Unintended Consequences

Joe Halpern
Cornell University
Computer Science Department

Evan Piermont
Royal Holloway, University of London
Department of Economics

Marie-Louise Vierø
Aarhus University
Department of Economics

Time, Uncertainties, and Strategies XI
The Eighteenth Day of December in the Year of Our Lord Twenty Twenty Five

ATTEMPTING *to dissuade pregnant women from consuming alcohol while pregnant, many US states have passed laws classifying drinking during pregnancy as child abuse/neglect. Such laws are robustly associated with an increase in adverse birth outcomes, i.e., low birth-weight or preterm births.*[1]

[1] Subbaraman, M., and Roberts, S. "Costs associated with policies regarding alcohol use during pregnancy: results from 1972-2015 Vital Statistics." PloS one 14.5 (2019).

SEEING *that the average sparrow eats 2kg of grain per year, the Maoist regime instituted the “eliminate sparrows campaign” encouraging the citizenry to hit noisy pots and pans so as to prevent sparrows from resting in their nests, with the goal of causing them to drop dead from exhaustion. After pushing sparrow populations to near extinction within China, crop yields plummeted.*[2]

[2] Source: copy/paste from wikipedia

DURING *company rule of colonial India, the Governors-General became concerned of the proliferation of cobras. Embracing some newfangled capitalism, he came up with what appeared an ingenious solution — pay a bounty to any man who brought forward a head of a dead cobra. This “cash for snakes” program worked for a while, but was followed by a marked uptick in the number of cobras plaguing the city.*[3]

[3] Source: probably fiction

- In each vignette, an agent misunderstands the true causal structure:
    - Alcoholic women, afraid of prosecution, stopped going to the prenatal doctors appointments
    - Sparrows eat locusts; locusts bad
    - Locals, enticed by the bounty, started breeding cobras
- An agent can be unaware of / inattentive to a variable

In this paper we

- model decision problems as causal structures (Pearl/Halpern style)
    - eventually, model games as well
- allow misperception the causal structure
- examine causal updating and the sustainability of beliefs

Given variables $\mathcal{V}$: a **causal model** $\mathbf{m} = (\mathbf{g}, \mathbf{f})$ is

- a directed acyclic graph, $\mathbf{g}$, of over $\mathcal{V}$
  - Say $Y \rightsquigarrow_{\mathbf{g}} X$ if edge from Y to X
  - Let $\mathcal{R}(X)$ collect the set of values $X \in \mathcal{V}$ can take
- a set of structural equations, $\mathbf{f}$, for each variable

$$\mathbf{f}_X : \prod_{\substack{Y \in \mathcal{V} \\ Y \rightsquigarrow_{\mathbf{g}} X}} \mathcal{R}(Y) \to \mathcal{R}(X)$$

Let $\mathcal{M}$ denote all models (over $\mathcal{V}$)

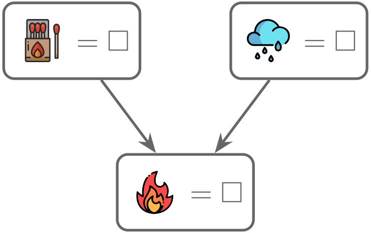
Example

- Three variables: $X_{\text{🌧}}, X_{\text{📦}}, X_{\text{🔥}}$

$$X_{\text{📦}} = 0 \qquad (\mathbf{f}_{X_{\text{📦}}})$$

$$X_{\text{🌧}} = 0 \qquad (\mathbf{f}_{X_{\text{🌧}}})$$

$$X_{\text{🔥}} = \min\{X_{\text{📦}}, 1 - X_{\text{🌧}}\} \qquad (\mathbf{f}_{X_{\text{🔥}}})$$



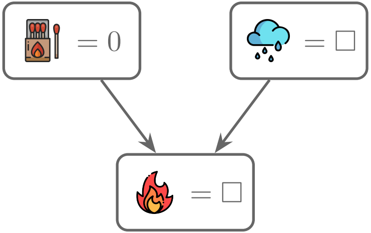
Example

- Three variables: $X_{\text{🌧}}, X_{\text{🧨}}, X_{\text{🔥}}$

$$X_{\text{🧨}} = 0 \qquad (\mathbf{f}_{X_{\text{🧨}}})$$

$$X_{\text{🌧}} = 0 \qquad (\mathbf{f}_{X_{\text{🌧}}})$$

$$X_{\text{🔥}} = \min\{\ 0\ \ , 1 - X_{\text{🌧}}\} \qquad (\mathbf{f}_{X_{\text{🔥}}})$$



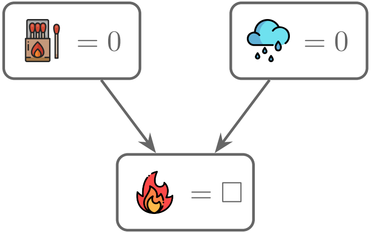
Example

- Three variables: $X_{🌧}, X_{🧨}, X_{🔥}$

$$X_{🧨} = 0 \qquad (\mathbf{f}_{X_{🧨}})$$

$$X_{🌧} = 0 \qquad (\mathbf{f}_{X_{🌧}})$$

$$X_{🔥} = \min\{\ 0\ \ , 1 -\ \ 0\ \ \} \qquad (\mathbf{f}_{X_{🔥}})$$



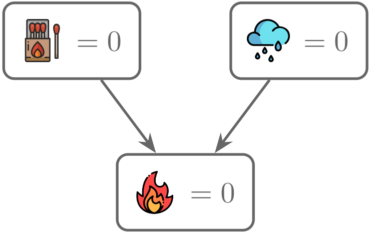
Example

- Three variables: $X_{🌧}, X_{🧨}, X_{🔥}$

$X_{🧨} = 0$ $(\mathbf{f}_{X_{🧨}})$

$X_{🌧} = 0$ $(\mathbf{f}_{X_{🌧}})$

$X_{🔥} = 0$ $(\mathbf{f}_{X_{🔥}})$



Notice:

- if X has no $\mathbf{g}$-ancestors, then $\mathbf{f}_X$ must be constant
- a model determines the value for all variables
- it is possible to add noise / stochastic effects (but not for the talk)

An **intervention** is an agentic assignment of variables:

- $\vec{X} = (X_1 \ldots X_n)$ is a vector of variables
- $\vec{x} = (x_1 \ldots x_n)$ is a vector of values, $x_i \in \mathcal{R}(X_i)$
- Then $[\vec{X} \leftarrow \vec{x}]$ is the intervention setting each X_i to value x_i
- Yields the *counterfactual model* $\mathbf{m}^{[\vec{X} \leftarrow \vec{x}]}$
  - same graph as $\mathbf{m}$, but $\mathbf{f}_{X_i}$ replaced by constant function x_i

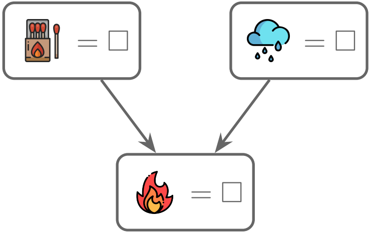
Example

◇ Intervene to set $X_{\text{📦}} \leftarrow 1$

$$X_{\text{📦}} = 0 \quad X_{\text{📦}} = 1 \qquad (\mathbf{f}_{X_{\text{📦}}})$$

$$X_{\text{🌧}} = 0 \qquad (\mathbf{f}_{X_{\text{🌧}}})$$

$$X_{\text{🔥}} = \min\{X_{\text{📦}}, 1 - X_{\text{🌧}}\} \qquad (\mathbf{f}_{X_{\text{🔥}}})$$



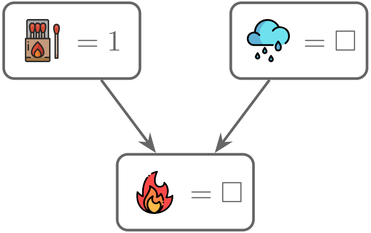
Example

- Intervene to set $X_{\text{🧨}} \leftarrow 1$

$$~~X_{\text{🧨}} = 0~~ \quad X_{\text{🧨}} = 1 \qquad (\mathbf{f}_{X_{\text{🧨}}})$$

$$X_{\text{🌧}} = 0 \qquad (\mathbf{f}_{X_{\text{🌧}}})$$

$$X_{\text{🔥}} = \min\{\ 1\ \ , 1 - X_{\text{🌧}}\} \qquad (\mathbf{f}_{X_{\text{🔥}}})$$



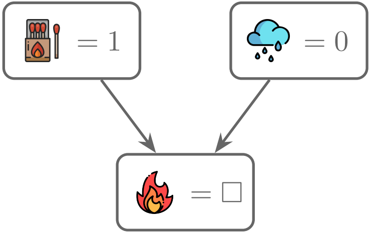
Example

- Intervene to set $X_{🎇} \leftarrow 1$

$$X_{🎇} = 0 \quad X_{🎇} = 1 \qquad (\mathbf{f}_{X_{🎇}})$$

$$X_{🌧} = 0 \qquad (\mathbf{f}_{X_{🌧}})$$

$$X_{🔥} = \min\{\ 1\ \ , 1 -\ \ 0\ \ \} \qquad (\mathbf{f}_{X_{🔥}})$$



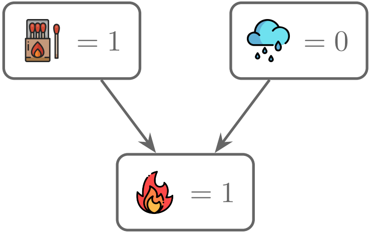
Example

- Intervene to set $X_{🧨} \leftarrow 1$

$X_{🧨} = 0 \quad X_{🧨} = 1$ $(\mathbf{f}_{X_{🧨}})$

$X_{🌧} = 0$ $(\mathbf{f}_{X_{🌧}})$

$X_{🔥} = 1$ $(\mathbf{f}_{X_{🔥}})$



A **causal decision problem** is

- utility function $u : \prod_{X \in \mathcal{V}} \mathcal{R}(X) \to \mathbb{R}$
  - We write $u(\mathbf{m})$ to indicate the utility of the outcome of model $\mathbf{m}$
- A probability μ over $\mathcal{M}$
- A set of actions A, where each action $a = ([\vec{X} \leftarrow \vec{x}], \vec{Y})$ is:
  - An intervention $[\vec{X} \leftarrow \vec{x}]$
  - A set of variables to observe $\vec{Y}$

The utility of action $a \in A$ is

$$U(a, \mu) = \sum_{\mathbf{m} \in \mathcal{M}} u(\mathbf{m}^a)\mu(\mathbf{m})$$

where $\mathbf{m}^a$ is the counterfactual model induced by action a

- ◇ i.e., if $a = ([\vec{X} \leftarrow \vec{x}], \vec{Y})$, then $\mathbf{m}^a = \mathbf{m}^{[\vec{X} \leftarrow \vec{x}]}$

Example: The Governors General

There are four relevant variables:

- $X_{🐍}$: cobra population — range $= \{0, 1, 2\}$
- $X_{🪙}$: there is a bounty — range $= \{0, 1\}$
- $X_{🔪}$: locals hunt snakes — range $= \{0, 1\}$
- $X_{🥚}$: locals breed snakes — range $= \{0, 1\}$

The gg’s utility is

$$u(\vec{X}) = -2X_{🐍} - (X_{🪙} \times X_{🔪})$$

Example

🐍	0	1	2	0	1	2	0	1	2	0	1	2
🪙	0	1	0	1	0	1	1	0	1	1	0	1
🔪	0	1	0	1	0	1	0	1	0	1	0	1
u	0	-3	-4	-1	-2	-5	0	-2	-4	-1	-2	-5

Example

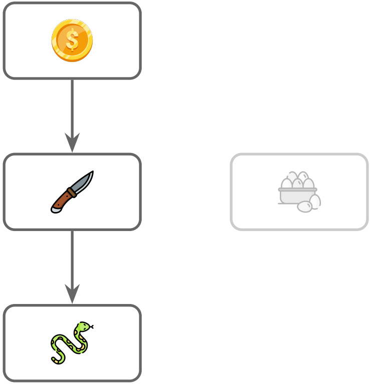
gg has two actions:

$$a_{\text{🪙}} = ([X_{\text{🪙}} \leftarrow 1], (X_{\text{🐍}}, X_{\text{🔪}}))$$

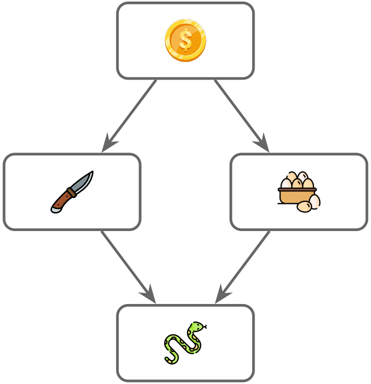
and

$$a_{\text{🚫}} = ([X_{\text{🪙}} \leftarrow 0], (X_{\text{🐍}}, X_{\text{🔪}}))$$

Example



Unaware Model: $\mathbf{m}^\dagger$



True Model: $\mathbf{m}^\star$

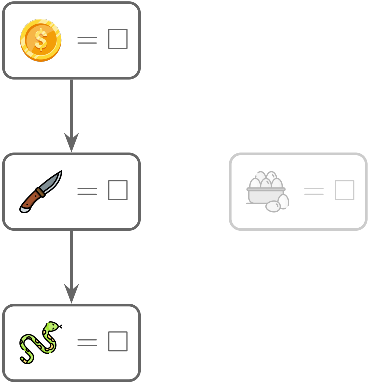
Example

- Structural Equations:

$X_{\text{🧺}} = 0$ $\qquad (\mathbf{f}_{X_{\text{🧺}}})$

$X_{\text{🔪}} = X_{\text{🪙}}$ $\qquad (\mathbf{f}_{X_{\text{🔪}}})$

$X_{\text{🐍}} = 1 - X_{\text{🔪}}$ $\qquad (\mathbf{f}_{X_{\text{🐍}}})$



Example

- Structural Equations:

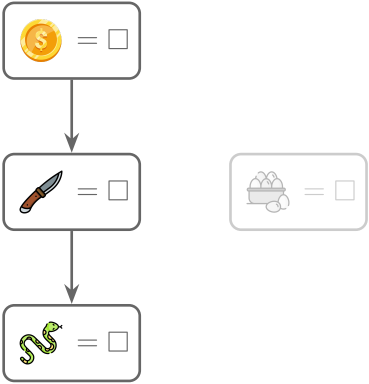
$$X_{\text{🧺}} = 0 \qquad (\mathbf{f}_{X_{\text{🧺}}})$$

$$X_{\text{🔪}} = X_{\text{🪙}} \qquad (\mathbf{f}_{X_{\text{🔪}}})$$

$$X_{\text{🐍}} = 1 - X_{\text{🔪}} \qquad (\mathbf{f}_{X_{\text{🐍}}})$$

- Action $a_{\text{🚫}}$

$$X_{\text{🪙}} = 0 \qquad (\mathbf{f}_{X_{\text{🪙}}})$$



Example

- Structural Equations:

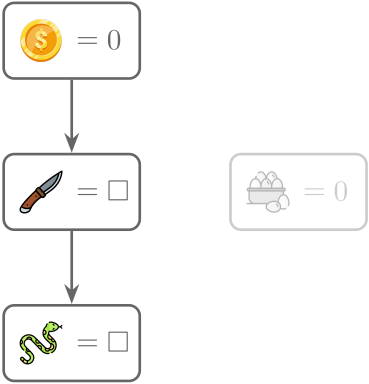
$$X_{\text{🥚}} = 0 \qquad (\mathbf{f}_{X_{\text{🥚}}})$$

$$X_{\text{🔪}} = 0 \qquad (\mathbf{f}_{X_{\text{🔪}}})$$

$$X_{\text{🐍}} = 1 - X_{\text{🔪}} \qquad (\mathbf{f}_{X_{\text{🐍}}})$$

- Action $a_{\text{🚫}}$

$$X_{\text{🪙}} = 0 \qquad (\mathbf{f}_{X_{\text{🪙}}})$$



Example

- Structural Equations:

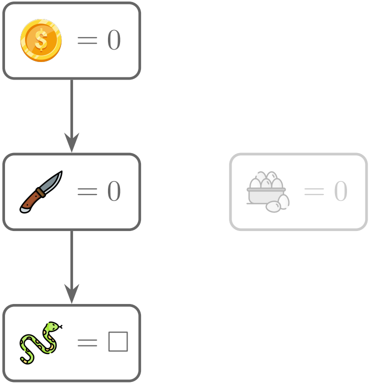
$X_{\text{🧺}} = 0$ $\quad$ $(\mathbf{f}_{X_{\text{🧺}}})$

$X_{\text{🔪}} = 0$ $\quad$ $(\mathbf{f}_{X_{\text{🔪}}})$

$X_{\text{🐍}} = 1 - 0$ $\quad$ $(\mathbf{f}_{X_{\text{🐍}}})$

- Action $a_{\text{🚫}}$

$X_{\text{🪙}} = 0$ $\quad$ $(\mathbf{f}_{X_{\text{🪙}}})$



Example

- Structural Equations:

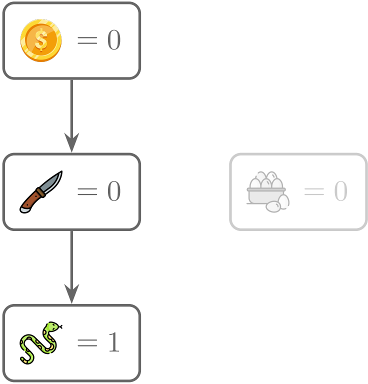
$$X_{🧺} = 0 \qquad (\mathbf{f}_{X_{🧺}})$$

$$X_{🔪} = 0 \qquad (\mathbf{f}_{X_{🔪}})$$

$$X_{🐍} = 1 \qquad (\mathbf{f}_{X_{🐍}})$$

- Action $a_{🚫}$

$$X_{🪙} = 0 \qquad (\mathbf{f}_{X_{🪙}})$$



Example

🐍	0	1	2	0	1	2	0	1	2	0	1	2
🪙	0	1	0	1	0	1	1	0	1	1	0	1
🔪	0	1	0	1	0	1	0	1	0	1	0	1
u	0	-3	-4	-1	-2	-5	0	-2	-4	-1	-2	-5

For $a_{🚫}$:

- gg expects a utility of -2
- gg expects to observe $X_{🐍} = 1$, $X_{🔪} = 0$

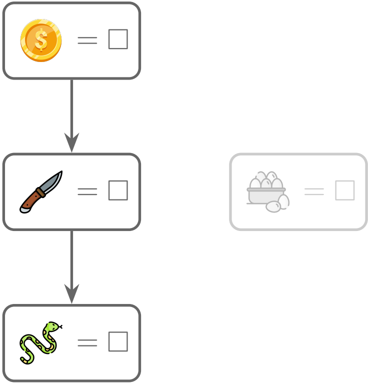
Example

- Structural Equations:

$X_{\text{🧺}} = 0$ $(\mathbf{f}_{X_{\text{🧺}}})$

$X_{\text{🔪}} = X_{\text{🪙}}$ $(\mathbf{f}_{X_{\text{🔪}}})$

$X_{\text{🐍}} = 1 - X_{\text{🔪}}$ $(\mathbf{f}_{X_{\text{🐍}}})$



Example

- Structural Equations:

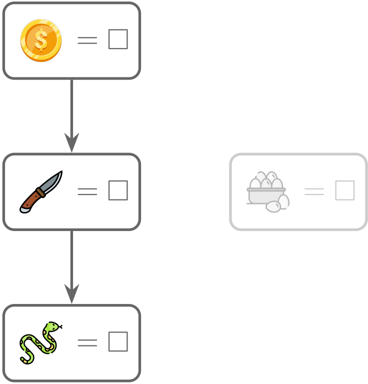
$X_{\text{🥚}} = 0$ $\quad(\mathbf{f}_{X_{\text{🥚}}})$

$X_{\text{🔪}} = X_{\text{🪙}}$ $\quad(\mathbf{f}_{X_{\text{🔪}}})$

$X_{\text{🐍}} = 1 - X_{\text{🔪}}$ $\quad(\mathbf{f}_{X_{\text{🐍}}})$

- Action $a_{\text{🪙}}$

$X_{\text{🪙}} = 1$ $\quad(\mathbf{f}_{X_{\text{🪙}}})$



Example

- Structural Equations:

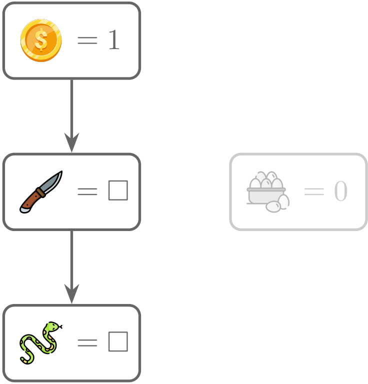
$X_{\text{🥚}} = 0$ $(\mathbf{f}_{X_{\text{🥚}}})$

$X_{\text{🔪}} = 1$ $(\mathbf{f}_{X_{\text{🔪}}})$

$X_{\text{🐍}} = 1 - X_{\text{🔪}}$ $(\mathbf{f}_{X_{\text{🐍}}})$

- Action $a_{\text{🪙}}$

$X_{\text{🪙}} = 1$ $(\mathbf{f}_{X_{\text{🪙}}})$



Example

- Structural Equations:

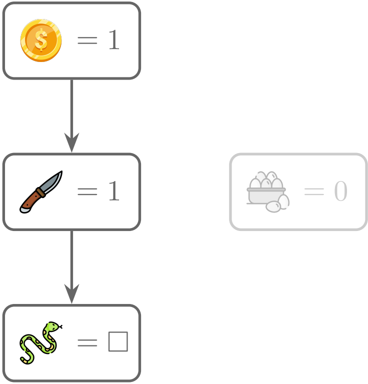
$X_{🧺} = 0$ $(\mathbf{f}_{X_{🧺}})$

$X_{🔪} = 1$ $(\mathbf{f}_{X_{🔪}})$

$X_{🐍} = 1 - 1$ $(\mathbf{f}_{X_{🐍}})$

- Action $a_{🪙}$

$X_{🪙} = 1$ $(\mathbf{f}_{X_{🪙}})$



Example

- Structural Equations:

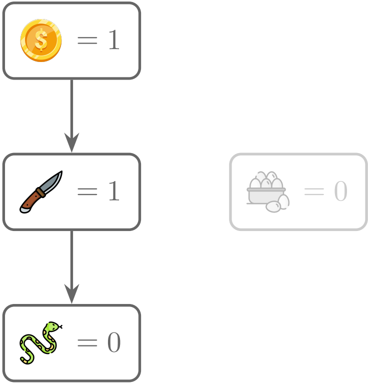
$X_{\text{🥚}} = 0$ $(\mathbf{f}_{X_{\text{🥚}}})$

$X_{\text{🔪}} = 1$ $(\mathbf{f}_{X_{\text{🔪}}})$

$X_{\text{🐍}} = 0$ $(\mathbf{f}_{X_{\text{🐍}}})$

- Action $a_{\text{🪙}}$

$X_{\text{🪙}} = 1$ $(\mathbf{f}_{X_{\text{🪙}}})$



Example

🐍	0	1	2	0	1	2	0	1	2	0	1	2
🪙	0	1	0	1	0	1	1	0	1	1	0	1
🔪	0	1	0	1	0	1	0	1	0	1	0	1
u	0	-3	-4	-1	-2	-5	0	-2	-4	-1	-2	-5

For $a_{\text{🪙}}$:

- gg expects a utility of -1
- gg expects to observe $X_{\text{🐍}} = 0$, $X_{\text{🔪}} = 1$

Example

If μ is certainty in $\mathbf{m}^{\dagger}$, then

$$U(a_{\text{🪙}}, \mu) = -1 > -2 = U(a_{\text{🚫}}, \mu)$$

- gg would choose $a_{\text{🪙}}$

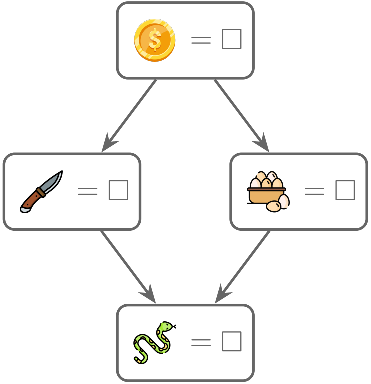
Example

- Structural Equations:

$$X_{\text{🥚}} = X_{\text{🪙}} \qquad (\mathbf{f}_{X_{\text{🥚}}})$$

$$X_{\text{🔪}} = X_{\text{🪙}} \qquad (\mathbf{f}_{X_{\text{🔪}}})$$

$$X_{\text{🐍}} = 1 - X_{\text{🔪}} + (2 \times X_{\text{🥚}}) \qquad (\mathbf{f}_{X_{\text{🐍}}})$$



Example

- Structural Equations:

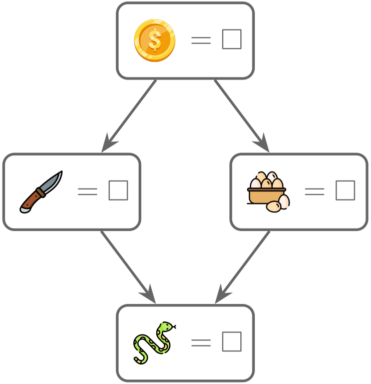
$$X_{\text{🥚}} = X_{\text{🪙}} \qquad (\mathbf{f}_{X_{\text{🥚}}})$$

$$X_{\text{🔪}} = X_{\text{🪙}} \qquad (\mathbf{f}_{X_{\text{🔪}}})$$

$$X_{\text{🐍}} = 1 - X_{\text{🔪}} + (2 \times X_{\text{🥚}}) \qquad (\mathbf{f}_{X_{\text{🐍}}})$$

- Action $a_{\text{🪙}}$

$$X_{\text{🪙}} = 1 \qquad (\mathbf{f}_{X_{\text{🪙}}})$$



Example

- Structural Equations:

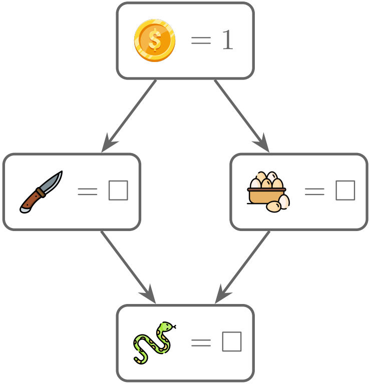
$$X_{🥚} = 1 \qquad (\mathbf{f}_{X_{🥚}})$$

$$X_{🔪} = 1 \qquad (\mathbf{f}_{X_{🔪}})$$

$$X_{🐍} = 1 - X_{🔪} + (2 \times X_{🥚}) \qquad (\mathbf{f}_{X_{🐍}})$$

- Action $a_{🪙}$

$$X_{🪙} = 1 \qquad (\mathbf{f}_{X_{🪙}})$$



Example

- Structural Equations:

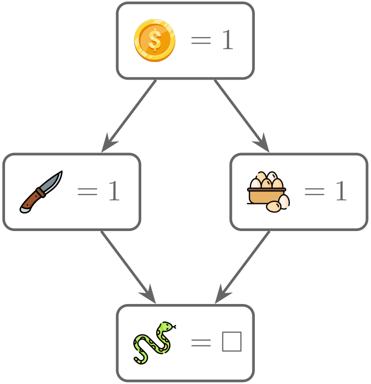
$$X_{🥚} = 1 \qquad (\mathbf{f}_{X_{🥚}})$$

$$X_{🔪} = 1 \qquad (\mathbf{f}_{X_{🔪}})$$

$$X_{🐍} = 1 - 1 + (2 \times 1) \qquad (\mathbf{f}_{X_{🐍}})$$

- Action $a_{🪙}$

$$X_{🪙} = 1 \qquad (\mathbf{f}_{X_{🪙}})$$



Example

- Structural Equations:

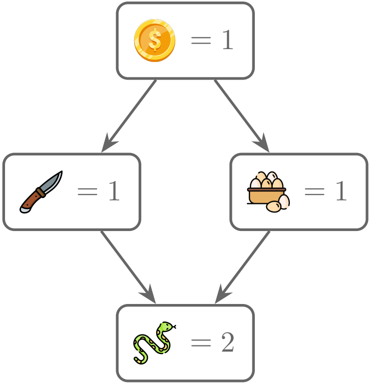
$$X_{\text{🥚}} = 1 \qquad (\mathbf{f}_{X_{\text{🥚}}})$$

$$X_{\text{🔪}} = 1 \qquad (\mathbf{f}_{X_{\text{🔪}}})$$

$$X_{\text{🐍}} = 2 \qquad (\mathbf{f}_{X_{\text{🐍}}})$$

- Action $a_{\text{🪙}}$

$$X_{\text{🪙}} = 1 \qquad (\mathbf{f}_{X_{\text{🪙}}})$$



Example

- gg expects to see $X_{🐍} = 0$ but instead sees $X_{🐍} = 2$
- this is an *unintended consequence*
- notice there is a ”projection consistency”:
  - When $X_{🧺} = 0$, $\mathbf{f}^{\star}$ reduces to $\mathbf{f}^{\dagger}$

$$\mathbf{f}^{\star}_{X_{🐍}} = 1 - X_{🔪} + 2X_{🧺} \quad \underset{X_{🧺}\to 0}{\Longrightarrow} \quad \mathbf{f}^{\dagger}_{X_{🐍}} = 1 - X_{🔪}$$

To model ‘continuation behavior,’ we need to consider:

- Evidence
- Reaction to evidence

Given an action $a = ([\vec{X} \leftarrow \vec{x}], \vec{Y})$ and a model $\mathbf{m}$:

$E^a_{\mathbf{m}} \subseteq \mathcal{M}$ is all models that generate the same observable after a as $\mathbf{m}$

- ◇ $\hat{\mathbf{m}} \in E^a_{\mathbf{m}}$ if and only if $\mathbf{m}^{[\vec{X} \leftarrow \vec{x}]}$ and $\hat{\mathbf{m}}^{[\vec{X} \leftarrow \vec{x}]}$ yield the same values for $\vec{Y}$

Instead of a single belief

- consider a lexicographic probability system: $\lambda = (\mu_0, \ldots, \mu_n)$
- λ evaluates probabilities according to μ_0
- However, given an event E define:

$$\lambda(\,\cdot\mid E\,) := (\mu_{k_0}(\,\cdot\mid E\,), \ldots, \mu_{k_m}(\,\cdot\mid E\,))$$

  - where k_i is the i^{th} smallest index such that $\mu_{k_i}(E) > 0$.
  - i.e., condition all elements of the l.p.s., dropping incompatible elements

A **dynamic causal decision problem** is

- utility function u
- A lexicographic probability system $\lambda = (\mu_0, \ldots, \mu_n)$ over $\mathcal{M}$
- A set of actions A
- A discount rate δ

The value of an action is then:

$$V(a,\lambda) = \sum_{\mathbf{m}\in\mathcal{M}} \Big(u(\mathbf{m}^a) + \delta \max_{a\in A} V(a, \lambda(\,\cdot \mid E^a_{\mathbf{m}})) \Big) \lambda(\mathbf{m})$$

- Classic exploration / exploitation trade-off
- The agent anticipates beliefs to evolve...
- ...but only insofar as her current belief $\mu_0 \in \lambda$

A **steady-state** is a triple $(\mathbf{m}^\star, \lambda, a)$ where:

(1) $a \in \arg\max_{a' \in A} V(a', \lambda)$

(2) $\lambda(\,\cdot\,) = \lambda(\,\cdot \mid E^a_{\mathbf{m}^\star})$

- ◇ $\mathbf{m}^\star$ is the true model
- ◇ a is optimal given current beliefs λ
- ◇ observed evidence $E^a_{\mathbf{m}^\star}$ does not change beliefs

Example

Let $\lambda = (\mu_0, \mu_1, \ldots)$ where $\mu_0(\mathbf{m}^\dagger) = 1$, $\mu_1(\mathbf{m}^\star) = 1$:

- The action $a_{\text{🪙}}$ is optimal given λ
- But, this is *not* a steady-state:
  - $\mathbf{m}^\star$ yields observation $o = (X_{\text{🐍}} = 2, X_{\text{🔪}} = 1)$
  - $\mathbf{m}^\dagger$ (counterfactually) yields observation $o = (X_{\text{🐍}} = 0, X_{\text{🔪}} = 1)$
- $\lambda' := \lambda(\,\cdot \mid E^{a_{\text{🪙}}}_{m^\star}) \neq \lambda$ (in fact, $\cong \mu_1$)
- $(\mathbf{m}^\star, \lambda', a_{\text{🚫}})$ is a steady-state

Example

Fix $\lambda^0 = (\mu_0, \mu_1, \ldots)$ and $\mathbf{m}^\star$. For $k \geq 1$ recursively define

- $a^k \in \arg\max_A V(\,\cdot\,, \lambda^{k-1})$
- $\lambda^k = \lambda^{k-1}(\,\cdot \mid E^{a^k}_{\mathbf{m}^\star})$

Then $(\mathbf{m}^\star, \lambda^k, a^k)$ converges to a steady-state

SINCE *time immemorial, members of the Aition tribe have woken before sunrise to preform an elaborate ritual to their God Apotelesma. They operate under the conviction that, without their daily votive, she would not allow the sun to rise. Everyday, in the daylight after their morning prayer, the Aitions conclude their model of the world must be correct as it accurately predicted the sunrise.*[4]

[4] Source: Hemingway, E. ”The Sun Also Rises” New York: Scribner’s (1926).

Example: The Aitions

There are two relevant variable's:

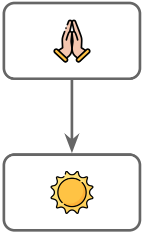
$X_{🙏}$: the Aitions make daily offering — range $= \{0, 1\}$

$X_{☀}$: the sun rises — range $= \{0, 1\}$

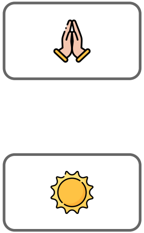
The Ations' utility is

$$u(\vec{X}) = 2X_{☀} - X_{🙏}$$

Example



Unaware Model: $\mathbf{m}^{\dagger}$



True Model: $\mathbf{m}^{\star}$

$X_{\text{🙏}} = 0 \qquad (\mathbf{f}_{X_{\text{🙏}}})$

$X_{\text{☀}} = X_{\text{🙏}} \qquad (\mathbf{f}_{X_{\text{☀}}})$

$X_{\text{🙏}} = 0 \qquad (\mathbf{f}_{X_{\text{🙏}}})$

$X_{\text{☀}} = 1 \qquad (\mathbf{f}_{X_{\text{☀}}})$

Example

The Aitions has two actions:

$$a_{\text{🚫}} = ([X_{\text{🙏}} \leftarrow 0], (X_{\text{🙏}}, X_{\text{☀}}))$$

$$a_{\text{🙏}} = ([X_{\text{🙏}} \leftarrow 1], (X_{\text{🙏}}, X_{\text{☀}}))$$

🙏	0	0	1	1
☀	0	1	0	1
u	0	2	-1	1

Example

The Aitions has two actions:

$$a_{\text{🚫}} = ([X_{\text{🙏}} \leftarrow 0], (X_{\text{🙏}}, X_{\text{☀️}}))$$

$$u =^{\dagger} 0, \quad X_{\text{🙏}} =^{\dagger} 0, \quad X_{\text{☀️}} =^{\dagger} 0$$

$$a_{\text{🙏}} = ([X_{\text{🙏}} \leftarrow 1], (X_{\text{🙏}}, X_{\text{☀️}}))$$

🙏	0	0	1	1
☀️	0	1	0	1
u	0	2	-1	1

Example

The Aitions has two actions:

$$a_{🚫} = ([X_{🙏} \leftarrow 0], (X_{🙏}, X_{☀}))$$

$$u =^{\dagger} 0, \quad X_{🙏} =^{\dagger} 0, \quad X_{☀} =^{\dagger} 0$$

$$u =^{\star} 2, \quad X_{🙏} =^{\star} 0, \quad X_{☀} =^{\star} 1$$

$$a_{🙏} = ([X_{🙏} \leftarrow 1], (X_{🙏}, X_{☀}))$$

🙏	0	0	1	1
☀	0	1	0	1
u	0	2	-1	1

Example

The Aitions has two actions:

$$a_{\text{🚫}} = ([X_{\text{🙏}} \leftarrow 0], (X_{\text{🙏}}, X_{\text{☀}}))$$

$$u =^{\dagger} 0, \quad X_{\text{🙏}} =^{\dagger} 0, \quad X_{\text{☀}} =^{\dagger} 0$$

$$u =^{\star} 2, \quad X_{\text{🙏}} =^{\star} 0, \quad X_{\text{☀}} =^{\star} 1$$

$$a_{\text{🙏}} = ([X_{\text{🙏}} \leftarrow 1], (X_{\text{🙏}}, X_{\text{☀}}))$$

$$u =^{\dagger} 1, \quad X_{\text{🙏}} =^{\dagger} 1, \quad X_{\text{☀}} =^{\dagger} 1$$

$$u =^{\star} 1, \quad X_{\text{🙏}} =^{\star} 1, \quad X_{\text{☀}} =^{\star} 1$$

🙏	0	0	1	1
☀	0	1	0	1
u	0	2	-1	1

Example

Let $\lambda = (\mu_0, \mu_1, \ldots)$ where $\mu_0(\mathbf{m}^\dagger) = 1$, $\mu_1(\mathbf{m}^\star) = 1$:

- The action $a_{\text{🙏}}$ is optimal given λ
- This *is* a steady-state:
    - $\mathbf{m}^\star$ yields the expected observation
    - there is no belief updating

- The Aitions are too confident — they do not explore enough
- How to model an agent who is:
    - Not fully aware of all causal structures, but
    - Introspective: acknowledge the above

We add the following ‘subjective components’ to the model:

- $\overline{V}(*) \in \mathbb{R}$ — the (lifetime) value to changing beliefs
- $\pi : A \to [0, 1]$ — $\pi(a)$ is prob that a will yield $*$
- $\tau : [0, 1]^A \times (A \times 2^{\mathcal{M}}) \to [0, 1]^A$, updating rule for π
  - write $\pi(\,\cdot \mid a, E\,)$ instead of $\tau(\pi, (a, E))$
  - is the probabilities after observing action a produce evidence E
  - Assume: $\pi(a \mid a, E\,) \leq \pi(\,a\,)$; to 0 as a chosen many times

The value of an action $\overline{V}(a,\lambda,\pi)$ is

$$\pi(a)\,\overline{V}(*) + (1-\pi(a))\sum_{\mathbf{m}\in\mathcal{M}}\Big(u(\mathbf{m}^a) + \delta\max_{a\in A}\overline{V}(a,\lambda',\pi')\Big)\lambda(\mathbf{m})$$

where $\lambda' = \lambda(\cdot \mid E^a_{\mathbf{m}})$ and $\pi' = \pi(\cdot \mid a, E^a_{\mathbf{m}})$.

Two levels of exploration / exploitation: Unawareness & Probability

An **introspective steady-state** is a tuple $(\mathbf{m}^\star, \lambda, a, \overline{V}(*), \pi)$ where:

(1) $a \in \arg\max_{a' \in A} \overline{V}(a', \mu)$

(2) $\lambda(\,\cdot\,) = \lambda(\,\cdot \mid E^a_{\mathbf{m}^\star})$

(3) $\pi(\,\cdot\,) = \pi(\,\cdot \mid a, E^a_{\mathbf{m}^\star})$

- For $\overline{V}(*)$ large, if $(\mathbf{m}^\star, \lambda, a, \overline{V}(*), \pi)$ is an iss, then $(\mathbf{m}^\star, \lambda, a)$ is a ss.
- For $\overline{V}(*)$ small, if $(\mathbf{m}^\star, \lambda, a)$ is a ss, then $(\mathbf{m}^\star, \lambda, a, \overline{V}(*), \pi)$ is an iss (for some π)

Example

Let $\lambda = (\mu_0, \mu_1, \ldots)$ where $\mu_0(\mathbf{m}^\dagger) = 1$; consider:

$$\xi = (\mathbf{m}^\star, \lambda, a_{\text{🙏}}, \overline{V}(*), \pi)$$

- If $\pi(a_{\text{🚫}})\,\overline{V}(*)$ is high, then ξ is not an is not a iss
- If $\pi(a_{\text{🚫}})\,\overline{V}(*)$ is low, then (if $\pi(a_{\text{🙏}}) = 0$) ξ is an iss

AFTER *visiting Paris for TUS last year, I developed a penchant for savory French tarts. When I got back to the America, I went into a diner and ordered a Quiche Lorraine. I was but two bites into it when a surly looking man came up and punched me in the face.*[5]

[5] Source: David Kreps' psychedelic visions.

Our eventual goal is use this framework to study game theory:

- Players’ actions are sets of interventions
- Equilibrium is mutual steady-states
- Allows for complete subjectivity of the game form
- Higher-order beliefs seems painful

Thanks