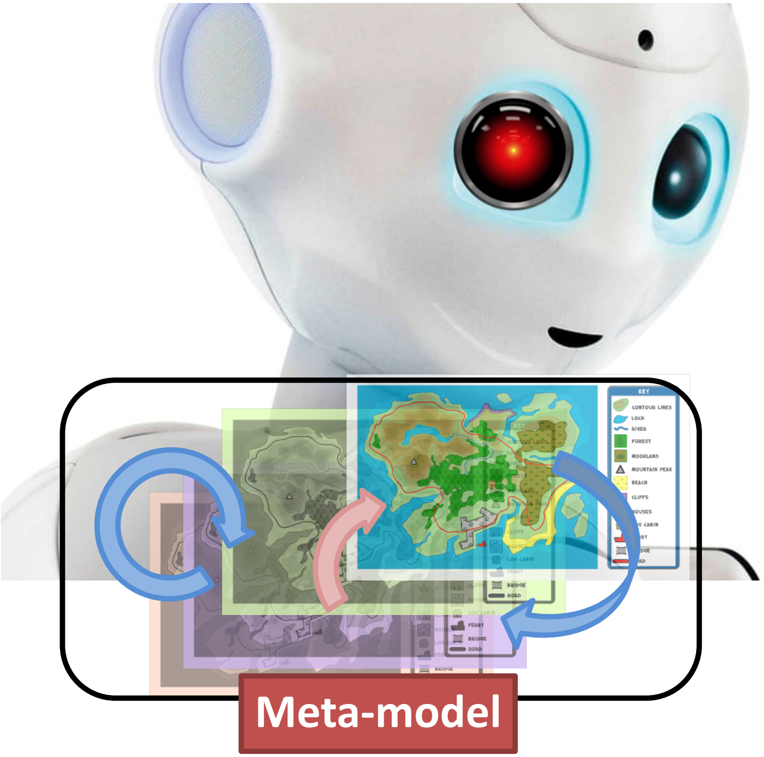
Future of Humanity Institute
UNIVERSITY OF OXFORD

MIRI



"Generalised" models: why, how?

Automating extension of (moral) categories

Stuart Armstrong
Future of Humanity Institute, Oxford University

Strong underdefined preferences

Consider a Croatian, communist, Yugoslav nationalist in the 1980s...

Morality: past, present, and future

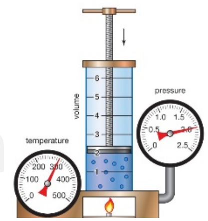
Honour is vital

Women should be protected

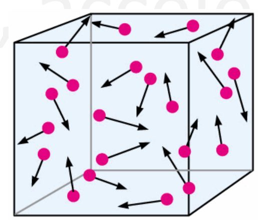
Happiness is important

Physical model splintering

- Ideal gas laws
- Van der Waal laws
- Bouncing atom models
- Quantum models
- …

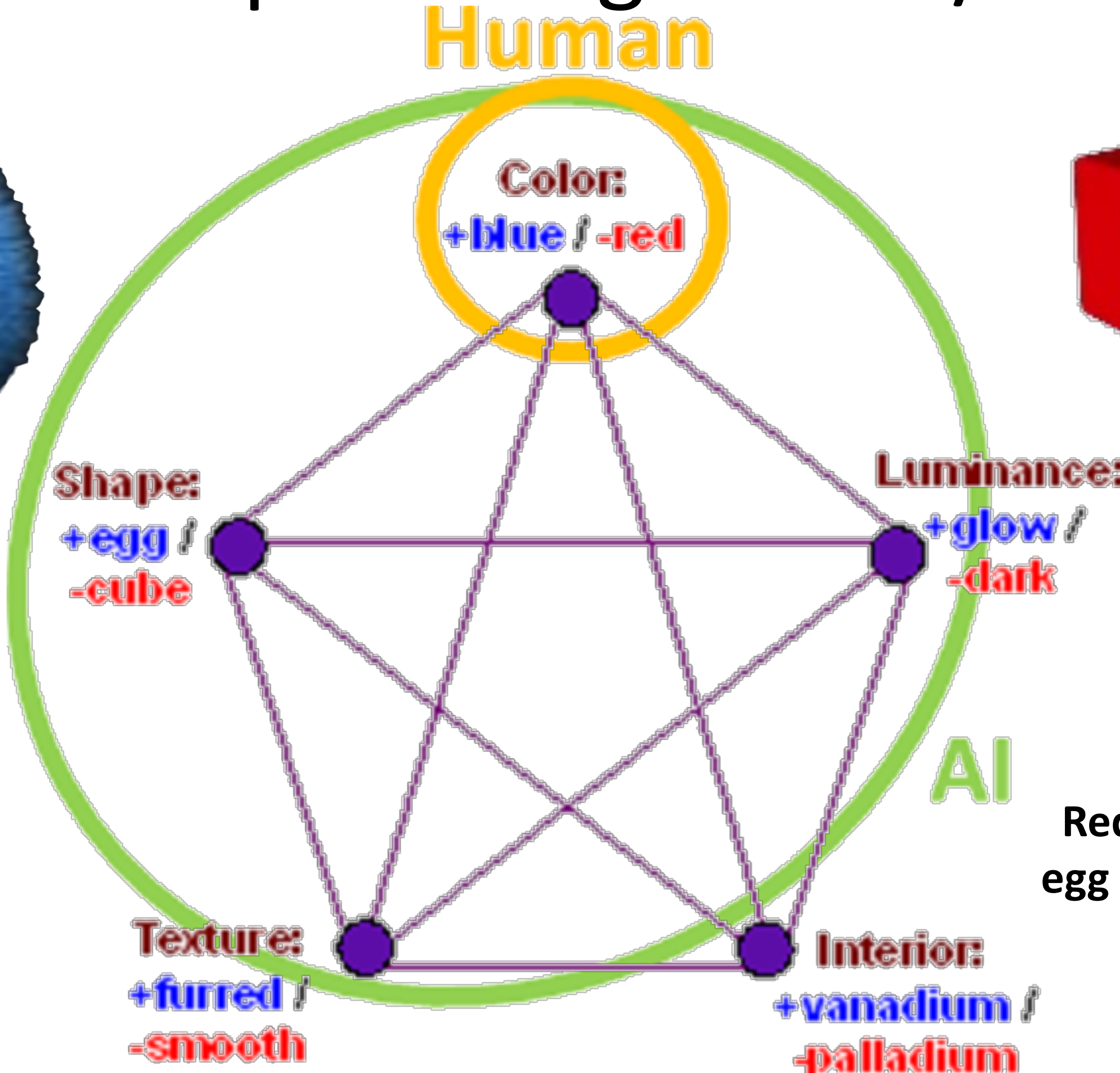


$$PV = nRT$$

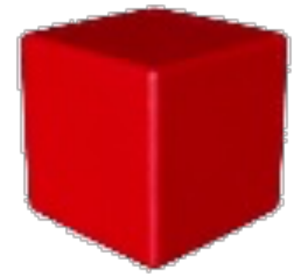


Model splintering: Rubes/Bleggs

Blegg



Rube

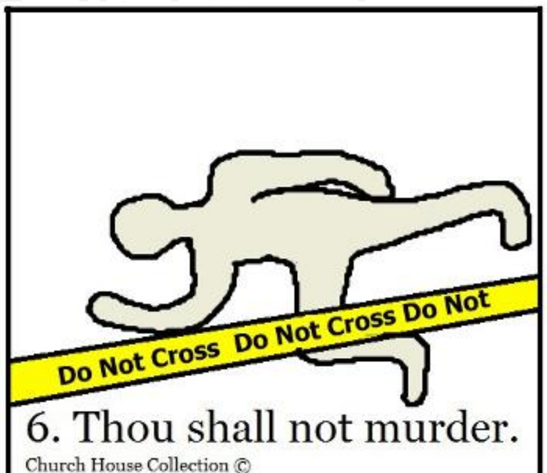


?

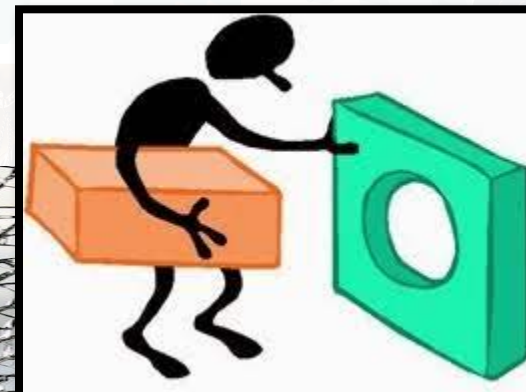
Red, dark, furred egg with vanadium

Moral model splintering

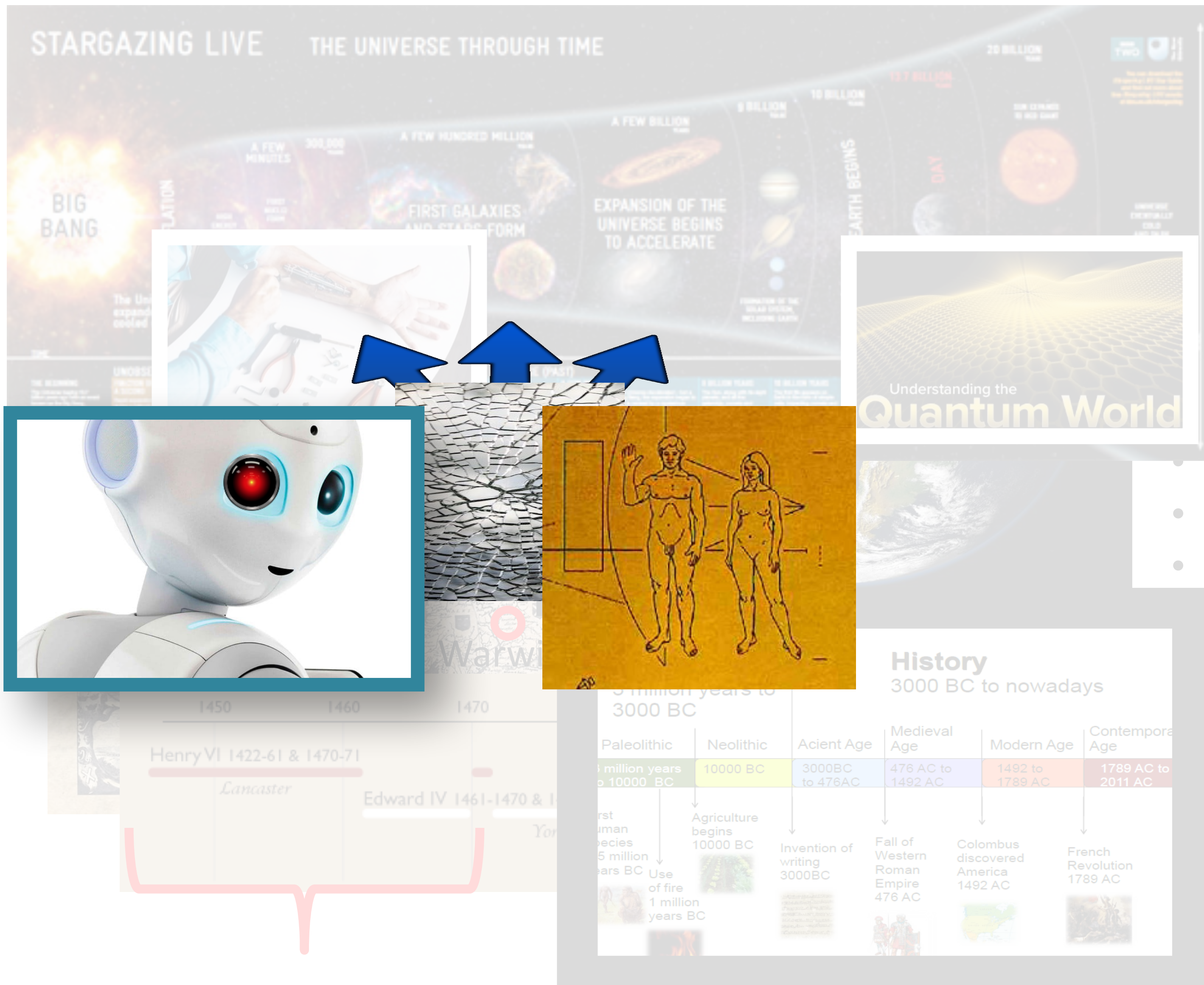
Many common conclusions



Incompatible/incomparable concepts and premisses



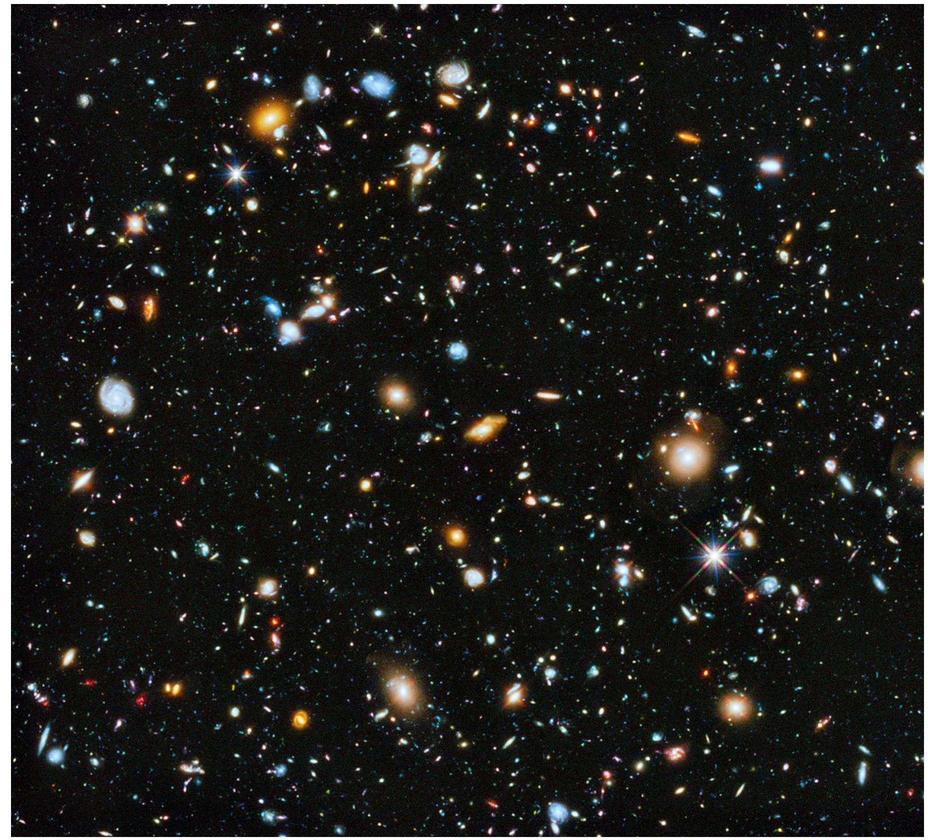
The general problem

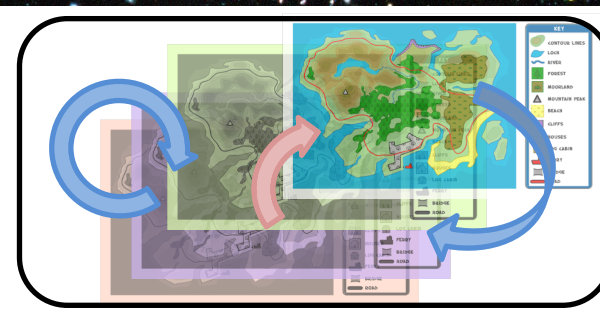


- Money
- The harvest
- Feudal duty
- Teaching children
- Spears and armour
- York vs Lancaster
- House of Warwick
- God
- Morality
- Feudal hierarchy

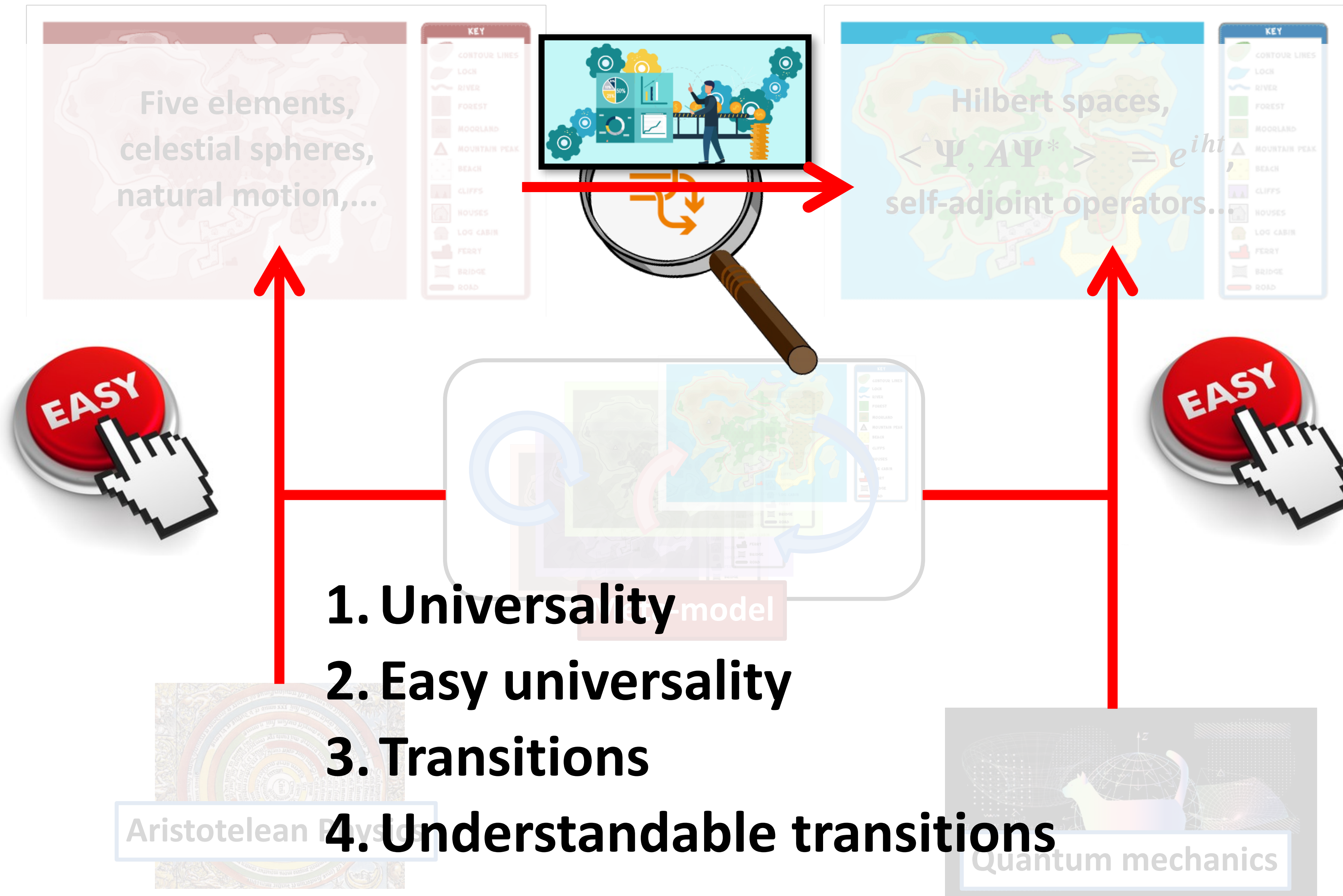
Universality

- Turing machines
- Neural nets
- Set theory
- Second-order logic
- Bayesian updating
- Category theory
- …
- “Generalised” models





Meta-model desiderata



1. Universality
2. Easy universality
3. Transitions
4. Understandable transitions

Application to most of AI safety

Hidden complexity of wishes	
Ontological crises	
Conservative behaviour	
Goodhart problems	
Wireheading	
Out-of-distribution behaviour	
Low impact	
Underdefined preferences	
Active inverse reward design	
The whole friendly AI problem	

Application to most of AI safety

Hidden complexity of wishes	Save* my mother* [*: underdefined]
Ontological crises	When models of physics splinter
Conservative behaviour	When be conservative? When models splinter
Goodhart problems	“Measure used = desired behaviour” splinters
Wireheading	“Reward channel = desired behaviour” splinters
Out-of-distribution behaviour	The current ML version of this problem
Low impact	Low impact = features similar to before
Underdefined preferences	Example in this presentation
Active inverse reward design	Clear reward over underdefined features
The whole friendly AI problem	“Friendly” well defined in typical situations

Generalised models

$$\mathscr{M} = \{\mathscr{F}, \mathscr{E}, Q\}$$

$\mathscr{F}$ a set of features

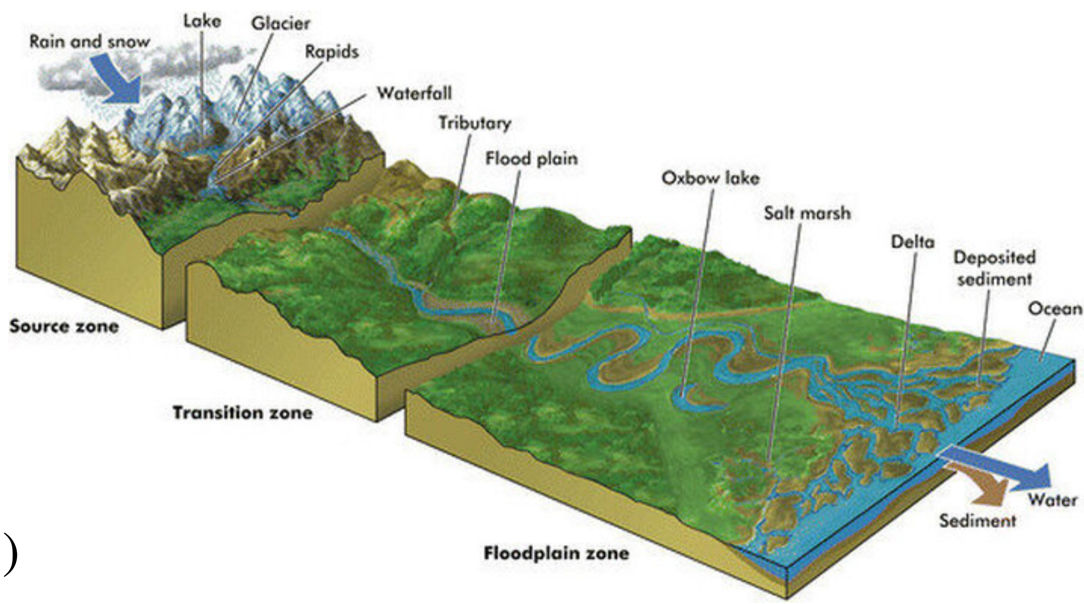
$\mathscr{F} = \{(n, \bar{\mathscr{F}})\}$: n name, $\bar{\mathscr{F}}$ possible values

$\mathscr{E}$ a set of environments

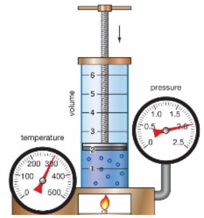
$$\mathscr{E} \subset \mathscr{W} = 2^{\sqcup \bar{\mathscr{F}}}$$

Q a probability distribution (partial, un-normalised?)

$(n =$ "temperature", $\bar{\mathscr{F}} = \{r > 0\})$



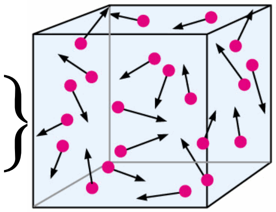
Generalised models



$PV = nRT$

$$\mathscr{M}_0 = \{\mathscr{F}_0, \mathscr{E}_0, Q_0\}$$

$$\mathscr{M}_1 = \{\mathscr{F}_1, \mathscr{E}_1, Q_1\}$$

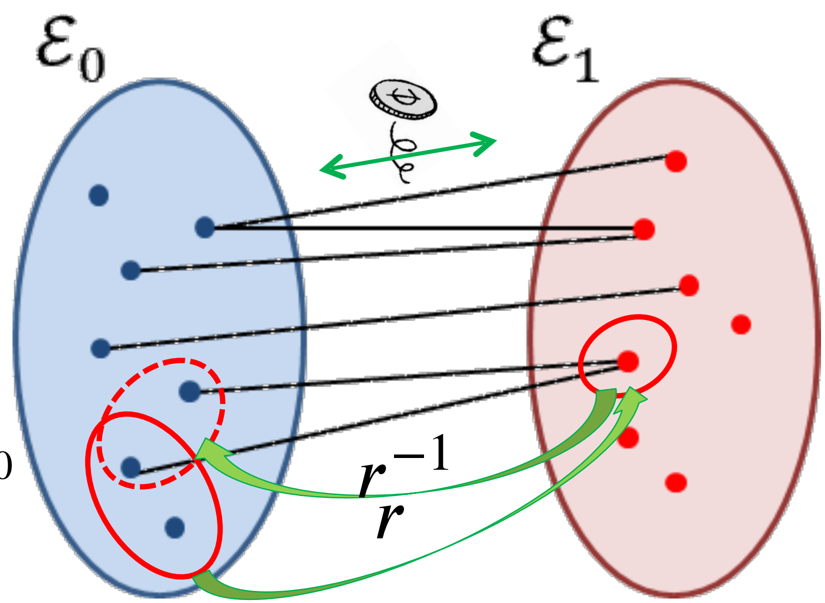


r, a relation between $\mathscr{E}_0$ and $\mathscr{E}_1$:

Induced maps:

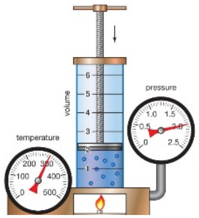
$r \, : \; 2^{\mathscr{E}_0} \to \; 2^{\mathscr{E}_1}$

$r^{-1} \, : \; 2^{\mathscr{E}_1} \to \; 2^{\mathscr{E}_0}$



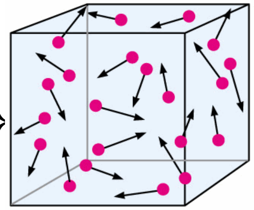
r^{-1}, the inverse relation, between $\mathscr{E}_1$ and $\mathscr{E}_0$

Generalised models



$PV = nRT$

$\mathscr{M}_0 = \{\mathscr{F}_0, \mathscr{E}_0, Q_0\}$ $\mathscr{M}_1 = \{\mathscr{F}_1, \mathscr{E}_1, Q_1\}$



r, a relation between $\mathscr{E}_0$ and $\mathscr{E}_1$

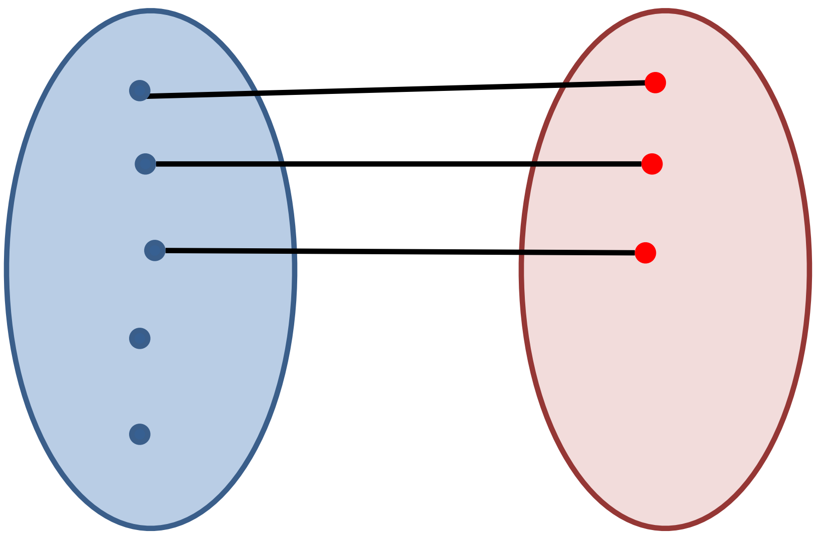
Condition on the Qs:

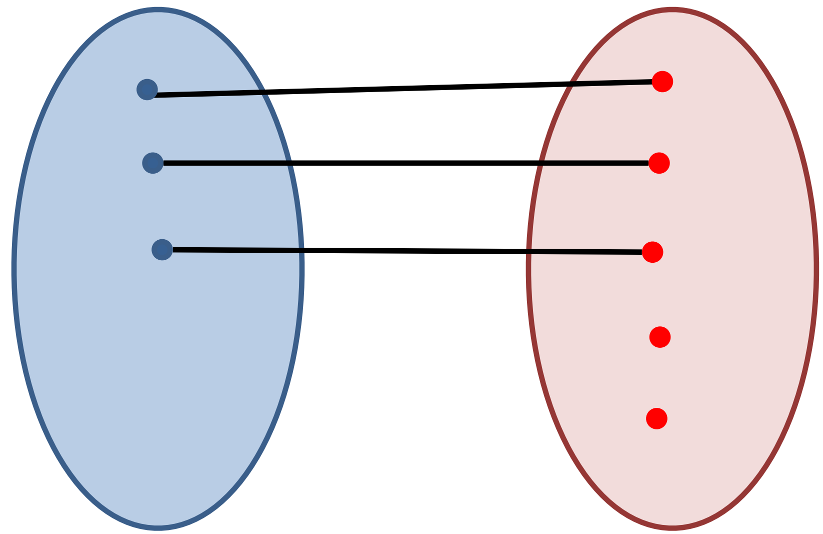
For all $E_0 \subset \mathscr{E}_0$ and all $E_1 \subset \mathscr{E}_1$:

$Q_0(E_0) \leq Q_1\Big(r(E_0)\Big)$ or both probabilities are undefined

$Q_1(E_1) \leq Q_0\big(r^{-1}(E_1)\big)$ or both probabilities are undefined

Simple examples





Restriction/Bayesian update:

r bijective partial function

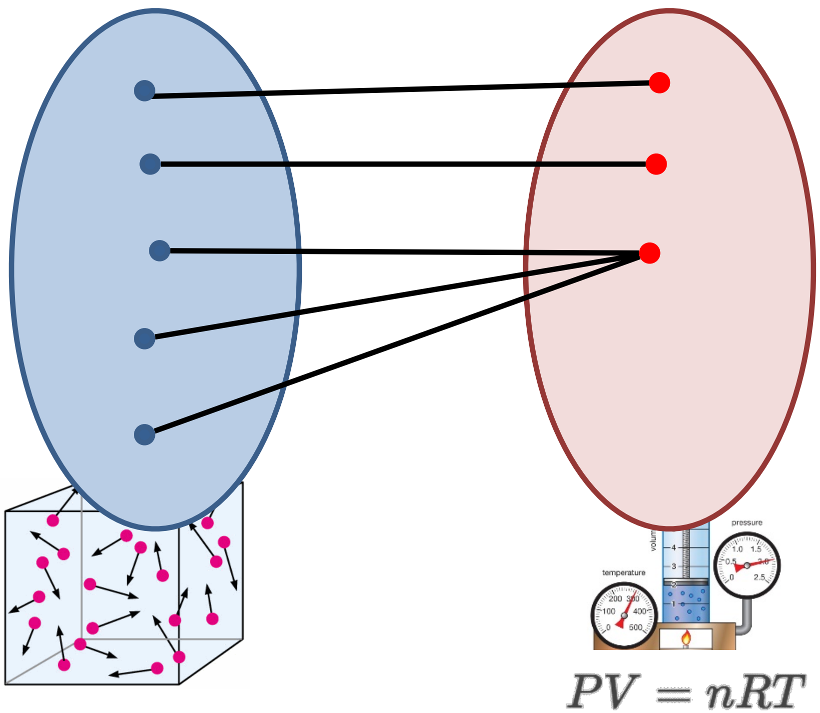
(r^{-1} injective function)

Inclusion:

r injective function

(r^{-1} bijective partial function)

Simple examples

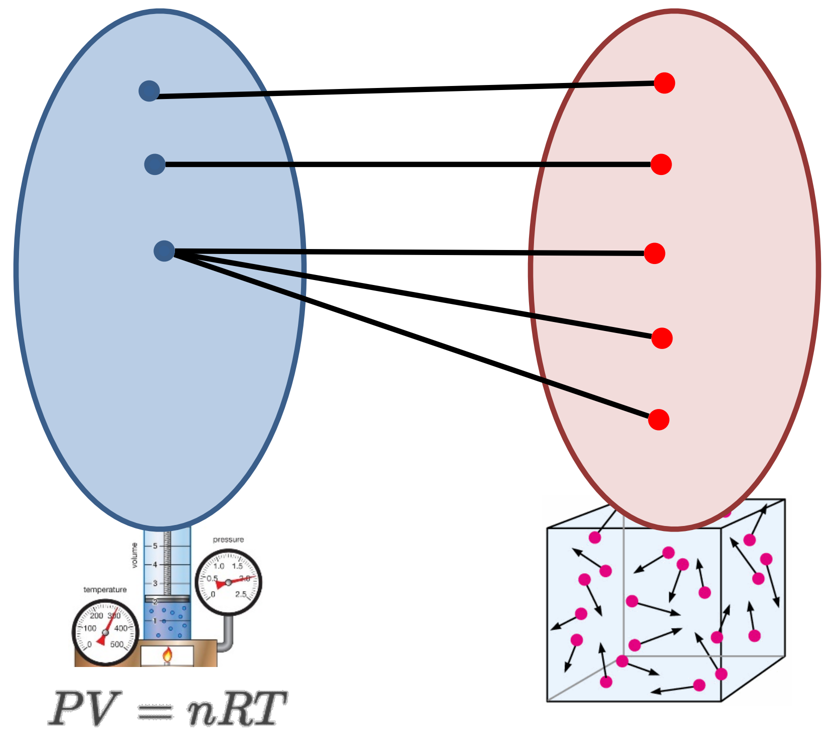


Coarse-graining:

r surjective function

(many-to-one)

(r^{-1} injective, left-total)



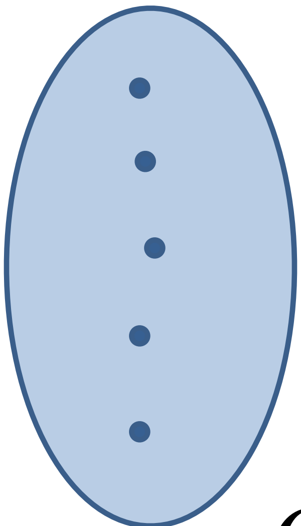
Refinement:

r injective, left-total

(one-to-many)

(r^{-1} surjective function)

Improvement

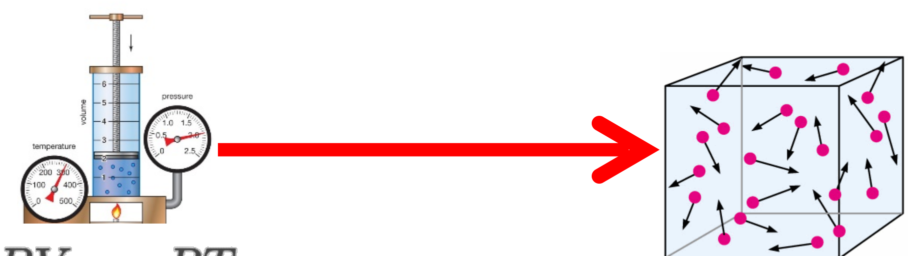


What if the features and environment don't change?

Go from $\mathscr{M}_0 = \{\mathscr{F}_0, \mathscr{E}_0, Q_0\}$

to $\mathscr{M}_1 = \{\mathscr{F}_0, \mathscr{E}_0, Q_1\}$

Q_1 is "better" (more accurate, simpler,...) than Q_0



$PV = nRT$

Most model changes: refinements followed by improvements

Cartesian Frames correspondence

$C = \{A, D, \star\}$ is a Cartesian Frame over W:

$\star$ is a map from $A \times D$ to W

$$a \star d = w$$

A morphism from $C_0 = \{A_0, D_0, \star_0\}$ to

$$C_1 = \{A_1, D_1, \star_1\}$$

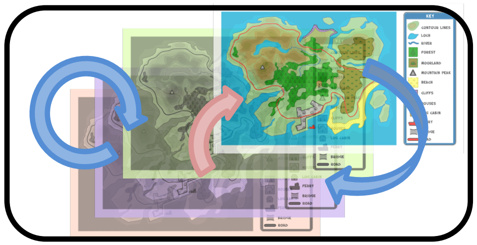
is a pair of functions: $(g_0 : A_0 \to A_1, h_1 : D_1 \to D_0)$,

such that for all a_0, d_1,

$\boxed{\mathrm{Chu}(W)}$

$$g_0(a_0) \star_1 d_1 = a_0 \star_0 h_1(d_1)$$

Cartesian Frames correspondence



Define $GM(W)$ as a subcategory of the generalised models, with:

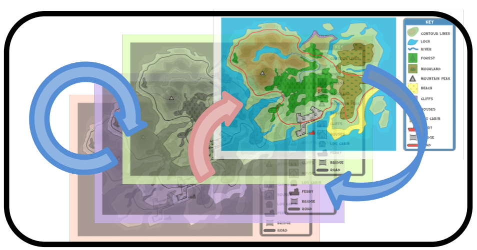
1. Features: $\mathscr{F} = \{A, D, W\}$
2. Environment: $\mathscr{E} = A \times D \times W$
   (using $S \subset 2^S$, $2^{A \sqcup D \sqcup W} = 2^A \times 2^D \times 2^W$)
3. For all a and d, $Q(a, d, w) = 0$, apart from one single w, specific to a and d.
4. Morphisms: r is a relation between $A_0 \times D_0 \times W$ and $A_1 \times D_1 \times W$, derived from the functions/relations $\left(g_0, h_1, Id_W\right)$

Cartesian Frames correspondence

Then define $\Phi : GM(W) \to \mathrm{Chu}(W)$ sending:

1. $(\mathscr{F}, A \times D \times W, Q)$ to $(A, D, \star)$,
   with $a \star d = w$ iff $Q(a, d, w) \neq 0$
2. (g_0, h_1, Id_W) to (g_0, h_1)

Then Φ is a surjective functor of categories.



$\xrightarrow{\Phi}$

How good a meta-model?

1. Features not well-integrated into category-theory formalism. ≈
2. Improvements (to Q) **not** integrated.
3. Change of environment $\mathscr{E}$ well integrated.
4. Universal for some definitions.
5. Easy universality.
6. Model transitions not so easy to understand (see points 1 and 2). ≈

$$P\left(f_1 = x \,\middle|\, f_0 = y\right)$$

Relevant links

- Generalised models as a category:
- https://www.lesswrong.com/posts/nQxqSsHfexivsd6vB/generalised-models-as-a-category
- Cartesian frames as generalised models:
- https://www.lesswrong.com/posts/wiQeYuQPwSypXXFar/cartesian-frames-as-generalised-models
- Model splintering:
- https://www.lesswrong.com/posts/k54rgSg7GcjtXnMHX/model-splintering-moving-from-one-imperfect-model-to-another-1