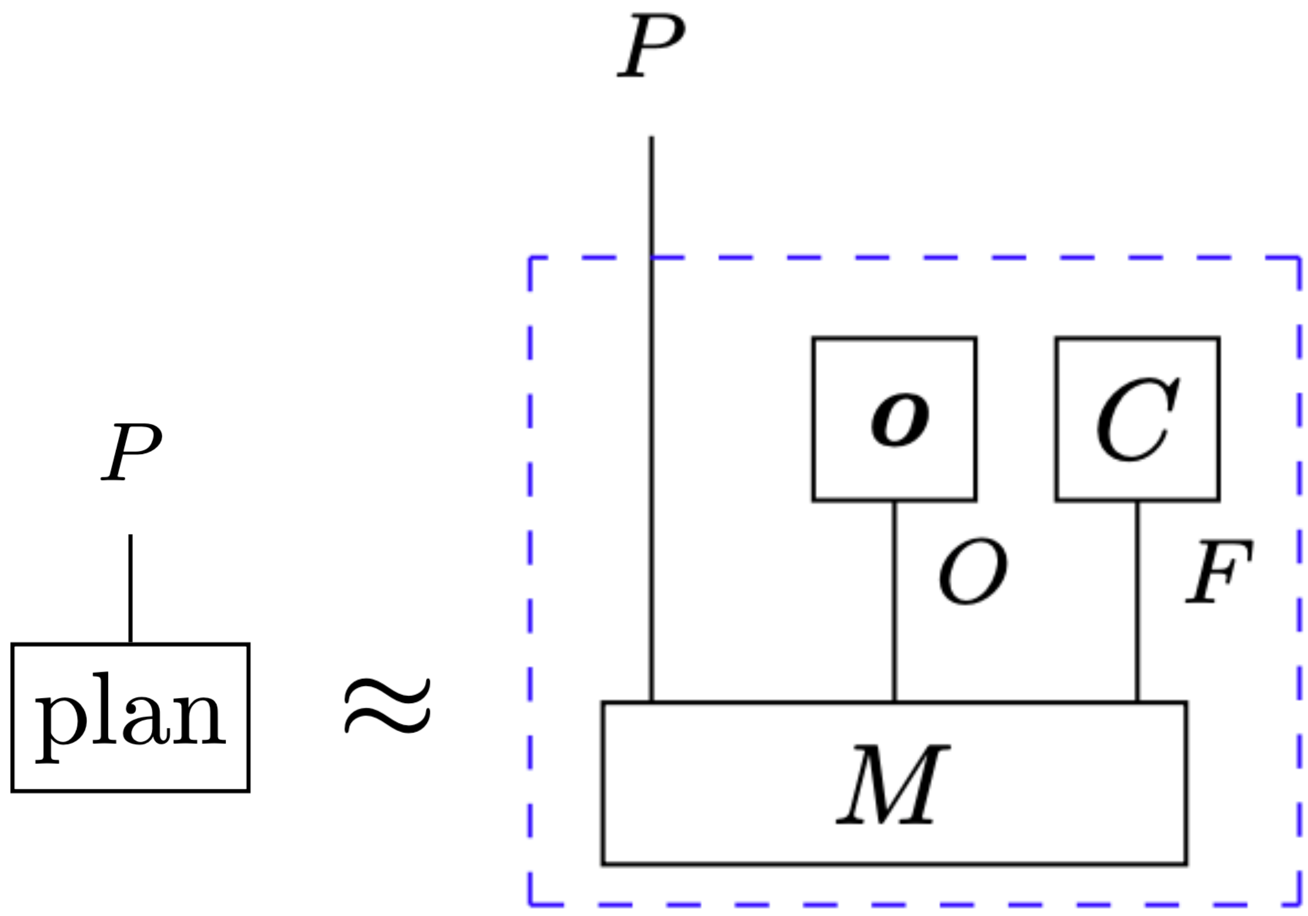
Free Energy Principle: can carry out approximately via

$$\text{plan}(\pi) := \sigma(\log E(\pi) - F(\pi) - G(\pi))$$

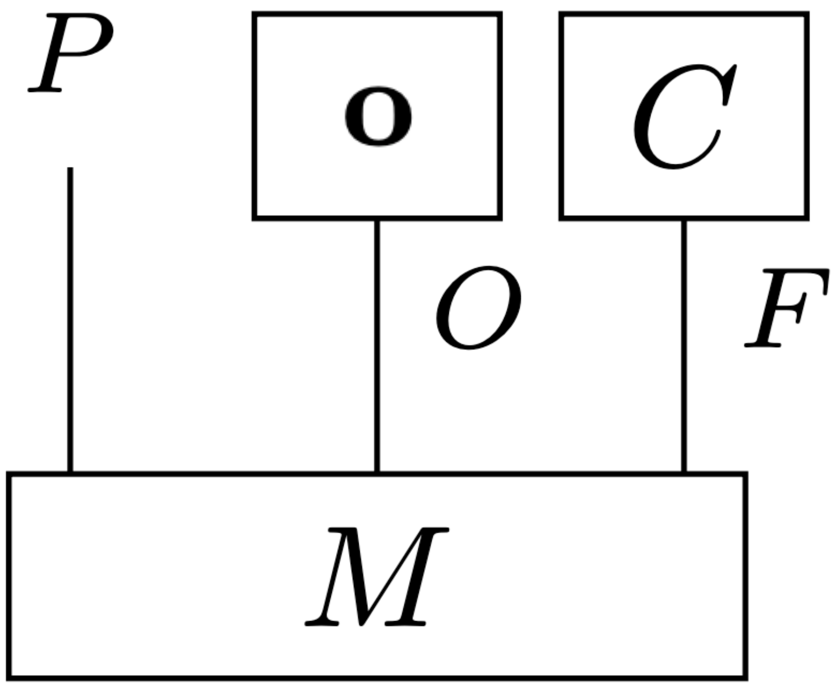
softmax habits VFE EFE

Let's derive this!

Active Inference

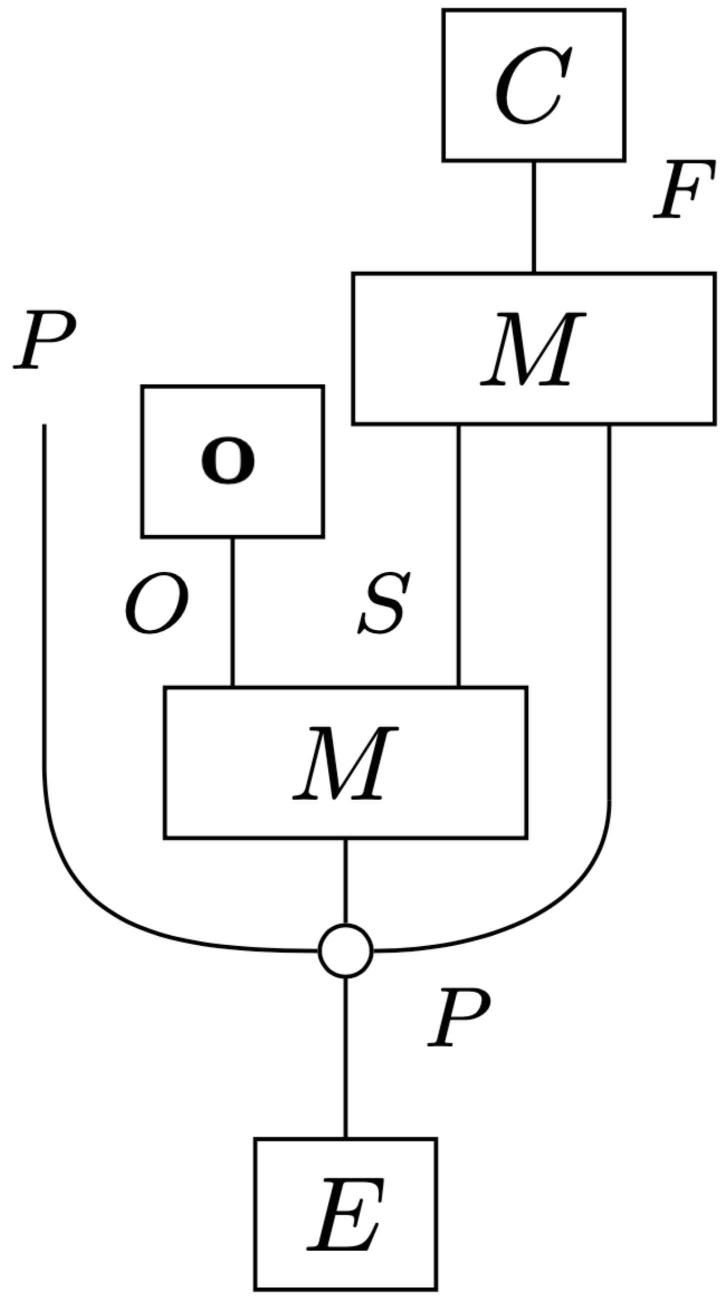


Active Inference



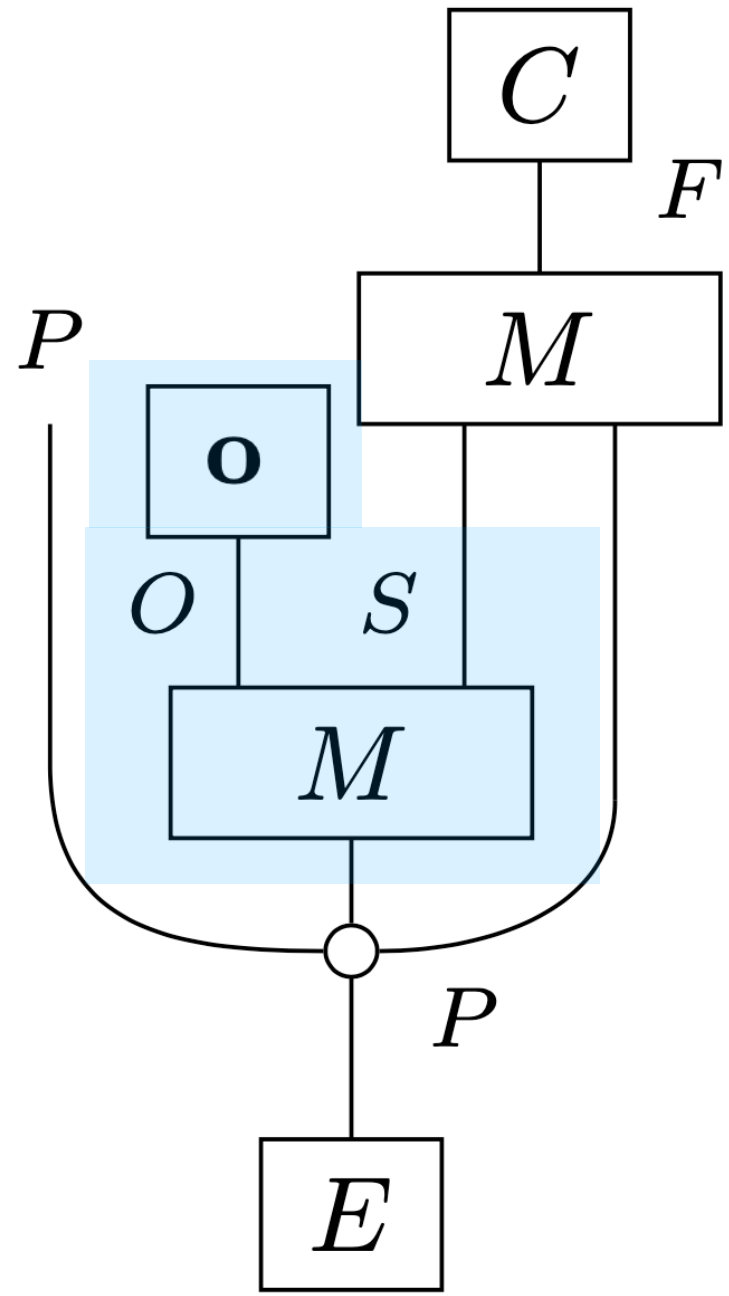
Active Inference

$=$

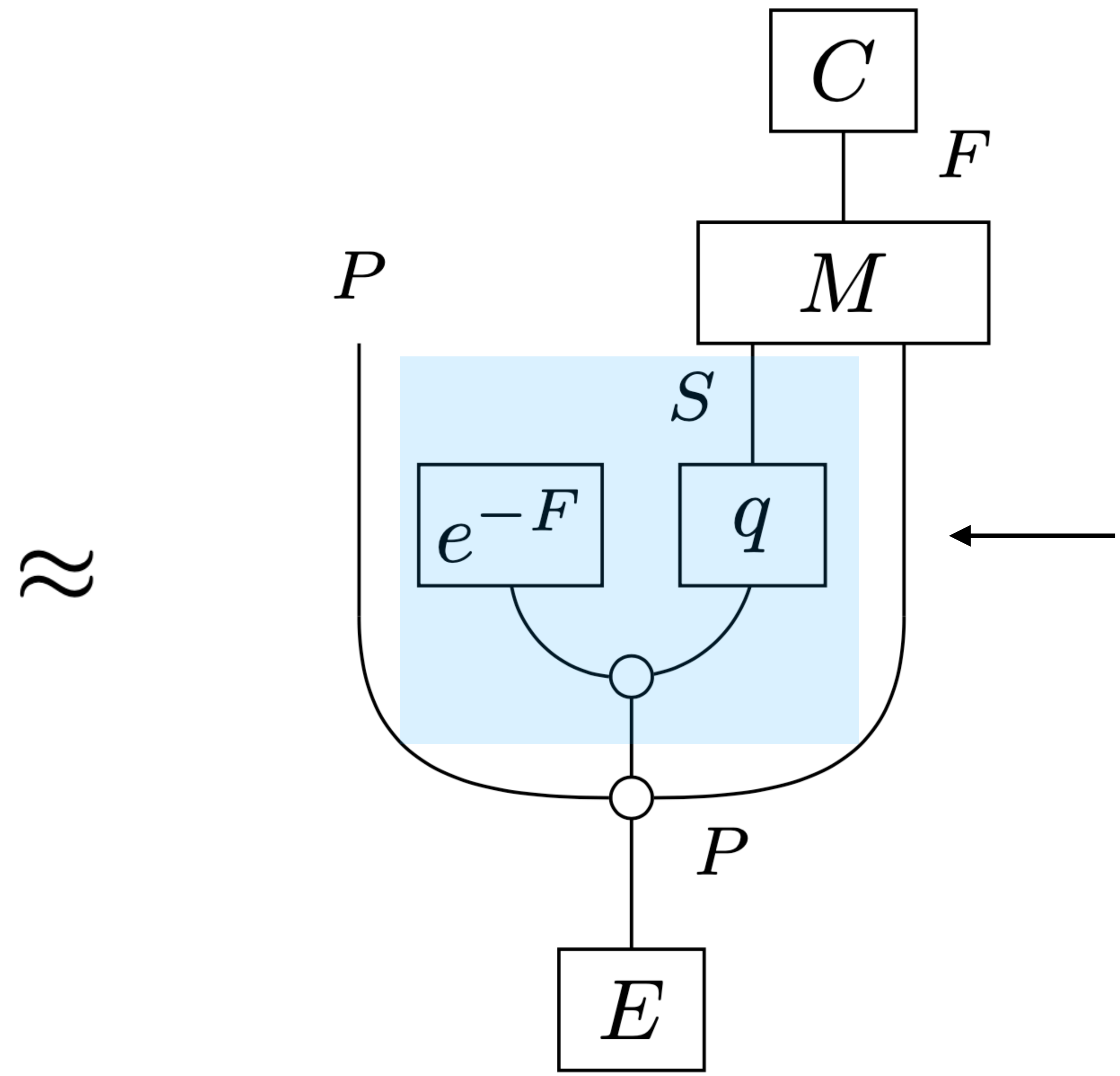


Active Inference

$=$

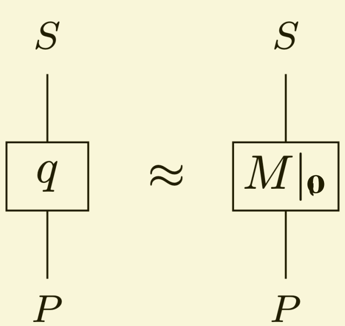


Active Inference



Perception

Approx inference $q(\pi)$ per $\pi \in P$

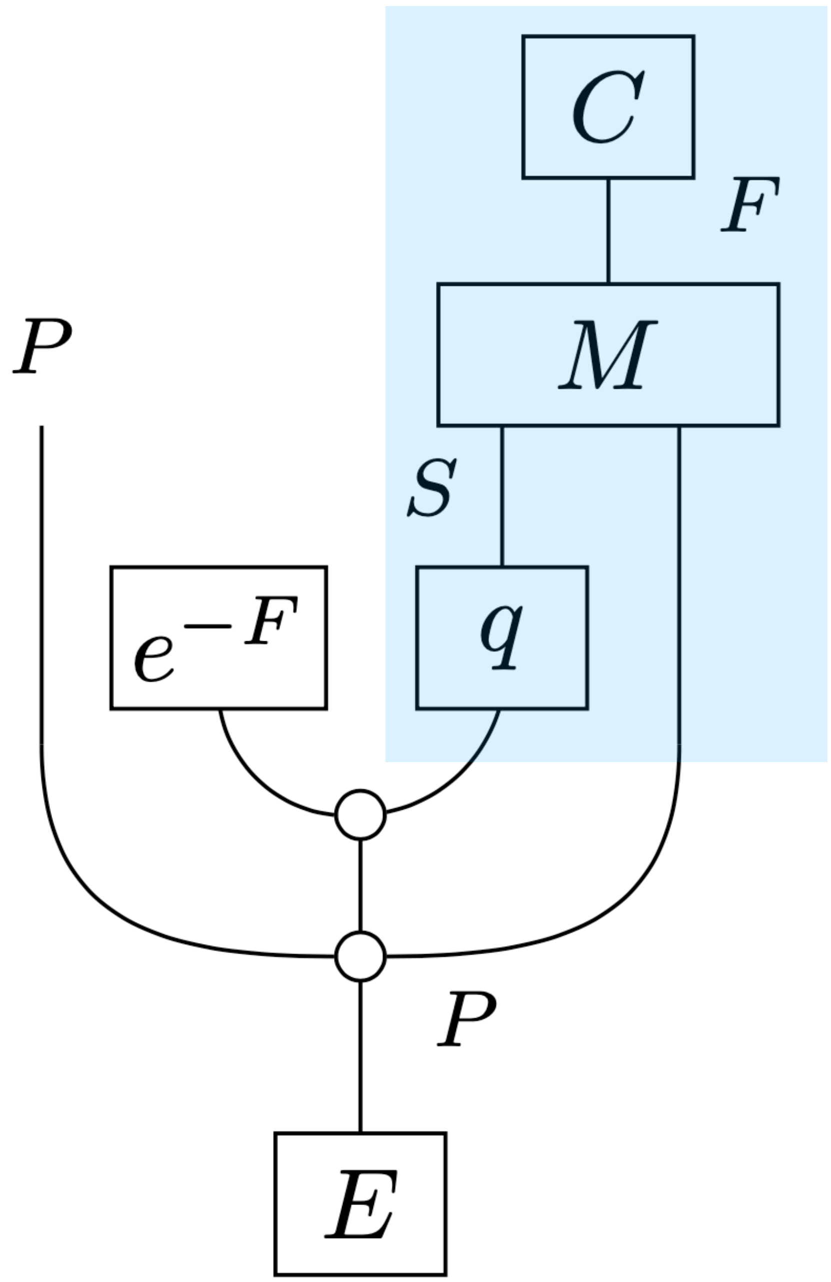


by minimising VFE $F(\pi) = F(q(\pi))$

(VFE updating)

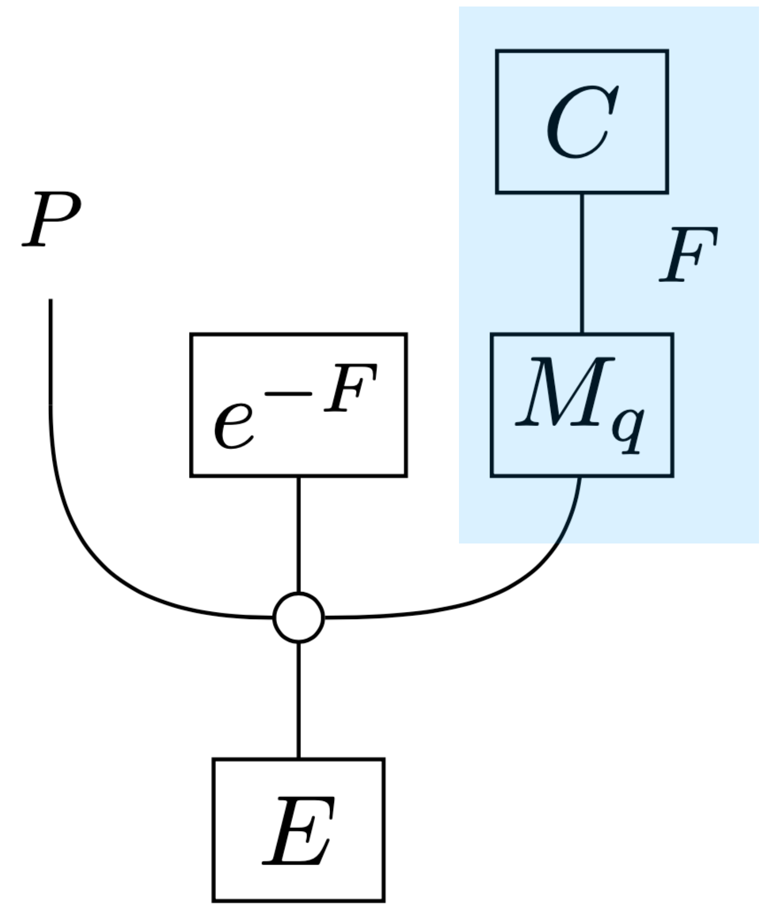
Active Inference

$\approx$

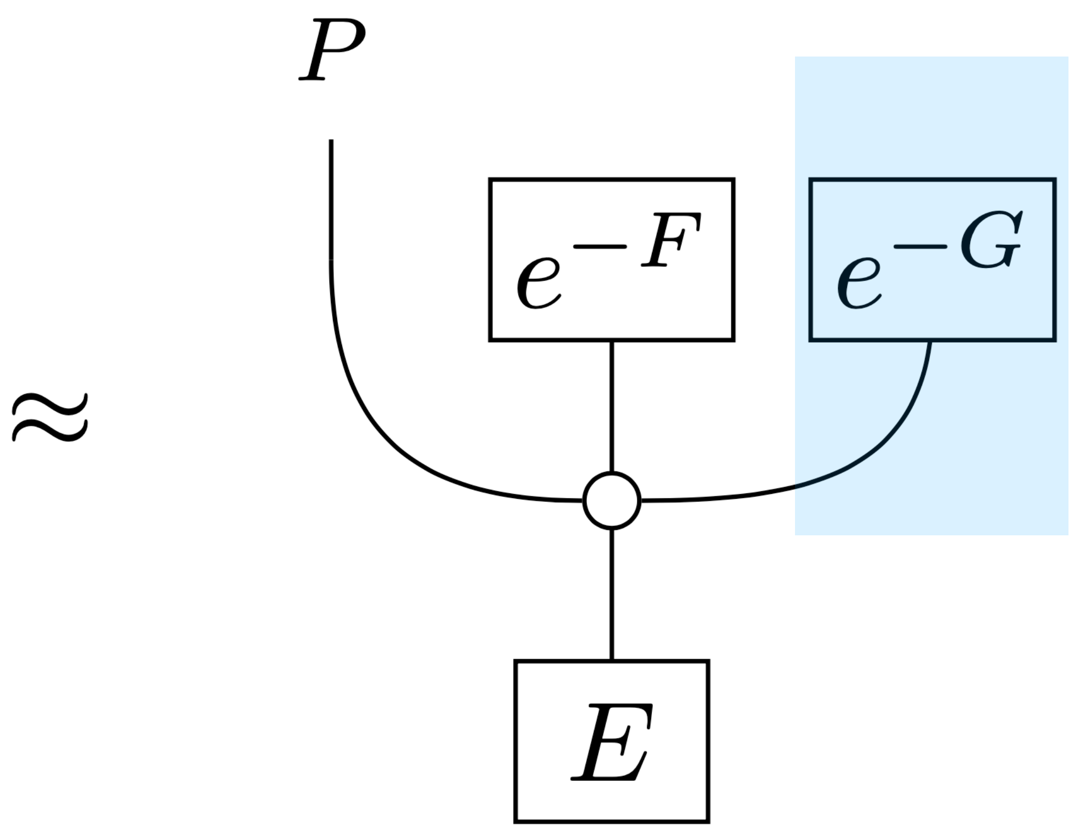


Active Inference

$=$



Active Inference

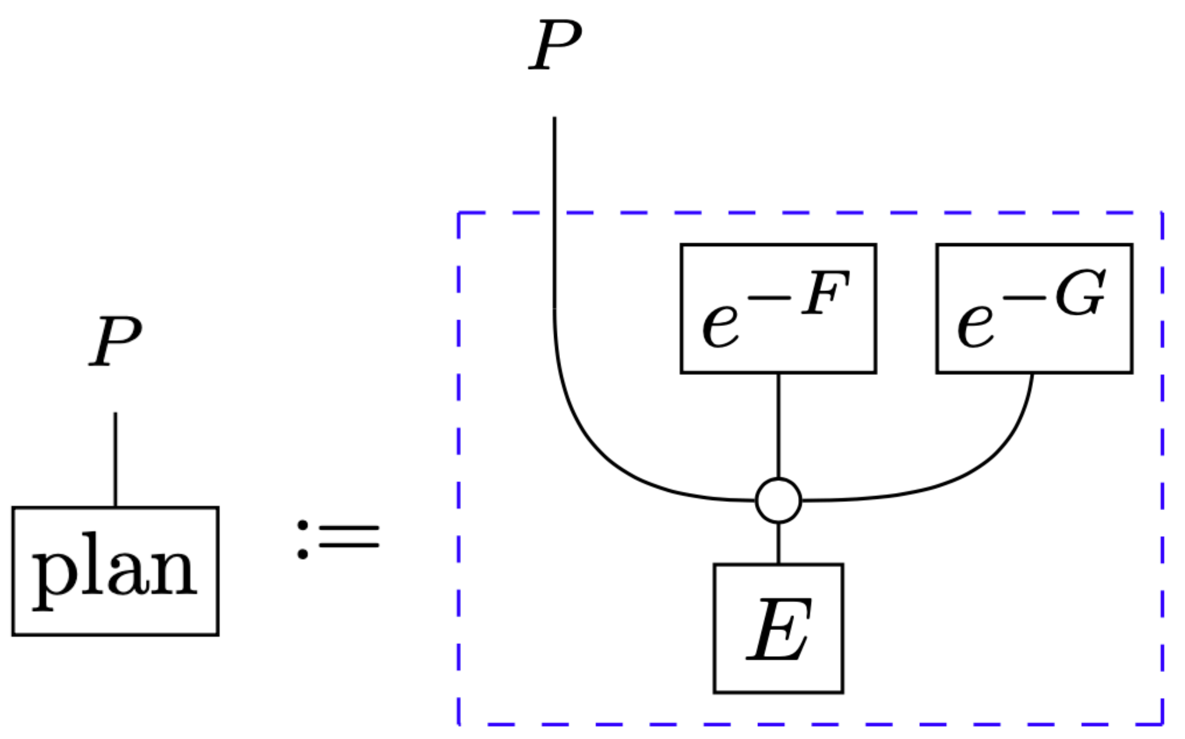


Prediction

Approximated via the EFE

$$G(\pi) = G(M_q(\pi), C)$$

Active Inference



Conclusion: we obtain the active inference scheme

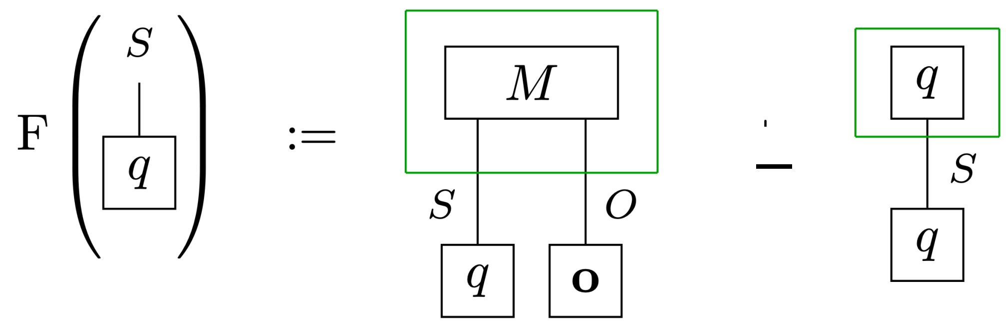
$$\text{plan}(\pi) := \sigma(\log E(\pi) - F(\pi) - G(\pi))$$

softmax habits VFE EFE

Compositionality of Free Energy

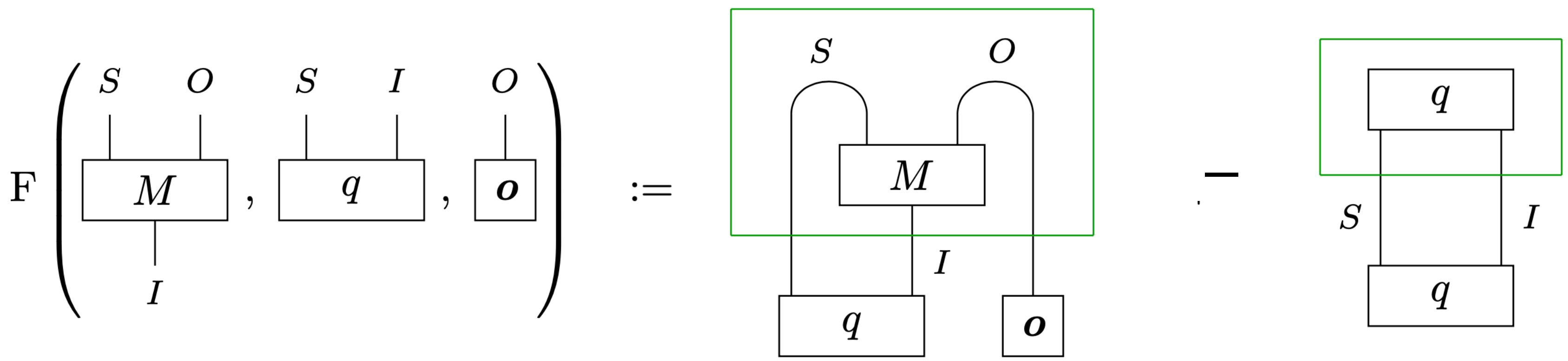
Compositionality of Free Energy

Recall the **VFE** is:



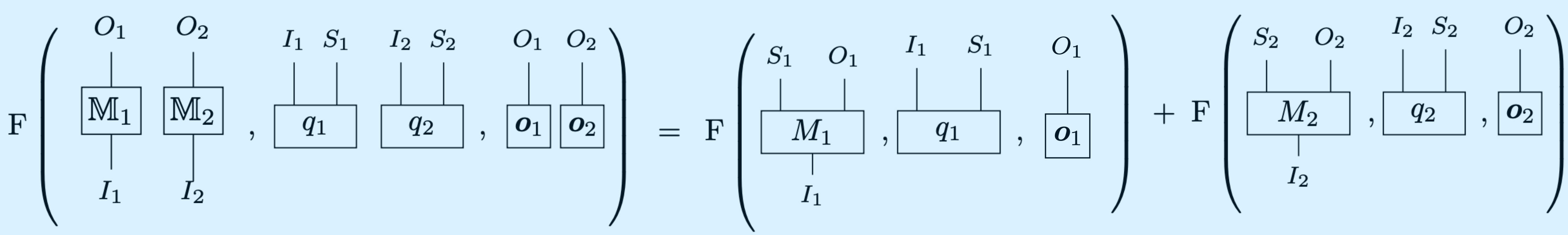
Compositionality of Free Energy

For an open generative model we define the **open VFE** as:



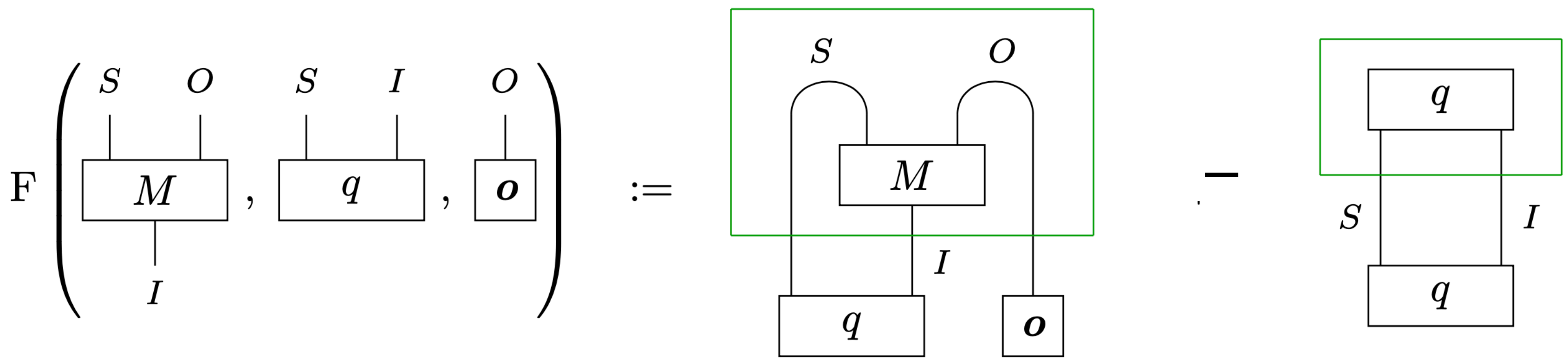
Theorem

Open VFE is **compositional** in that:



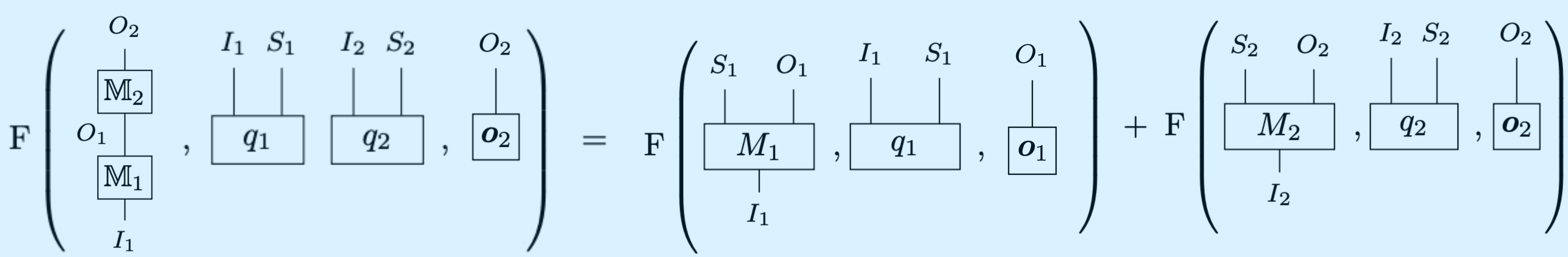
Compositionality of Free Energy

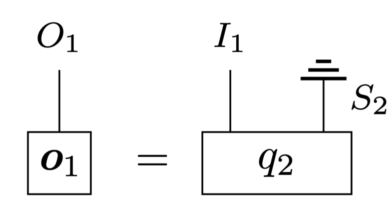
For an open generative model we define the **open VFE** as:



Theorem

Open VFE is **compositional** in that:



where 

Outlook

Outlook

String diagrams provide a natural language for active inference!
This includes generative models, free energy, updating…

Future work:

- Interpretation of our notion of ‘**Open VFE**’
- Diagrammatic account of **message passing**
- Pearl vs Jeffrey vs VFE **updating in cognition**
- Connections to **compositional intelligence**, **categorical cybernetics** and **consciousness.**

Thanks!