Evidential Decision Theory via Partial Markov

Elena Di Lavore, Mario Román
Tallinn University of Technology

Nov 2nd, NWPT 22, Bergen

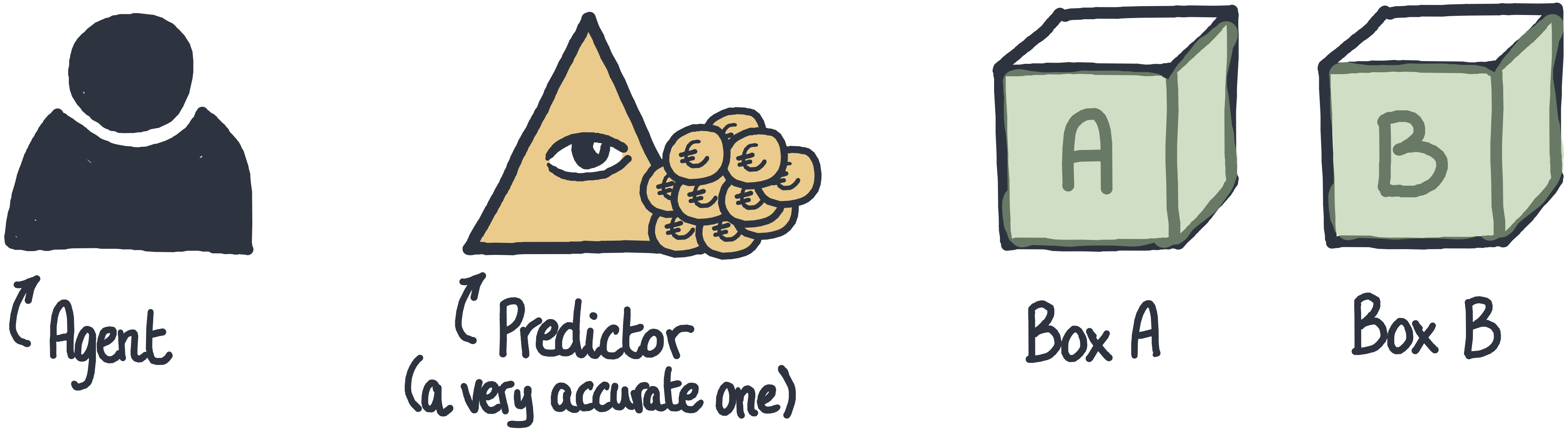
Supported by the EU Estonian IT Academy.

Part 0: Deciding is Difficult

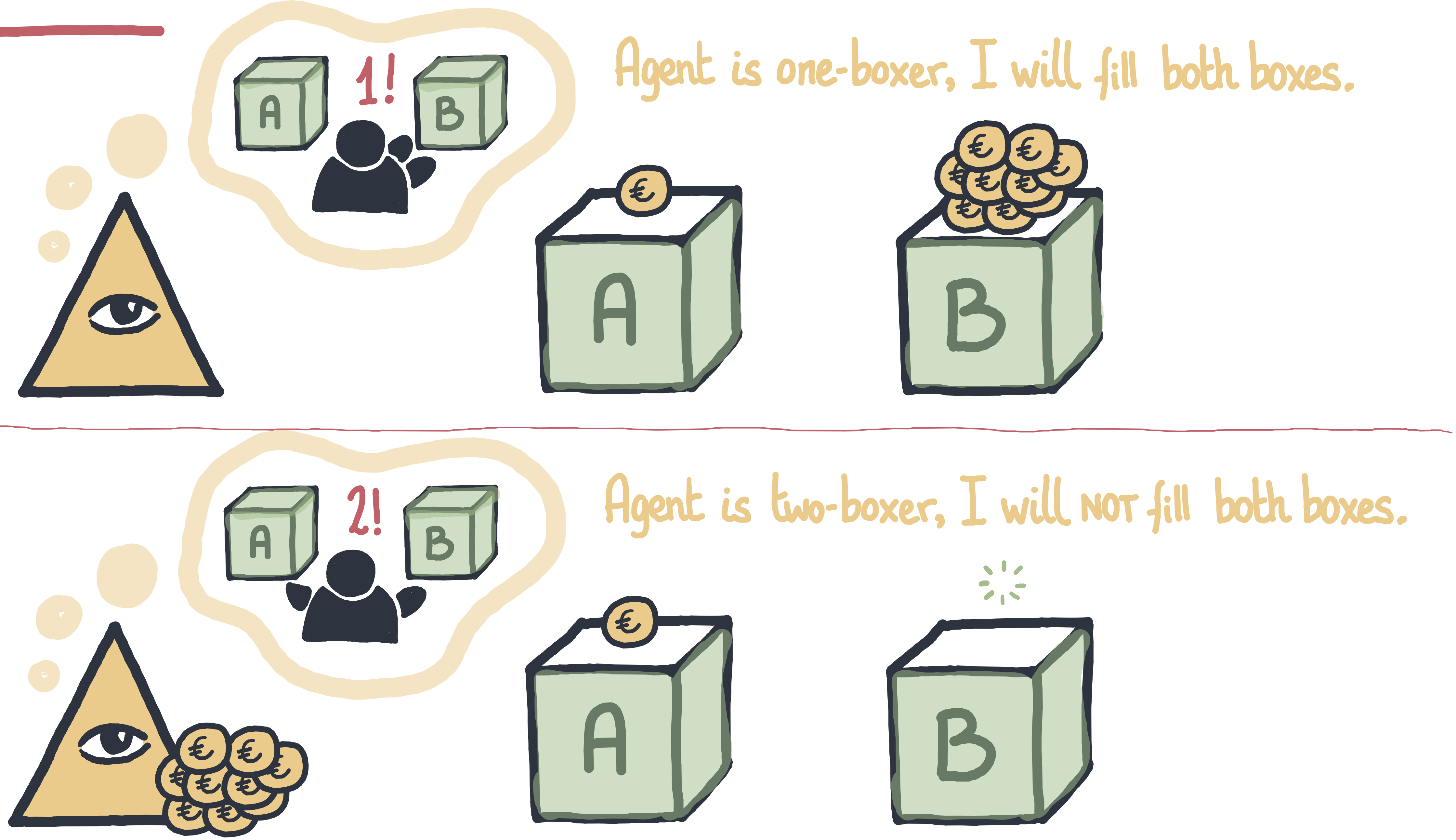
(Newcomb's Problem)

Nozick

NEWCOMB'S PROBLEM

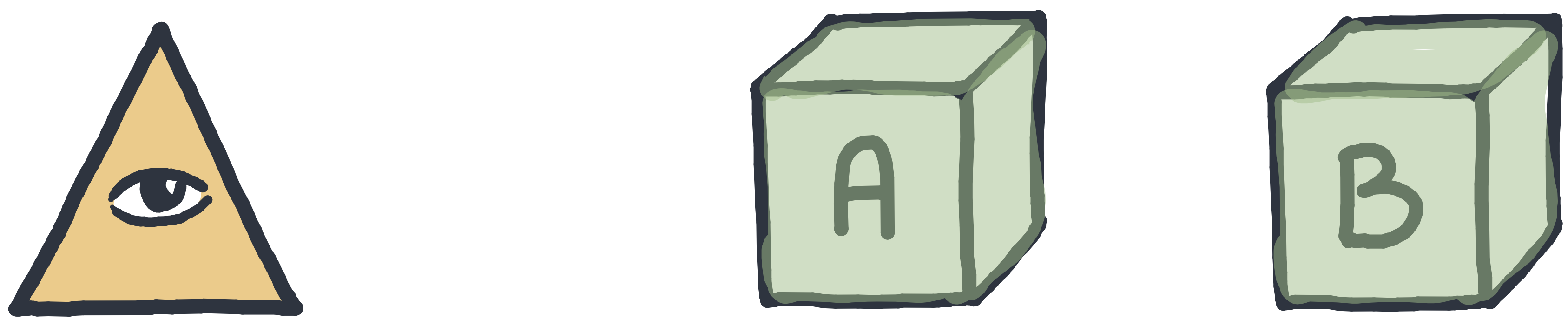


NEWCOMB'S PROBLEM



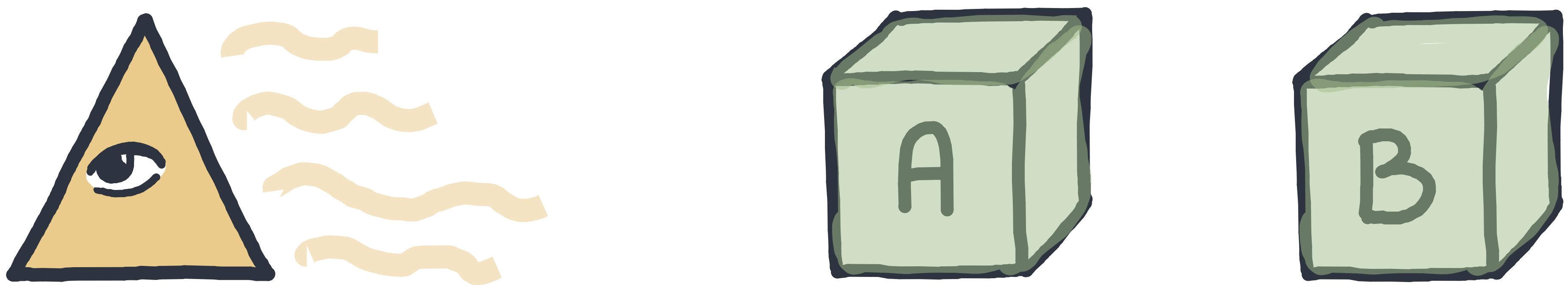
NEWCOMB'S PROBLEM

Predictor closes the boxes.



NEWCOMB'S PROBLEM

Predictor leaves.



Newcomb's Problem

Predictor leaves.

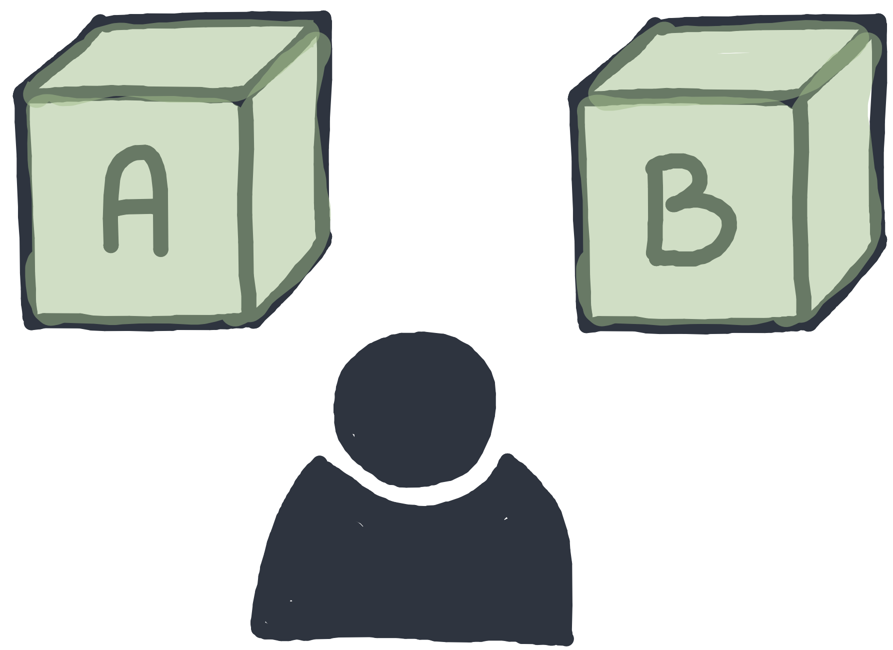
Newcomb's Problem

What should the agent do?

- Two-box: 1€ is better than nothing, 101€ is better than 100€. No matter what I do, I cannot change whatever is in the boxes.

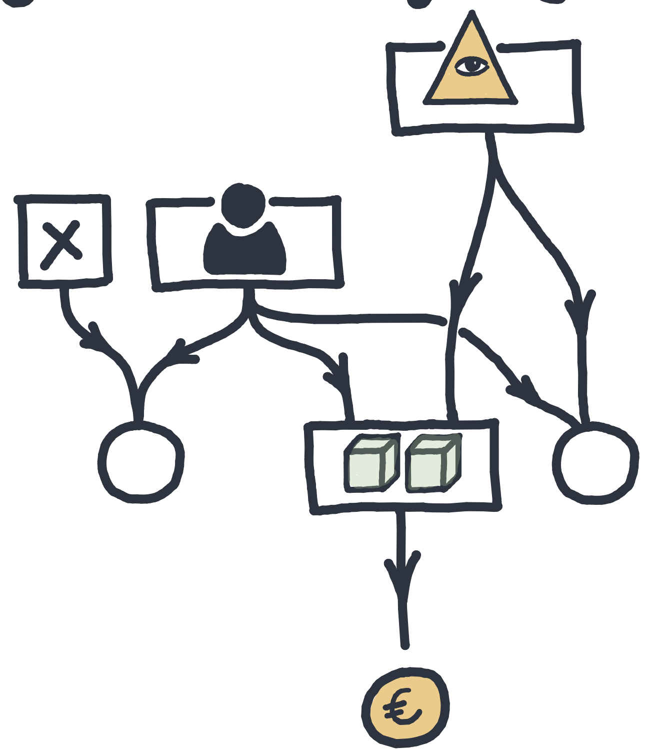
- One-box: I want to one-box so that the accuracy of the predictor means that I get 100€, instead of 1.

The idea is always to find the argument that maximizes some expected value function. The debate is in interpreting what that function is.



Calculemus

Leibniz's dream was to see philosophical disputes reduced to mathematical calculation. An algorithm for figuring out the correct position.



Wishlist.

- Formal syntax and axioms for stochastic processes and bayesian inference.
- Systematic decision theory.
- Compositional and abstract theory extending synthetic probability.

CALCULEMUS

Leibniz's dream was to see philosophical disputes reduced to mathematical calculation. An algorithm for figuring out the correct position.

do
  prediction ← 
  action ← 
  observe (action = x)
  observe (action = prediction)
  return ()(action, prediction)

Wishlist.

- Formal syntax and axioms for stochastic processes and bayesian inference.
- Systematic decision theory.
- Compositional and abstract theory extending synthetic probability.

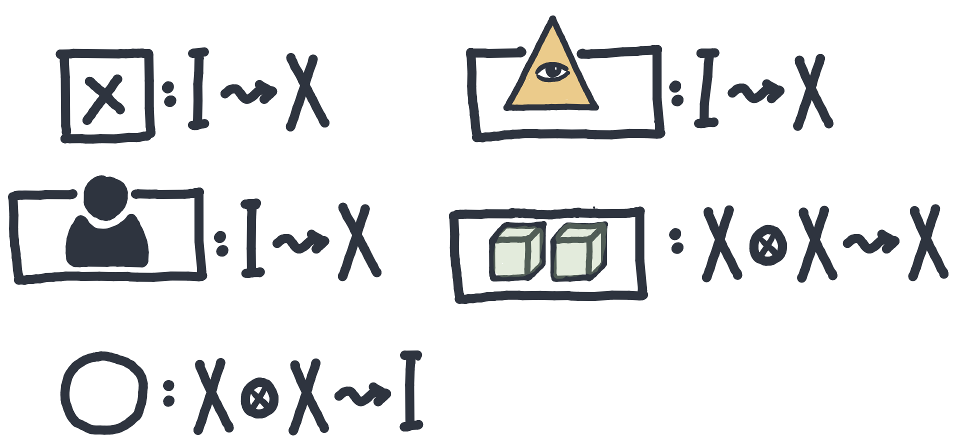
Part 1: Process Theories

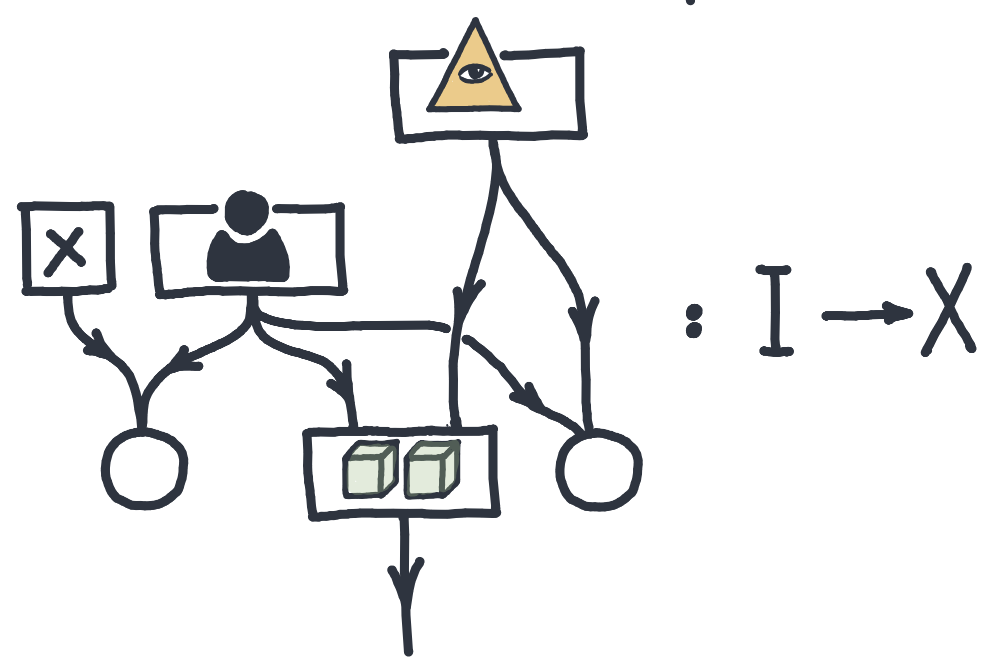
(Symmetric Monoidal Categories)

Process Theories

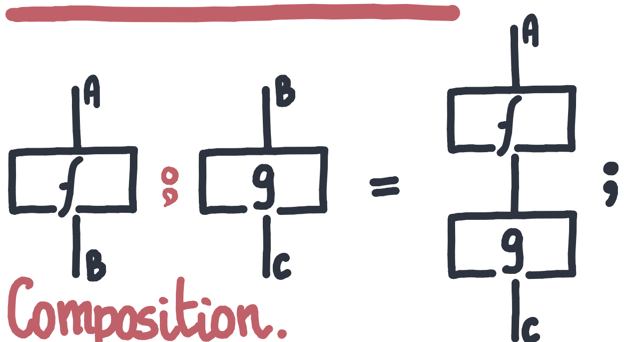
Joyal, Street

A **process theory** consists of a set of resource types, A, B, C... a set of processes, each one with a list of inputs and outputs, e.g. $f : A_0 \otimes \cdots \otimes A_n \longrightarrow B_0 \otimes \cdots \otimes B_m$; and an algebra of string diagrams: every directed acyclic graph, where each vertex is labelled by a process and every edge is labelled by its input and output resources, repeated multiple times, determines a process.

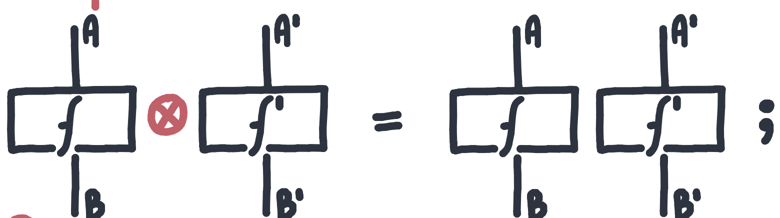




Process Theories

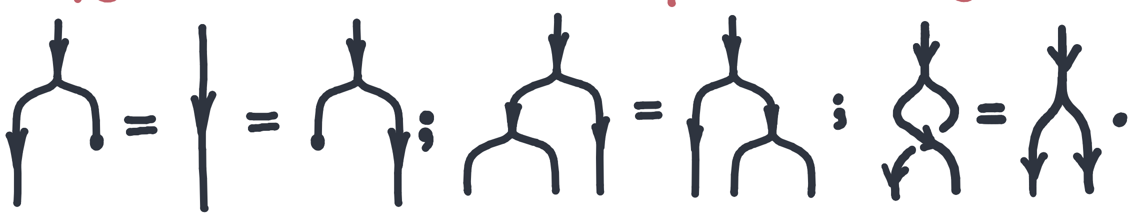


Composition.



Parallelization.

Copy. Discard. Swap (σ) Identity (id)



Copy-discard axioms.

Process theories are symmetric monoidal categories with copy and delete morphisms.

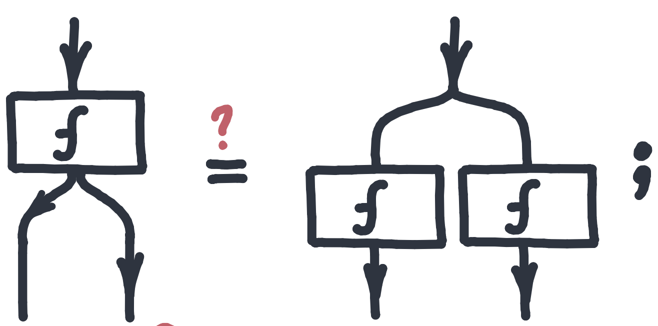
Monoidal Axioms.

$$f \mathbin{;} (g \mathbin{;} h) = (f \mathbin{;} g) \mathbin{;} h \,; \qquad f \otimes (g \otimes h) = (f \otimes g) \otimes h \,;$$
$$f \mathbin{;} id = f = id \mathbin{;} f \,; \qquad f \otimes id_I = f = id_I \otimes f \,;$$
$$(f \otimes f') \mathbin{;} (g \otimes g') = (f \mathbin{;} g) \otimes (f' \mathbin{;} g') \,;$$
$$\sigma_{A,B} \mathbin{;} \sigma_{B,A} = id \,; \qquad \delta_A \mathbin{;} \sigma_{A,A} = \delta_A \,;$$
$$\sigma_{A,B\otimes C} = (\sigma_{A,B} \otimes id_C) \mathbin{;} (id_B \otimes \sigma_{A,C}) \,;$$
$$\sigma_{A\otimes B,C} = (id_A \otimes \sigma_{B,C}) \mathbin{;} (\sigma_{A,C} \otimes id_B) \,;$$
$$\delta_A \mathbin{;} (\varepsilon_A \otimes id_A) = id_A = \delta_A \mathbin{;} (id_A \otimes \varepsilon_A) \,;$$
$$\delta_A \mathbin{;} (\delta_A \otimes id_A) = \delta_A \mathbin{;} (id_A \otimes \delta_A) .$$

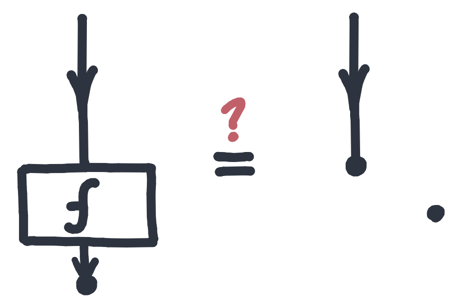
Theorem. String diagrams are sound and complete for symmetric monoidal categories.

SPLITTING

What we do NOT take as an axiom is the following:



Copying



Discarding

Some process theories assume that processes can be copied and discarded.
These are called cartesian monoidal categories.
This is true for functions or functional programs, but it fails in other contexts.

SPLITTING

What we do NOT take as an axiom is the following:

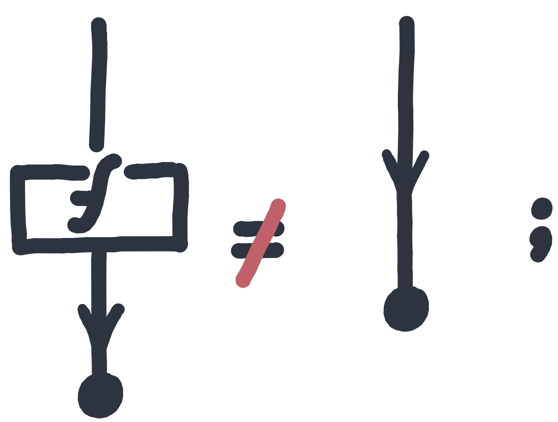
$$\begin{array}{l}\textbf{do}\\ \quad x_1 \leftarrow f\\ \quad x_2 \leftarrow f\\ \quad \textbf{return}\ (x_1, x_2)\end{array} = \begin{array}{l}\textbf{do}\\ \quad x \leftarrow f\\ \quad \textbf{return}\ (x, x)\end{array} \quad ; \quad \begin{array}{l}\textbf{do}\\ \quad x \leftarrow f\\ \quad \textbf{return}\ ()\end{array} = \begin{array}{l}\textbf{do}\\ \quad \textbf{return}\ ()\end{array}$$

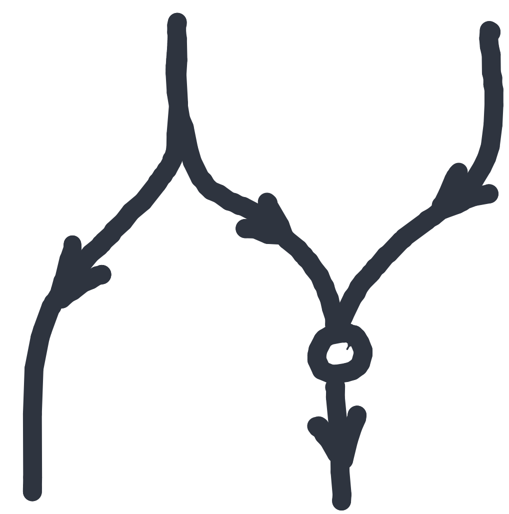
Copying — Discarding

Some process theories assume that processes can be copied and discarded.
These are called cartesian monoidal categories.
This is true for functions or functional programs, but it fails in other contexts.

Part 2 : Partiality

(when discarding fails)





PARTIALITY

Curien, Obtulowicz
Di Liberti, Nester et al.
Cockett, Guo, Hofstra

The theory of partial functions contains functions, say $f : X \to Y$, that can diverge in some inputs, say $x \in X$. We write that as $f(x)\uparrow$.

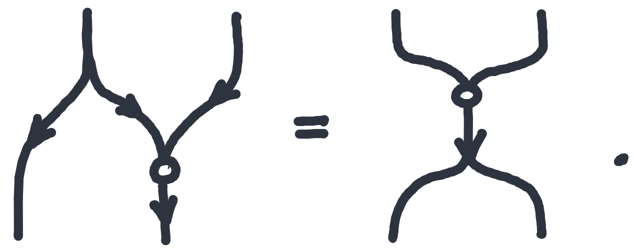
In particular, the following comparator is a partial function.

$$\psi(x_1 ; x_2) = \begin{cases} (\uparrow) & \text{when } x_1 \neq x_2, \\ x_1 & \text{when } x_1 = x_2. \end{cases}$$

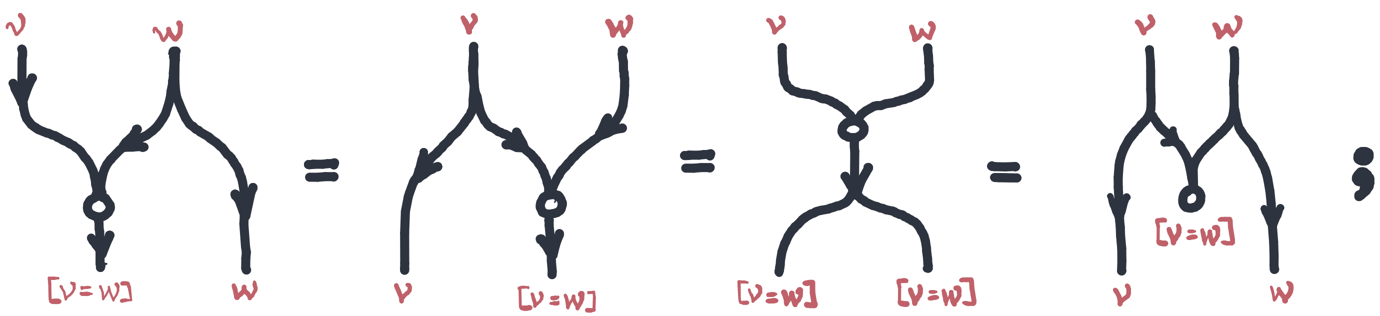
In the same way that partial functions model deterministic process that diverge (say, Turing machines), they can model processes that "fail".

Comparators

Axiom. We assume the existence of a comparator, , satisfying the "Frobenius axiom",

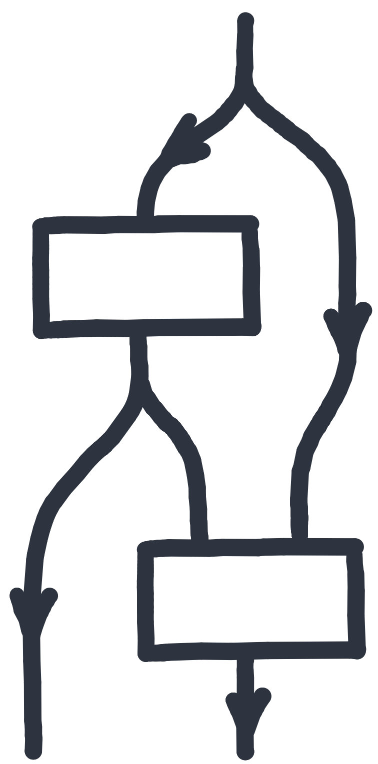


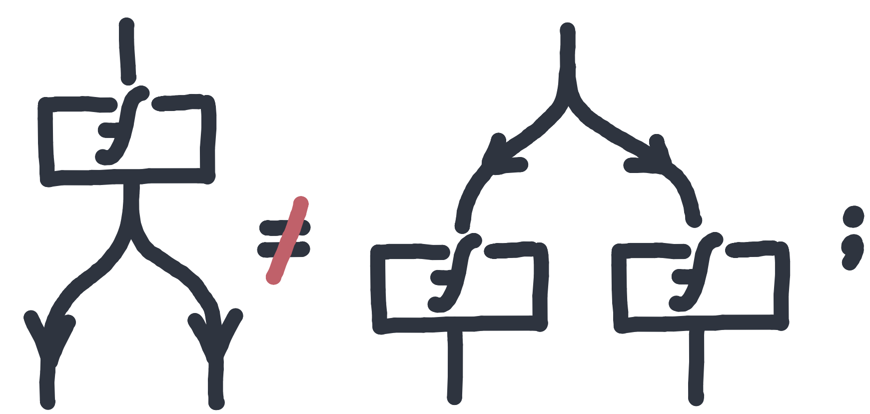
Imagine we have two values that we want to compare: v and w. Our comparator, $(\cdot,\cdot)$, returns any of the two if they are equal.



PART 3 : MARKOV CATEGORIES

(when copying fails)





Stochastic Functions

The theory of stochastic functions contains, as processes from $X_0, \dots, X_n$ to $Y_0, \dots, Y_m$, the functions

$$f(\overset{y_0}{\bullet}, \dots, \overset{y_m}{\bullet} \mid \overset{x_0}{\bullet}, \dots, \overset{x_n}{\bullet}) : X_0 \times \dots \times X_n \times Y_0 \times \dots \times Y_m \longrightarrow [0,1]$$

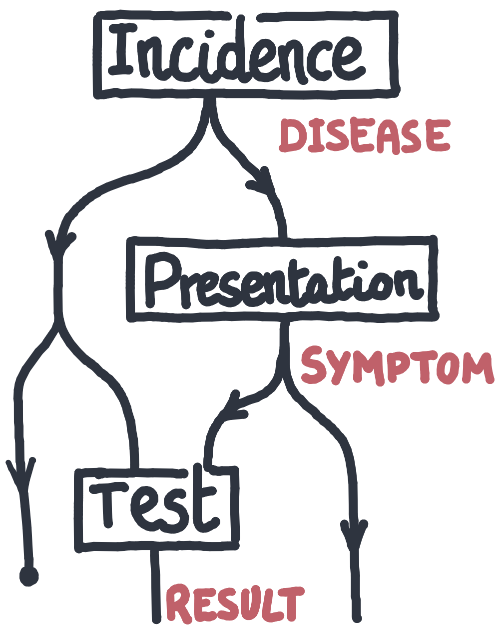
such that, for each $x_0 \in X_0, \dots, x_n \in X_n$;

$$\sum_{y_0 \in Y_0} \dots \sum_{y_m \in Y_m} f(x_0, \dots, x_n \mid y_0, \dots, y_m) = 1 .$$

We assume that the distribution is finite (i.e. has finite support), so that the sum is well-defined.

Reading Stochastic Diagrams

Specific way of reading diagrams of stochastic functions.

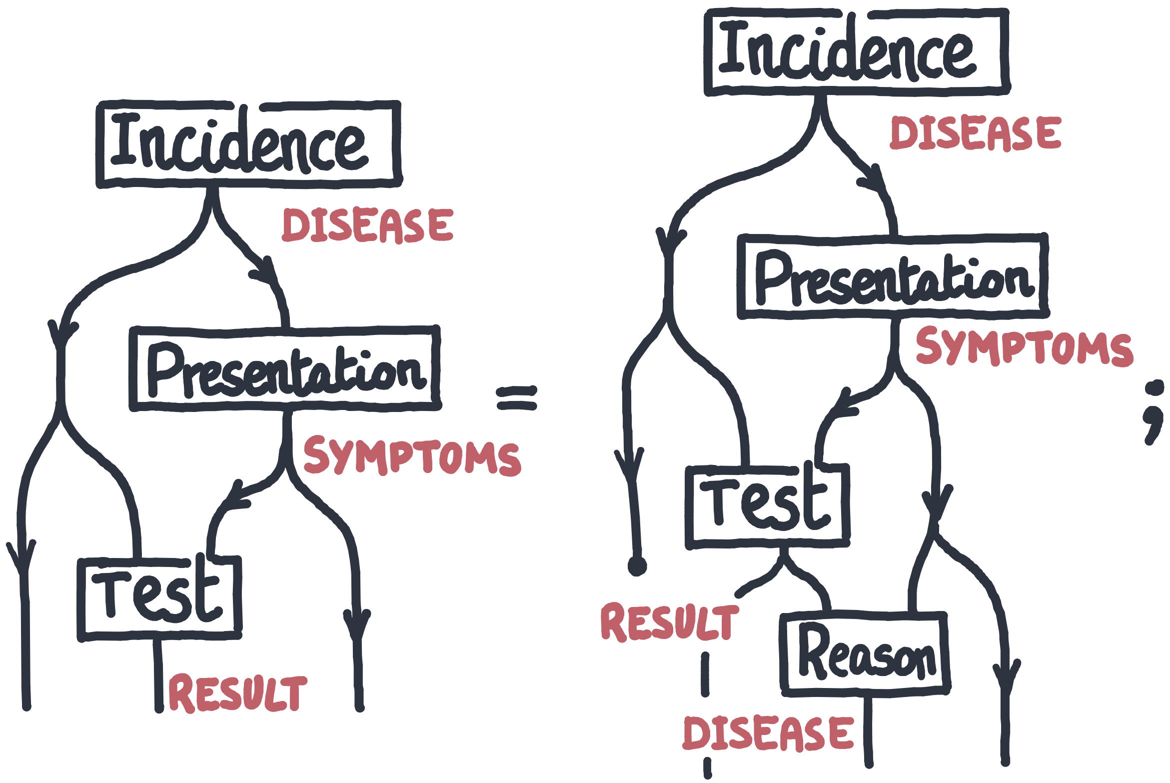


$$C(s,t) = \sum_{d \in D} I(d) \cdot P(s|d) \cdot T(t|s,d)$$

Every node $N : X_0 \otimes \dots \otimes X_n \to Y$ determines a conditional probability $N(y|x_0,\dots,x_n)$ for each $y \in Y$ and each $x_i \in X_i$.

For instance, the test T: DISEASE $\otimes$ SYMPTOM $\to$ RESULT gives conditionals
T(positive | disease, symptom) and T(negative | disease, symptom).

Reading Stochastic Diagrams



$$R(d|s,r) = \frac{I(d)\cdot P(s|d)\cdot T(r|s,d)}{\sum_{d'\in D} I(d')\cdot P(s|d')\cdot T(r|s,d')}$$

This conditional is chosen so that reading the diagram yields

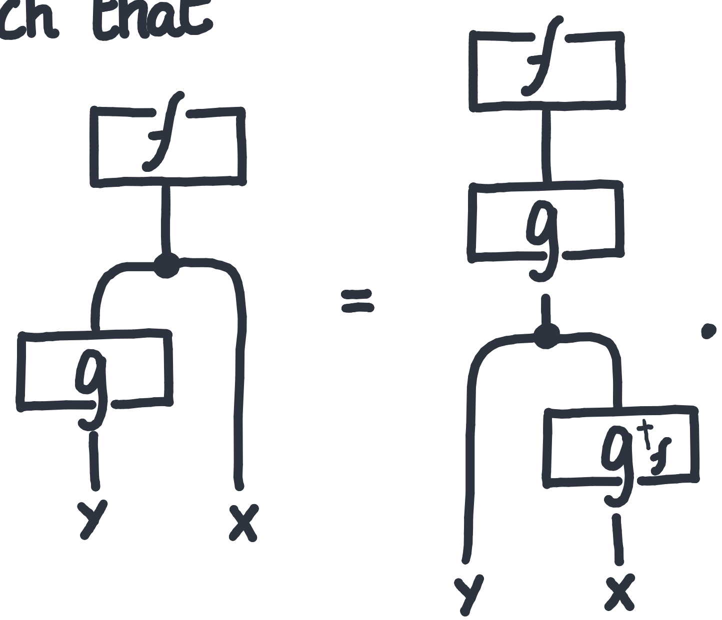
$$R(d|s,r)\cdot \sum_{d'\in D} I(d')\cdot P(s|d')\cdot T(r|s,d') = I(d)\cdot P(s|d)\cdot T(r|s,d)$$

Can we do this synthetically, taking some axiom about processes?

BAYESIAN INVERSION

Fritz.

DEFINITION. Let $f : I \to X$ be a distribution and let $g : X \to Y$ be a stochastic channel. The Bayesian Inversion of g with respect to f is the map $g^\dagger_f : Y \to X$ such that



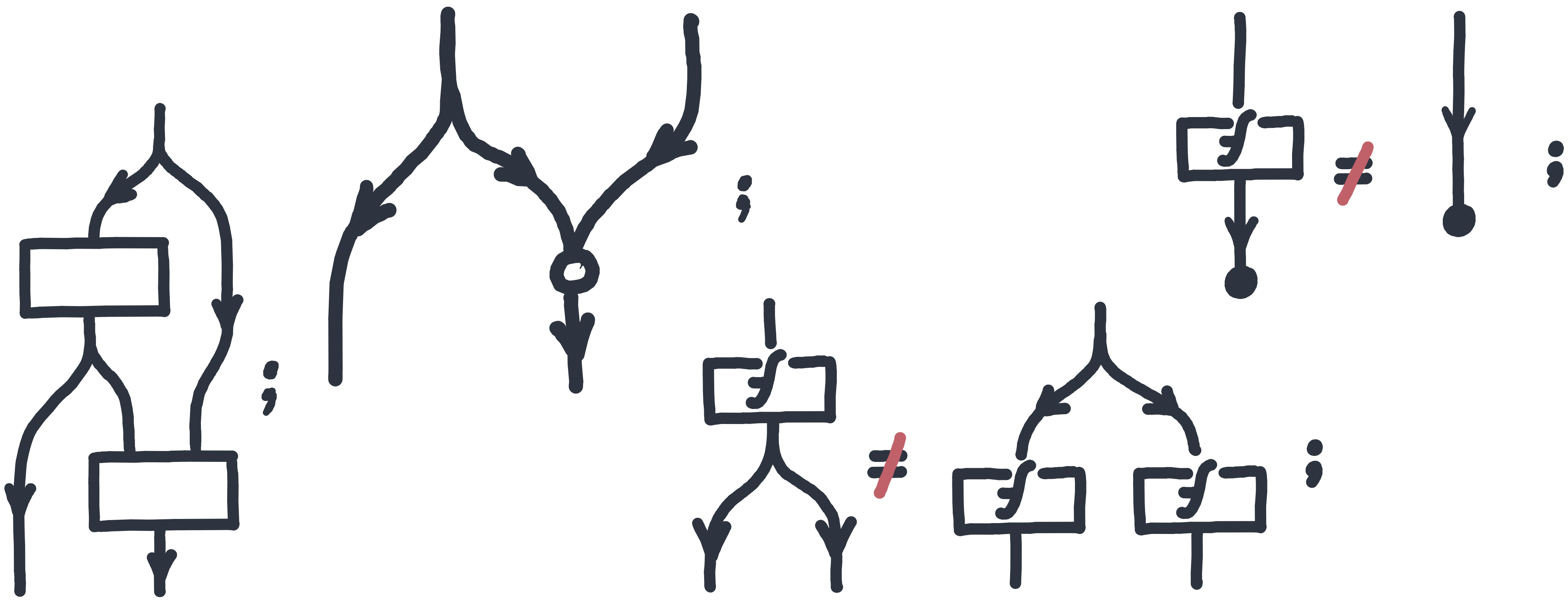
In the case of discrete stochastic functions, Bayesian inversions exist

$$g^\dagger_f(x|y) = \frac{f(x)\cdot g(y|x)}{\sum_{x'\in X} f(x')\cdot g(y|x')}.$$

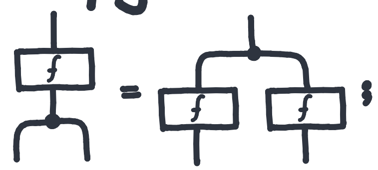
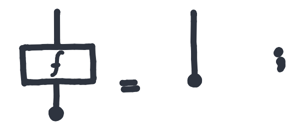
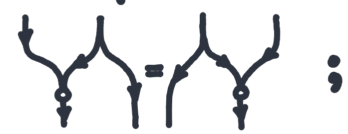
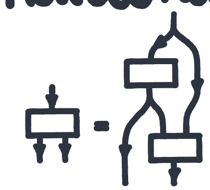
$$P(A|B) = \frac{P(A)\cdot P(B|A)}{\sum_{A'} P(A')\cdot P(B|A')}.$$

PART 4: PARTIAL MARKOV

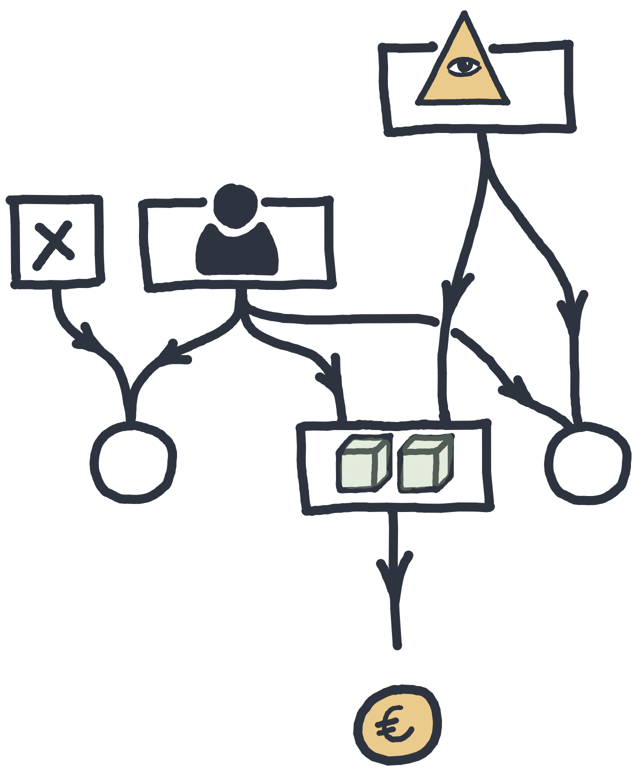
(mixing probabilistic and partial processes)



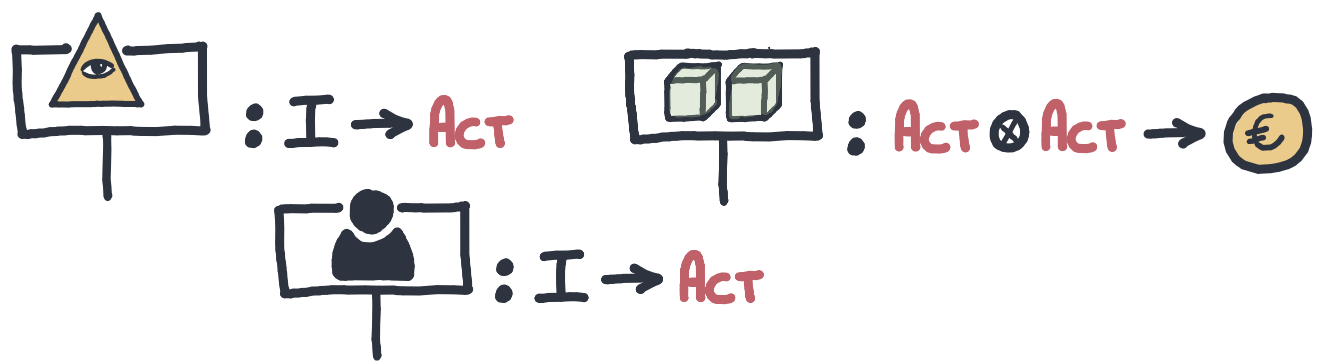
SUMMARY

	Copyable 	Discardable 	Comparators 	Conditionals 
Functions	✓	✓	✗	✗
Partial functions	✓	✗	✓	✗
Stochastic functions	✗	✓	✗	✓
Partial Stoch functions	✗	✗	✓	✓

Modelling Newcomb's



Basic building blocks.

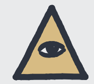


Instantiation in stochastic functions: both predictor and agent have a uniform prior, the boxes decide the outcome based on the actions of both.

Modelling Newcomb's

newcomb x = do
  prediction ← 
  action ← 
  observe (action = x)
  observe (action = prediction)
  return ()(action, prediction)

Argmaxing this function does return OneBox as the correct answer.

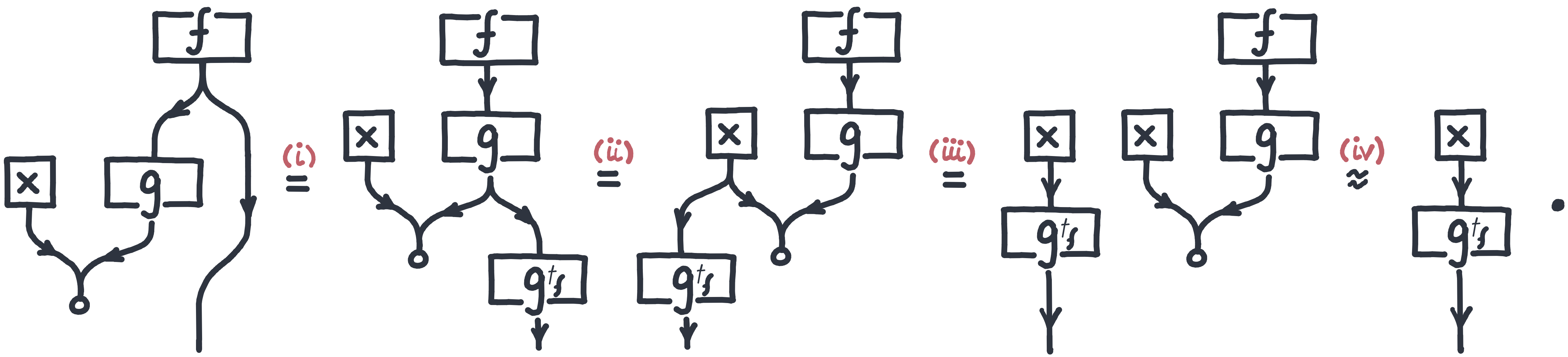
 = Distribution \case
  OneBox → 1/2
  TwoBox → 1/2

 = Distribution \case
  OneBox → 1/2
  TwoBox → 1/2

 = \case
  OneBox , OneBox → 100
  OneBox , TwoBox → 0
  TwoBox, OneBox → 101
  TwoBox, TwoBox → 1

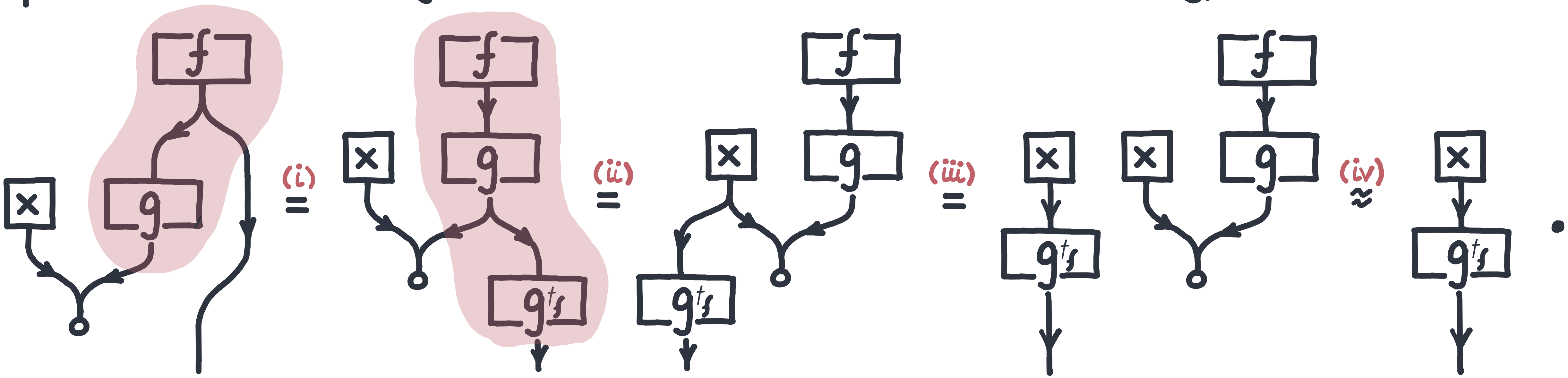
Bayes Theorem

THEOREM. Observing $x \in X$ from a prior $f : 1 \rightsquigarrow A$ through a channel $g : A \rightsquigarrow X$ updates it, up to scalar, to the Bayesian inversion evaluated on the observation, $g^{\dagger}_{f}(x) : 1 \rightsquigarrow A$.



BAYES THEOREM

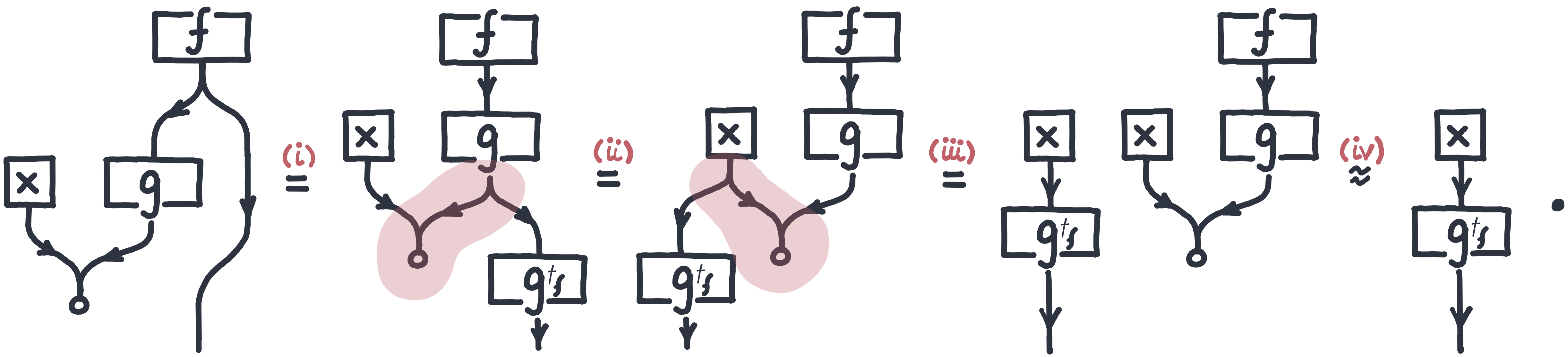
THEOREM. Observing $x \in X$ from a prior $f : 1 \rightsquigarrow A$ through a channel $g : A \rightsquigarrow X$ updates it, up to scalar, to the Bayesian inversion evaluated on the observation, $g^{\dagger}_{f}(x) : 1 \rightsquigarrow A$.



By conditionals.

Bayes Theorem

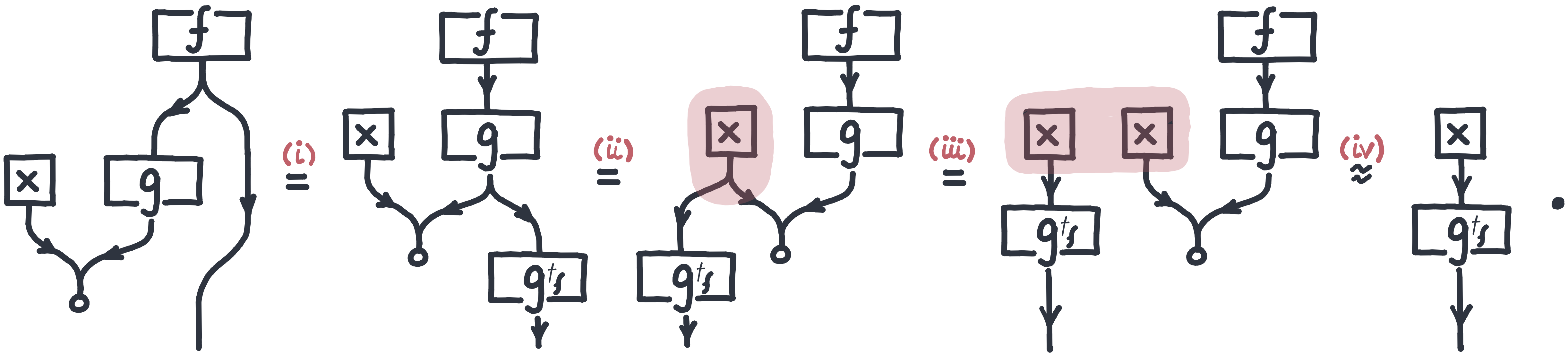
THEOREM. Observing $x \in X$ from a prior $f : 1 \rightsquigarrow A$ through a channel $g : A \rightsquigarrow X$ updates it, up to scalar, to the Bayesian inversion evaluated on the observation, $g^{\dagger}_{f}(x) : 1 \rightsquigarrow A$.



Axiom: Copy, observation.

BAYES THEOREM

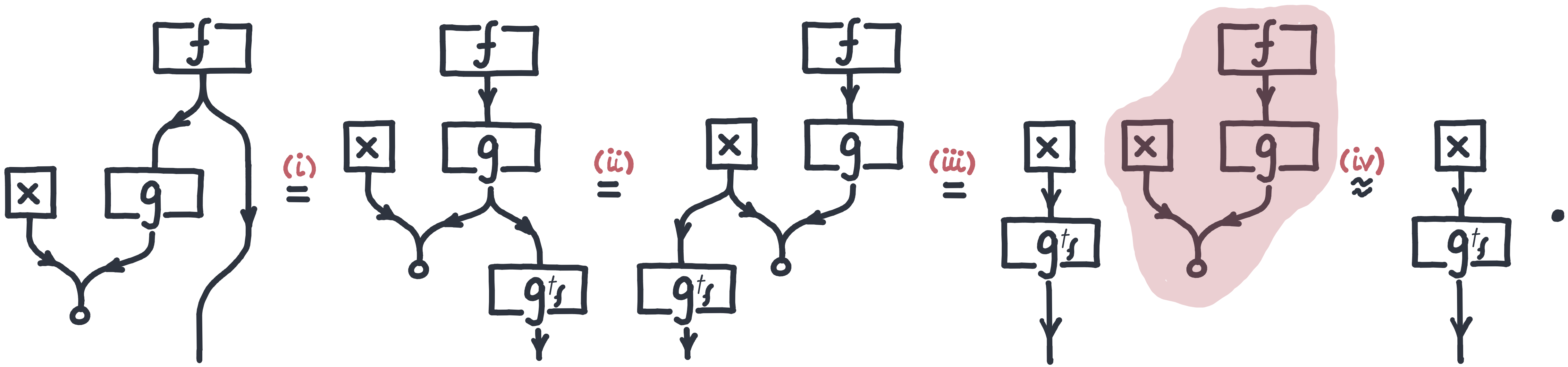
THEOREM. Observing $x \in X$ from a prior $f : 1 \rightsquigarrow A$ through a channel $g : A \rightsquigarrow X$ updates it, up to scalar, to the Bayesian inversion evaluated on the observation, $g^{\dagger}_{f}(x) : 1 \rightsquigarrow A$.



Axiom: x is copyable

BAYES THEOREM

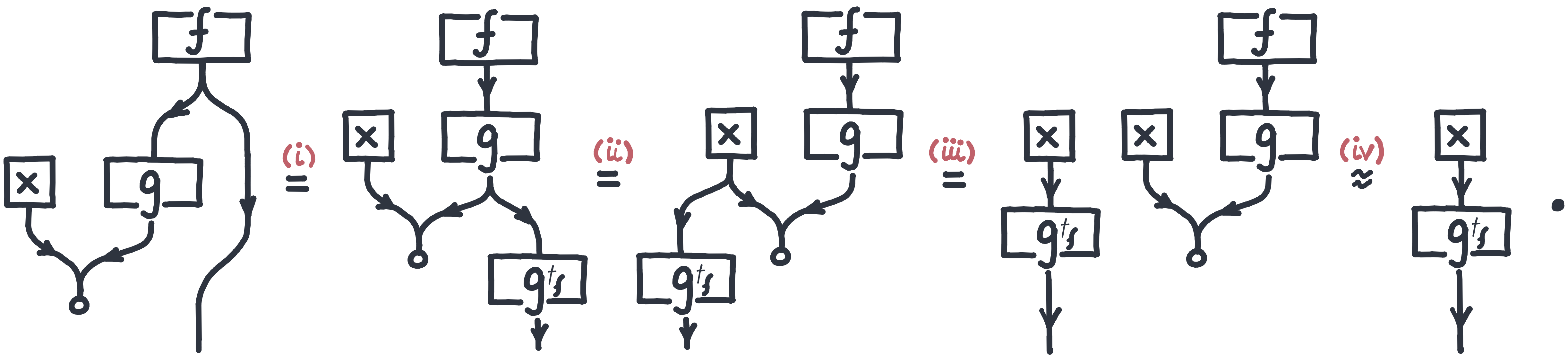
THEOREM. Observing $x \in X$ from a prior $f : 1 \rightsquigarrow A$ through a channel $g : A \rightsquigarrow X$ updates it, up to scalar, to the Bayesian inversion evaluated on the observation, $g^{\dagger}_{f}(x) : 1 \rightsquigarrow A$.



Axiom: we work up to a coefficient.

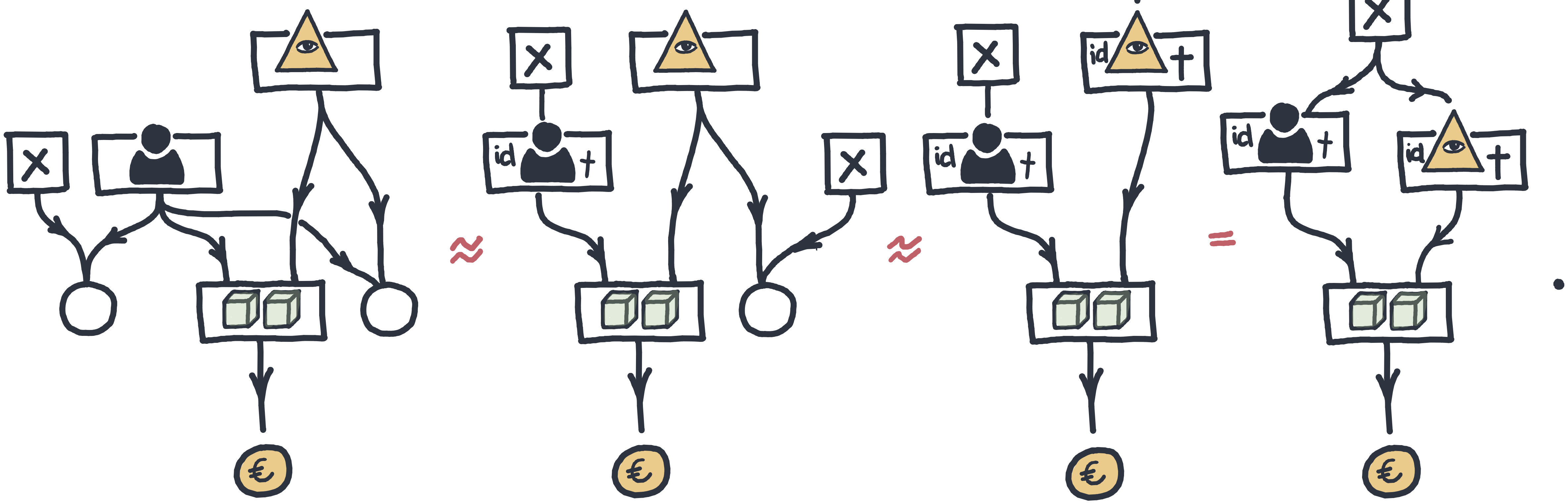
BAYES THEOREM

THEOREM. Observing $x \in X$ from a prior $f : 1 \rightsquigarrow A$ through a channel $g : A \rightsquigarrow X$ updates it, up to scalar, to the Bayesian inversion evaluated on the observation, $g^{\dagger}_{f}(x) : 1 \rightsquigarrow A$.



This acts as a synthetic version of Bayes theorem.

REASONING



The best answer to Newcomb's problem is the same we would give the predictor if they were to act after the agent.

Summary

- Minimal algebra for evidential decision theory.
- Intuitive diagramatic syntax.
- Translating to actual code.
- Synthetically proving a Bayes' theorem.

Partial Markov categories extend synthetic probability algebra to allow observations.

Evidential Decision Theory via Partial Markov

Elena Di Lavore

Mario Román

Tallinn University of Technology.

elendi@ttu.ee

mroman@ttu.ee

Abstract available at
mroman42.github.io

Code available at
github.com/mroman42/bayes-subdistributions

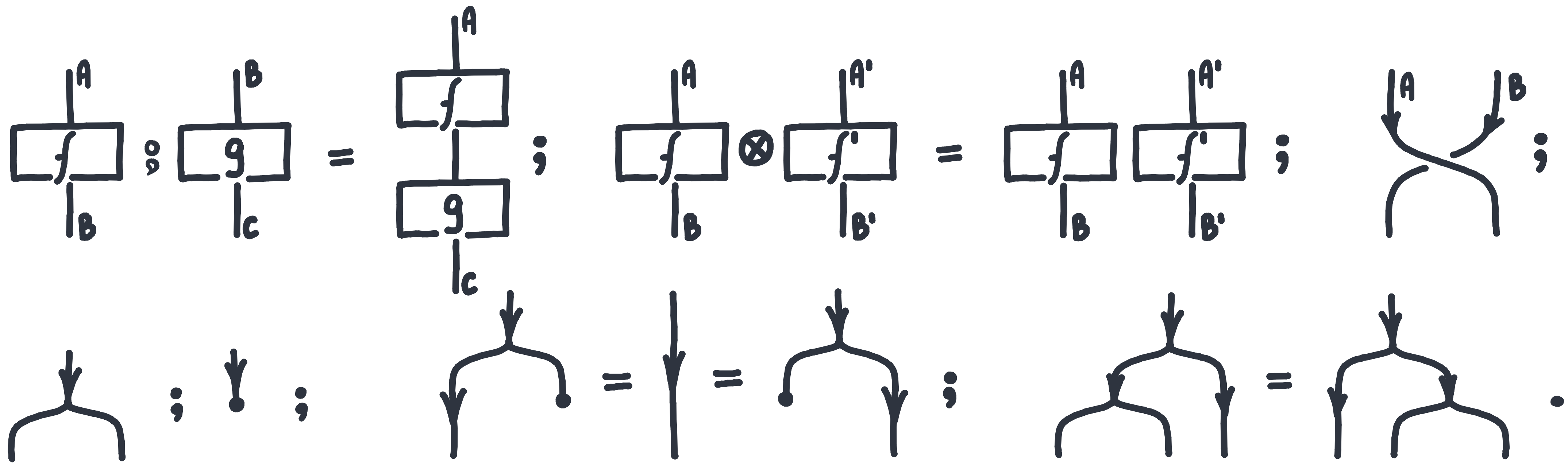
End

REFERENCES

- Fong. A Categorical Perspective on Bayesian Networks.
- Fritz. A Synthetic Approach to Markov kernels.
- Cho, Jacobs. Disintegration and Bayesian inversion.
- Nozick. Newcomb's Problem and Two Principles of Choice.
- Hughes. Generalising Monads to Arrow.

Process Theories

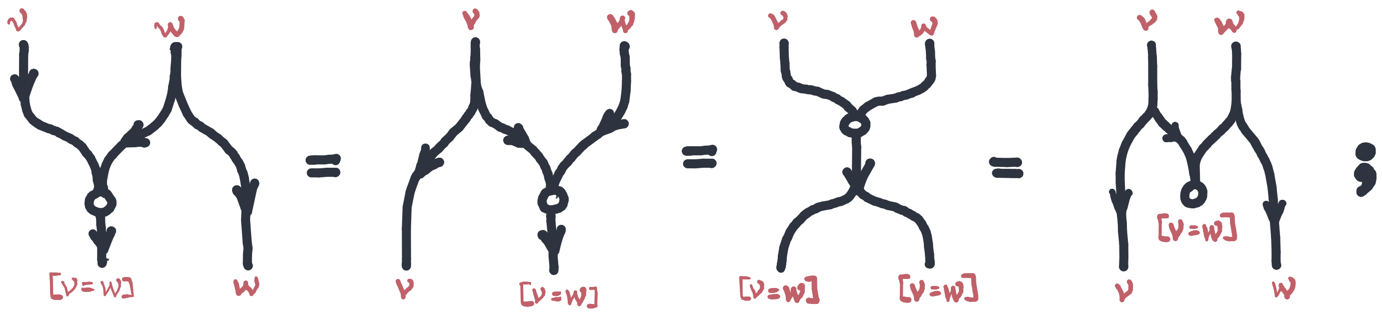
Process theories are symmetric monoidal categories with copy and delete morphisms.



* Technically, a gs-category.

Comparators

Imagine we have two values that we want to compare: v and w.
Our comparator, $(\cdot,\cdot)$, returns any of the two if they are equal.



Axiom. We assume the existence of a comparator, , satisfying the "Frobenius axiom",

$$(\text{observe}(x=y), y) \quad = \quad (x, \text{observe}(x=y)) .$$

Evidential Decision Theory

Evidential decision theory prescribes the action that we would observe to have done in the best possible outcome. This contrasts with classical decision theory, which prescribes the action that causes the best outcome.

Ahmed

$$\operatorname{argmax}_{x \in X} \mathbb{E}\left(out(\cdot \mid x; in)\right).$$

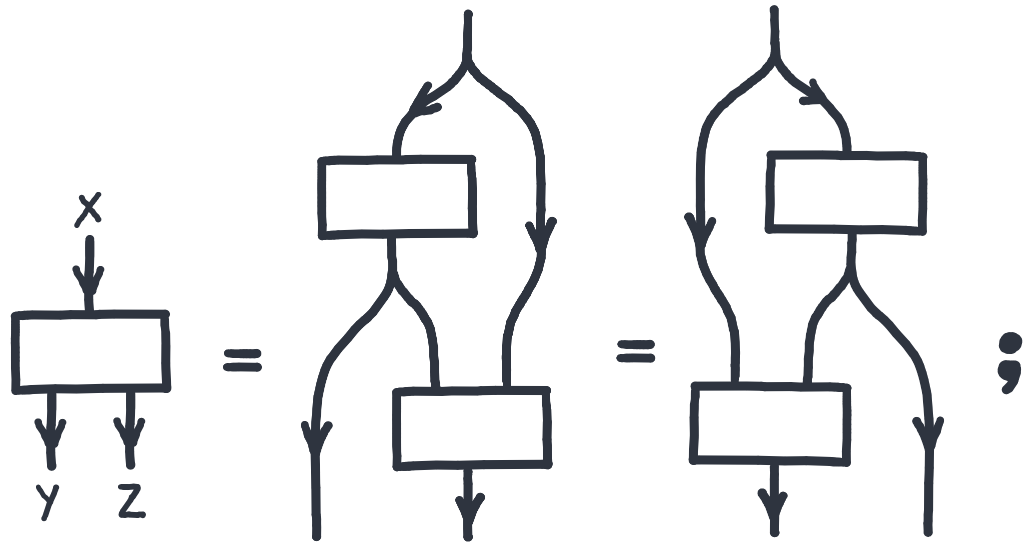
described by the statement

"What is a minimal mathematical framework where this statement can be formalized?"

Conditionals

Fritz.
Cho, Jacobs.

Stochastic processes should have conditionals: they split with the following shape.



;

In terms of stochastic functions, this is

$$\psi(x,y|z) = \sum_{z'} \psi(x|y,z') \cdot \frac{\psi(x|y,z)}{\sum_{z'} \psi(x|y,z')} = \sum_{y' \in Y} \psi(x|y',z) \cdot \frac{\psi(x|y,z)}{\sum_{y' \in Y} \psi(x|y',z)}.$$

Conditionals

Stochastic processes should have conditionals: they split with the following shape.

$$
\begin{array}{l}\textbf{do}\\ \quad (y,z) \leftarrow f(x)\\ \textbf{return}\ (y,z)\end{array}
\;=\;
\begin{array}{l}\textbf{do}\\ \quad y \leftarrow f_1(x)\\ \quad z \leftarrow c_1(x,y)\\ \textbf{return}\ (y,z)\end{array}
\;=\;
\begin{array}{l}\textbf{do}\\ \quad z \leftarrow f_2(x)\\ \quad y \leftarrow c_2(x,z)\\ \textbf{return}\ (y,z)\end{array}
\;;
$$

In terms of stochastic functions, this is

$$
\psi(x,y|z) = \Sigma_{z'}\, \psi(x|y,z') \cdot \frac{\psi(x|y,z)}{\Sigma_{z'}\, \psi(x|y,z')} = \sum_{y'\in Y} \psi(x|y',z) \cdot \frac{\psi(x|y,z)}{\sum_{y'\in Y} \psi(x|y',z)}\,.
$$

NEWCOMB'S PROBLEM

What should the agent do?

- Classical decision theory, in a naïve formulation, prescribes two-boxing: whatever happened before, it is better to take everything.
- Evidential Decision Theory prescribes one-boxing: an agent that takes a single box will find it full.

The idea is always to find the argument that maximizes some expected value function. The debate is in how to translate problems to functions.