$$
\begin{array}{cccccccccccc}
w_0 & w_0 & w_0 & w_0 & w_0 & w_0 & w_1 & w_1 & w_1 & w_1 & w_1 & w \\
w_2 & w_2 & w_3 & w_3 & w_3 & w_3 & w_2 & w_2 & w_2 & w_2 & w_3 & w \\
w_5 & w_5 & w_4 & w_4 & w_5 & w_5 & w_4 & w_4 & w_5 & w_5 & w_4 & w \\
w_6 & w_7 & w_6 & w_7 & w_6 & w_7 & w_6 & w_7 & w_6 & w_7 & w_6 & w
\end{array}
$$

Cartesian Frames

Introduction to Cartesian Frames

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the first post in a sequence on **Cartesian frames**, a new way of modeling agency that has recently shaped my thinking a lot.

Traditional models of agency have some problems, like:

- They treat the "agent" and "environment" as primitives with a simple, stable input-output relation. (See "Embedded Agency.")
- They assume a particular way of carving up the world into variables, and don't allow for switching between different carvings or different levels of description.

Cartesian frames are a way to add a first-person perspective (with choices, uncertainty, etc.) on top of a third-person "here is the set of all possible worlds," in such a way that many of these problems either disappear or become easier to address.

The idea of Cartesian frames is that we take as our basic building block a binary function which combines a choice from the agent with a choice from the environment to produce a world history.

We don't think of the agent as having inputs and outputs, and we don't assume that the agent is an object persisting over time. Instead, we only think about a set of possible choices of the agent, a set of possible environments, and a function that encodes what happens when we combine these two.

This basic object is called a Cartesian frame. As with dualistic agents, we are given a way to separate out an “agent” from an “environment." But rather than being a basic feature of the world, this is a “frame” — a particular way of conceptually carving up the world.

We will use the combinatorial properties of a given Cartesian frame to derive versions of inputs, outputs and time. One goal here is that by making these notions derived rather than basic, we can make them more amenable to approximation and thus less dependent on exactly how one draws the Cartesian boundary. Cartesian frames also make it much more natural to think about the world at multiple levels of description, and to model agents as having subagents.

Mathematically, Cartesian frames are exactly Chu spaces. I give them a new name because of my specific interpretation about agency, which also highlights different mathematical questions.

Using Chu spaces, we can express many different relationships between Cartesian frames. For example, given two agents, we could talk about their sum ($\oplus$), which can choose from any of the choices available to either agent, or we could talk about their tensor ($\otimes$), which can accomplish anything that the two agents could accomplish together as a team.

Cartesian frames also have duals ($-^*$) which you can get by swapping the agent with the environment, and $\oplus$ and $\otimes$ have De Morgan duals ($\&$ and $⅋$ respectively), which represent taking a sum or tensor of the environments. The category also has an internal hom, $\multimap$, where $C \multimap D$ can be thought of as "D with a C-shaped hole in it." These operations are very directly analogous to those used in linear logic.

1. Definition

Let W be a set of possible worlds. A Cartesian frame C over W is a triple $C = (A, E, \cdot)$, where A represents a set of possible ways the agent can be, E represents a set of possible ways the environment can be, and $\cdot : A \times E \to W$ is an evaluation function that returns a possible world given an element of A and an element of E.

We will refer to A as the agent, the elements of A as possible agents, E as the environment, the elements of E as possible environments, W as the world, and elements of W as possible worlds.

Definition: A Cartesian frame C over a set W is a triple $(A, E, \cdot)$, where A and E are sets and $\cdot : A \times E \to W$. If $C = (A, E, \cdot)$ is a Cartesian frame over W, we say Agent(C) = A, Env(C) = E, World(C) = W, and Eval(C) = $\cdot$.

A finite Cartesian frame is easily visualized as a matrix, where the rows of the matrix represent possible agents, the columns of the matrix represent possible environments, and the entries of the matrix are possible worlds:

$$\begin{array}{cc} & \begin{array}{ccc} & E & \\ e_1 & e_2 & e_3 \end{array} \\ A \begin{array}{c} a_1 \\ a_2 \\ a_3 \end{array} & \begin{pmatrix} w_1 & w_2 & w_3 \\ w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 \end{pmatrix}. \end{array}$$

E.g., this matrix tells us that if the agent selects a_3 and the environment selects e_1, then we will end up in the possible world w_7.

Because we're discussing an agent that has the freedom to choose between multiple possibilities, the language in the definition above is a bit overloaded. You can think of A as representing the agent before it chooses, while a particular $a \in A$ represents the agent's state after making a choice.

Note that I'm specifically *not* referring to the elements of A as "actions" or "outputs"; rather, the elements of A are possible ways the agent can choose to be.

Since we're interpreting Cartesian frames as first-person perspectives tacked onto sets of possible worlds, we'll also often phrase things in ways that identify a Cartesian frame C with its agent. E.g., we will say "C_0 is a subagent of C_1" as a shorthand for "C_0's agent is a subagent of C_1's agent."

We can think of the environment E as representing the agent's uncertainty about the set of counterfactuals, or about the game that it's playing, or about "what the world is as a function of my behavior."

A Cartesian frame is effectively a way of factoring the space of possible world histories into an agent and an environment. Many different Cartesian frames can be put on the same set of possible worlds, representing different ways of doing this factoring. Sometimes, a Cartesian frame will look like a subagent of another Cartesian frame. Other times, the Cartesian frames may look more like independent agents playing a game with each other, or like agents in more complicated relationships.

2. Normal-Form Games

When viewed as a matrix, a Cartesian frame looks much like the normal form of a game, but with possible worlds rather than pairs of utilities as entries.

In fact, given a Cartesian frame over W, and a function from W to a set V, we can construct a Cartesian frame over V by composing them in the obvious way. Thus, if we had a Cartesian frame $(A, E, \cdot)$ and a pair of utility functions $U_A : W \to R$ and $U_E : W \to R$, we could construct a Cartesian frame over R^2, given by $(A, E, \star)$,

where $a \star e := (U_A(a \cdot e), U_E(a \cdot e))$. This Cartesian frame will look exactly like the normal form of a game.

(Although it is a bit weird to think of the environment set as having a utility function.)

We can use this connection with normal-form games to illustrate three features of the ways in which we will use Cartesian frames.

2.1. Coarse World Models

First, note that we can talk about a Cartesian frame over R^2, even though one would not normally think of R^2 as a set of possible worlds.

In general, we will often want to talk about Cartesian frames over "coarse" models of the world, models that leave out some details. We might have a world model W that fully specifies the universe at the subatomic level, while also wanting to talk about Cartesian frames over a set V of high-level descriptions of the world.

We will construct Cartesian frames over V by composing Cartesian frames over W with the function from W to V that sends more refined, detailed descriptions of the universe to coarser descriptions of the same universe.

In this way, we can think of an element of $(r_1, r_2) \in R^2$ as the coarse, high-level possible world given by

"Those possible worlds for which $U_A = r_1$ and $U_E = r_2$."

Definition: Given a Cartesian frame $C = (A, E, \cdot)$ over W, and a function $f : W \to V$, let $f^\circ(C)$ denote the Cartesian frame over V, $f^\circ(C) = (A, E, \star)$, where $a \star e = f(a \cdot e)$.

2.2. Symmetry

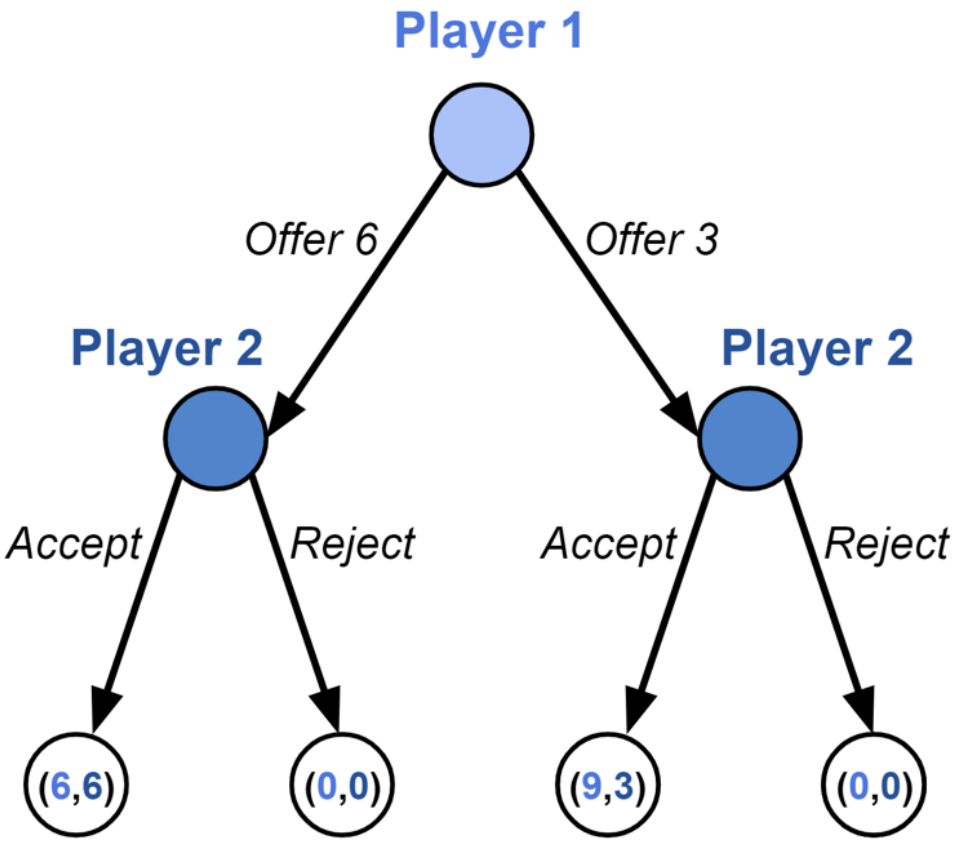
Second, normal-form games highlight the symmetry between the players.

We do not normally think about this symmetry in agent-environment interactions, but this symmetry will be a key aspect of Cartesian frames. Every Cartesian frame $C = (A, E, \cdot)$ has a dual which swaps A and E and transposes the matrix.

2.3. Relation to Extensive-Form Games

Third, much of what we'll be doing with Cartesian frames in this sequence can be summarized as "trying to infer extensive-form games from normal-form games" (ignoring the "games" interpretation and just looking at what this would entail formally).

Consider the ultimatum game. We can represent this game in extensive form:



Given any game in extensive form, we can then convert it to a game in normal form. In this case:

	Offer 6	Offer 3
Accept 6, Accept 3	6, 6	9, 3
Accept 6, Reject 3	6, 6	0, 0
Reject 6, Accept 3	0, 0	9, 3
Reject 6, Reject 3	0, 0	0, 0

The strategies in the normal-form game are the policies in the extensive-form game.

If we then delete the labels, so now we just have a bunch of combinatorial structure about which things send you to the same place, I want to know when we can infer properties of the original extensive-form game, like time and information states.

Although we've used games to note some features of Cartesian frames, we should be clear that Cartesian frames aren't about utilities or game-theoretic rationality. We are not trying to talk about what the agent does, or what the agent should do. In fact, we are effectively taking as our most fundamental building block that an agent can freely choose from a set of available actions.

The theory of Cartesian frames is trying to understand what agents' options are. Utility functions and facts about what the agent actually does can possibly later be placed on top of the Cartesian frame framework, but for now we will be focusing on building up a calculus of what the agent *could* do.

3. Controllables

We would like to use Cartesian frames to reconstruct ideas like "an agent persisting over time," inputs (or "what the agent can learn"), and outputs (or "what the agent can do"), by taking as basic:

1. an agent's ability to "freely choose" between options;
2. a collection of possible ways those options can correspond to world histories; and
3. a notion of when world histories are considered the same in some coarse world model.

In this way, we hope to find new ways of thinking about partial and approximate versions of these concepts.

Instead of thinking of the agent as an object with outputs, I expect a more embedded view to think of all the facts about the world that the agent can force to be true or false.

This includes facts of the form "I output foo," but it also includes facts that are downstream from immediate outputs. Since we're working with "what can I make happen?" rather than "what is my output?", the theory becomes less dependent on precisely answering questions like "Is my output the way I move my mouth, or is it the words that I say?"

We will call the analogue of outputs in Cartesian frames **controllables**. The types of our versions of "outputs" and "inputs" are going to be subsets of W, which we can think of as properties of the world. E.g., S might be the set of worlds in which woolly mammoths exist; we could then think of "controlling S" as "controlling whether or not mammoths exist."

We'll define what an agent can control as follows. First, given a Cartesian frame $C = (A, E, \cdot)$ over W, and a subset S of W, we say that S is *ensurable* in C if there exists an $a \in A$ such that for all $e \in E$, we have $a \cdot e \in S$. Equivalently, we say that S is ensurable in C if at least one of the rows in the matrix only contains elements of S.

Definition: $\text{Ensure}(C) = \{S \subseteq W \mid \exists a \in A,\ \forall e \in E,\ a \cdot e \in S\}$.

If an agent can ensure S, then regardless of what the environment does — and even if the agent doesn't know what the environment does, or its behavior isn't a function of what the environment does — the agent has some strategy which makes sure that the world ends up in S. (In the degenerate case where the agent is empty, the set of ensurables is empty.)

Similarly, we say that S is *preventable* in C if at least one of the rows in the matrix contains *no* elements of S.

Definition: $\text{Prevent}(C) = \{S \subseteq W \mid \exists a \in A,\ \forall e \in E,\ a \cdot e \notin S\}$.

If S is both ensurable and preventable in C, we say that S is controllable in C.

Definition: $\text{Ctrl}(C) = \text{Ensure}(C) \cap \text{Prevent}(C)$.

3.1. Closure Properties

Ensurability and preventability, and therefore also controllability, are closed under adding possible agents to A and removing possible environments from E.

Claim: If $A' \supseteq A$ and $E' \subseteq E$, and if for all $a \in A$ and $e \in E'$ we have $a \star e = a \cdot e$, then $\text{Ctrl}(A', E', \star) \supseteq \text{Ctrl}(A, E, \cdot)$.

Proof: Trivial. □

Ensurables are also trivially closed under supersets. If I can ensure some set of worlds, then I can ensure some larger set of worlds representing a weaker property (like "mammoths exist *or* cave bears exist").

Claim: If $S_1 \subseteq S_2 \subseteq W$, and $S_1 \in \text{Ensure}(C)$, then $S_2 \in \text{Ensure}(C)$.

Proof: Trivial. □

Prevent(C) is similarly closed under subsets. Ctrl(C) need not be closed under subsets or supersets.

Since Ensure(C) and Prevent(C) will often be large, we will sometimes write them using a minimal set of generators.

Definition: Let $\langle S_1, \dots, S_n \rangle_\supseteq$ denote the the closure of $\{S_1, \dots, S_n\}$ under supersets. Let $\langle S_1, \dots, S_n \rangle_\subseteq$ denote the closure of $\{S_1, \dots, S_n\}$ under subsets.

3.2. Examples of Controllables

Let us look at some simple examples. Consider the case where there are two possible environments, r for rain, and s for sun. The agent independently chooses between two options, u for umbrella, and n for no umbrella. $A = \{u, n\}$ and $E = \{r, s\}$. There are four possible worlds, $W = \{ur, us, nr, ns\}$. We interpret ur as the world where the agent has an umbrella and it is raining, and similarly for the other worlds. The Cartesian frame, C_1, looks like this:

$$C_1 = \begin{array}{c} \\ u \\ n \end{array} \begin{array}{c} \begin{array}{cc} r & s \end{array} \\ \begin{pmatrix} ur & us \\ nr & ns \end{pmatrix} \end{array}.$$

$\text{Ensure}(C_1) = \langle \{ur, us\}, \{nr, ns\} \rangle_\supseteq$, or

$$\{ \{ u r, u s \}, \{ n r, n s \}, \{ u r, u s, n r \}, \{ u r, u s, n s \}, \{ n r, n s, u r \}, \{ n r, n s, u s \}, W \},$$

and $\text{Prevent}(C_1) = \langle \{ur, us\}, \{nr, ns\} \rangle_\subseteq$, or

$$\{ \{ u r, u s \}, \{ n r, n s \}, \{ u r \}, \{ u s \}, \{ n r \}, \{ n s \}, \{ \} \}.$$

Therefore $\text{Ctrl}(C_1) = \{\{ur, us\}, \{nr, ns\}\}$.

The elements of $\text{Ctrl}(C_1)$ are not actions, but subsets of W: rather than assuming a distinction between "actions" and other events, we just say that the agent can guarantee that the actual world is drawn from the set of possible worlds in which it has an umbrella ($\{ur, us\}$), and it can guarantee that the actual world is drawn from the set of possible worlds in which it doesn't have an umbrella ($\{nr, ns\}$).

Next, let's modify the example to let the agent see whether or not it is raining before choosing whether or not to carry an umbrella. The Cartesian frame will now look like this:

$$C_2 = \begin{array}{c} \\ u \\ n \\ u \leftrightarrow r \\ u \leftrightarrow s \end{array} \begin{array}{c} \begin{array}{cc} r & s \end{array} \\ \begin{pmatrix} ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{pmatrix} \end{array}.$$

The agent is now larger, as there are two new possibilities: it can carry the umbrella if and only if it rains, or it can carry the umbrella if and only if it is sunny. Ctrl(C_2) will also be larger than Ctrl(C_1).

Ctrl(C_2) = {{ur, us}, {nr, ns}, {ur, ns}, {nr, us}}.

Under one interpretation, the new options u ↔ r and u ↔ s feel different from the old ones u and n. It feels like the agent's basic options are to either carry an umbrella or not, and the new options are just incorporating u and n into more complicated policies.

However, we could instead view the agent's "basic options" as a choice between "I want my umbrella-carrying to match when it rains" and "I want my umbrella-carrying to match when it's sunny." This makes u and n feel like the conditional policies, while u ↔ r and u ↔ s feel like the more basic outputs. Part of the point of the Cartesian frame framework is that we are not privileging either interpretation.

Consider now a third example, where there is a third possible environment, m, for meteor. In this case, a meteor hits the earth before the agent is even born, and there isn't a question about whether the agent has an umbrella. There is a new possible world, which we will also call m, in which the meteor strikes. The Cartesian frame will look like this:

$$C_3 = \begin{array}{c} \\ u \\ n \\ u \leftrightarrow r \\ u \leftrightarrow s \end{array} \begin{array}{c} \begin{array}{ccc} r & s & m \end{array} \\ \begin{pmatrix} ur & us & m \\ nr & ns & m \\ ur & ns & m \\ nr & us & m \end{pmatrix} \end{array}.$$

Ensure(C_3) = ⟨{ur, us, m}, {nr, ns, m}, {ur, ns, m}, {nr, us, m}⟩$_{\supset}$, and

Prevent(C_3) = ⟨{ur, us}, {nr, ns}, {ur, ns}, {nr, us}⟩$_{\subset}$. As a consequence, Ctrl(C_3) = {}.

This example illustrates that nontrivial agents may be unable to control the world's state. Because the agent can't prevent the meteor, the agent in this case has no controllables.

This example also illustrates that agents may be able to ensure or prevent some things, even if there are possible worlds in which the agent was never born. While the agent of C_3 cannot ensure that it exists, the agent can ensure that *if* there is no meteor, then it carries an umbrella ({ur, us, m}).

If we wanted to, we could instead consider the agent's ensurables (or its ensurables and preventables) its "outputs." This lets us avoid the counter-intuitive result that agents have no outputs in worlds where their existence is contingent.

I put the emphasis on controllables because they have other nice features; and as we'll see later, there is an operation called "assume" which we can use to say: "The agent, *under the assumption that there's no meteor*, has controllables."

4. Observables

The analogue of inputs in the Cartesian frame model is **observables**. Observables can be thought of as a closure property on the agent. If an agent is able to observe S, then the agent can take policies that have different effects depending on S.

Formally, let S be a subset of W. We say that the agent of a Cartesian frame $C = (A, E, \cdot)$ is able to observe whether S if for every pair $a_0, a_1 \in A$, there exists a single element $a \in A$ which implements the conditional policy that copies a_0 in possible worlds in S (i.e., for every $e \in E$, if $a \cdot e \in S$, then $a \cdot e = a_0 \cdot e$) and copies a_1 in possible worlds outside of S.

When a implements the conditional policy "if S then do a_0, and if not S then do a_1" in this way, we will say that a is in the set $\text{if}(S, a_0, a_1)$.

Definition: Given $C = (A, E, \cdot)$, a Cartesian frame over W, S a subset of W, and $a_0, a_1 \in A$, let $\text{if}(S, a_0, a_1)$ denote the set of all $a \in A$ such that for all $e \in E$, $(a \cdot e \in S) \to (a \cdot e = a_0 \cdot e)$ and$(a \cdot e \notin S) \to (a \cdot e = a_1 \cdot e)$.

Agents in this setting observe events, which are true or false, not variables in full generality. We will say that C's observables, Obs(C), are the set of all S such that C's agent can observe whether S.

Definition: $\text{Obs}(C) = \{S \subseteq W \mid \forall a_0, a_1 \in A, \ \exists a \in A, \ a \in \text{if}(S, a_0, a_1)\}$.

Another option for talking about what the agent can observe would be to talk about when C's agent can distinguish between two disjoint subsets S and T. Here, we would say that the agent of $C = (A, E, \cdot)$ can distinguish between S and T if for all $a_0, a_1 \in A$, there exists an $a \in A$ such that for all $e \in E$, either $a \cdot e = a_0 \cdot e$ or $a \cdot e = a_1 \cdot e$, and whenever $a \cdot e \in S$, $a \cdot e = a_0 \cdot e$, and whenever $a \cdot e \in T$, $a \cdot e = a_1 \cdot e$. This more general definition would treat our observables as the special case $T = W \setminus S$. Perhaps at some point we will want to use this more general notion, but in this sequence, we will stick with the simpler version.

4.1. Closure Properties

Claim: Observability is closed under Boolean combinations, so if $S, T \in \text{Obs}(C)$ then $W \setminus S$, $S \cup T$, and $S \cap T$ are also in Obs(C).

Proof: Assume $S, T \in \text{Obs}(C)$. We can see easily that $W \setminus S \in \text{Obs}(C)$ by swapping a_0 and a_1. It suffices to show that $S \cup T \in \text{Obs}(C)$, since an intersection can be constructed with complements and union.

Given a_0 and a_1, since $S \in \text{Obs}(C)$, there exists an $a_2 \in A$ such that for all $e \in E$, we have $a_2 \in \text{if}(S, a_0, a_1)$. Then, since $T \in \text{Obs}(C)$, there exists an $a_3 \in A$ such that for all $e \in E$, we have $a_3 \in \text{if}(T, a_0, a_2)$. Unpacking

and combining these, we get for all $e \in E$, $a_3 \in if(S \cup T, a_0, a_1)$. Since we could construct such an a_3 from an arbitrary $a_0, a_1 \in A$, we know that $S \cup T \in Obs(C)$. □

This highlights a key difference between our version of "inputs" and the standard version. Agent models typically draw a strong distinction between the agent's immediate sensory data, and other things the agent might know. Observables, on the other hand, include all of the information that *logically follows* from the agent's observations.

Similarly, agent models typically draw a strong distinction between the agent's immediate motor outputs, and everything else the agent can control. In contrast, if an agent can ensure an event S, it can also ensure everything that logically follows from S.

Since Obs(C) will often be large, we will sometimes write it using a minimal set of generators under union.

Since Obs(C) is closed under Boolean combinations, such a minimal set of generators will be a partition of W (assuming W is finite).

Definition: Let $\langle S_1, \ldots, S_n \rangle_\cup$ denote the the closure of $\{S_1, \ldots, S_n\}$ under union (including $\{\}$, the empty union).

Just like what's controllable, what's observable is closed under removing possible environments.

Claim: If $E' \subseteq E$, and if for all $a \in A$ and $e \in E'$ we have $a \star e = a \cdot e$, then $Obs(A, E', \star) \supseteq Obs(A, E, \cdot)$.

Proof: Trivial. □

It is interesting to note, however, that what's observable is not closed under adding possible agents to A.

4.2. Examples of Observables

Let's look back at our three examples from earlier. The first example, C_1, looked like this:

$$C_1 = \begin{array}{c} \\ u \\ n \end{array} \begin{array}{c} \begin{array}{cc} r & s \end{array} \\ \begin{pmatrix} ur & us \\ nr & ns \end{pmatrix} \end{array}.$$

$Obs(C_1) = \langle W \rangle_\cup = \{\{\}, W\}$. This is the smallest set of observables possible. The agent can act, but it can't change its behavior based on knowledge about the world.

The second example looked like:

$$C_2 = \begin{array}{c} \\ u \\ n \\ u \leftrightarrow r \\ u \leftrightarrow s \end{array} \begin{array}{c} \begin{array}{cc} r & s \end{array} \\ \begin{pmatrix} ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{pmatrix} \end{array}.$$

Here, $Obs(C_2) = \langle \{ur, nr\}, \{us, ns\} \rangle_\cup = \{\{\}, \{ur, nr\}, \{us, ns\}, W\}$. The agent can observe whether or not it's raining. One can verify that for any pair of rows, there is a third row (possibly equal to one or both of the first

two) that equals the first if it is ur or nr, and equals the second otherwise.

The third example looked like:

$$C_3 = \begin{array}{c} \\ u \\ n \\ u \leftrightarrow r \\ u \leftrightarrow s \end{array} \begin{array}{c} \begin{array}{cccc} r & s & & m \end{array} \\ \begin{pmatrix} ur & us & m \\ nr & ns & m \\ ur & ns & m \\ nr & us & m \end{pmatrix} \end{array}.$$

Here, $\mathrm{Obs}(C_3) = \langle \{ur, nr\}, \{us, ns\}, \{m\} \rangle_\cup$, which is

$$\{ \{ \}, \{ u r, n r \}, \{ u s, n s \}, \{ m \}, \{ u r, n r, u s, n s \}, \{ u r, n r, m \}, \{ u s, n s, m \}, W \}.$$

This example has an odd feature: the agent is said to be able to "observe" whether the meteor strikes, even though the agent is never instantiated in worlds in which it strikes. Since the agent has no control when the meteor strikes, the agent can vacuously implement conditional policies.

Let's look at two more examples. First, let's modify C_1 to represent the point of view of a powerless bystander:

$$C_4 = \begin{array}{c} \\ 1 \end{array} \begin{array}{c} \begin{array}{cccc} ur & nr & us & ns \end{array} \\ \begin{pmatrix} ur & nr & us & ns \end{pmatrix} \end{array}.$$

Here, the agent has no decisions, and everything is in the hands of the environment.

Alternatively, we can modify C_1 to represent the point of view of the agent from C_1 and environment from C_1 together. The resulting frame looks like this:

$$C_5 = \begin{array}{c} \\ ur \\ nr \\ us \\ ns \end{array} \begin{array}{c} \begin{array}{c} 1 \end{array} \\ \begin{pmatrix} ur \\ nr \\ us \\ ns \end{pmatrix} \end{array}.$$

$\mathrm{Ensure}(C_4) = \langle W \rangle_\supset$ and $\mathrm{Prevent}(C_4) = \langle \{\} \rangle_\subset$, so $\mathrm{Ctrl}(C_4) = \{\}$. Meanwhile, $\mathrm{Obs}(C_4) = \langle \{ur\}, \{nr\}, \{us\}, \{ns\} \rangle_\cup$.

On the other hand, $\mathrm{Obs}(C_5) = \langle W \rangle_\cup$, $\mathrm{Ensure}(C_5)$ and $\mathrm{Prevent}(C_5)$ are the closure of $\{\{ur\}, \{nr\}, \{us\}, \{ns\}\}$ under supersets and subsets respectively, and $\mathrm{Ctrl}(C_5) = 2^W \setminus \{\{\}, W\}$.

In the first case, the agent's ability to observe the world is maximal and its ability to control the world is minimal; while in the second case, observability is minimal and controllability is maximal. An agent with full control over what happens will not be able to observe anything, while an agent that can observe everything can change nothing.

This is perhaps counter-intuitive. If $S \in \mathrm{Obs}(C)$ meant "I can go look at something to check whether we're in an S world," then one might look at C_5 and say: "This agent is all-powerful. It can do *anything*. Shouldn't we then think of it as all-seeing and all-knowing, rather than saying it 'can't observe anything'?" Similarly, one

might look at C_4 and say: "This agent's choices can't change the world at all. But then it seems bizarre to say that everything is 'observable' to the agent. Shouldn't we rather say that this agent is powerless *and* blind?"

The short answer is that, when working with Cartesian frames, we are in a very "What choices can you make?" paradigm, and in that kind of paradigm, the thing closest to an "input" is "What can I condition my choices on?" (Which is a closure property on the agent, rather than a discrete action like "turning on the Weather Channel.")

In that context, an agent with only one option automatically has maximal "inputs" or "knowledge," since it can vacuously implement every conditional policy. At the same time, an agent with too many options can't have any "inputs," since it could then use its high level of control to diagonalize against the observables it is conditioning on and make them false.

5. Controllables and Observables Are Disjoint

A maximally observable frame has minimal controllables, and vice versa. This turns out to be a special case of our first interesting result about Cartesian frames: an agent can't observe what it controls, and can't control what it observes.

To see this, first consider the following frame:

$$C_6 = \begin{array}{cc} & 1 \\ a_0 & w_0 \\ a_1 & w_1 \end{array}.$$

Here, if $a \in \text{if}(\{w_1\}, a_0, a_1)$, then $a \cdot 1$ would not be able to be either w_0 or w_1. If it were w_1, then it would have to copy a_0, and $a_0 \cdot 1 = w_0$. But if it were w_0, then it would have to copy a_1, and $a_1 \cdot 1 = w_1$. So $\text{if}(S, a_0, a_1)$ is empty in this case.

Notice that in this example, $\text{if}(S, a_0, a_1)$ isn't empty merely because our A lacks the right a to implement the conditional policy. Rather, the conditional policy is impossible to implement even in principle.

Fortunately, before checking whether C's agent can observe S, we can perform a simpler check to rule out these problematic cases. It turns out that if $S \in \text{Obs}(C)$, then every column in C consists either entirely of elements of S or entirely of elements outside of S. (This is a necessary condition for being observable, not a sufficient one.)

Definition: Given a Cartesian frame $C = (A, E, \cdot)$ over W, and a subset S of W, let E_S denote the subset $\{e \in E \mid \forall a \in A, e \cdot a \in S\}$.

Lemma: If $S \in \text{Obs}((A, E, \cdot))$, then for all $e \in E$, it is either the case that $e \in E_S$ or $e \in E_{W \setminus S}$.

Proof: Take $S \in \text{Obs}((A, E, \cdot))$, and assume for contradiction that there exists an $e \in E$ in neither E_S nor $E_{W \setminus S}$. Thus, there exists an $a_0 \in A$ such that $a_0 \cdot e \notin S$ and an $a_1 \in A$ such that $a_1 \cdot e \in S$. Since $S \in \text{Obs}((A, E, \cdot))$, there must exist an $a \in A$ such that $a \in \text{if}(S, a_0, a_1)$. Consider whether or not $a \cdot e \in S$. If $a \cdot e \in S$, then $a \cdot e = a_0 \cdot e \notin S$. However, if $a \cdot e \notin S$, then $a \cdot e = a_1 \cdot e \in S$. Either way, this is a contradiction. □

This lemma immediately gives us the following theorem showing that in nontrivial Cartesian frames, observables and controllables are disjoint.

Theorem: Let C be a Cartesian frame over W, with Env(C) nonempty. Then,Ctrl(C) ∩ Obs(C) = {}.

Proof: Let e ∈ Env(C), and suppose for contradiction that S ∈ Ctrl(C) ∩ Obs(C). Since S ∈ Prevent(C), there exists an $a_0 \in A$ such that $a_0 \cdot e \notin S$. Since S ∈ Ensure(C), there exists an $a_1 \in A$ such that $a_1 \cdot e \in S$. This contradicts our lemma above. □

5.1. Properties That Are Both Observable and Ensurable Are Inevitable

We also have a one-sided result showing that if S is both observable and ensurable in C, then S must be inevitable — i.e., the entire matrix must be contained in S.

We'll first define a Cartesian frame's image, which is the subset of W containing every possible world that is actually hit by the evaluation function — the set of worlds that show up in the matrix.

Definition: $\text{Image}(C) = \{w \in W \mid \exists a \in A,\ \exists e \in E \quad \text{s.t.} \quad a \cdot e = w\}$.

Image(C) ⊆ S can be thought of as a degenerate form of either S ∈ Ensure(C) or S ∈ Obs(C), where in the first case, the agent must make it the case that S, and in the second case the agent can do conditional policies because the $a \cdot e \notin S$ condition is never realized.[1] Conversely, if an agent can both observe and ensure S, then the observability and the ensurability must both be degenerate.

Theorem: S ∈ Ensure(C) ∩ Obs(C) if and only if Image(C) ⊆ S and Agent(C) is nonempty.

Proof: Let $C = (A, E, \cdot)$ be a Cartesian frame over W. First, if Image(C) ⊆ S, then S ∈ Obs(C), since $a_0 \in \text{if}(S, a_0, a_1)$ for all $a_0, a_1 \in A$. If A is also nonempty, then S ∈ Ensure(C), there exists an a ∈ A, and for all e ∈ E, $a \cdot e \in S$.

Conversely, if A is empty, Ensure(C) is empty, so S ∉ Ensure(C) ∩ Obs(C). If Image(C) ⊈ S , then there exist $a_0 \in A$ and e ∈ E such that $a_0 \cdot e \notin S$. Then S ∉ Ensure(C) ∩ Obs(C), since if S ∈ Ensure(C), there exists an a_1 such that in particular $a_1 \cdot e \in S$, so e is in neither E_S nor $E_{W \setminus S}$, which implies S ∉ Obs(C). □

Corollary: If Agent(C) is nonempty, Ensure(C) ∩ Obs(C) = $\langle \text{Image}(C) \rangle_{\supset}$.

Proof: Trivial. □

5.2. Controllables and Observables in Degenerate Frames

All of the results so far have shown that an agent's observables and controllables cannot simultaneously be too large. We also have some results that in some extreme cases, Obs(C) and Ctrl(C) cannot be too small. In particular, if there are few possible agents, observables must be large, and if there are few possible environments, controllables must be large.

Claim: If $|\text{Agent}(C)| \leq 1$, $\text{Obs}(C) = 2^W$.

Proof: If Agent(C) is empty, $S \in \text{Obs}(C)$ for all $S \subseteq W$ vacuously. If Agent(C) = {a} is a singleton, then $S \in \text{Obs}(C)$ for all $S \subseteq W$, because $a \in \text{if}(S, a, a)$. □

Claim: If Agent(C) is nonempty and Env(C) is empty, then $\text{Ctrl}(C) = \text{Ensure}(C) = 2^W$. If Agent(C) is nonempty and Env(C) is a singleton, $\text{Ensure}(C) = \{S \subseteq W \mid S \cap \text{Image}(C) \neq \{\}\}$ and $\text{Ctrl}(C) = \{S \subseteq W \mid S \cap \text{Image}(C) \neq \{\}, W \setminus S \cap \text{Image}(C) \neq \{\}\}$.

Proof: If Agent(C) is nonempty and Env(C) is empty, $S \in \text{Ensure}(C)$ for all $S \subseteq W$ vacuously.

If Agent(C) is nonempty and Env(C) = {e} is a singleton, every $S \subseteq W$ that intersects Image(C) nontrivially is in Ensure(C), since if $w \in S \cap \text{Image}(C)$, there must be some $a \in A$ such that $a \cdot e = w$, this a satisfies $a \cdot e' \in S$ for all $e' \in E$. Conversely, if S and Image(C) are disjoint, no $a \in A$ can satisfy this property. The result for Ctrl then follows trivially from the result for Ensure. □

5.3. A Suggestion of Time

Cartesian frames as we've been discussing them are agnostic about time. Possible agents, environments, and worlds could represent snapshots of a particular moment in time, or they could represent lengthy processes.

The fact that an agent's controllables and observables are disjoint, however, suggests a sort of arrow of time, where facts an agent can observe must be “before” the facts that agent has control over. This hints that we may be able to use Cartesian frames to formally represent temporal relationships.

One reason it would be nice to represent time is that we could model agents that repeatedly learn things, expanding their set of observables. Suppose that in some frame C, Agent(C) includes choices the agent makes over an entire calendar year. Agent(C)'s observables would only include the facts the agent can condition on at the start of the year, when it's first able to act; we haven't defined a way to formally represent the agent learning new facts over the course of the year.

It turns out that this additional temporal structure *can* be elegantly expressed using Cartesian frames. We will return to this topic in the very last post in this sequence. For now, however, we only have this hint that particular Cartesian frames have something like a "before" and "after."

6. Why Cartesian Frames?

The goal of this sequence will be to set up the language for talking about problems using Cartesian frames.

Concretely, I'm writing this sequence because:

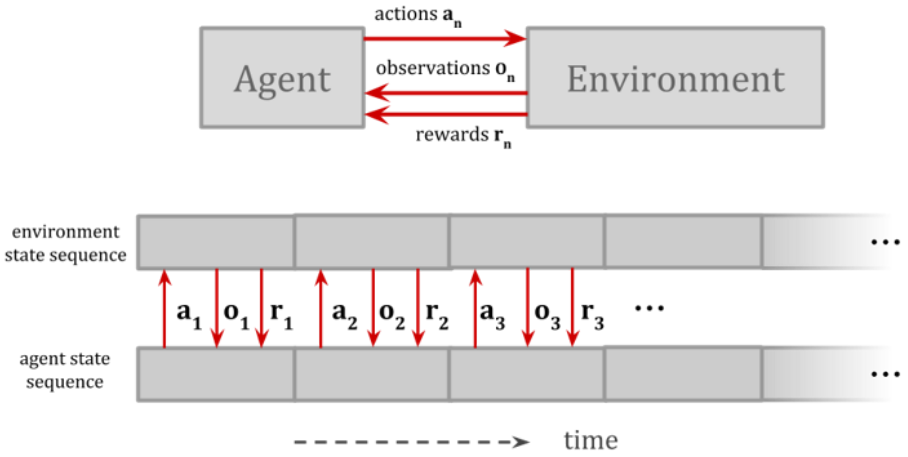
- I've recently found that I have a new perspective to bring to a lot of other MIRI researchers' work. This perspective seems to me to be captured in the mathematical structure of Cartesian frames, but it's the new perspective rather than the mathematical structure per se that seems important to me. I want to try sharing this mathematical object and the accompanying philosophical interpretation, to see if it successfully communicates the perspective.
- I want collaborators to work with on Cartesian frames. If you're a math person who finds the things in this sequence exciting, I'd be interested in talking about it more. You can comment, PM, or email me.

- I want help with paradigm-building, but I also want there to be an ecosystem where people do normal science within this paradigm. I would consider it a good outcome if there existed a decent-sized group of people on the AI Alignment Forum and LessWrong for whom it makes just as much sense to pull out the Cartesian frames paradigm as it makes to pull out the cybernetic agent paradigm.

Below, I will say more about the cybernetic agent model and other ideas that helped motivate Cartesian frames, and I will provide an overview of upcoming posts in the sequence.

6.1. Cybernetic Agent Model

The cybernetic agent model describes an agent and an environment interacting over time:



In "Embedded Agency," Abram Demski and I noted that cybernetic agents like Marcus Hutter's AIXI are dualistic, whereas real-world agents will be embedded in their environment. Like a Cartesian soul, AIXI is crisply separated from its environment.

The dualistic model is often useful, but it's clearly a simplification that works better in some contexts than in others. One thing it would be nice to have is a way to capture the useful things about this simplification, while treating it as a high-level approximation with known limitations — rather than treating it as ground truth.

Cartesian frames carve up the world into a separate "agent" and "environment," and thereby adopt the basic conceit of dualistic Hutter-style agents. However, they treat this as a "frame" imposed on a more embedded, naturalistic world.[2]

Cartesian frames serve the same sort of intellectual function as the cybernetic agent model, and are intended to supersede this model. Our hope is that a less discrete version of ideas like "agent," "action," and "observation" will be better able to tolerate edge cases. E.g., we want to be able to model weirder, loopier versions of “inputs” that operate across multiple levels of description.

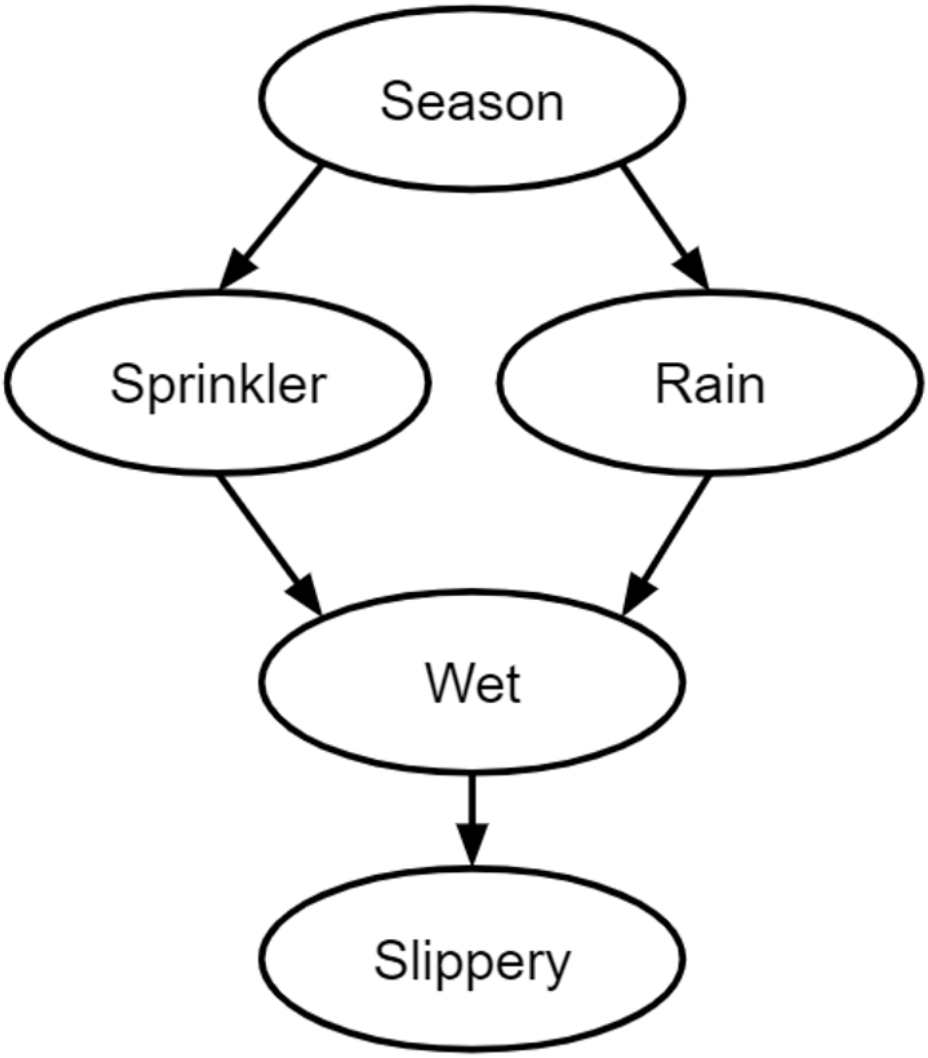
We will also devote special attention in this sequence to subagents, which are very difficult to represent in traditional dualistic models. In game theory, for example, we carve the world into different "agent" and "non-agent" parts, but we can't represent nontrivial agents that intersect other agents. A large part of the theory in this sequence will be giving us a language for talking about subagents.

6.2. Deriving Functional Structure

Another way of summarizing this sequence is that we'll be applying *reasoning* like Pearl's to *objects* like game theory's, with a *motivation* like Hutter's.

In Judea Pearl's causal models, you are given a bunch of variables, and an enormous joint distribution over the variables. The joint distribution is a large object that has a relational structure as opposed to a functional structure.

You then deduce something that looks like time and causality out of the combinatorial properties of the joint distribution. This takes the form of causal diagrams, which give you functions and counterfactuals.



This has some similarities to how we'll be using Cartesian frames, even though the formal objects we'll be working with are very different from Pearl's. We want a model that can replace the cybernetic agent model with something more naturalistic, and our plan for doing this will involve deriving things like time from the combinatorial properties of possible worlds.

We can imagine the real world as an enormous static object, and we can imagine zooming in on different levels of the physical world and sometimes seeing things that look like local functions. ("Ah, no matter what the rest of the world looks like, I can compute the state of Y from the state of X, relative to my uncertainty.") Switching which part of the world we're looking at, or switching which things we're lumping together versus splitting, can then change which things look like functions.

Agency itself, as we normally think about it, is functional in this way: there are multiple "possible" inputs, and whichever "option" we pick yields a deterministic result.

We want an approach to agency that treats this functional behavior less like a unique or fundamental feature of the world, and more like a special case of the world's structure in general — and one that may depend on what we're splitting or lumping together.

"We want to apply Pearl-like methods to Cartesian frames" is also another way of saying "we want to do the formal equivalent of inferring extensive-form games from normal-form games," our summary from before. The analogy is:

	base information	derived information
causality	joint probability distribution	causal diagram
games	normal-form game	extensive-form game
agency	Cartesian frame	control, observation, subagents, time, etc.

The game theory analogy is more relevant formally, while the Pearl analogy better explains why we're interested in this derivation.

Just as notions of time and information state are basic in causal diagrams and extensive-form games, so are they basic in the cybernetic agent model; and we want to make these aspects of the cybernetic agent model derived rather than basic, where it's possible to derive them. We also want to be able to represent things like subagents that are entirely missing from the cybernetic agent model.

Because we aren't treating high-level categories like "action" or "observation" as primitives, we can hope to end up with an agent model that will let us model more edge cases and odd states of the system. A more derived and decomposable notion of time, for example, might let us better handle settings where two agents are both trying to reach a decision based on their model of the other agent's future behavior.

We can also hope to distinguish features of agency that are more description-invariant from features that depend strongly on how we carve up the world.

One philosophical difference between our approach and Pearl's is that we will avoid the assumption that the space of possible worlds factors nicely into variables that are given to the agent. We want to instead just work with a space of possible worlds, and derive the variables for ourselves; or we may want to work in an ontology that lets us reason with multiple incompatible factorizations into variables.[3]

6.3. Contents

The rest of the sequence will cover these topics:

2. Additive Operations on Cartesian Frames - We talk about the category of Chu spaces, and introduce two additive operations one can do on Cartesian frames: sum $\oplus$, and product &. We talk about how to interpret these operations philosophically, in the context of agents making choices to affect the world. We also introduce the small Cartesian frame 0, and its dual $0^* = \top$.

3. Biextensional Equivalence - We define homotopy equivalence $\simeq$ for Cartesian frames, and introduce the small Cartesian frames null, 1_S, and $\bot_S$.

4. Controllables and Observables, Revisited - We use our new language to redefine controllables and observables.

5. Functors and Coarse Worlds - We show how to compare frames over a detailed world model W and frames over a coarse version of that world model V. We demonstrate that observability is a function not only of the observer and the observed, but of the level of description of the world.

6. Subagents of Cartesian Frames - We introduce the notion of a frame C whose agent is the subagent of a frame D, written $C \triangleleft D$. A subagent is an agent playing a game whose stakes are another agent's possible choices. This notion turns out to yield elegant descriptions of a variety of properties of agents.

7. Multiplicative Operations on Cartesian Frames - We introduce three new binary operations on Cartesian frames: tensor $\otimes$, par $\parr$, and lollipop $\multimap$.

8. Sub-Sums and Sub-Tensors - We discuss spurious environments, and introduce variants of sum, ⊞, and tensor, ⊠, that can remove some (but not too many) spurious environments.

9. Additive and Multiplicative Subagents - We discuss the difference between additive subagents, which are like future versions of the agent after making some commitment; and multiplicative subagents, which are like agents acting within a larger agent.

10. Committing, Assuming, Externalizing, and Internalizing - We discuss the additive notion of producing subagents and sub-environments by *committing* or *assuming*, and the multiplicative notion of *externalizing* (moving part of the agent into the environment) and *internalizing* (moving part of the environment into the agent).

11. Eight Definitions of Observability - We use our new tools to provide additional definitions and interpretations of observables. We talk philosophically about the difference between defining what's observable using product and defining what's observable using tensor, which corresponds to the difference between updateful and updateless observations.

12. Time in Cartesian Frames - We show how to formalize temporal relations with Cartesian frames.

I'll be releasing new posts most non-weekend days between now and November 11.

As Ben noted in his announcement post, I'll be giving talks and holding office hours this Sunday at 12-2pm PT and the following three Sundays at 2-4pm PT, to answer questions and discuss Cartesian frames. Everyone is welcome.

The online talks, covering much of the content of this sequence, will take place **this Sunday at 12pm PT** (~~Zoom link~~ added: recording of the talk) and **next Sunday at 2pm PT**.

This sequence is communicating ideas I have been developing slowly over the last year. Thus, I have gotten a lot of help from conversation with many people. Thanks to Alex Appel, Rob Bensinger, Tsvi Benson-Tilsen, Andrew Critch, Abram Demski, Sam Eisenstat, David Girardo, Evan Hubinger, Edward Kmett, Alexander Gietelink Oldenziel, Steve Rayhawk, Nate Soares, and many others.

Footnotes

1. This assumes a non-empty Agent(C). Otherwise, Image(C) could be empty and therefore a subset of S, even though S is not ensurable (because you need an element of Agent(C) in order to ensure anything). ↩

2. This is one reason for the name "Cartesian frames." Another reason for the name is to note the connection to Cartesian products. In linear algebra, a frame of an inner product space is a generalization of a basis of a vector space to sets that may be linearly dependent. With Cartesian frames, then, we have a Cartesian product that projects onto the world, not necessarily injectively. (Cartesian frames aren't actually "frames" in the linear-algebra sense, so this is only an analogy.) ↩

3. This, for example, might let us talk about a high-level description of a computation being "earlier" in some sort of logical time than the exact details of that same computation.

Problems like agent simulates predictor make me think that we shouldn't treat the world as factorizing into a single "true" set of variables at all, though I won't attempt to justify that claim here. ↩

Additive Operations on Cartesian Frames

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

The mathematical object (but not the philosophical interpretation) of a Cartesian Frame is studied under the name "Chu space."

(In category theory, Chu spaces are usually studied in the special case of $W = 2$. To learn more about Chu spaces, see Vaughan Pratt's guide to papers and *n*Lab's page on the Chu construction.)

In this post and the next one, I'll mostly be discussing standard facts about Chu spaces. I'll also discuss how to interpret the standard definitions as statements about agency.

Chu spaces form a category as a special case of the Chu construction. You may notice a strong similarity between operations on Cartesian frames and operations in linear logic, coming from the fact that the Chu construction is also intimately related to linear logic, and is used in the semantics for linear logic.

Linear logic has a large number of operations—additive conjunction (&), multiplicative conjunction (⊗), and so on—and many of those symbols will turn out to have interpretations for Cartesian frames, and they're actually going to be meaningful interpretations in this setting. For that reason, we'll be stealing much of our notation from linear logic, though this sequence won't assume familiarity with linear logic.

Definition: Chu(W) is the category whose objects are Cartesian frames over W, whose morphisms from $C = (A, E, \cdot)$ to $D = (B, F, \star)$ are pairs of functions $(g : A \to B, h : F \to E)$, such that $a \cdot h(f) = g(a) \star f$ for all $a \in A$ and $f \in F$, and whose composition of morphisms is given by $(g_1, h_1) \circ (g_0, h_0) = (g_1 \circ g_0, h_0 \circ h_1)$.

The composition of two morphisms $C_0 \to C_1$ and $C_1 \to C_2$, then, sends the agent of C_0 to C_2 and sends the environment of C_2 to C_0.

Claim: Chu(W) is a category.

Proof: It suffices to show that composition is well-defined and associative and there exist identity morphisms. For identity, (id_A, id_E) is clearly an identity on $C = (A, E, \cdot)$,

where id_X is the identity map from X to itself.

The composition $(g_1, h_1) \circ (g_0, h_0)$ of $(g_0, h_0) : C_0 \to C_1$ with $(g_1, h_1) : C_1 \to C_2$ is $(g_1 \circ g_0, h_0 \circ h_1) : C_0 \to C_2$. To verify that this is a morphism, we just need that $a_0 \cdot_0 h_0(h_1(e_2)) = g_1(g_0(a_0)) \cdot_2 e_2$ for all $a_0 \in \mathrm{Agent}(C_0)$, and $e_2 \in \mathrm{Env}(C_2)$, where $\cdot_i = \mathrm{Eval}(C_i)$. Indeed,

$$\begin{aligned} a_0 \cdot_0 h_0(h_1(e_2)) &= g_0(a_0) \cdot_1 h_1(e_2) \\ &= g_1(g_0(a_0)) \cdot_2 e_2, \end{aligned}$$

since each component is a morphism.

Associativity of the composition follows from the fact that it is just a pair of compositions of functions on sets, and composition is associative for sets. □

1. What Do These Morphisms Represent?

1.1. Morphisms as Interfaces

A Cartesian frame is a first-person perspective. The agent A finds itself in a certain situation or game, where it expects to encounter an environment E. The morphisms in Chu(W) allow the agent of one Cartesian frame to play the game of another Cartesian frame.

We can think of the morphisms from $C = (A, E, \cdot)$ to $D = (B, F, \star)$ as ways of fitting the agent of C into the environment of D. Indeed, for every morphism $(g, h) : C \to D$, one can construct a Cartesian frame $(A, F, \diamond)$, whose agent matches C's agent, and whose environment matches D's environment, with $\diamond$ given by $a \diamond f = a \cdot h(f) = g(a) \star f$. (The morphism from C to D can actually be viewed as the composition of $(\mathrm{id}_A, h) : C \to (A, F, \diamond)$ with $(g, \mathrm{id}_F) : (A, F, \diamond) \to D$.)

Two random large Cartesian frames will typically have no morphisms between them. When there *is* a morphism, the morphism functions as an interface that allows the agent A to interact with some other environment F. However, we aren't just randomly

throwing A and F together. A's interaction with F factors through the function $h : F \to E$, so A can in a sense still be thought of as using an interface where it interacts with E. It just interacts with an $e \in E$ that is of the form h(f) for some $f \in F$. But this is happening simultaneously with the dual view in which F can be thought of as still interacting with B!

Since a Cartesian frame is a first-person perspective, you can imagine A having the internal experience of interacting with E, while F has the "experience" of interacting with B. The morphism's job is to be the translation interface that allows this A and F to interact with each other, while preserving their respective internal experiences in such a way that they feel like they're interacting with E and B respectively. A gets to play B 's game, while still thinking that it is playing its own game.

1.2. Morphisms as Differences in Agents' Strength

We can also interpret the existence of a morphism from $C = (A, E, \cdot)$ to $D = (B, F, \star)$ as saying something like "D's agent is at least as strong as C's agent."

This is easiest to see for a morphism $(g, h) : C \to D$ where g and h are both injective. In this case, it is as though $A \subseteq B$ and $F \subseteq E$, so D's agent has more options to choose between and fewer environments it has to worry about.

Since some of the environments in $E \setminus F$ might have been good for the agent, the agent isn't necessarily strictly better off in D; but in a zero-sum game, the agent will indeed be strictly better off. I think this justifies saying that C's agent is in some sense weaker than D's agent.

If g or h is not injective, we could duplicate elements of B and E to make it injective, so the interpretation "C's agent is no stronger than D's agent" is reasonable in that case as well. In particular, the existence of a morphism from C to D implies that Ensure(C) ⊆ Ensure(D) (and thus Ctrl(C) ⊆ Ctrl(D)).

However, the existence of a morphism is stronger than just saying the set of ensurables is larger. The morphism from C to D can be thought of as telling D's agent how to strategy-steal from C's agent, and thus do anything that C's agent can do.

We now provide a few examples to illustrate morphisms between Cartesian frames. (If you're ready to forge ahead, skip to §2 instead.)

1.3. Simple Examples of Morphisms

Imagine a student who is deciding between staying up late studying for a test (a_s) or ignoring the test (a_i). We will represent the student with a Cartesian frame over letter grades, where $W = \{A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-, F\}$.

If the student doesn't study, her final grade is always a C+, represented by the possible world C+. If she does study, she may oversleep and get a bad grade (represented by the environment selecting e_o and putting her in D-). If she studies and doesn't oversleep, she is uncertain about whether her teacher is typical (e_t, resulting in A-) or unusually demanding (e_d, resulting in B+). We represent this with the frame

$$C_T = \begin{array}{c} \\ a_s \\ a_i \end{array} \begin{array}{c} \begin{array}{ccc} e_t & e_d & e_o \end{array} \\ \begin{pmatrix} A- & B+ & D- \\ C+ & C+ & C+ \end{pmatrix} \end{array}.$$

Let us also suppose that yesterday, the student had the extra option of copying another student's answers on test day to get a sure A+. However, she decided not to cheat. We represent the student's options yesterday, prior to precommitting, with the frame

$$C_Y = \begin{array}{c} \\ b_s \\ b_i \\ b_c \end{array} \begin{array}{c} \begin{array}{ccc} f_t & f_d & f_o \end{array} \\ \begin{pmatrix} A- & B+ & D- \\ C+ & C+ & C+ \\ A+ & A+ & A+ \end{pmatrix} \end{array}.$$

There is a morphism from the student's frame today to her frame yesterday, representing the fact that Agent(C_T) can be plugged into Agent(C_Y)'s game, or that the student was "stronger" yesterday than she is today.

Let us also suppose that the student's teacher *is* in fact demanding. If the student today knew this fact, we would instead represent her perspective with the frame

$$C_{T'} = \begin{array}{c} \\ a_s' \\ a_i' \end{array} \begin{array}{c} \begin{array}{cc} e_d' & e_o' \end{array} \\ \begin{pmatrix} B+ & D- \\ C+ & C+ \end{pmatrix} \end{array}.$$

Here, we have a morphism from the student today (C_T) to her perspective if she had an additional promise from the environment ($C_{T'}$). This represents the fact that $C_{T'}$ can strategy-steal from a version of herself who knows strictly less.

Given two Cartesian frames C_0 and C_1, I am not aware of an efficient universal method for determining whether there exists a morphism from C_0 to C_1. Indeed, I conjecture that this problem might be NP-complete. In the above cases, however, we can see that there exist morphisms from C_T to the other two frames by observing that C_T is effectively C_Y with a row deleted, or $C_{T'}$ with a column added.

While Agent(C_Y) and Agent($C_{T'}$) are both stronger than Agent(C_T), we have no morphisms between C_Y and $C_{T'}$; their options are different enough that we can't compare their strength directly.

1.4. Examples of Morphisms Going Both Ways

Every Cartesian frame has an identity morphism pointing to itself; and as we'll discuss in the next post, whenever two Cartesian frames C and D are equivalent (in a sense to be defined), there will be a morphism going from C to D and another going from D to C. But not all pairs of Cartesian frames with morphisms going both ways are equivalent. Consider, for example,

$$C_1 = \begin{array}{c} \\ a_0 \\ a_1 \end{array} \begin{array}{c} \begin{array}{cc} e_0 & e_1 \end{array} \\ \begin{pmatrix} w_0 & w_0 \\ w_0 & w_1 \end{pmatrix} \end{array} \text{ and } D_1 = \begin{array}{c} \\ b_0 \end{array} \begin{array}{c} \begin{array}{c} f_0 \end{array} \\ \begin{pmatrix} w_0 \end{pmatrix} \end{array}.$$

In $C_1 = (A, E, \cdot)$, the default outcome is w_0, but the agent and environment can handshake to produce w_1. In $D_1 = (B, F, \star)$, there are no choices, and there's only one

possible world, w_0.

It turns out that there is a morphism $(g, h) : C_1 \to D_1$, where g is the constant function b_0 and h is the constant function e_0; and there is a second morphism $(g', h') : D_1 \to C_1$, where g' is the constant function a_0 and h' is the constant function f_0. We can interpret these like so:

- There is a morphism $C_1 \to D_1$ because D_1 is effectively C_1 plus a promise from the environment "I'll choose e_0." The agent in D_1 is "stronger" in the sense that it has fewer possible environments to worry about. There is less the environment can do to interfere with the agent's choices.
- There is a morphism $D_1 \to C_1$ because C_1's agent has strictly more options than D_1's agent: moving from D_1 to C_1 lets you retain the option to produce w_0 if you want, but it also lets you try for w_1.

So we can view the smaller matrix as the larger matrix plus a promise from the environment "I'll choose e_0," or we can view it as the larger matrix plus a commitment from the agent "I'll choose a_0."

This example demonstrates that my intuitive statement "wherever there's a morphism from C to D, D is at least as strong as C" conflates two different notions of "stronger." These notions often go together, but come apart in situations such as the handshake example. Like the hypothetical student in $C_{T'}$, the agent of D_1 is "stronger" in the sense that the environment can't do as much to get in the way. But like the not-yet-precommitted student in C_Y, the agent of C_1 is "stronger" in the sense that it has more options.

2. Self-Duality

A key property of Chu(W) is that it is self-dual.

Definition: Let $-^* : \mathrm{Chu}(W) \to \mathrm{Chu}(W)^{\mathrm{op}}$ be the functor given by $(A, E, \cdot)^* = (E, A, \star)$, where $e \star a = a \cdot e$, and $(g, h)^* = (h, g)$.

The more standard notation for dual in linear logic would be $-^{\perp}$, but this is horrible notation.[1]

Claim: $-^*$ is an isomorphism between Chu(W) and Chu(W)^{op}.

Proof: First, we show $-^*$ is a functor. The objects in Chu(W)^{op} are the same as in Chu(W), the morphisms from D to C in Chu(W)^{op} are the morphisms from C to D in Chu(W), and composition is the same, but with the order reversed. $-^*$ clearly preserves identity morphisms. To show that $-^*$ preserves composition, we have

$$\begin{aligned}(g_0, h_0)^* \circ^{op} (g_1, h_1)^* &= (h_1, g_1) \circ (h_0, g_0) \\ &= (h_1 \circ h_0, g_0 \circ g_1) \\ &= ((g_0, h_0) \circ (g_1, h_1))^*.\end{aligned}$$

To see that it is an isomorphism, we need a left and right inverse. We will abuse notation and also write $-^*$ for the functor from Chu(W)^{op} to Chu(W) given by $(E, A, \star)^* = (A, E, \cdot)$, where $a \cdot e = e \star a$, and $(h, g)^* = (g, h)$. Clearly, we have $-^* : \text{Chu}(W) \to \text{Chu}(W)^{op}$ and $-^* : \text{Chu}(W)^{op} \to \text{Chu}(W)$ composing to the identity in both orders, so $-^*$ is an isomorphism. □

Going back to our visualization of Cartesian frames as matrices, $-^*$ just takes the transpose of the matrix, swapping agent with environment. "Chu(W) is self-dual" is another way of saying that transposing a Cartesian frame always gives you another Cartesian frame.

Philosophically, depending on our interpretation, this may be doing something weird. We talk about possible agents and possible environments, but we may mean something different by "possible" in those two cases.

Since we are imagining events from the point of view of the agents, "possible agents" is referring to all of the ways the agent can choose to be by exercising its "free will." We could think of "possible environments" similarly, or we could think of possible environments as representing the agent's uncertainty.

Under the view where possible environments represent uncertainty, $-^*$ is pointing to an interesting duality that swaps choices with uncertainty, swaps the "could" of "I could do X" with the "could" of "The world could have property Y," and (if we add probability to the mix) swaps mixed strategies with probabilistic uncertainty. "What

will I do?" becomes "What game am I playing?", or "What is the world-as-a-function-of-my-action like?"

I will introduce many operations on Cartesian frames, so it will help to highlight even the basic properties as I go. Here, I'll note:

Claim: For any Cartesian frame C, $(C^*)^* = C$.

Proof: Trivial. □

3. Sums of Cartesian Frames

The first binary operation on Cartesian frames I want to introduce is the sum, $\oplus$.

Definition: For Cartesian frames $C = (A, E, \cdot)$ and $D = (B, F, \star)$ over W, $C \oplus D$ is the Cartesian frame $(A \sqcup B, E \times F, \diamond)$, where $a \diamond (e, f) = a \cdot e$ if $a \in A$, and $a \diamond (e, f) = a \star f$ if $a \in B$.

The sum takes the disjoint union of the agents and the Cartesian product of the environments, and does the obvious thing with the evaluation function. The agent can choose any strategy from A or from B, and the environment has to respond to that strategy. We can interpret this as an agent that can choose between two different first-person perspectives: it can decide to interact with the environment as the agent of C, or as the agent of D.

Maybe "Rebecca the chess player" is considering which chess opening to employ, whereas "Rebecca the food-eater" is considering putting her plate down on the chess board and having lunch instead. "Rebecca the agent that can choose between playing chess and having lunch" is the sum of the other two Rebeccas.

If Rebecca tunnel-visions on the chess game, she may not consider her other options. Likewise if she tunnel-visions on lunch. If she inhabits the perspective of the third Rebecca, she can instead decide between chess moves *and* decide whether she wants to be playing chess at all.

Meanwhile, the environment must use a policy that selects an option from E if the agent chooses from A, and selects an option from F if the agent chooses from B.

In the chess example: The environment must be able to respond to different chess moves, but it must also be able to respond to Rebecca deciding to play a different game.

To give a formal example, let $C_2 = (A, E, \cdot)$ and $D_2 = (B, F, \star)$ be given by the matrices

$$C_2 = \begin{array}{c} \\ a_0 \\ a_1 \end{array} \begin{array}{c} \begin{array}{cc} e_0 & e_1 \end{array} \\ \begin{pmatrix} w_0 & w_1 \\ w_2 & w_3 \end{pmatrix} \end{array} \text{ and } D_2 = \begin{array}{c} \\ b_0 \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{ccc} f_0 & f_1 & f_2 \end{array} \\ \begin{pmatrix} w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 \\ w_{10} & w_{11} & w_{12} \end{pmatrix} \end{array}.$$

Here, $C_2 \oplus D_2$ is given by

$$C_2 \oplus D_2 = \begin{array}{c} \\ a_0 \\ a_1 \\ b_0 \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{cccccc} e_0f_0 & e_0f_1 & e_0f_2 & e_1f_0 & e_1f_1 & e_1f_2 \end{array} \\ \begin{pmatrix} w_0 & w_0 & w_0 & w_1 & w_1 & w_1 \\ w_2 & w_2 & w_2 & w_3 & w_3 & w_3 \\ w_4 & w_5 & w_6 & w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 & w_7 & w_8 & w_9 \\ w_{10} & w_{11} & w_{12} & w_{10} & w_{11} & w_{12} \end{pmatrix} \end{array}.$$

If we wish to interpret $C_2 \oplus D_2$ temporally, we can say: The agent first chooses what game to play. The environment then, as a function of which game was chosen, "chooses" what it does; and the agent simultaneously chooses its own move within the game it picked.

Definition: Let 0 be given by the Cartesian frame $0 = (\{\}, \{e\}, \cdot)$, where Agent(0) is the empty set, $Env(0) = \{e\}$ is any singleton set, and Eval(0) is trivial, since it has empty domain.

Claim: $\oplus$ is commutative and associative, and 0 is the identity of $\oplus$ (up to isomorphism).

Proof: Trivial. □

Returning to our interpretation of morphisms as differences in agents' strength: The agent of $C \oplus D$ can choose between being the agent from C or the agent from D, and so is stronger than either. Indeed, we can think of $C \oplus D$'s agent as the weakest agent

that is stronger than both C's agent and D's agent. Mathematically, this translates to $\oplus$ being the categorical coproduct in Chu(W).

Theorem: $C_0 \oplus C_1$ is the coproduct of C_0 and C_1 in Chu(W), and 0 is initial in Chu(W).

Proof: First, we show that 0 is initial. We want to show that there exists a unique morphism from 0 to a given C. Indeed, a morphism from 0 to $C = (A, E, \cdot)$ is a function from $\{\}$ to A along with a function from E to $\{e\}$, and there is always exactly one such pair of functions, regardless of what A and E are. It is also easy to see that this pair of functions is a morphism, since the condition for morphism is empty, since Agent(0) is empty. Thus 0 is initial.

Let $C_i = (A_i, E_i, \cdot_i)$, and let $C_0 \oplus C_1 = (A_0 \sqcup A_1, E_0 \times E_1, \diamond)$. We want to show that there exist inclusion morphisms $\iota_0 : C_0 \to C_0 \oplus C_1$ and $\iota_1 : C_1 \to C_0 \oplus C_1$ such that for any Cartesian frame $D = (B, F, \star)$, and any pair of morphisms $\phi_0 : C_0 \to D$ and $\phi_1 : C_1 \to D$, we have that there exists a unique morphism $\phi : C_0 \oplus C_1 \to D$ such that $\phi \circ \iota_0 = \phi_0$ and $\phi \circ \iota_1 = \phi_1$.

First, we need to specify $\iota_i : (A_i, E_i, \cdot_i) \to (A_0 \sqcup A_1, E_0 \times E_1, \diamond)$. We let $\iota_i = (j_i, k_i)$, where $j_i : A_i \to A_0 \sqcup A_1$ is just the the obvious inclusion of A_i into $A_0 \sqcup A_1$, and $k_i : E_0 \times E_1 \to E_i$ is just the obvious projection. This is clearly a morphism.

Given $\phi_0 = (g_0, h_0) : C_0 \to D$ and $\phi_1 = (g_1, h_1) : C_1 \to D$, we let $\phi = (g, h)$, where $g : A_0 \sqcup A_1 \to B$ is given by $g(a) = g_i(a)$ where i is such that $a \in A_i$, and $h : F \to E_0 \times E_1$ is given by $h(f) = (h_0(f), h_1(f))$. This is a morphism because for all $a \in A_0 \sqcup A_1$ and $f \in F$, we have

$$\begin{aligned} a \diamond h(f) &= a \cdot_i h_i(f) \\ &= g_i(a) \star f \\ &= g(a) \star f, \end{aligned}$$

where i is such that $a \in A_i$. It is clear from the definitions that $\phi \circ \iota_i = \phi_i$.

Finally, we need to show the uniqueness of this ϕ. Let $\phi' = (g', h') : C_0 \oplus C_1 \to D$ be a morphism such that $\phi' \circ \iota_i = \phi_i$ for both $i = 1, 2$. This means that $g'(a) = g_i(a)$ when $a \in A_i$, so $g'(a) = g(a)$ for all $a \in A_0 \sqcup A_1$. Similarly, $h'(f)$ must project to $h_0(f)$ and $h_1(f)$, so

$$\begin{aligned} h'(f) &= (h_0(f), h_1(f)) \\ &= h(f) \end{aligned}$$

for all $f \in F$. Thus $\phi' = \phi$. □

4. Products of Cartesian Frames

Dual to sum, we have the product operation, &. This operation is a product. It is also in the section on additive operations. There are many counterintuitive things about the notation of Chu spaces and linear logic.

Definition: For Cartesian frames $C = (A, E, \cdot)$ and $D = (B, F, \star)$ over W, C&D is the Cartesian frame $(A \times B, E \sqcup F, \diamond)$, where $(a, b) \diamond e = a \cdot e$ if $e \in E$, and $(a, b) \diamond e = b \star e$ if $e \in F$.

C&D means that the agent might have to be the agent of C, and might have to be the agent of D, but does not get to decide which one. Thus, it will have to choose a pair, (a, b), where a says how to behave in a C situation, and b says how to behave in a D situation. The environment will "choose" to either be C's environment or D's environment. When the agent and environment interact, the agent uses the component of its pair that matches the environment's choice.

Instead of thinking of the agent as choosing a pair, we could again think about the situation temporally. C&D is equivalent to an interaction where the environment first chooses which Cartesian frame, C or D, to play; then the agent observes this choice, and the agent and environment simultaneously behave as though they were in the chosen frame, either C or D.

(In fact, if Image(C) and Image(D) are disjoint, we can see this interpretation in the formalism by noting that Image(C) ∈ Obs(C&D)—that is, the agent can change its behavior on the basis of whether the environment selected from C or from D.)

For example, if we let C_2 and D_2 be as the example in §3,

$$C_2 = \begin{array}{c} \\ a_0 \\ a_1 \end{array} \begin{array}{c} \begin{array}{cc} e_0 & e_1 \end{array} \\ \begin{pmatrix} w_0 & w_1 \\ w_2 & w_3 \end{pmatrix} \end{array} \quad \text{and} \quad D_2 = \begin{array}{c} \\ b_0 \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{ccc} f_0 & f_1 & f_2 \end{array} \\ \begin{pmatrix} w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 \\ w_{10} & w_{11} & w_{12} \end{pmatrix} \end{array},$$

then $C_2\&D_2$ is given by

$$C_2\&D_2 = \begin{array}{c} \\ a_0b_0 \\ a_0b_1 \\ a_0b_2 \\ a_1b_0 \\ a_1b_1 \\ a_1b_2 \end{array} \begin{array}{c} \begin{array}{ccccc} e_0 & e_1 & f_0 & f_1 & f_2 \end{array} \\ \begin{vmatrix} w_0 & w_1 & w_4 & w_5 & w_6 \\ w_0 & w_1 & w_7 & w_8 & w_9 \\ w_0 & w_1 & w_{10} & w_{11} & w_{12} \\ w_2 & w_3 & w_4 & w_5 & w_6 \\ w_2 & w_3 & w_7 & w_8 & w_9 \\ w_2 & w_3 & w_{10} & w_{11} & w_{12} \end{vmatrix} \end{array}.$$

A second example: Suppose that we have two Cartesian frames, C_3 and D_3. C_3 is a frame in which it's raining, and the agent chooses whether to carry an umbrella. D_3 is a frame in which it's sunny, and the agent chooses whether to carry an umbrella.

$$C_3 = \begin{array}{c} \\ u \\ n \end{array} \begin{array}{c} \begin{array}{c} r \end{array} \\ \begin{pmatrix} u\,r \\ n\,r \end{pmatrix} \end{array} \quad \text{and} \quad D_3 = \begin{array}{c} \\ u \\ n \end{array} \begin{array}{c} \begin{array}{c} s \end{array} \\ \begin{pmatrix} u\,s \\ n\,s \end{pmatrix} \end{array}$$

It turns out that the second example we provided in "Introduction to Cartesian Frames" §3.2 (Examples of Controllables) is exactly equal to the product of these two Cartesian frames,

$$\begin{array}{rc} & \begin{array}{cc} r & s \end{array} \\ \begin{array}{r} uu = u \\ nn = n \\ un = u \leftrightarrow r \\ nu = u \leftrightarrow s \end{array} & \begin{pmatrix} ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{pmatrix}. \end{array}$$

The environment is the disjoint union of the rain and sun environments, and the policies of the agent can be viewed as "I get to choose what to do as a function of what game we're playing," where "what game we're playing" is "what the weather is."

Definition: Let $\top$ be given by the Cartesian frame $\top = (\{a\}, \{\}, \cdot)$, where Agent($\top$) is a singleton, Env($\top$) is the empty set, and Eval($\top$) is trivial, since it has empty domain.

Claim: & is commutative and associative, and $\top$ is the identity of & (up to isomorphism).

Proof: Trivial. □

& is essentially just $\oplus$ from the point of view of the environment. Thus, since $-^*$ swaps agent and environment, we can express & using $\oplus$ and $-^*$.

Claim: $C\&D = (C^* \oplus D^*)^*$, $\top = 0^*$, $C \oplus D = (C^*\&D^*)^*$, and $0 = \top^*$.

Proof: Trivial. □

In other words, $\oplus$ and & are De Morgan dual with respect to $-^*$.

In the same way that we interpreted $C \oplus D$ as having the weakest agent that is stronger than the agents of C and D, we can interpret C&D's agent as the *strongest* agent that is *weaker* than the agents of C and D.

Theorem: $C_0\&C_1$ is the product of C and D in Chu(W), and $\top$ is terminal in Chu(W).

Proof: Since $\oplus$ is the coproduct in Chu(W), it is the product in $\text{Chu}(W)^{\text{op}}$. Since $-^*$ is an isomorphism between Chu(W) and $\text{Chu}(W)^{\text{op}}$, we can take a product in Chu(W) of

C_0 and C_1 by sending them to $\mathrm{Chu}(W)^{\mathrm{op}}$ via this isomorphism, taking a product, and sending them back. Thus $(C_0^* \oplus C_1^*)^* = C_0 \& C_1$ is the product in $\mathrm{Chu}(W)$ of C_0 and C_1.

Similarly, since 0 is initial in $\mathrm{Chu}(W)$, it is terminal in $\mathrm{Chu}(W)^{\mathrm{op}}$. Thus, $0^* = \top$ is terminal in $\mathrm{Chu}(W)$. □

Our next post will discuss equivalence relations between Cartesian frames. We will introduce a homotopy equivalence on Cartesian frames, and employ these relations to classify small Cartesian frames up to homotopy.

Footnotes

1. One important reason $-^\perp$ is bad notation for dual is that A^B normally represents $B \to A$, where $\to$ is your category's internal hom functor. For Chu spaces, $\to$ is $\multimap$. Since $\perp$ will be the name for an object in our category, one would reasonably expect $C^\perp$ to represent $\perp \multimap C$, but it doesn't. Worse still, C^* *does* happen to be equivalent to $C \multimap \perp$, and this will be an important fact to understand. To minimize confusion, we instead use the common notation $-^*$ for dual. ↩

Biextensional Equivalence

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the third post in the Cartesian frames sequence. Read the first post here.

This post will introduce the standard equivalence relations we'll be using for Cartesian frames. Our primary interest will be in homotopy equivalence, which will allow us to classify frames according to their agents' and environments' effect on possible worlds.

1. Isomorphism

Before defining homotopy equivalence, I want to define isomorphism between Cartesian frames.

Definition: A morphism $(g, h) : C \to D$ is an *isomorphism* if both g and h are bijective.

If there is an isomorphism between C and D, we say $C \cong D$.

Claim: $\cong$ is an equivalence relation.

Proof: Reflexivity is trivial because the identity is an isomorphism. For symmetry, we have that if (g, h) is an isomorphism from C to D, then (g^{-1}, h^{-1}) is an isomorphism from D to C. Transitivity follows from the fact that bijection is transitive. □

Claim: $C \cong D$ if and only if there is a pair of morphisms $\phi : C \to D$ and $\psi : D \to C$ that compose to the identity morphism in both orders.

Proof: If $C \cong D$, we have $(g, h) : C \to D$ with both g and h bijective, and we can take $\phi = (g, h)$ and $\psi = (g^{-1}, h^{-1})$.

Conversely, if $(g_1, h_1) \circ (g_0, h_0)$ is the identity morphism on $C = (A, E, \cdot)$, then $g_1 \circ g_0$ is the identity on A, so g_0 must be injective. Similarly, $h_0 \circ h_1$ is the identity on E, so h_0 must be surjective. Surjectivity of g_0 and injectivity of h_0 follow similarly from the fact that $(g_0, h_0) \circ (g_1, h_1)$ is the identity on D. □

Isomorphism is pretty intuitive. It is basically saying that it doesn't matter what the possible agents and possible environments are, other than how they interact with the

evaluation function.

We will basically always be working up to at least isomorphism. For example, in the last post ("Additive Operations on Cartesian Frames"), we noted that ⊕ and & are commutative and associative up to isomorphism.

2. Homotopy Equivalence

2.1. Homotopic Morphisms

Our initial definition of homotopy equivalence will be devoid of interpretation, but the meaning will become clear later.

We say that two morphisms from C to D are homotopic if you can take the first function from the first morphism and the second function from the second morphism, and the resulting object is still a morphism.

Definition: Two morphisms $(g_0, h_0), (g_1, h_1) : C \to D$ with the same source and target are called *homotopic* if (g_0, h_1) is also a morphism.

Note that the mere existence of two morphisms from C to D doesn't entail that those morphisms are homotopic. Consider the frame $C_0 = (A, E, \cdot)$ given by

$$C_0 = \begin{array}{c} \\ a_0 \\ a_1 \end{array} \begin{array}{c} \begin{array}{cc} e_0 & e_1 \end{array} \\ \begin{pmatrix} w_1 & w_0 \\ w_0 & w_1 \end{pmatrix} \end{array}.$$

There are two morphisms from C_0 to itself: the identity morphism $(id_A, id_E) : C_0 \to C_0$, and a morphism $(g, h) : C_0 \to C_0$ that flips C_0's rows and columns, sending a_0 to a_1, a_1 to a_0, e_0 to e_1, and e_1 to e_0. These two morphisms are not homotopic, because neither (id_A, h) nor (g, id_E) is a morphism.

Being homotopic is an equivalence relation on morphisms. (As such, in the above example it would have been superfluous to demonstrate that both (id_A, h) and (g, id_E) aren't morphisms.)

Claim: Homotopic is an equivalence relation.

Proof: Let $(g_i, h_i) : (A, E, \cdot) \to (B, F, \star)$.

Reflexivity is trivial. For symmetry, we want to show that if (g_0, h_0), (g_1, h_1), and (g_0, h_1) are all morphisms, then so is (g_1, h_0). Indeed, for all $a \in A$ and $f \in F$,

$$\begin{aligned} g_1(a) \star f &= a \cdot h_1(f) \\ &= g_0(a) \star f \\ &= a \cdot h_0(f). \end{aligned}$$

For transitivity, we want to show that if (g_0, h_0), (g_0, h_1), (g_1, h_1), (g_1, h_2), and (g_2, h_2) are all morphisms, then so is (g_0, h_2). Indeed, for all $a \in A$ and $f \in F$,

$$\begin{aligned} g_0(a) \star f &= a \cdot h_1(f) \\ &= g_1(a) \star f \\ &= a \cdot h_2(f). \end{aligned}$$

□

Being homotopic is also respected by composition.

Claim: If $\phi_0 : C_0 \to C_1$ is homotopic to $\phi_1 : C_0 \to C_1$, and $\psi_0 : C_1 \to C_2$ is homotopic to $\psi_0 : C_1 \to C_2$, then $\psi_0 \circ \phi_0$ is homotopic to $\psi_1 \circ \phi_1$.

Proof: Let $C_i = (A_i, E_i, \cdot_i)$, let $\phi_i = (g_i, h_i)$, and let $\psi_i = (j_i, k_i)$. We want to show that $(j_0 \circ g_0, h_1 \circ k_1)$ is a morphism. Indeed, since (g_0, h_1) and (j_0, k_1) are morphisms,

$$\begin{aligned} j_0(g_0(a)) \cdot_2 f &= g_0(a) \cdot_1 k_1(f) \\ &= a \cdot_0 h_1(k_1(f)). \end{aligned}$$

□

Next, we define when two Cartesian frames are homotopy equivalent in the standard way.

2.2. Homotopy Equivalence

Homotopy equivalence relies on the existence of morphisms between C and D that we can compose in either order and end up with something that is homotopic to the identity morphism.

Definition: C is *homotopy equivalent* to D, written $C \simeq D$, if there exists a pair of morphisms $\phi : C \to D$ and $\psi : D \to C$ such that $\psi \circ \phi$ is homotopic to the identity on C and $\phi \circ \psi$ is homotopic to the identity on D.

Claim: $\simeq$ is an equivalence relation.

Proof: Reflexivity is trivial, by taking $\psi = \phi$ to be the identity. Symmetry is also trivial by swapping ϕ and ψ.

For transitivity, assume that for $i = 0, 1$, we have $\phi_i : C_i \to C_{i+1}$ and $\psi_i : C_{i+1} \to C_i$ such that $\phi_i \circ \psi_i$ and $\psi_i \circ \phi_i$ are homotopic to the identity. It suffices to show that $\phi_0 \circ \phi_1 \circ \psi_1 \circ \psi_0$ and $\psi_1 \circ \psi_0 \circ \phi_0 \circ \phi_1$ are both homotopic to the identity. In both cases, since composition respects what is homotopic, we have that the inner pair of morphisms cancels, and then the outer pair of morphisms cancels. □

Note that homotopy equivalence is weaker than isomorphism.

Claim: If $C \cong D$, then $C \simeq D$.

Proof: Trivial. □

We now have the homotopy equivalence relation on Cartesian frames, but no real philosophical interpretation. To understand the meaning of $\simeq$, we will first need to define biextensional collapse.

3. Biextensional Equivalence

3.1. Biextensionality

Definition: A Cartesian frame $C = (A, E, \cdot)$ is called *biextensional* if whenever $a_0, a_1 \in A$ are such that $a_0 \cdot e = a_1 \cdot e$, for all $e \in E$, we have $a_0 = a_1$, and whenever $e_0, e_1 \in E$ are such that $a \cdot e_0 = a \cdot e_1$, for all $a \in A$, we have $e_0 = e_1$.

This is basically saying that a Cartesian frame is biextensional if all of its possible agents and possible environments are distinct when viewed as functions from environment to world and functions from agent to world respectively. The agent doesn't have two options that invariably produce the same outcomes as each other, nor does the environment.

Viewed as a matrix, C is biextensional if all of its rows and columns are distinct.

We have the following lemma that hints at the relationship between biextensionality and homotopy equivalence.

Lemma: Let C and D be biextensional Cartesian frames. Then, $C \simeq D$ if and only if $C \cong D$.

Proof: The "if" direction is trivial. For the "only if" direction, let $C = (A, E, \cdot)$ and $D = (B, F, \star)$ be two biextensional Cartesian frames, and let $C \simeq D$. Thus, there is a pair of morphisms $(g, h) : C \to D$ and $(j, k) : D \to C$ such that $(j \circ g, h \circ k)$ is homotopic to the identity on C, and $(g \circ j, k \circ h)$ is homotopic to the identity on D. Thus $(j \circ g, id_E)$ and $(g \circ j, id_F)$ are both morphisms, where id_S is the identity on the set S. This means that $j(g(a)) \cdot e = a \cdot e$ for all a and e, which since C is biextensional implies that $j \circ g$ is the identity on A. Similarly, since $(g \circ j, id_F)$ is a morphism, we have that $g \circ j$ is the identity on B. Thus $g : A \to B$ is a bijection.

By the symmetry of homotopic, we also have that $(id_A, k \circ h)$ and $(id_B, h \circ k)$, which similarly gives us that $k \circ h$ is the identity on E and $h \circ k$ is the identity of F, so $h : F \to E$ is a bijection. Thus, $C \cong D$. □

Thus, we now understand how to interpret homotopy equivalence for biextensional Cartesian frames: it is equivalent to isomorphism.

To understand homotopy equivalence in general, we will first show how to collapse any Cartesian frame into a biextensional one.

3.2. Biextensional Collapse

Given a Cartesian frame $C = (A, E, \cdot)$, we can define an equivalence relation on A that says two possible agents are equivalent if they implement the same function from E

to W; and we can similarly say that two elements of E are equivalent if they implement the same function from A to W.

Definition: Given a Cartesian frame $C = (A, E, \cdot)$, for $a_0, a_1 \in A$, we say $a_0 \sim a_1$ if $a_0 \cdot e = a_1 \cdot e$ for all $e \in E$. For $e_0, e_1 \in E$, we say that $e_0 \sim e_1$ if $a \cdot e_0 = a \cdot e_1$ for all $a \in A$.

Claim: $\sim$ is an equivalence relation on A and on E .

Proof: Trivial. □

Definition: Given a Cartesian frame $C = (A, E, \cdot)$, for $a \in A$, let $\hat{a}$ denote the equivalence class of a up to $\sim$. Let $\hat{A}$ denote the set of equivalence classes of $\sim$ in A. Similarly, for $e \in E$, let $\hat{e}$ denote the equivalence class of e up to $\sim$, and let $\hat{E}$ denote the set of equivalence classes of $\sim$ in E.

Definition: Given a Cartesian frame $C = (A, E, \cdot)$, the *biextensional collapse* of C, denoted $\hat{C}$, is the Cartesian frame $(\hat{A}, \hat{E}, \hat{\cdot})$, where $\hat{a} \hat{\cdot} \hat{e} = a \cdot e$.

Claim: $\hat{C}$ is well-defined.

Proof: We need to show that $\hat{\cdot}$ is well defined, meaning we need to show that for all $a_0 \sim a_1$ and $e_0 \sim e_1$, we have that $a_0 \cdot e_0 = a_1 \cdot e_1$. This is immediate from the definition of $\sim$. □

Viewed as a matrix, $\hat{C}$ is basically formed from C by deleting any duplicate rows and any duplicate columns. It doesn't matter whether you delete duplicate rows or duplicate columns first. After doing both, you will end up with a matrix with no duplicates.

Claim: $\hat{C}$ is biextensional for all Cartesian frames C.

Proof: Let $C = (A, E, \cdot)$. We want to show that for all $\hat{a_0} \neq \hat{a_1} \in \hat{A}$, there exists an $\hat{e} \in \hat{E}$ such that $\hat{a_0} \hat{\cdot} \hat{e} \neq \hat{a_1} \hat{\cdot} \hat{e}$. Indeed, since $\hat{a_0} \neq \hat{a_1}$, we have that $a_0 \not\sim a_1$, so there

exists an $e \in E$ such that $a_0 \cdot e \neq a_1 \cdot e$, which gives us that $\hat{a_0} \hat{\cdot} \hat{e} \neq \hat{a_1} \hat{\cdot} \hat{e}$. Similarly, for all $\hat{e_0} \neq \hat{e_1} \in \hat{E}$, there exists an $\hat{e} \in \hat{E}$ such that $\hat{a} \hat{\cdot} \hat{e_0} \neq \hat{a} \hat{\cdot} \hat{e_1}$. □

Claim: C is biextensional if and only if $C \cong \hat{C}$.

Proof: If C is biextensional, then all equivalence classes up to $\sim$ on both A and E are singletons. Thus, the morphism $(g, h) : C \to \hat{C}$ given by $g(a) = \hat{a}$ and $h(\hat{e}) = e$ is well-defined, and both g and h are bijective, so $C \cong \hat{C}$.

Conversely, if $C \cong \hat{C}$, then C is isomorphic to a biextensional Cartesian frame, and since biextensionality is clearly preserved by isomorphism, C is also biextensional. □

3.3. Biextensional Equivalence

We can now (finally) use biextensional collapse to give an intuitive meaning to homotopy equivalence.

Claim: $C \simeq D$ if and only if $\hat{C} \cong \hat{D}$.

Proof: It suffices to show that $C \simeq \hat{C}$ for all Cartesian frames C. Then, we will have that if $C \simeq D$, then $\hat{C} \simeq C \simeq D \simeq \hat{D}$. Since homotopy equivalence is the same as isomorphism on biextensional Cartesian frames, this gives $\hat{C} \cong \hat{D}$. And conversely, if $\hat{C} \cong \hat{D}$ then $C \simeq \hat{C} \cong \hat{D} \simeq D$, so $C \simeq D$.

Let $C = (A, E, \cdot)$. We want to show that $C \simeq \hat{C}$. We do this by constructing a pair of morphisms $(g, h) : C \to \hat{C}$, and $(j, k) : \hat{C} \to C$. We will define $g : A \to \hat{A}$ by $a \mapsto \hat{a}$, and $k : E \to \hat{E}$ by $e \mapsto \hat{e}$. For $h : \hat{E} \to E$, and $j : \hat{A} \to A$, we can send each equivalence class to any one member of that class. The choice does not matter.

Now, we want to show that $(g \circ j, k \circ h)$ is homotopic to the identity on $\hat{C}$, and that $(j \circ g, h \circ k)$ is homotopic to the identity on C. The first case is trivial, since $g \circ j$ and

$k \circ h$ are the identity on $\hat{A}$ and $\hat{E}$ respectively. $j \circ g$ and $h \circ k$ need not be the identity on A and E, but $j(g(a)) \sim a$ and $h(k(e)) \sim e$ for all $a \in A$ and $e \in E$. To show that $(j \circ g, h \circ k)$ is homotopic to the identity on C, we just need to show that $(j \circ g, id_E)$ is a morphism, where id_E is the identity on E. However, this just means that $j(g(a)) \cdot e = a \cdot e$ for all $a \in A$ and $e \in E$, which follows from the fact that $j(g(a)) \sim a$. □

We now have that two Cartesian frames are homotopy equivalent if and only if their biextensional collapses are isomorphic. Thus, when C and D are homotopy equivalent, we will also call them biextensionally equivalent.

Definition: We say C and D are *biextensionally equivalent* if $C \simeq D$.

When working up to biextensional equivalence, we are basically saying that we are ignoring any multiplicity in the space of possible worlds and possible environments.

Claim: Each biextensional equivalence class contains a unique biextensional Cartesian frame.

Proof: Each biextensional equivalence class has at least one element, C, and $\hat{C}$ is in the same equivalence class as C and is biextensional, so there must be at least one biextensional Cartesian frame in the class. If there were two biextensional Cartesian frames, they would have to be isomorphic, because isomorphic is equivalent to biextensional equivalence on biextensional Cartesian frames. □

From my perspective, the value of this equivalence relation is that it lets us be less realist about possible agents and possible environments, and instead just care about differences between possible worlds.

This fits well with our general approach in this sequence. Cartesian frames are particular ways of looking at the world and mentally carving it up into an agent component and an environment component, but we allow many different carvings, and we do not give any one carving privileged status as the "true" carving. Thus, we put less weight on our conception of the agent and environment, and more weight on the worlds themselves.

Giving less realism to possible agents/environments also fits with the fact that "worlds" may include details about the agent and environment, "possible agents" may specify features of the agent beyond its "actions," and so on.

Imagine an agent with two unrelated choices: which color to think about (green G, or red R) and whether to go for a walk or stay home (W or H). This yields the possible

agents $A = \{GH, GW, RH, RW\}$. The environment either is safe or has bears: $E = \{S, B\}$. If we represent this scenario with the Cartesian frame

$$C_0 = \begin{array}{c} \\ GH \\ GW \\ RH \\ RW \end{array} \begin{array}{c} \begin{array}{cc} S & B \end{array} \\ \begin{pmatrix} w_0 & w_1 \\ w_2 & w_3 \\ w_4 & w_5 \\ w_6 & w_7 \end{pmatrix} \end{array},$$

then the possible worlds w_0 and w_4 differ only in which *thought* the agent is thinking; likewise w_1 and w_5, etc.

We could have instead described a frame

$$C_1 = \begin{array}{c} \\ GH \\ GW \\ RH \\ RW \end{array} \begin{array}{c} \begin{array}{cc} S & B \end{array} \\ \begin{pmatrix} w_8 & w_9 \\ w_{10} & w_{11} \\ w_8 & w_9 \\ w_{10} & w_{11} \end{pmatrix} \end{array},$$

in which case we would not be treating the agent's thoughts as a relevant difference between possible worlds.[1] But we have the *option* of fully representing "agent-internal" properties using possible worlds, just the same as "environment-internal" properties. As such, we don't need to separately reify possible agents or possible environments.

3.4. Example

One reason there are two definitions here is because the homotopy definition is easier to work with categorically, while the biextensionality definition is easier to work with directly with matrices.

Let C_0 and C_1 be the Cartesian frames given by:

$$C_0 \cong \begin{pmatrix} w_0 & w_1 & w_1 \\ w_2 & w_3 & w_3 \\ w_0 & w_1 & w_1 \end{pmatrix} \text{ and } C_1 \cong \begin{pmatrix} w_2 & w_3 & w_2 \\ w_0 & w_1 & w_0 \\ w_2 & w_3 & w_2 \end{pmatrix}.$$

Note that when working up to isomorphism, there is no need to label the rows or columns.

We can then see that $C_0 \simeq C_1$ because

$$\widehat{C}_0 \cong \widehat{C}_1 \cong \begin{pmatrix} w_0 & w_1 \\ w_2 & w_3 \end{pmatrix}.$$

To verify the equivalence using the the homotopy definition would be far more tedious.

3.5. Relationship to Additive Operations

Since we will often want to work with Cartesian frames up to biextensional equivalence, it will be helpful to know that all of our additive operations respect biextensional equivalence.

Claim: If $C_0 \simeq C_1$ and $D_0 \simeq D_1$, then $C_0^* \simeq C_1^*$, $C_0 \oplus D_0 \simeq C_1 \oplus D_1$, and $C_0 \& D_0 \simeq C_1 \& D_1$.

Proof: It is clear from the definition of biextensional collapse that $\widehat{-}$ commutes with $-^*$. Thus since $\widehat{C_0} \cong \widehat{C_1}$, we have $\widehat{C_0^*} \cong \widehat{C_1^*}$, so $C_0^* \simeq C_1^*$.

For the rest, it suffices to show that if $C_0 \simeq C_1$, then $C_0 \oplus D \simeq C_1 \oplus D$. Then, since $\oplus$ is symmetric up to isomorphism, we have

$$\begin{aligned} C_0 \oplus D_0 &\simeq C_1 \oplus D_0 \\ &\cong D_0 \oplus C_1 \\ &\simeq D_1 \oplus C_1 \\ &\cong C_1 \oplus D_1, \end{aligned}$$

and using the fact that $\oplus$ and $\&$ are De Morgan dual, we have

$$C_0 \,\&\, D_0 \cong (C_0^* \oplus D_0^*)^* \simeq (C_1^* \oplus D_1^*)^* \cong C_1 \,\&\, D_1 .$$

We will use the homotopy equivalence definition. Let $C_i = (A_i, E_i, \cdot_i)$ and let $D = (B, F, \star)$. Let $(g_0, h_0) : C_0 \to C_1$ and $(g_1, h_1) : C_1 \to C_0$ compose to something homotopic to the identity in both orders. We want to construct a $(g_0', h_0') : C_0 \oplus D \to C_1 \oplus D$ and $(g_1', h_1') : C_1 \oplus D \to C_0 \oplus D$, that similarly compose to something homotopic to the identity in both orders. We will take $g_i' : A_i \sqcup B \to A_{1-i} \sqcup B$ to be given by $g_i'(a) = g_i(a)$ if $a \in A_i$, and $g_i'(a) = a$ if $a \in B$. Similarly, we will take $h_i' : E_{1-i} \times F \to E_i \times F$ to be given by $h_i'(e, f) = (h_i(e), f)$.

Without loss of generality, it suffices to show that (g_0', h_0') is a morphism and that $(g_1', h_1') \circ (g_0', h_0')$ is homotopic to the identity on $C_0 \oplus D$. The fact that (g_1', h_1') is a morphism and $(g_0', h_0') \circ (g_1', h_1')$ is homotopic to the identity will follow symmetrically. Let $\diamond_i = \mathrm{Eval}(C_i \oplus D)$.

To show that (g_0', h_0') is a morphism, observe that for all $a \in A_0$ and $(e, f) \in E_1 \times F$, we have

$$\begin{aligned} g_0'(a) \diamond_1 (e, f) &= g_0(a) \cdot_1 e \\ &= a \cdot_0 h_0(e) \\ &= a \diamond_0 (h_0(e), f) \\ &= a \diamond_0 h_0'(e, f). \end{aligned}$$

Similarly, for all $a \in B$ and $(e, f) \in E_1 \times F$, we have

$$\begin{aligned} g_0'(a) \diamond_1 (e, f) &= a \star f \\ &= a \diamond_0 (h_0(e), f) \\ &= a \diamond_0 h_0'(e, f). \end{aligned}$$

To show that $(g_1', h_1') \circ (g_0', h_0')$ is homotopic to the identity on $C_0 \oplus D$, we just need that for all $a \in A_0 \sqcup B$ and all $(e, f) \in E_0 \times F$, we have $a \diamond_0' (e, f) = g_1'(g_0'(a)) \diamond_0' (e, f)$. Indeed, if $a \in B$, then $a = g_1'(g_0'(a))$, and if $a \in A_0$, then

$$\begin{aligned} a \diamond_0' (e, f) &= a \cdot_0 e \\ &= g_1(g_0(a)) \cdot_0 e \\ &= g_1'(g_0'(a)) \diamond_0' (e, f). \end{aligned}$$

□

Image is also clearly preserved by biextensional equivalence.

Claim: If $C \simeq D$, then Image(C) = Image(D).

Proof: Trivial from the biextensional collapse definition. □

4. Some Small Cartesian Frames

We will now classify all biextensional Cartesian frames (and thus biextensional equivalence classes of Cartesian frames) in which the agent's size is at most one

and/or the environment's size is at most one.

Definition: null is the Cartesian frame $(\{\}, \{\}, \cdot)$ with empty agent, empty environment, and empty evaluation function.

If you have an empty Cartesian frame—one with no image, no elements of W—then it must be biextensionally equivalent to either null, 0, or $\top$.

Claim: If |Agent(C)| = 0 and |Env(C)| ≠ 0, then C ≃ 0. If |Env(C)| = 0 and |Agent(C)| ≠ 0, then C ≃ $\top$. If |Agent(C)| = |Env(C)| = 0, then C ≃ null.

Proof: If |Agent(C)| = 0 and |Env(C)| ≠ 0, then all environments are equivalent up to ∼, so $\hat{C}$ has one possible environment and no possible agents, so $\hat{C} \cong 0$, so C ≃ 0. Similarly, if |Env(C)| = 0 and |Agent(C)| ≠ 0, all agents are equivalent up to ∼, so $\hat{C} \cong \top$ and C ≃ $\top$. If |Agent(C)| = |Env(C)| = 0, then C is already equal to null. □

Claim: The only three biextensional Cartesian frames C with Image(C) = {} are 0, $\top$, and null.

Proof: A Cartesian frame has empty image if and only if it has empty agent or empty environment. All three of 0, $\top$, and null are clearly biextensional, and any other Cartesian frame with empty image is biextensionally equivalent to one of them, and so cannot be biextensional. □

We now understand all biextensional Cartesian frames with empty agent or empty environment. Let's look at the case where either the agent or environment is a singleton.

1_S is the biextensional Cartesian frame you get when the agent has only one option, and the frame's image is some set of possible worlds S. Since Env(1_S) will be in bijective correspondence with S = Image(1_S) and the labels on Env(1_S) don't matter, we will identify Env(1_S) with S.

Definition: Given $S \subseteq W$, 1_S is the Cartesian frame $1_S = (\{a\}, S, \star)$, where $a \star s = s$ for all $s \in S$. 1 is the Cartesian frame 1_W.

We can think of 1_S as the perspective of a bystander who has no control, and is just observing which world the environment brings about.

$\bot_S$ is the transpose of 1_S, where the environment has only one option and the agent's options are S. You can think of $\bot_S$ as a powerful agent facing no obstacles, beyond being constrained to S: it gets to choose exactly what world we're in.

Definition: Given $S \subseteq W$, $\bot_S$ is the Cartesian frame $\bot_S = (S, \{e\}, \star)$, where $s \star e = s$ for all $s \in S$. $\bot$ is the Cartesian frame $\bot_W$.

The names 1 and $\bot$ will make more sense later, when we define multiplicative operations on Cartesian frames.[2]

We can think of 1 as a powerless, all-knowing agent, and 1_S as 1 with a promise from the environment that the world will be in S. Similarly, we can think of $\bot$ as an all-powerful agent, and $\bot_S$ as $\bot$ with a commitment to do S.

The class of frames where the agent has only one option, 1_S, contains 1 at one extreme (where S = W) and $\top$ at the other extreme (where S = {}). Meanwhile, the class of frames where the environment has only one option, $\bot_S$, contains $\bot$ at one extreme (where S = W) and 0 at the other (where S = {}).

Claim: $1^* = \bot$, $\bot^* = 1$, $\overset{*}{1_S} = \bot_S$, $\overset{*}{\bot_S} = 1_S$, $1_{\{\}} = \top$, $\bot_{\{\}} = 0$.

Proof: Trivial. □

Claim: If |Agent(C)| = 1, then $C \simeq 1_S$, where S = Image(C). If |Env(C)| = 1, then $C \simeq \bot_S$, where S = Image(C).

Proof: If Agent(C) = {a}, then equivalence classes of environments are given by where they send a. There will be one such equivalence class for each $s \in$ Image(C), and it will send a to s. Thus $\hat{C} = 1_S$, so $C \simeq 1_S$. The |Env(C)| = 1 case is the same with agent and environment swapped. □

Now that we have built up language for talking about Cartesian frames categorically, we are ready to revisit controllables and observables and interpret them through the lens of category theory. This will be the focus of our next post.

Footnotes

1. Similarly, we could have decided that we don't care about certain things about the environment. For example, if we only care whether there are bears in possible worlds where the agent went for a walk and might therefore encounter them, then we could construct a frame

$$C_2 = \begin{array}{c} \\ GH \\ GW \\ RH \\ RW \end{array} \begin{array}{c} \begin{array}{cc} S & B \end{array} \\ \begin{pmatrix} w_{12} & w_{12} \\ w_{13} & w_{14} \\ w_{15} & w_{15} \\ w_{16} & w_{17} \end{pmatrix} \end{array}.$$

↩

2. Indeed, this section on small Cartesian frames would make more sense as part of our discussion of multiplicative operations on Cartesian frames; our motivation for discussing these objects will be provided there. I'm introducing these objects early because they will be useful in a few contexts before we get to multiplicative operations. ↩

Controllables and Observables, Revisited

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the fourth post in the Cartesian frames sequence. Read the first post here.

Previously, we defined controllables as the sets of possible worlds an agent can both ensure and prevent, and we defined observables as the sets of possible worlds such that the agent can implement all conditional policies.

Now that we have built up more language, we can redefine controllables and observables more categorically.

1. Controllables

1.1. Ensurables and Preventables

The categorical definition of ensurables is very simple.

Definition: Ensure(C) is the set of all $S \subseteq W$ such that there exists a morphism

$\phi : 1_S \to C$.

As an example, let $C_0 = (A, E, \cdot)$. Recall that $1_S = (\{b\}, S, \star)$, where $b \star s = s$ for all

$s \in S$. E.g., if $S_0 = \{w_0, w_3, w_4\}$, then

$$1_{S_0} = \begin{matrix} & w_0 & w_3 & w_4 \\ b & (w_0 & w_3 & w_4) \end{matrix}.$$

If there is a morphism (g, h) from 1_{S_0} to C_0, this means that:

- There is a function g from 1_{S_0}'s agent $\{b\}$ to C_0's agent A, i.e., a function that always outputs a specific $a \in A$.
- There is a function h from C_0's environment E to 1_{S_0}'s environment S_0.
- The specific $a \in A$ picked out by g exactly implements that function $h : E \to S_0$.

That function $h : E \to S_0$ is exactly like the function you get by looking at that row, so a morphism $(g, h) : 1_{S_0} \to C_0$ is like a row in C_0 that is entirely contained in S_0. If there are multiple such rows, then there will be multiple distinct morphisms $1_{S_0} \to C_0$ picking out different $a \in A$.

In "Biextensional Equivalence," we noted that 1_S is like a passive observer who has a promise from the environment that the world will be in S. The existence of a morphism $1_S \to C$ means that there's an interface that allows a powerless bystander who has been promised S to play C's game. Since Agent(1_S) only has one option, this interface must send that one option to some option for C's agent that is compatible with this promise.

Claim: This definition is equivalent to the one in "Introduction to Cartesian Frames": Ensure(C) = $\{S \subseteq W \mid \exists a \in A, \forall e \in E, a \cdot e \in S\}$.

Proof: Let $C = (A, E, \cdot)$ and let $1_S = (\{b\}, S, \star)$, where $b \star s = s$ for all $s \in S$. First, assume there exists a morphism $(g, h) : 1_S \to C$. Here, $g : \{b\} \to A$ and $h : E \to S$. Consider the element $g(b) \in A$. It suffices to show that $g(b) \cdot e \in S$ for all $e \in E$. Indeed, $g(b) \cdot e = b \star h(e) \in S$.

Conversely, assume that there exists an $a \in A$, such that $a \cdot e \in S$ for all $e \in E$. Then, there is a morphism $(g, h) : 1_S \to C$ given by $g(b) = a$, and $h(e) = a \cdot e$. This is a morphism because

$$\begin{aligned} g(b) \cdot e &= a \cdot e \\ &= h(e) \\ &= b \star h(e) \end{aligned}$$

for all $b \in \{b\}$ and $e \in E$. □

Definition: Prevent(C) is the set of all $S \subseteq W$ such that there exists a morphism $\phi_1 : 1_{W \setminus S} \to C$.

Claim: This definition is equivalent to the one in "Introduction to Cartesian Frames": $\text{Prevent}(C) = \{S \subseteq W \mid \exists a \in A, \forall e \in E, a \cdot e \notin S\}$.

Proof: This follows from the proof for Ensure(C), substituting $W \setminus S$ for S. □

Our categorical definition gives us a bunch of facts about how ensurability interacts with various operations on Cartesian frames. First, ensurability is monotonic in the existence of morphisms.

Claim: If there exists a morphism $\phi : C \to D$, then $\text{Ensure}(C) \subseteq \text{Ensure}(D)$.

Proof: If $S \in \text{Ensure}(C)$, there exists a morphism $\psi : 1_S \to C$, so we have $\phi \circ \psi : 1_S \to D$, so $S \in \text{Ensure}(D)$. □

This fact justifies our interpretation of the existence of a morphism from C to D as saying that "D is at least as strong as C."

We also have that ensurables interact very strongly with sums and products. The ensurables of a product are the intersection of the original two agents' ensurables, and the ensurables of a sum are (usually) the union of the original two agents' ensurables.

This makes sense when we think of $C \oplus D$ as "there are two games, and the agent gets to choose which one we play," and C&D as "there are two games, and the environment gets to choose which one we play." The agent of $C \oplus D$ can make sure something happens if either C or D's agent could, whereas the agent of C&D can only ensure things that are ensurable across both games.

Claim: $\text{Ensure}(C\&D) = \text{Ensure}(C) \cap \text{Ensure}(D)$.

Proof: Since & is a categorical product, if there exists a morphism from 1_S to C and a morphism from 1_S to D, there must exist a morphism from 1_S to C&D. Thus $\text{Ensure}(C\&D) \supseteq \text{Ensure}(C) \cap \text{Ensure}(D)$. Conversely, since & is a categorical product, there exist projection morphisms from C&D to C and from C&D to D, so $\text{Ensure}(C\&D) \subseteq \text{Ensure}(C) \cap \text{Ensure}(D)$. □

Claim: If $C \neq \text{null}$ and $D \neq \text{null}$, then $\text{Ensure}(C \oplus D) = \text{Ensure}(C) \cup \text{Ensure}(D)$.

Proof: Let $C = (A, E, \cdot)$, and let $D = (B, F, \star)$. Since C is not null, if E were empty, then A would have to be nonempty, and Ensure(C) would be the full set 2^W. $C \oplus D$ would also have empty environment and nonempty agent, and so Ensure($C \oplus D$) would also be the full set 2^W. Thus, we are done in the case where E is empty. Similarly, we are done in the case where F is empty. Assume E and F are nonempty.

It is clear that Ensure($C \oplus D$) $\supseteq$ Ensure(C) $\cup$ Ensure(D) because $C \oplus D$ is a coproduct, so there are canonical injection morphisms from C to $C \oplus D$ and from D to $C \oplus D$.

Conversely, if $S \in$ Ensure($C \oplus D$), then there is a morphism $(g, h) : 1_S \to C \oplus D$. Since $g : \text{Agent}(1_S) \to A \sqcup B$, and $\text{Agent}(1_S) = \{b'\}$ is a singleton, the image of g must be entirely in A or in B. Without loss of generality, assume it is in A. Then, let f be any element of F. There is a morphism $(g', h') : 1_S \to C$ given by $g'(b') = g(b')$, and $h'(e) = h(e, f)$. This is a morphism because

$$\begin{aligned} b' \diamond h'(e) &= b' \diamond h(e, f) \\ &= g(b') \bullet (e, f) \\ &= g'(b') \cdot e, \end{aligned}$$

where $\diamond = \text{Eval}(1_S)$ and $\bullet = \text{Eval}(C \oplus D)$. Thus, $S \in$ Ensure(C). $\square$

The condition that C and D are not null is annoying. It is also not very informative. It is reasonable to just think of null as not a real Cartesian frame, and not worry about it. To see what concretely goes wrong, let $C = (\{a\}, \{e\}, \cdot)$, where $a \cdot e = w$, and let D = null. Ensure(C) = $\langle \{w\} \rangle_{\supseteq}$ and Ensure(D) = {}. However, $C \oplus D = (\{a\}, \{\}, \star)$, so Ensure($C \oplus D$) = 2^W, since the a ensures everything. The null brought us into the degenerate case where there are no possible environments, but since it had no possible agents, it had no ensurables until it was combined with C.

We also have that ensurables are preserved by biextensional equivalence.

Claim: If $C \simeq D$, Ensure(C) = Ensure(D).

Proof: From the homotopy equivalence definition, we have that if $C \simeq D$, there exist morphisms from C to D and vice versa. □

Finally, we have that there is a tradeoff between a Cartesian frame's ability to ensure things and its dual's ability to prevent things.

Claim: Ensure(C) ∩ Prevent(C^*) = {}.

Proof: Trivial. □

1.2. Controllables

Controllables also have a simple categorical definition.

Definition: Let 2_S denote the Cartesian frame $1_S \oplus 1_{W \setminus S}$.

Again, $C \oplus D$ represents a game where the agent chooses whether to play C or D, and the environment must be able to respond in either case. While 1_S has one possible agent, 2_S has two possible agents, representing the choice between S and $W \setminus S$.

2_S's environments are all possible pairs of exactly one $s \in S$ and exactly one $s \in W \setminus S$. For example, if $S_0 = \{w_0, w_3, w_4\}$ (as in our earlier example of a 1_S) and $W = \{w_0, w_1, w_2, w_3, w_4\}$, then

$$2_{S_0} = \begin{array}{c} \\ b_{S_0} \\ b_{W \setminus S_0} \end{array} \begin{array}{c} \begin{array}{cccccc} w_0w_1 & w_0w_2 & w_3w_1 & w_3w_2 & w_4w_1 & w_4w_2 \end{array} \\ \begin{pmatrix} w_0 & w_0 & w_3 & w_3 & w_4 & w_4 \\ w_1 & w_2 & w_1 & w_2 & w_1 & w_2 \end{pmatrix} \end{array}.$$

So the agent decides whether we're in S, and the environment picks a strategy for S and another for the complement of S.

Definition: Ctrl(C) is the set of all $S \subseteq W$ such that there exists a morphism $\phi : 2_S \to C$.

Claim: This definition is equivalent to the one in "Introduction to Cartesian Frames": Ctrl(C) = Ensure(C) ∩ Prevent(C).

Proof: Since ⊕ is the categorical coproduct, there exists a morphism from $2_S = 1_S \oplus 1_{W \setminus S}$ to C if and only if there exist a pair of morphisms $\phi_0 : 1_S \to C$ and $\phi_1 : 1_{W \setminus S} \to C$, which by the above definitions of ensurables and preventables is true if and only if S is in both Ensure(C) and Prevent(C). □

Since Ctrl(C) is the set of all $S \subseteq W$ with both S and $W \setminus S$ in Ensure(C), we immediately have that the following closure properties on ensurables also apply to controllables.

Claim: If there exists a morphism $\phi : C \to D$, then Ctrl(C) ⊆ Ctrl(D).

Proof: Trivial. □

Claim: Ctrl(C&D) = Ctrl(C) ∩ Ctrl(D).

Proof: Trivial. □

Claim: If C ≃ D, then Ctrl(C) = Ctrl(D).

Proof: Trivial. □

Note that although Ensure(C ⊕ D) = Ensure(C) ∪ Ensure(D) is usually true, there isn't a corresponding result for controllables.

2. Observables

We also have a new definition of observables, but it is not nearly as trivial as the definition of controllables.

Definition: Obs(C) is the set of all $S \subseteq W$ such that there exist C_0 and C_1 with Image(C_0) ⊆ S and Image(C_1) ⊆ $W \setminus S$ such that $C \simeq C_0 \& C_1$.

Claim: This definition is equivalent to the one in "Introduction to Cartesian Frames": $\text{Obs}(C) = \{S \subseteq W \mid \forall a_0, a_1 \in A, \exists a \in A, a \in \text{if}(S, a_0, a_1)\}$.

Proof: Throughout the proof, we will let $\mathrm{Obs}_{\mathrm{old}}(-)$ refer to observables as they were originally defined, and let $\mathrm{Obs}(-)$ refer to observables under our categorical definition.

The proof will be broken into three parts:

- First, we will show that $\mathrm{Obs}_{\mathrm{old}}(-)$ is closed under biextensional equivalence.
- Second, we will show that if $\mathrm{Image}(C_0) \subseteq S$ and $\mathrm{Image}(C_1) \subseteq W \setminus S$, then $S \in \mathrm{Obs}_{\mathrm{old}}(C_0 \& C_1)$. In combination with the first part, this implies that $\mathrm{Obs}(C) \subseteq \mathrm{Obs}_{\mathrm{old}}(C)$.
- Third, we will show that if $S \in \mathrm{Obs}_{\mathrm{old}}(A, E, \cdot)$, then $(A, E, \cdot) \simeq (A, E_S, \cdot) \& (A, E_{W \setminus S}, \cdot)$.

This gives us that $\mathrm{Obs}_{\mathrm{old}}(C) \subseteq \mathrm{Obs}(C)$, by taking $C_0 = (A, E_S, \cdot)$ and $C_1 = (A, E_{W \setminus S}, \cdot)$.

Part 1.

We want to show that $\mathrm{Obs}_{\mathrm{old}}(-)$ is closed under biextensional equivalence.

Let $C = (A, E, \cdot)$, let $D = (B, F, \star)$, and let $C \simeq D$. We will use the homotopy equivalence definition, so let $(g_0, h_0) : C \to D$ and $(g_1, h_1) : D \to C$ be such that $(g_0, h_0) \circ (g_1, h_1)$ and $(g_1, h_1) \circ (g_0, h_0)$ are both homotopic to the identity. Given $S \in \mathrm{Obs}_{\mathrm{old}}(C)$, we want to show that $S \in \mathrm{Obs}_{\mathrm{old}}(D)$.

Given $b_0, b_1 \in B$, we want to show that there exists a $b \in B$ such that for all $f \in F$, if $b \star f \in S$, then $b \star f = b_0 \star f$, and otherwise $b \star f = b_1 \star f$. Letting $a_i = g_1(b_i)$, the fact that $S \in \mathrm{Obs}_{\mathrm{old}}(C)$ gives that there exists an $a \in A$, such that for all $e \in E$, if $a \cdot e \in S$, then $a \cdot e = a_0 \cdot e$, and otherwise, $a \cdot e = a_1 \cdot e$. We will take $b = g_0(a)$.

For all $f \in F$, we have that

$$\begin{aligned} b \star f &= g_0(a) \star f \\ &= a \cdot h_0(f). \end{aligned}$$

Further, since $(g_0, h_0) \circ (g_1, h_1)$ is homotopic to the identity, we also have that for all $f \in F$,

$$\begin{aligned} b_i \star f &= g_0(g_1(b_i)) \star f \\ &= g_1(b_i) \cdot h_0(f) \\ &= a_i \cdot h_0(f). \end{aligned}$$

Together these give that if $b \star f \in S$, then $a \cdot h_0(f) \in S$, so

$$\begin{aligned} b \star f &= a \cdot h_0(f) \\ &= a_0 \cdot h_0(f) \\ &= b_0 \star f, \end{aligned}$$

and if $b \star f \notin S$, then $a \cdot h_0(f) \notin S$, so

$$\begin{aligned} b \star f &= a \cdot h_0(f) \\ &= a_1 \cdot h_0(f) \\ &= b_1 \star f. \end{aligned}$$

Thus $S \in \mathrm{Obs}_{\mathrm{old}}(D)$, so $\mathrm{Obs}_{\mathrm{old}}(-)$ is closed under biextensional equivalence.

Part 2.

We want to show that if $\mathrm{Image}(C_0) \subseteq S$ and $\mathrm{Image}(C_1) \subseteq W \setminus S$, then $S \in \mathrm{Obs}_{\mathrm{old}}(C_0 \& C_1)$.

Let $C_0 = (A, E, \cdot)$, let $C_1 = (B, F, \star)$, and let $C_0 \& C_1 = (A \times B, E \sqcup F, \diamond)$. Given $(a_0, b_0), (a_1, b_1) \in A \times B$, we want to show that there exists an $(a, b) \in A \times B$ such that if $(a, b) \diamond e \in S$, then $(a, b) \diamond e = (a_0, b_0) \diamond e$, and otherwise $(a, b) \diamond e = (a_1, b_1) \diamond e$. We will take $(a, b) = (a_0, b_1)$.

For all $f \in F$, $(a_0, b_1) \diamond f = b_1 \star f \notin S$, since $\mathrm{Image}(C_1) \subseteq W \setminus S$. Thus, if $(a_0, b_1) \diamond e \in S$, then $e \in E$, so

$$(a_0, b_1) \diamond e = a_0 \cdot e$$
$$= (a_0, b_0) \diamond e.$$

Similarly, if $(a_0, b_1) \diamond e \notin S$, then $e \in F$, so

$$(a_0, b_1) \diamond e = b_1 \cdot e$$
$$= (a_1, b_1) \diamond e.$$

Thus, $S \in \mathrm{Obs}_{\mathrm{old}}(C_0 \& C_1)$.

Part 3.

We want to show that if $S \in \mathrm{Obs}_{\mathrm{old}}(A, E, \cdot)$, then $(A, E, \cdot) \simeq (A, E_S, \cdot) \& (A, E_{W \setminus S}, \cdot)$.

Let $C = (A, E, \cdot)$, let $C_0 = (A, E_S, \cdot)$, and let $C_1 = (A, E_{W \setminus S}, \cdot)$. (Here the $\cdot$ in C_0 and C_1 is the restriction of $\cdot$ in C to the respective domain.) Let $\star = \mathrm{Eval}(C_0 \& C_1)$.

First, let's quickly deal with the degenerate case where A is empty. In this case $E_S = E_{W \setminus S} = E$. If E is also empty, then $C \simeq \mathrm{null} \simeq \mathrm{null} \& \mathrm{null} \simeq C_0 \& C_1$. If E is nonempty, then $C \simeq 0 \simeq 0 \& 0 \simeq C_0 \& C_1$. Thus, we can restrict our attention to the case where A is nonempty. Note that in this case, E_S and $E_{W \setminus S}$ are disjoint, and as we saw before they cover E, so $E = E_S \sqcup E_{W \setminus S}$.

We need to construct a $(g_0, h_0) : C \to C_0 \& C_1$ and a $(g_1, h_1) : C_0 \& C_1 \to C$, which compose to something homotopic to the identity in both orders. Since $E = E_S \sqcup E_{W \setminus S}$, we can just take h_0 and h_1 to be the identity on E. We will take $g_0 : A \to A \times A$ to be the diagonal given by $g_0(a) = (a, a)$. Finally, for $g_1 : A \times A \to A$, we will use the fact that $S \in \mathrm{Obs}_{\mathrm{old}}(C)$. We will let $g_1(a_0, a_1)$ be chosen such that $g_1(a_0, a_1) \cdot e = a_0 \cdot e$ if $g_1(a_0, a_1) \cdot e \in S$, and $g_1(a_0, a_1) \cdot e = a_1 \cdot e$ otherwise. We can always choose such a $g_1(a_0, a_1)$, by the definition of $\mathrm{Obs}_{\mathrm{old}}(C)$.

To see that (g_0, h_0) is a morphism, observe that for all $a \in A$ and $e \in E_S \sqcup E_{W \setminus S}$, we have

$$\begin{aligned} g_0(a) \star e &= (a, a) \star e \\ &= a \cdot e \\ &= a \cdot h_0(e), \end{aligned}$$

regardless of which half e is in.

To see that (g_1, h_1) is a morphism, observe that for all $(a_0, a_1) \in A \times A$ and $e \in E$, if $e \in E_S$, then

$$\begin{aligned} g_1(a_0, a_1) \cdot e &= a_0 \cdot e \\ &= (a_0, a_1) \star e \\ &= (a_0, a_1) \star h_1(e), \end{aligned}$$

while if $e \in E_{W \setminus S}$, then

$$\begin{aligned} g_1(a_0, a_1) \cdot e &= a_1 \cdot e \\ &= (a_0, a_1) \star e \\ &= (a_0, a_1) \star h_1(e). \end{aligned}$$

Finally, the fact that (g_0, h_0) and (g_1, h_1) compose to something homotopic to the identity in both orders is trivial, since $h_0 \circ h_1$ and $h_1 \circ h_0$ are both the identity, so trivially $a \cdot e = a \cdot h_0(h_1(e))$, and $(a_0, a_1) \star e = (a_0, a_1) \star h_1(h_0(e))$. (Technically, this is verifying that the identity is homotopic each composition, but since being homotopic is symmetric, this is fine.)

<u>Putting it together.</u>

If $S \in \mathrm{Obs}(C)$, then $C \simeq C_0 \& C_1$, where $\mathrm{Image}(C_0) \subseteq S$ and $\mathrm{Image}(C_1) \subseteq W \setminus S$. By part 2, $S \in \mathrm{Obs}_{\mathrm{old}}(C_0 \& C_1)$, which by part 1, means that $S \in \mathrm{Obs}_{\mathrm{old}}(C)$. Conversely, if $S \in \mathrm{Obs}_{\mathrm{old}}(A, E, \cdot)$, then by part 3, $(A, E, \cdot) \simeq (A, E_S, \cdot) \& (A, E_{W \setminus S}, \cdot)$, which since $\mathrm{Image}(A, E_S, \cdot) \subseteq S$ and $\mathrm{Image}(A, E_{W \setminus S}, \cdot) \subseteq W \setminus S$, implies that $S \in \mathrm{Obs}(A, E, \cdot)$. $\square$

Note that from the above proof, if $S \in \text{Obs}(C)$, we know how to construct the C_0 and C_1 such that $C \simeq C_0 \& C_1$. In particular, every column of C must be entirely contained in S or entirely outside of S, and C_0 just takes the subset of columns in S while C_1 takes the subset of columns outside of S.

One thing to like about this new definition is that it shows that when an agent can observe S, you can actually break it up into two different agents. The first agent chooses how to behave in worlds in S and is promised that the world will in fact be in S, and the second does the same for the worlds not in S. These two agents combine using & to form the original agent.

Observables are much less well-behaved than controllables, so there is much less to say about them at this point. We do have that observability is preserved under biextensional equivalence, which is trivial under the new definition and was proven within the previous proof for the old definition.

Claim: If $C \simeq D$, then $\text{Obs}(C) = \text{Obs}(D)$.

Proof: Trivial. □

3. Controllables and Observables Are Still Disjoint

To become more used to our new definitions, let us reprove the incompatibility theorems from before. First, a lemma.

Lemma: Let $C \simeq C_0 \& C_1$, with $\text{Image}(C_0) \subseteq S$ and $\text{Image}(C_1) \subseteq W \setminus S$. If $S \in \text{Ensure}(C)$, then $C_1 \simeq \top$. If $S \in \text{Prevent}(C)$, then $C_0 \simeq \top$.

Proof: If $S \in \text{Ensure}(C)$, there exists a morphism from 1_S to C, so there exists a morphism from 1_S to $C_0 \& C_1$. Composing this with the canonical projection from $C_0 \& C_1$ to C_1 gives a morphism $(g, h) : 1_S \to C_1$. Let $C_1 = (A, E, \cdot)$, and let $1_S = (\{b\}, S, \star)$. If there were an $e \in E$, then $g(b) \cdot e = b \star h(e)$ would be in both S and $W \setminus S$, a

contradiction. Therefore C_1 has empty environment. Also, since $g(b) \in A$, C_1 has nonempty agent. Therefore $C_1 \simeq \top$.

Symmetrically, if $S \in \text{Prevent}(C)$, then $W \setminus S \in \text{Ensure}(C)$, so $C_0 \simeq \top$. □

Now we can reprove (a slightly stronger version of) our main incompatibility theorem.

Theorem: If $C \not\simeq \top$, then $\text{Ctrl}(C) \cap \text{Obs}(C) = \{\}$.

Proof: We prove the contrapositive. Assume $S \in \text{Ctrl}(C) \cap \text{Obs}(C)$. Let $C \simeq C_0 \& C_1$, with $\text{Image}(C_0) \subseteq S$ and $\text{Image}(C_1) \subseteq W \setminus S$. By the above lemma, both $C_0 \simeq \top$ and $C_1 \simeq \top$. Thus $C \simeq \top \& \top \cong \top$. □

We also reprove (the important direction of) the one-sided result.

Theorem: If $S \in \text{Ensure}(C) \cap \text{Obs}(C)$, then $\text{Image}(C) \subseteq S$.

Proof: If $S \in \text{Ensure}(C) \cap \text{Obs}(C)$, then $C \simeq C_0 \& C_1$, with $\text{Image}(C_0) \subseteq S$ and $\text{Image}(C_1) \subseteq W \setminus S$. By the above lemma, $C_1 \simeq \top$, so $C \simeq C_0 \& \top \cong C_0$, so $\text{Image}(C) \subseteq S$. □

In our next post, we will move to discussing Cartesian frames over different worlds, or different world models. E.g., W might be the set of all possible microphysical states of a room, while V is the smaller set of all possible arrangements of macroscopic objects in the room. We will describe how to translate between frames over W and frames over V.

In the process, we will note some surprising facts about coarser and more refined models of the world, as they relate to observables.

Functors and Coarse Worlds

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the fifth post in the Cartesian frames sequence. Read the first post here.

Up until this point, we have only been working with Cartesian frames over a fixed world W. Now, we are going to start talking about Cartesian frames over different worlds.

1. Functors from Functions Between Worlds

In the Cartesian frames framework, a world is a set of possible worlds w that can all potentially occur in the same frame.

I find it useful to think about "different worlds" W and V in the case where W and V are different world *models* that carve up a situation in two different ways. W might be a refined world model, one that describes a situation in more detail; while V is a coarser model of the same situation that elides some distinctions in W.

Returning to an example from "Biextensional Equivalence,"
$W = \{w_0, w_1, w_2, w_3, w_4, w_5, w_6, w_7\}$ could be a world model that includes details about what the agent is thinking (G for a thought about the color green, R for red), as shown in

$$C_0 = \begin{array}{c} \\ GH \\ GW \\ RH \\ RW \end{array} \begin{array}{c} \begin{array}{cc} S & B \end{array} \\ \begin{pmatrix} w_0 & w_1 \\ w_2 & w_3 \\ w_4 & w_5 \\ w_6 & w_7 \end{pmatrix} \end{array},$$

while $V = \{w_8, w_9, w_{10}, w_{11}\}$ could be a world model that leaves out this information, representing the same real-world situation with the frame

$$C_1 = \begin{array}{c} \\ GH \\ GW \\ RH \\ RW \end{array} \begin{array}{c} \begin{array}{cc} S & B \end{array} \\ \begin{pmatrix} w_8 & w_9 \\ w_{10} & w_{11} \\ w_8 & w_9 \\ w_{10} & w_{11} \end{pmatrix} \end{array}.$$

To move between frames like C_0 and C_1 and compare their properties, we will need a way to send agents and environments of frames defined over one world, to agents and environments of frames over an entirely different world. Functors will allow us to do this.

Definition: Given two sets W and and V, and a function $p : W \to V$, let $p^\circ : \text{Chu}(W) \to \text{Chu}(V)$ denote the functor that sends the object $(A, E, \cdot) \in \text{Chu}(W)$ to the object $(A, E, \star) \in \text{Chu}(V)$, where $a \star e = p(a \cdot e)$, and sends the morphism (g, h) to the morphism with the same underlying functions, (g, h).

To visualize this functor, you can imagine Chu(W) as a graph, with matrices as nodes (in the finite case) and arrows representing morphisms. Chu(V) is another graph made of matrices and arrows. To move each frame C from Chu(W) to Chu(V), we use p to entrywise replace the possible worlds in C's matrix with elements of V, without changing the functional properties of the rows and columns; and then we move all the arrows from Chu(W) to Chu(V), which is possible because no functional properties of the original matrices were lost. (Frames and morphisms may or may not be *added* when we move to Chu(V).)

In the cases where we say "W is a refined version of V" or "V is a coarse version of W," all we mean is that the function $p : W \to V$ is surjective.

Claim: p° is well-defined.

Proof: We need to show that p° actually sends objects and morphisms of Chu(W) to objects and morphisms of Chu(V), and that it preserves identity morphisms and composition. p° clearly sends objects to objects. To see that p° sends morphisms to

morphisms, observe that if $(g, h) : (A_0, E_0, \cdot_0) \to (A_1, E_1, \cdot_1)$, and $p^\circ(A_i, E_i, \cdot_i) = (A_i, E_i, \star_i)$, then for all $a \in A_0$ and $e \in E_1$,

$$\begin{aligned} g(a) \star_1 e &= p(g(a) \cdot_1 e) \\ &= p(a \cdot_0 h(e)) \\ &= a \star_0 h(e), \end{aligned}$$

so $p^\circ(g, h) = (g, h)$ is a morphism. It is clear that p° preserves identity and composition, since it has no effect on morphisms. □

We also have that p° preserves all of our additive operations.

Claim: $p^\circ(C \oplus D) = p^\circ(C) \oplus p^\circ(D)$, $p^\circ(C \& D) = p^\circ(C) \& p^\circ(D)$, $p^\circ(C^*) = p^\circ(C)^*$, $p^\circ(0) = 0$, $p^\circ(\top) = \top$, and $p^\circ(\text{null}) = \text{null}$.

Proof: Trivial. □

Our new functor's relationship with 1 and $\bot$ is more interesting. In particular, we can define 1_S and $\bot_S$ from 1 and $\bot$ using functors.

Claim: Let $S \subseteq W$ and let $\iota : S \to W$ be the inclusion of S in W. Then $1_S = \iota^\circ(1)$ and $\bot_S = \iota^\circ(\bot)$. (Here, the 1 and $\bot$ are from Chu(S), not Chu(W).)

Proof: Trivial. □

This gives us a more categorical definition of 1_S and $\bot_S$ from 1 and $\bot$. We will give a more categorical definition of 1 and $\bot$ later, when we talk about multiplicative operations.

p° also preserves biextensional equivalence in one direction. (Two equivalent frames in W will always be equivalent in V, but two inequivalent frames in W won't necessarily be inequivalent in V.)

Claim: If $C \simeq D$, then $p^\circ(C) \simeq p^\circ(D)$.

Proof: Let $C = (A, E, \cdot)$ and let $D = (B, F, \star)$. Let $(g_0, h_0) : C \to D$ and $(g_1, h_1) : D \to C$ compose to something homotopic to the identity in both orders. We want to show that $(g_0, h_0) : p^\circ(C) \to p^\circ(D)$ and $(g_1, h_1) : p^\circ(D) \to p^\circ(C)$ compose to something homotopic to the identity in both orders. Indeed $p(g_1(g_0(a)) \cdot e) = p(a \cdot e)$ for all $a \in A$ and $e \in E$, and $p(g_0(g_1(b)) \star f) = p(b \star f)$ for all $b \in B$ and $f \in F$. □

We also have that p° preserves what's ensurable, where we transition from subsets of W to subsets of V in the obvious way.

Claim: Let $p : W \to V$, and let $p(S) = \{v \in V \mid \exists w \in S, p(w) = v\}$. If $S \in \text{Ensure}(C)$, then $p(S) \in \text{Ensure}(p^\circ(C))$.

Proof: Trivial from the original definition of ensurables. □

We also get a stronger result when dealing with subsets of W and V that correspond exactly.

Claim: Let $p : W \to V$, and let $S \subseteq W$ and $T \subseteq V$ be such that for all $w \in W$, we have $p(w) \in T$ if and only if $w \in S$. Then $S \in \text{Ensure}(C)$ if and only if $T \in \text{Ensure}(p^\circ(C))$, and $S \in \text{Ctrl}(C)$ if and only if $T \in \text{Ctrl}(p^\circ(C))$.

Proof: Trivial from the original definitions of ensurables and controllables. □

The relationship between observability and functors is quite interesting. We will devote the next section to discussing this relationship and its philosophical consequences.

2. What's Observable is Relative to a Coarse World Model

Since observability is not closed under supersets, we can only really hope to get a result for observables in the stronger case where $S \subseteq W$ and $T \subseteq V$ correspond exactly; but interestingly, even then, the preservation result for observables is only one-directional.

Claim: Let $p : W \to V$ and let $S \subseteq W$ and $T \subseteq V$ be such that for all $w \in W$, we have $p(w) \in T$ if and only if $w \in S$. Then if $S \in \text{Obs}(C)$, then $T \in \text{Obs}(p^\circ(C))$.

Proof: If $C \simeq C_0 \& C_1$, with $\text{Image}(C_0) \subseteq S$ and $\text{Image}(C_1) \subseteq W \setminus S$, then

$p^\circ(C) \simeq p^\circ(C_0) \& p^\circ(C_1)$, and $\text{Image}(p^\circ(C_0)) = p(\text{Image}(C_0)) \subseteq p(S) \subseteq T$, while

$\text{Image}(p^\circ(C_1)) = p(\text{Image}(C_1)) \subseteq p(W \setminus S) \subseteq V \setminus T$. □

The most interesting thing here is that the converse is not also true. There are examples where $T \in \text{Obs}(p^\circ(C))$, even though $S \notin \text{Obs}(C)$.

When p is surjective, we think of V as a coarse world model that forgets some details from W. Sometimes, an agent can be able to observe S relative to a coarse description of the world, but not in the more refined description, even in cases where S is definable in both the coarse and refined descriptions.

2.1. Example

Let us look at an example. In this example, the agent is an AI that will be given a number and asked whether it is prime or not. There are two possible environments $E = \{\text{Prime}, \text{Nonprime}\}$.

The agent A has six strategies:

- It can compute whether the number is prime and answer correctly (AccurateHot), or it can compute whether the number is prime and answer incorrectly (InaccurateHot). In both of these cases, it has to use a lot of energy and will become hotter.
- It can also just not think and say that the number is prime (PrimeCool), or it can not think and say that the number is not prime (NonprimeCool).
- Finally, it can deliberately waste a lot of energy for no reason and say the number is prime or waste a lot of energy and say the number is not prime (PrimeHot and NonprimeHot respectively).

Finally, $W = \{\text{PAH}, \text{PAC}, \text{PIH}, \text{PIC}, \text{NAH}, \text{NAC}, \text{NIH}, \text{NIC}\}$, where the first letter indicates whether the AI was given a prime or nonprime number, the second letter indicates whether the AI's answer was accurate or inaccurate, and the third letter indicates whether the AI is hot. The Cartesian frame, C, looks like this.

$$C = \begin{array}{c} \\ \text{AccurateHot} \\ \text{InaccurateHot} \\ \text{PrimeCool} \\ \text{NonprimeCool} \\ \text{PrimeHot} \\ \text{NonprimeHot} \end{array} \begin{array}{c} \begin{array}{cc} \text{Prime} & \text{Nonprime} \end{array} \\ \begin{pmatrix} \text{PAH} & \text{NAH} \\ \text{PIH} & \text{NIH} \\ \text{PAC} & \text{NIC} \\ \text{PIC} & \text{NAC} \\ \text{PAH} & \text{NIH} \\ \text{PIH} & \text{NAH} \end{pmatrix} \end{array}$$

We will let V be the coarse description of the world in which we only pay attention to the input/output behavior of the AI and ignore whether or not it becomes hot.
$V = \{PA, PI, NA, NI\}$, and we will let $p : W \to V$ be the function that deletes the third letter.

This gives us the following for $p^\circ(C)$.

$$p^\circ(C) = \begin{array}{c} \\ \text{AccurateHot} \\ \text{InaccurateHot} \\ \text{PrimeCool} \\ \text{NonprimeCool} \\ \text{PrimeHot} \\ \text{NonprimeHot} \end{array} \begin{array}{c} \begin{array}{cc} \text{Prime} & \text{Nonprime} \end{array} \\ \begin{pmatrix} \text{PA} & \text{NA} \\ \text{PI} & \text{NI} \\ \text{PA} & \text{NI} \\ \text{PI} & \text{NA} \\ \text{PA} & \text{NI} \\ \text{PI} & \text{NA} \end{pmatrix} \end{array} \simeq \begin{array}{c} \\ \text{Accurate} \\ \text{Inaccurate} \\ \text{Prime} \\ \text{Nonprime} \end{array} \begin{array}{c} \begin{array}{cc} \text{Prime} & \text{Nonprime} \end{array} \\ \begin{pmatrix} \text{PA} & \text{NA} \\ \text{PI} & \text{NI} \\ \text{PA} & \text{NI} \\ \text{PI} & \text{NA} \end{pmatrix} \end{array}$$

The important thing to notice here is that $\{PA, PI\} \in \text{Obs}(p^\circ(C))$—when we ignore heat, the agent can base conditional strategies on whether the number is prime—but $\{PAH, PAC, PIH, PIC\} \notin \text{Obs}(C)$.

In particular, $p^\circ(C) \simeq C_0 \& C_1$, where

$$C_0 = \begin{array}{c} \\ \text{Accurate} \\ \text{Inaccurate} \end{array} \begin{array}{c} \text{Prime} \\ \begin{pmatrix} \text{PA} \\ \text{PI} \end{pmatrix} \end{array} \text{ and } C_1 = \begin{array}{c} \\ \text{Accurate} \\ \text{Inaccurate} \end{array} \begin{array}{c} \text{Nonprime} \\ \begin{pmatrix} \text{NA} \\ \text{NI} \end{pmatrix} \end{array},$$

while it is easy to see that $\{\text{PAH}, \text{PAC}, \text{PIH}, \text{PIC}\} \notin \text{Obs}(C)$, because there is no

$a \in \text{if}(\{\text{PAH}, \text{PAC}, \text{PIH}, \text{PIC}\}, \text{PrimeCool}, \text{NonprimeCool})$.

2.2. Discussion

The above example illustrates something interesting about observables. It shows that what's observable is not only a function of the observing agent and the thing that is observed. It is also a function of the level of description of the world!

This makes sense because we are thinking of observation as the ability to implement conditional policies. To implement a conditional policy is to be indistinguishable from the constant policy a_0 in worlds in S and indistinguishable from the constant policy a_1 in

worlds outside of S. This indistinguishability makes observables relative to the level of description of the world.

There is something internal to the agent that is different between the world where it implements a conditional policy and the world where it implements a constant policy. However, when we talk of S being an observable for the agent, we are working relative to a level of description that does not track that internal difference.

3. Functors from Cartesian Frames

When $p : W \to V$ is surjective, p° will send Cartesian frames over the more refined W to

Cartesian frames over the less refined V. What if we want to go in the other direction?

While there is a unique function from less refined worlds to more refined worlds, there are many functions in the other direction. Luckily, we have an object that lets us deal with many functions at once.

Definition: Let $C = (V, E, \cdot)$ be a Cartesian frame over W, with $\text{Agent}(C) = V$. Then

$C^\circ : \text{Chu}(V) \to \text{Chu}(W)$ is the functor that sends $(B, F, \star)$ to $(B, F \times E, \diamond)$, where

$b \diamond (f, e) = (b \star f) \cdot e$, and sends the morphism (g, h) to (g, h'), where $h'(f, e) = (h(f), e)$.

(Notice how this definition looks a bit like currying.)

Claim: C° is well-defined.

Proof: We need to show that C° actually sends objects and morphisms of $\text{Chu}(V)$ to

objects and morphisms of $\text{Chu}(W)$, and that it preserves identity morphisms and

composition.

C° clearly sends objects to objects. To see that it sends morphisms to morphisms, let $(g, h) : (B_0, F_0, \star_0) \to (B_1, F_1, \star_1)$ be a morphism in Chu(V), let $(B_i, F_i \times E, \diamond_i) = C^\circ(B_i, F_i, \star_i)$, and let $(g, h') = C^\circ(g, h)$.

We want to show that $(g, h') : (B_0, F_0 \times E, \diamond_0) \to (B_1, F_1 \times E, \diamond_1)$ is a morphism, which is true because

$$\begin{aligned} g(b) \diamond_1 (f, e) &= (g(b) \star_1 f) \cdot e \\ &= (b \star_0 h(f)) \cdot e \\ &= b \diamond_0 (h(f), e) \\ &= b \diamond_0 h'(f, e) \end{aligned}$$

for all $b \in B_0$ and $(f, e) \in F_1 \times E$. C° clearly preserves identity morphisms and composition. □

The coarse-to-refined functor C° preserves &, $\top$, and null, but not $\oplus$, 0, or $-^*$, which make sense, since C° is violating the symmetry between agent and environment.

Claim: $C^\circ(\top) = \top$, and $C^\circ(\text{null}) = \text{null}$.

Proof: Trivial. □

Claim: $C^\circ(D_0 \& D_1) = C^\circ(D_0) \& C^\circ(D_1)$.

Proof: Let $C = (V, E, \cdot)$ and let $D_i = (B_i, F_i, \star_i)$. We have that

$C^\circ(D_0 \& D_1) = (B_0 \times B_1, (F_0 \sqcup F_1) \times E, \diamond)$ and

$C^\circ(D_0) \& C^\circ(D_1) = (B_0 \times B_1, (F_0 \times E) \sqcup (F_1 \times E), \bullet)$. The agent and environment are the same, so we just need to check that $\diamond = \bullet$.

Take $(b_0, b_1) \in B_0 \times B_1$ and $(f, e) \in (F_0 \sqcup F_1) \times E = (F_0 \times E) \sqcup (F_1 \times E)$. Without loss of generality, assume $f \in F_0$. Observe that

$$(b_0, b_1) \diamond (f, e) = (b_0 \star_0 f) \cdot e$$
$$= (b_0, b_1) \bullet (f, e).$$

□

One way to see that C° does not preserve $\oplus$ is to see that the environments are different, since $C^\circ(D_0 \oplus D_1)$ has one copy of E in the environment, while $C^\circ(D_0) \oplus C^\circ(D_1)$ has two copies.

We also have that C° preserves biextensional equivalence.

Claim: if $D_0 \simeq D_1$, then $C^\circ(D_0) \simeq C^\circ(D_1)$.

Proof: Let $D_i = (B_i, F_i, \star_i)$, and let $C^\circ(D_i) = (B_i, F_i \times E, \diamond_i)$. Let $(g_0, h_0) : D_0 \to D_1$ and $(g_1, h_1) : D_1 \to D_0$ compose to something homotopic to the identity in both orders. It suffices to show that $C^\circ(g_1, h_1) \circ C^\circ(g_0, h_0)$ is homotopic to the identity on $C^\circ(D_0)$, since the other composition will be symmetric. Indeed

$$g_1(g_0(b)) \diamond_0 (f, e) = g_1(g_0(b)) \star_0 f$$
$$= b \star_0 f$$
$$= b \diamond_0 (f, e)$$

for all $b \in B_0$, and $(f, e) \in F_0 \times E$. □

Before we talk about the relationship between functors from functions and functors from Cartesian frames, I want to pause to talk about how to view Cartesian frames as sets of functions.

4. Cartesian Frames as Sets of Functions

One way to view (some) Cartesian frames is as sets of functions.

Definition: Given a set P of functions from E to W, let $CF(P)$ denote the Cartesian frame over W given by $(P, E, \cdot)$, where $p \cdot e = p(e)$.

Claim: $CF(P)$ is well-defined.

Proof: Trivial. □

Not every Cartesian Frame is expressible this way: every Cartesian frame is biextensionally equivalent to a Cartesian frame with duplicate columns and rows, and these uncollapsed frames are excluded because sets do not allow multiplicity.

Claim: For every Cartesian frame C over W, there exists a set of functions $P : \text{Env}(C) \to W$, such that $C \simeq \text{CF}(P)$.

Proof: Take $C = (A, E, \cdot)$, and take P to be the set of all $p : E \to W$ such that there exists an $a \in A$ such that for all $e \in E$, $p(e) = a \cdot e$. Take $(g_0, h_0) : C \to \text{CF}(P)$ and $(g_1, h_1) : \text{CF}(P) \to C$, given as follows: $h_0 = h_1$ is the identity on E, $g_0(a)$ is the function $e \mapsto a \cdot e$, and $g_1(p)$ is some $a \in A$ such that $p(-) = a \cdot -$. These are both clearly morphisms, and they compose to something homotopic to the identity, since $h_0 \circ h_1$ and $h_1 \circ h_0$ are both the identity. □

This give us an alternate definition of Cartesian frames up to biextensional equivalence. This almost gives a complete alternate definition of Cartesian frames; if we instead took P to be a multiset, then we could identify the Cartesian frame $\text{CF}(P)$ with the multiset P.

Note that this is not as symmetric as our original definition of Cartesian frames. The "sets of functions" approach here thinks of a Cartesian frame as a set of functions from the environment to the world, but we could instead think of it as a set of functions from the agent to the world.

Definition: Given a set P of functions for A to W, let $\text{CF}^*(P)$ denote the Cartesian frame over W given by $(A, P, \cdot)$, where $a \cdot p = p(a)$.

Claim: $\text{CF}^*(P) = (\text{CF}(P))^*$.

Proof: Trivial. □

Thinking of Cartesian frames in this way is not particularly different from our original definition. It is just thinking about a function with two inputs as a parameterized function with one input and one parameter. However, this way of understanding Cartesian frames will allow us to more easily relate functors from functions to functors from Cartesian frames.

5. Relationship Between the Two Functor Definitions

Functors from functions are a special case of functors from Cartesian frames. Indeed, they correspond when Env(C) is a singleton.

Claim: For any $p : V \to W$, $p^\circ = (CF^*(\{p\}))^\circ$. Conversely, if $C = (V, \{e\}, \cdot)$ is a Cartesian frame over W with singleton environment, then $C^\circ = p^\circ$, where $p(v) = v \cdot e$.

Proof: Observe that $CF^*(\{p\}) = (V, \{e\}, \cdot)$, where $v \cdot e = p(e)$. That $p^\circ = (CF^*(\{p\}))^\circ$ is trivial from considering the definition of C° in the special case where E is a singleton. □

However, we can do a lot more with functors from Cartesian frames. In the case where $p : W \to V$ is a surjection, p° shows how to send Cartesian frames over the more refined W to the less refined V. We want to go in the other direction using an inverse of p.

Since p is a surjection, it has a right inverse, but it might have many right inverses. If we want to go from Cartesian frames over V to Cartesian frames over W, we could pick any right inverse to p, but since we have functors from Cartesian frames, we don't have to.

Claim: For any surjective $p : W \to V$, let Q be the set of all $q : V \to W$ such that $p \circ q$ is the identity on V. Then for any Cartesian frame C over V, $(p^\circ \circ (CF^*(Q))^\circ)(C) \simeq C$. Thus $(CF^*(Q))^\circ$ is right inverse to p° up to biextensional equivalence.

Proof: Let $C = (A, E, \cdot)$. Then $(CF^*(Q))^\circ(C) = (A, E \times Q, \star)$, where $a \star (e, q) = q(a \cdot e)$, and $(p^\circ \circ (CF^*(Q))^\circ)(C) = (A, E \times Q, \diamond)$, where

$$\begin{aligned} a \diamond (e, q) &= p(q(a \cdot e)) \\ &= a \cdot e. \end{aligned}$$

(Viewed as a matrix, $(p^\circ \circ (CF^*(Q))^\circ)(C)$ is isomorphic to C with $|Q|$ copies of each column.)

To explicitly see the homotopy equivalence, take $(g_0, h_0) : (A, E, \cdot) \to (A, E \times Q, \diamond)$ by $g_0(a) = a$ and $g_1(e, q) = e$, and take $(g_1, h_1) : (A, E \times Q, \diamond) \to (A, E, \cdot)$ by $g_1(a) = a$ and

$h_1(e) = (e, q)$ for some fixed $q \in Q$. These are clearly morphisms and clearly compose to something homotopic to the identity in both orders, since the g_i are the identity. Note that we used the surjectivity of p when we said "for some fixed $q \in Q$," since the surjectivity of p is what makes Q nonempty. □

Functors from Cartesian frames will prove useful in the next section, when we finally introduce the concept of subagent.

Subagents of Cartesian Frames

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Here, we introduce and discuss the concept of a subagent in the Cartesian Frames paradigm.

Note that in this post, as in much of the sequence, we are generally working up to biextensional equivalence. In the discussion, when we informally say that a frame has some property or is some object, what we'll generally mean is that this is true of its biextensional equivalence class.

1. Definitions of Subagent

1.1. Categorical Definition

Definition: Let C and D be Cartesian frames over W. We say that C's agent is a subagent of D's agent, written $C \triangleleft D$, if for every morphism $\phi : C \to \bot$ there exists a pair of morphisms $\phi_0 : C \to D$ and $\phi_1 : D \to \bot$ such that $\phi = \phi_1 \circ \phi_0$.

Colloquially, we say that every morphism from C to $\bot$ factors through D. As a shorthand for "C's agent is a subagent of D's agent," we will just say "C is a subagent of D."

At a glance, it probably isn't clear what this definition has to do with subagents. We'll first talk philosophically about what we mean by "subagent", and then give an alternate definition that will make the connection more clear.

When I say "subagent," I am actually generalizing over two different relationships that may not immediately seem like they belong together.

First, there is the relationship between the component and the whole. One football player is a subagent of the entire football team.

Second, there is the relationship between an agent before and after making a precommitment or a choice. When I precommit not to take a certain action, I am effectively replacing myself with a weaker agent that has fewer options. The new agent with the commitment is a subagent of the original agent.

These are the two notions I am trying to capture with the word "subagent". I am making the philosophical claim that we should think of them primarily as one concept, and am partially backing up this claim by pointing to the simplicity of the above definition. In a future post, we will discuss the formal differences between these two

kinds of subagent, but I think it is best to view them as two special cases of the one simple concept.

(My early drafts of the "Embedded Agency" sequence used the word "subagent" in the title for both the Subsystem Alignment and Robust Delegation sections.)

1.2. Currying Definition

Definition: Let C and D be Cartesian frames over W. We say that $C \triangleleft D$ if there exists a Cartesian frame Z over Agent(D) such that $C \simeq D^{\circ}(Z)$.

Assume for this discussion that we only care about frames up to biextensional equivalence. In effect, the above definition is saying that "C is a subagent of D" means "C's agent is playing a game, Z, where the stakes are to help decide what D's agent does." (And this game may or may not have multiple players, and may or may not fully cover all the options of D's agent.)

Letting $C = (A, E, \cdot)$ and $D = (B, F, \star)$, it turns out (as we will see later) that we can explicitly construct Z. $Z = (A, X, \diamond)$, where X is the set of all morphisms from C to D, and $\diamond : A \times X \to B$ is given by $a \diamond (g, h) = g(a)$.

We will later prove the categorical and currying definitions equivalent, but let's first interpret this definition using examples.

Z is a Cartesian frame whose agent is the agent of C and whose world is the agent of D. This seems like the kind thing we would have when C is a subagent of D.

Thinking about the football example: We have the football player A as the agent in a Cartesian frame C over the world W. We also have the football team B as the agent in a Cartesian frame D over the same world W.

Z is a Cartesian frame *over the football team*; and the agent of this frame is again the football player A. X, the environment of Z, represents the rest of the football team: the player's effect on the team as a whole (here treated as the player's world) is a function of what the player chooses and what the rest of the team chooses. We can think of Z as representing a "zoomed-in" picture of A interacting with its local environment (the team), while C represents a "zoomed-out" picture of A interacting with its teammates *and* the larger world (rival teams, referees, etc.).

$D^\circ(Z) = (A, X \times F, \bullet)$, so E is equivalent to $X \times F$, which is saying that the environment for the football player in its original frame (C) is equivalent to the Cartesian product of the rest of the team X with the *team's* environment F.

Thinking about the precommitment example: C has made a precommitment, so there is an inclusion morphism $\iota : A \to B$, which shows that C's agent's options are a subset of D's agent's options. Z is just $CF^*(\{\iota\})$, so $X = \{\iota\}$ is a singleton. $D^\circ(Z) = (A, X \times F, \bullet)$, so E is equivalent to $X \times F = F$, so here A is a subset of B and E is equivalent to F.

Although the word "precommitment" suggests a specific (temporal, deliberative) interpretation, formally, precommitment just looks like deleting rows from a matrix (up to biextensional equivalence), which can represent a variety of other situations.

A Cartesian frame $Z = (A, X, \diamond)$ over B is like a nondeterministic function from A to B, where X represents the the nondeterministic bits. When changing our frame from $(B, F, \star)$ to $(A, E, \cdot) \simeq (A, X \times F, \bullet)$, we are identifying with A and externalizing the nondeterministic bits X into the environment.

1.3. Covering Definition

The categorical definition is optimized for elegance, while the currying definition is optimized to be easy to understand in terms of agency. We have a third definition, the covering definition, which is optimized for ease of use.

Definition. Let $C = (A, E, \cdot)$ and $D = (B, F, \star)$ be Cartesian frames over W. We say that $C \triangleleft D$ if for all $e \in E$, there exists an $f \in F$ and a $(g, h) : C \to D$ such that $e = h(f)$.

We call this the covering definition because the morphisms from C to D cover the set E.

2. Equivalence of Definitions

2.1. Equivalence of Categorical and Covering Definitions

The equivalence of the categorical and covering definitions follows directly from the fact that the morphisms from C to $\bot$ are exactly the elements of Env(C).

Claim: The categorical and covering definitions of subagent are equivalent.

Proof: Let $C = (A, E, \cdot)$ and let $D = (B, F, \star)$. First, observe that the morphisms from C to $\bot$ correspond exactly to the elements of E. For each $e \in E$, it is easy to see that $(g, h) : C \to (W, \{j\}, \diamond)$, given by $h(j) = e$ and $g(a) = a \cdot e$, is a morphism, and every morphism is uniquely defined by h(j), so there are no other morphisms. Let ϕ_e denote the morphisms with $h(j) = e$.

Similarly, the morphisms from D to $\bot$ correspond to the elements of F. Let ψ_f denote the morphisms corresponding to $f \in F$.

Thus, the categorical definition can be rewritten to say that for every morphism $\phi_e : C \to \bot$, there exist morphisms $(g, h) : C \to D$ and $\psi_f : D \to \bot$, such that $\phi_e = \psi_f \circ (g, h)$. However, $\psi_f \circ (g, h) : C \to (W, \{j\}, \diamond)$ sends j to h(f), and so equals ϕ_e if and only if $e = h(f)$. Thus the categorical definition is equivalent to the covering definition. $\square$

2.2. Equivalence of Covering and Currying Definitions

Claim: The covering definition of subagent implies the currying definition of subagent.

Proof: Let $C = (A, E, \cdot)$ and $D = (B, F, \star)$ be Cartesian frames over W. Assume that $C \triangleleft D$ according to the covering definition.

Let X be the set of all morphisms from C to D, and let $Z = (A, X, \diamond)$ be a Cartesian frame over B, with $\diamond$ given by $a \diamond (g, h) = g(a)$. We have that $D^\circ(Z) = (A, X \times F, \bullet)$, with

$$\begin{aligned} a \bullet ((g, h), f) &= (a \diamond (g, h)) \star f \\ &= g(a) \star f \end{aligned}$$

for all $a \in A$, $(g, h) \in X$, and $f \in F$.

To show that $C \simeq D^\circ(Z)$, we need to construct morphisms $g_0, h_0 : C \to D^\circ(Z)$ and $g_1, h_1 : D^\circ(Z) \to C$ which compose to something homotopic to the identity in both orders.

We will let g_0 and g_1 be the identity on A, and we let $h_0 : X \times F \to E$ be given by $h_0((g, h), f) = h(f)$. Finally, we let $h_1(e) = ((g, h), f)$ such that $h(f) = e$. We can always choose such a $(g, h) \in X$ and $f \in F$ by the covering definition of subagent.

We have that (g_0, h_0) is a morphism, since

$$\begin{aligned} g_0(a) \bullet ((g, h), f) &= a \bullet ((g, h), f) \\ &= g(a) \star f \\ &= a \cdot h(f) \\ &= a \cdot h_0((g, h), f). \end{aligned}$$

Similarly, we have that (g_1, h_1) is a morphism since $h_1(e) = ((g, h), f)$, where $h(f) = e$, so

$$\begin{aligned} g_1(a) \cdot e &= a \cdot e \\ &= a \cdot h(f) \\ &= g(a) \star f \\ &= a \bullet ((g, h), f) \\ &= a \bullet h_1(e). \end{aligned}$$

It is clear that (g_0, h_0) and (g_1, h_1) compose to something homotopic to the identity in both orders, since g_0 and g_1 are the identity on A. Thus, $C \simeq D^\circ(Z)$. □

Claim: The currying definition of subagent implies the covering definition of subagent.

Proof: Let $C = (A, E, \cdot)$ and $D = (B, F, \star)$ be Cartesian frames over W. Let $Z = (Y, X, \diamond)$ be a Cartesian frame over B, and let $C \simeq D^\circ(Z)$. Our goal is to show that for every

$e \in E$, there exists a $(g, h) : C \to D$ and $f \in F$ such that $e = h(f)$. We will start with the special case where $C = D^\circ(Z)$.

We have that $D^\circ(Z) = (Y, X \times F, \bullet)$, where $y \bullet (x, f) = (y \diamond x) \star f$. First, note that for every $x \in X$, there exists a morphism $(g_x, h_x) : D^\circ(Z) \to D$ given by $g_x(y) = y \diamond x$, and $h_x(f) = (x, f)$. To see that this is a morphism, observe that

$$\begin{aligned} g_x(y) \star f &= (y \diamond x) \star f \\ &= y \bullet (x, f) \\ &= f \bullet h_x(f) \end{aligned}$$

for all $y \in Y$ and $f \in F$.

To show that $D^\circ(Z) \triangleleft D$ according to the covering definition, we need that for all $(x, f) \in X \times F$, there exists an $f' \in F$ and a $(g, h) : D^\circ(Z) \to D$ such that $h(f') = (x, f)$. Indeed we can take $(g, h) = (g_x, h_x)$ and $f' = f$.

Now, we move to the case where $C \simeq D^\circ(Z)$, but $C \neq D^\circ(Z)$. It suffices to show that under the covering definition of subagent, if $C_0 \triangleleft D$, and $C_1 \simeq C_0$, then $C_1 \triangleleft D$.

Let $C_i = (A_i, E_i, \cdot_i)$, and let $(g_0, h_0) : C_0 \to C_1$ and $(g_1, h_1) : C_1 \to C_0$ compose to something homotopic to the identity in both orders. Assume that $C_0 \triangleleft D$. To show that $C_1 \triangleleft D$, let the possible environment $e \in E_1$ be arbitrary.

$h_0(e) \in E_0$, so there exists an $f \in F$ and $(g, h) : C_0 \to D$ such that $h(f) = h_0(e)$. Consider the morphism $(g', h') : C_1 \to D$, where $g' = g \circ g_1$, and $h'(f) = e$ and $h'(f') = (h_1 \circ h)(f')$ on all $f' \neq f$. To see that this is a morphism, observe that for all $a \in A_1$, we have

$$\begin{aligned} g'(a) \star f &= g(g_1(a)) \star f \\ &= a \cdot_1 h_1(h(f)) \\ &= a \cdot_1 h_1(h_0(e)) \\ &= a \cdot_1 e \\ &= a \cdot_1 h'(f), \end{aligned}$$

while for $f' \in F$, $f' \neq f$, we have

$$\begin{aligned} g'(a) \star f' &= g(g_1(a)) \star f' \\ &= a \cdot_1 h_1(h(f')) \\ &= a \cdot_1 h'(f'). \end{aligned}$$

Now, notice that for our arbitrary $e \in E_1$, $(g', h') : C_1 \to D$ and $f \in F$ satisfy $h'(f) = e$, so $C_1 \triangleleft D$ according to the to the covering definition.

Thus, whenever $C \simeq D^\circ(Z)$, we have $C \triangleleft D$ according to the covering definition, so the currying definition implies the covering definition of subagent. □

3. Mutual Subagents

The subagent relation is both transitive and reflexive. Surprisingly, this relation is not anti-symmetric, even up to biextensional equivalence.

Claim: $\triangleleft$ is reflexive. Further, if $C \simeq D$, then $C \triangleleft D$.

Proof: Let $C = (A, E, \cdot)$ and $D = (B, F, \cdot)$ be Cartesian Frames over W, with $C \simeq D$.

Consider the Cartesian frame Z over B given by $Z = (B, \{x\}, \diamond)$, where $b \diamond x = b$.

Observe that $D \cong D^\circ(Z)$. Thus $C \simeq D^\circ(Z)$, so $C \triangleleft D$, according to the currying definition.

□

Claim: $\triangleleft$ is transitive.

Proof: We will use the categorical definition. Let $C_0 \triangleleft C_1$ and $C_1 \triangleleft C_2$. Given a morphism, $\phi_0 : C_0 \to \bot$, since $C_0 \triangleleft C_1$, we know that $\phi_0 = \phi_1 \circ \phi_2$ with $\phi_1 : C_1 \to \bot$ and $\phi_2 : C_0 \to C_1$. Further, since $C_1 \triangleleft C_2$, we know that $\phi_1 = \phi_3 \circ \phi_4$ with $\phi_3 : C_2 \to \bot$ and $\phi_4 : C_1 \to C_2$. Thus,

$$\begin{aligned} \phi_0 &= (\phi_3 \circ \phi_4) \circ \phi_2 \\ &= \phi_3 \circ (\phi_4 \circ \phi_2), \end{aligned}$$

with $\phi_3 : C_2 \to \bot$ and $\phi_4 \circ \phi_2 : C_0 \to C_2$, so $C_0 \triangleleft C_2$. □

As a corollary, we have that subagents are well-defined up to biextensional equivalence.

Corollary: If $C_0 \simeq C_1$, $D_0 \simeq D_1$, and $C_0 \triangleleft D_0$, then $C_1 \triangleleft D_1$.

Proof: $C_1 \triangleleft C_0 \triangleleft D_0 \triangleleft D_1$. □

Sometimes, there are Cartesian frames $C \not\simeq D$ with $C \triangleleft D$ and $D \triangleleft C$. We can use this fact to define a third equivalence relation on Cartesian frames over W, weaker than both $\cong$ and $\simeq$.

Definition: For Cartesian frames C and D over W, we say $C \bowtie D$ if $C \triangleleft D$ and $D \triangleleft C$.

Claim: $\bowtie$ is an equivalence relation.

Proof: Reflexivity and transitivity follow from reflexivity and transitivity of $\triangleleft$. Symmetry is trivial. □

This equivalence relation is less natural than $\cong$ and $\simeq$, and is not as important. We discuss it mainly to emphasize that two frames can be mutual subagents without being biextensionally equivalent.

Claim: $\bowtie$ is strictly weaker than $\simeq$, which is strictly weaker than $\cong$.

Proof: We already know that $\simeq$ is weaker than $\cong$. To see that $\bowtie$ is weaker than $\simeq$, observe that if $C \simeq D$, then $C \triangleleft D$ and $D \triangleleft C$, so $C \bowtie D$.

To see that $\simeq$ is *strictly* weaker than $\cong$, observe that $\top \oplus \top \simeq \top$ (both have empty environment and nonempty agent), but $\top \oplus \top \ncong \top$ (the agents have different size).

To see that $\bowtie$ is strictly weaker than $\simeq$, observe that $\top \bowtie \text{null}$ (vacuous by covering definition), but $\top \not\simeq \text{null}$ (there are no morphisms from null to $\top$). $\square$

I do not have a simple description of exactly when $C \bowtie D$, but there are more cases than just the trivial ones like $C \simeq D$ and vacuous cases like $\top \bowtie \text{null}$. As a quick example:

$$\begin{matrix} x \\ (y) \end{matrix} \bowtie \begin{pmatrix} x & x \\ y & y \\ x & y \end{pmatrix}.$$

To visualize this, imagine an agent that is given the choice between cake and pie. This agent can be viewed as a team consisting of two subagents, Alice and Bob, with Alice as the leader.

Alice has three choices. She can choose cake, she can choose pie, or she can delegate the decision to Bob. We represent this with a matrix where Bob is in Alice's environment, and the third row represents Alice letting the environment make the call:

$$\begin{pmatrix} x & x \\ y & y \\ x & y \end{pmatrix}.$$

If we instead treat Alice-and-Bob as a single superagent, then their interaction across the agent-environment boundary becomes agent-internal deliberation, and their functional relationship to possible worlds just becomes a matter of "What does the group decide?". Thus, Alice is a subagent of the Alice-and-Bob team:

$$\begin{pmatrix} x & x \\ y & y \\ x & y \end{pmatrix} \triangleleft \begin{matrix} x \\ (y) \end{matrix}.$$

However, Alice also has the ability to commit to not delegating to Bob. This produces a future version of Alice that doesn't choose the third row. This new agent is a precommitment-style subagent of the original Alice, but using biextensional collapse, we can also see that this new agent is equivalent to the smaller matrix. Thus:

$$\begin{pmatrix} x \\ y \end{pmatrix} \simeq \begin{pmatrix} x & x \\ y & y \end{pmatrix}^{\triangleleft} \begin{pmatrix} x & x \\ y & y \\ x & y \end{pmatrix}.$$

It is also easy to verify formally that these are mutual subagents using the covering definition of subagent.

I'm reminded here of the introduction and deletion of mixed strategies in game theory. The third row of Alice's frame is a mix of the first two rows, so we can think of Bob as being analogous to a random bit that the environment cannot see. I informally conjecture that for finite Cartesian frames, $C \bowtie D$ if and only if you can pass between C and D by doing something akin to deleting and introducing mixed strategies for the agent.

However, this informal conjecture is not true for infinite Cartesian frames:

$$\begin{pmatrix} & & & & & \\ & & & & & \\ & & & & & \\ z & x & z & z & z & \cdots \\ z & y & z & z & z & \cdots \\ z & z & x & z & z & \cdots \\ z & z & y & z & z & \cdots \\ z & z & z & x & z & \cdots \\ z & z & z & y & z & \cdots \\ z & z & z & z & x & \cdots \\ z & z & z & z & y & \cdots \\ & & & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \bowtie \begin{pmatrix} & & & & & \\ & & & & & \\ z & y & z & z & z & \cdots \\ z & z & x & z & z & \cdots \\ z & z & y & z & z & \cdots \\ z & z & z & x & z & \cdots \\ z & z & z & y & z & \cdots \\ z & z & z & z & x & \cdots \\ z & z & z & z & y & \cdots \\ & & & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

We can see that these frames are mutual subagents by noting that one can transition back and forth by repeatedly committing not to take the top row.

I do not know of any examples of $\bowtie$ that look qualitatively different from those discussed here, but I do not have a good understanding of exactly what the equivalence classes look like.

4. Universal Subagents and Superagents

We can view $\top$ as a universal subagent and $\bot$ as a universal superagent.

Claim: $\top \triangleleft C \triangleleft \bot$ for all Cartesian frames C.

Proof: We use the categorical definition. That $\top \triangleleft C$ is vacuous, since there is no morphism from $\top$ to $\bot$. That $C \triangleleft \bot$ is also trivial, since any $\phi : C \triangleleft \bot$ is equal to $\phi \circ id_{\bot}$. $\square$

Since null $\bowtie \top$, we also have null $\triangleleft C$ for all C.

We also have a that $\bot_S$ is a superagent of all Cartesian frames with image in S.

Claim: $C \triangleleft \bot_S$ if and only if Image(C) $\subseteq$ S.

Proof: Let $C = (A, E, \cdot)$, and let $\bot_S = (S, \{f\}, \star)$, with $s \star f = s$.

First, assume Image(C) $\subseteq$ S. We will use the covering definition. Given an $e \in E$, let $(g, h) : C \to \bot_S$ be given by $g(a) = a \cdot e$ and $h(f) = e$. We have that g is well-defined because Image(C) $\subseteq$ S, and (g, h) is a morphism because for all $a \in A$,

$$\begin{aligned} a \cdot h(f) &= a \cdot e \\ &= g(a) \\ &= g(a) \star f. \end{aligned}$$

Thus, there is a morphism $(g, h) : C \to \bot_S$ and an element $f \in \{f\}$ such that $h(f) = e$ for an arbitrary $e \in E$, so $C \triangleleft \bot_S$.

Conversely, assume Image(C) $\not\subseteq$ S, so let $a \in A$ and $e \in E$ be such that $a \cdot e \notin S$. If we assume for contradiction that $C \triangleleft \bot_S$, then by the covering definition, there must be a morphism $(g, h) : C \to \bot_S$ such that $h(f) = e$. But then we have that

$$\begin{aligned} a \cdot e &= a \cdot h(f) \\ &= g(a) \star f \\ &= g(a) \end{aligned}$$

must be both inside and outside of S, a contradiction. □

Convention: We will usually write $C \triangleleft \perp_S$ instead of Image(C) ⊆ S, as it is shorter.

Corollary: S ∈ Obs(C) if and only if $C \simeq C_0 \& C_1$ for some $C_0 \triangleleft \perp_S$ and $C_1 \triangleleft \perp_{W \setminus S}$.

Proof: This is just rewriting our definition of observables from "Controllables and Observables, Revisited." □

In the coming posts, we will introduce multiplicative operations on Cartesian frames, and use these to distinguish between additive and multiplicative subagents and superagents.

Multiplicative Operations on Cartesian Frames

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the seventh post in the Cartesian frames sequence.

Here, we introduce three new binary operations on Cartesian frames, and discuss their properties.

1. Tensor

Our first multiplicative operation is the tensor product, $\otimes$.

One way we can visualize our additive operations from before, $\oplus$ and &, is to imagine two robots (say, a mining robot Agent(C) and a drilling robot Agent(D)) that have an override mode allowing an AI supervisor to take over that robot's decisions.

- $C \oplus D$ represents the supervisor deciding which robot to take control of, then selecting that robot's action. (The other robot continues to run autonomously.)
- C&D represents something in the supervisor's environment (e.g., its human operator) deciding which robot the supervisor will take control of. Then the supervisor selects that robot's action (while the other robot runs autonomously).

$C \otimes D$ represents an AI supervisor that controls both robots simultaneously. This lets Agent($C \otimes D$) direct Agent(C) and Agent(D) to work together as a team.

Definition: Let $C = (A, E, \cdot)$ and $D = (B, F, \star)$ be Cartesian frames over W. The tensor product of C and D, written $C \otimes D$, is given by $C \otimes D = (A \times B, \hom(C, D^*), \diamond)$, where $\hom(C, D^*)$ is the set of morphisms $(g, h) : C \to D^*$ (i.e., the set of all pairs $(g : A \to F, h : B \to E)$ such that $b \star g(a) = a \cdot h(b)$ for all $a \in A$, $b \in B$), and $\diamond$ is given by $(a, b) \diamond (g, h) = b \star g(a) = a \cdot h(b)$.

Let us meditate for a moment on why this definition represents two agents working together on a team. The following will be very informal.

Let Alice be an agent with Cartesian frame $C = (A, E, \cdot)$, and let Bob be an agent with Cartesian frame $D = (B, F, \star)$. The team consisting of Alice and Bob should have agent $A \times B$, since the team's choices consist of deciding what Alice does and also deciding what Bob does.

The environment is a bit more complicated. Starting from Alice, to construct the team, we want to internalize Bob's choices: instead of just being choices in A's environment, Bob's choices will now be additional options for the team $A \times B$.

To do this, we want to first see Bob as being embedded in Alice's environment. This embedding is given by a function $h : B \to E$, which extends each $b \in B$ to a full environment $e \in E$. We will view Alice's possible environments as being constructed by combining a choice by Bob (that is, a $b \in B$) with a function from Bob's choices to possible environments ($h : B \to E$). Then, we will move the B part across the Cartesian boundary into the agent.

Now, the agent looks like $A \times B$, while the environment looks like $B \to E$. However, we must have been able to do this starting from Bob as well, so a possible environment can also be viewed as function $g : A \to F$.

Since we should get the same world regardless of whether we think of the team as starting with Alice or with Bob, these functions g and h should agree with each other.

This looks a bit like currying. The environment for an Alice-Bob team should be able to take in a Bob to create an environment for Alice, and it should also be able to take in an Alice to create an environment for Bob.

1.1. Example

We will illustrate this new operation using a simple formal example.

Jack, Kate, and Luke are simultaneously casting votes on whether to have a party. Each agent can vote for or against the party. The possible worlds are encoded as strings listing which people vote for the party, $W = \{\varepsilon, J, K, L, JK, JL, KL, JKL\}$. Jack's perspective is given by the frame

$$C_J = \begin{pmatrix} J & JK & JL & JKL \\ \varepsilon & K & L & KL \end{pmatrix},$$

Kate's perspective is given by the frame

$$C_K = \begin{pmatrix} K & JK & KL & JKL \\ \varepsilon & J & L & JL \end{pmatrix},$$

and Luke's perspective is given by the frame

$$C_L = \begin{pmatrix} L & JL & KL & JKL \\ \varepsilon & J & K & JK \end{pmatrix}.$$

Since Luke's environment can be thought of as the team consisting of Jack and Kate, one might expect that $C_J \otimes C_K \cong C_L^*$. Indeed, we will show this is the case.

Let $C_J = (A, E, \cdot)$, and let $C_K = (B, F, \star)$. We label the elements of A, E, B, and F as follows:

$$C_J = \begin{matrix} a_J \\ a_\varepsilon \end{matrix} \begin{pmatrix} e_\varepsilon & e_K & e_L & e_{KL} \\ J & JK & JL & JKL \\ \varepsilon & K & L & KL \end{pmatrix}, \text{ and } C_K = \begin{matrix} b_K \\ b_\varepsilon \end{matrix} \begin{pmatrix} f_\varepsilon & f_J & f_L & f_{JL} \\ K & JK & KL & JKL \\ \varepsilon & J & L & JL \end{pmatrix}.$$

We will first enumerate all of the morphisms from C_J to C_K^*. A morphism $(g, h) : C_J \to C_K^*$ consists of a function $g : A \to F$ and a function $h : B \to E$. There are 16 functions from A to F and 16 functions from B to E, but most of the 256 pairs do not form morphisms.

Let us break the possibilities into cases based on $g(a_J)$. Observe that

$b_K \star g(a_J) = a_J \cdot h(b_K)$: the possible worlds where (from Kate's perspective) Kate votes for the party and Jake-interfacing-with-Kate's-perspective votes for the party too, are the same as the possible worlds where (from Jake's perspective) Jake votes for the party and Kate-interfacing-with-Jake's-perspective does too. These possible worlds must have a J in them, so $g(a_J)$ must be either f_J or f_{JL}.

If $g(a_J) = f_J$, then

$$\begin{aligned} a_J \cdot h(b_K) &= b_K \star g(a_J) \\ &= JK, \end{aligned}$$

so $h(b_K) = e_K$. Similarly,

$$\begin{aligned} a_J \cdot h(b_\varepsilon) &= b_\varepsilon \star g(a_J) \\ &= J, \end{aligned}$$

so $h(b_\varepsilon) = e_\varepsilon$, and

$$\begin{aligned} b_K \star g(a_\varepsilon) &= a_\varepsilon \cdot h(b_K) \\ &= K, \end{aligned}$$

so $g(a_\varepsilon) = f_\varepsilon$.

Similarly, if $g(a_J) = f_{JL}$, then $h(b_K) = e_{KL}$, $h(b_\varepsilon) = e_L$, and $g(a_\varepsilon) = f_L$.

Thus, there are only two candidate morphisms:

- The first, which we will call $\phi_\varepsilon = (g_\varepsilon, h_\varepsilon)$, is given by $g_\varepsilon(a_\varepsilon) = f_\varepsilon$, $g_\varepsilon(a_J) = f_J$, $h_\varepsilon(b_\varepsilon) = e_\varepsilon$, and $h_\varepsilon(b_K) = e_K$.
- The second, $\phi_L = (g_L, h_L)$, is given by $g_L(a_\varepsilon) = f_L$, $g_L(a_J) = f_{JL}$, $h_L(b_\varepsilon) = e_L$, and $h_L(b_K) = e_{KL}$.

It is easy to see that these are both indeed morphisms, by checking the definition of morphism on all four pairs in $A \times B$.

Thus, $\mathrm{Env}(C_J \otimes C_K) = \{\phi_\varepsilon, \phi_L\}$, and $\mathrm{Agent}(C_J \otimes C_K) = A \times B$, and we can compute $\mathrm{Eval}(C_J \otimes C_K)$ from the definitions of the morphisms. The result is as follows:

$$C_J \otimes C_K = \begin{array}{c} \\ (a_J, b_K) \\ (a_J, b_\varepsilon) \\ (a_\varepsilon, b_K) \\ (a_\varepsilon, b_\varepsilon) \end{array} \begin{array}{c} \begin{array}{cc} \phi_\varepsilon & \phi_L \end{array} \\ \begin{pmatrix} JK & JKL \\ J & JL \\ K & KL \\ \varepsilon & L \end{pmatrix} \end{array}.$$

This is clearly C_L^*, up to reordering and relabeling rows and columns.

2. Properties of Tensor

Tensor introduces a lot of categorical structure to Chu spaces, in fact giving us a star-autonomous category. This post and the ones to come will be ignoring connections to larger topics in category theory, but only because my time and my familiarity with category theory are limited, not because these connections are unimportant.

I encourage the interested reader to learn more about the structure of Chu spaces on the excellent category theory wiki *n*Lab, beginning with their article on the Chu construction.

2.1. Commutativity, Associativity, and Identity

Claim: $\otimes$ is commutative and associative, and 1 is the identity of $\otimes$ (up to isomorphism).

Proof: Commutativity is clear from the definition of $\otimes$, once one unpacks the definition of $\hom(C, D^*)$.

To see that 1 is the identity of $\otimes$, let $C = (A, E, \cdot)$, let $1 = (\{b\}, W, \star)$, and let $C \otimes 1 = (A \times \{b\}, \hom(C, 1^*), \diamond)$.

Consider the isomorphism $(\iota_0, \iota_1) : C \to C \otimes 1$ given by $\iota_0(a) = (a, b)$ and $\iota_1(g, h) = h(b)$. We need to show that (ι_0, ι_1) is a morphism, and that both ι_0 and ι_1 are bijective. To see that (ι_0, ι_1) is a morphism, observe that for all $a \in A$ and $(g, h) : C \to 1^*$,

$$\begin{aligned} \iota_0(a) \diamond (g, h) &= a \cdot h(b) \\ &= a \cdot \iota_1(g, h). \end{aligned}$$

Clearly, ι_0 is a bijection, so all that remains to show is that ι_1 is bijective.

To see that ι_1 is injective, observe that if $\iota_1(g_0, h_0) = \iota_1(g_1, h_1)$, then $h_0(b) = h_1(b)$, so $h_0 = h_1$, and

$$\begin{aligned} g_0(a) &= b \star g_0(a) \\ &= a \cdot h_0(b) \\ &= a \cdot h_1(b) \\ &= b \star g_1(a) \\ &= g_1(a) \end{aligned}$$

for all $a \in A$, so $g_0 = g_1$.

To see that ι_1 is surjective, observe that for every $e \in E$, there exists a morphism $(g_e, h_e) : C \to 1^*$, given by $h_e(b) = e$ and $g_e(a) = a \cdot e$. This is clearly a morphism, since

$$\begin{aligned} b \star g_e(a) &= g_e(a) \\ &= a \cdot e \\ &= a \cdot h_e(b), \end{aligned}$$

and $\iota_1(g_e, h_e) = e$.

Next, we need to show that $\otimes$ is associative, which will be much more tedious. Let $C_i = (A_i, E_i, \cdot)$. Since we have already established commutativity, it suffices to show that $(C_0 \otimes C_1) \otimes C_2 \cong (C_0 \otimes C_2) \otimes C_1$.

Let $D = (A_0 \times A_1 \times A_2, F, \star)$, where F is the set of all triples of functions $(g_0 : A_1 \times A_2 \to E_0, g_1 : A_0 \times A_2 \to E_1, g_2 : A_0 \times A_1 \to E_2)$, such that for all $a_i \in A_i$, we have

$$\begin{aligned} a_0 \cdot_0 g_0(a_1, a_2) &= a_1 \cdot_1 g_1(a_0, a_2) \\ &= a_2 \cdot_2 g_2(a_0, a_1), \end{aligned}$$

and $\star$ is given by

$$\begin{aligned} (a_0, a_1, a_2) \star (g_0, g_1, g_2) &= a_0 \cdot_0 g_0(a_1, a_2) \\ &= a_1 \cdot_1 g_1(a_0, a_2) \\ &= a_2 \cdot_2 g_2(a_0, a_1). \end{aligned}$$

We will show that $(C_0 \otimes C_1) \otimes C_2 \cong D$, and since the definition of D is symmetric in swapping C_1 and C_2, it will follow that $(C_0 \otimes C_2) \otimes C_1 \cong D$, so $(C_0 \otimes C_1) \otimes C_2 \cong (C_0 \otimes C_2) \otimes C_1$.

We construct a morphism (ι_0, ι_1) from $(C_0 \otimes C_1) \otimes C_2$ to D as follows. ι_0 is just the identity on $A_0 \times A_1 \times A_2$. We will let $\iota_1(g_0, g_1, g_2)$ be the morphism $(g_2, h) : C_0 \otimes C_1 \to C_2^*$, where $h : A_2 \to \hom(C_0, C_1^*)$ is given by $h(a_2) = (h_0^{a_2}, h_1^{a_2}) : C_0 \to C_1^*$, where $h_0^{a_2}(a_0) = g_1(a_0, a_2)$, and $h_1^{a_2}(a_1) = g_0(a_1, a_2)$.

First, we need to show that ι_1 is well-defined, by showing that $h(a_2)$ is a morphism from C_0 to C_1^*, and that (g_2, h) is a morphism from $C_0 \otimes C_1 \to C_2^*$. To see that $h(a_2) = (h_0^{a_2}, h_1^{a_2})$ is a morphism, observe that for $a_0 \in A_0$ and $a_1 \in A_1$,

$$\begin{aligned} a_1 \cdot_1 h_0^{a_2}(a_0) &= a_1 \cdot_1 g_1(a_0, a_2) \\ &= a_0 \cdot_0 g_0(a_1, a_2) \\ &= a_0 \cdot_0 h_1^{a_2}(a_1). \end{aligned}$$

To see that (g_2, h) is a morphism, observe for all $(a_0, a_1) \in A_0 \times A_1$ and all $a_2 \in A_2$,

$$\begin{aligned} a_2 \cdot_2 g_2(a_0, a_1) &= a_0 \cdot_0 g_0(a_1, a_2) \\ &= a_0 \cdot_0 h_1^{a_2}(a_1) \\ &= (a_0, a_1) \diamond (h_0^{a_2}, h_1^{a_2}) \\ &= (a_0, a_1) \diamond h(a_2), \end{aligned}$$

where $\diamond = \mathrm{Eval}(C_0 \otimes C_1)$.

Now that we know ι_1 is well-defined, we need to show that (ι_0, ι_1) is a morphism.

Indeed, for all $(a_0, a_1, a_2) \in A_0, A_1, A_2$, and for all $(g_0, g_1, g_2) \in F$, we have

$$\begin{aligned} \iota_0(a_0, a_1, a_2) \star (g_0, g_1, g_2) &= a_2 \cdot_2 g_2(a_0, a_1) \\ &= (a_0, a_1, a_2) \bullet (g_2, h) \\ &= (a_0, a_1, a_2) \bullet \iota_1(g_0, g_1, g_2), \end{aligned}$$

where $\bullet = \mathrm{Eval}((C_0 \otimes C_1) \otimes C_2)$.

Finally, to show that (ι_0, ι_1) is an isomorphism, we need to show that ι_0 and ι_1 are bijective. ι_0 is trivial, since it is the identity, so it suffices to show that ι_1 is bijective.

To see that ι_1 is surjective, let (g, h) be a morphism from $C_0 \otimes C_1$ to C_2^*, so $g : A_0 \times A_1 \to E_2$, and $h : A_2 \to \hom(C_0, C_1^*)$. Again, let $h(a_2) = (h_0^{a_2}, h_1^{a_2})$. We will define (g_0, g_1, g_2) by $g_2 = g$, $g_1(a_0, a_2) = h_0^{a_2}(a_0)$, and $g_0(a_1, a_2) = h_1^{a_2}(a_1)$.

We need to show that $(g_0, g_1, g_2) \in F$, by showing that for all $(a_0, a_1, a_2) \in A_0 \times A_1 \times A_2$, we have

$$\begin{aligned} a_0 \cdot_0 g_0(a_1, a_2) &= a_1 \cdot_1 g_1(a_0, a_2) \\ &= a_2 \cdot_2 g_2(a_0, a_1). \end{aligned}$$

Observe that since (g, h) is a morphism,

$$\begin{aligned} a_2 \cdot_2 g_2(a_0, a_1) &= a_2 \cdot_2 g(a_0, a_1) \\ &= (a_0, a_1) \star h(a_2) \\ &= (a_0, a_1) \star (h_0^{a_2}, h_1^{a_2}), \end{aligned}$$

where $\star = \mathrm{Eval}(C_0 \otimes C_1)$. Also, by the definition of $C_0 \otimes C_1$, we have that

$$(a_0, a_1) \star (h_0^{a_2}, h_1^{a_2}) = a_0 \cdot_0 h_1^{a_2}(a_1) = a_0 \cdot_0 g_0(a_1, a_2),$$

and similarly

$$(a_0, a_1) \star (h_0^{a_2}, h_1^{a_2}) = a_1 \cdot_1 h_0^{a_2}(a_1) = a_1 \cdot_1 g_1(a_0, a_2).$$

Thus,

$$a_0 \cdot_0 g_0(a_1, a_2) = a_1 \cdot_1 g_1(a_0, a_2) = a_2 \cdot_2 g_2(a_0, a_1),$$

so $(g_0, g_1, g_2) \in F$. Finally, observe that $\iota_1(g_0, g_1, g_2)$ is in fact (g, h).

To show that ι_1 is injective, assume $\iota_1(g_0, g_1, g_2) = \iota_1(g_0', g_1', g_2') = (g, h)$, and given an $a_2 \in A_2$, let $h(a_2) = (h_0^{a_2}, h_1^{a_2})$. Clearly, this means $g_2 = g = g_2'$. Further, for all $a_0 \in A_0$, $a_1 \in A_1$, and $a_2 \in A_2$,

$$g_0(a_1, a_2) = h_1^{a_2}(a_1) = g_0'(a_1, a_2)$$

and

$$g_1(a_0, a_2) = h_0^{a_2}(a_0) = g_1'(a_0, a_2).$$

Thus $(g_0, g_1, g_2) = (g_0', g_1', g_2')$. Thus, ι_1 is bijective, so (ι_0, ι_1) is an isomorphism, so $(C_0 \otimes C_1) \otimes C_2 \cong D \cong (C_0 \otimes C_2) \otimes C_1$. □

2.2. Biextensional Equivalence

Since many of our intuitions about Cartesian frames are up to biextensional equivalence, we should verify that tensor is well-defined up to biextensional equivalence.

Claim: If $C_0 \simeq C_1$ and $D_0 \simeq D_1$, then $C_0 \otimes D_0 \simeq C_1 \otimes D_1$.

Proof: It suffices to show that for all D, $C_0 \otimes D \simeq C_1 \otimes D$. Then, by commutativity of tensor,

$$\begin{aligned} C_0 \otimes D_0 &\simeq C_0 \otimes D_1 \\ &\cong D_1 \otimes C_0 \\ &\simeq D_1 \otimes C_1 \\ &\equiv C_1 \otimes D_1 . \end{aligned}$$

Let $C_i = (A_i, E_i, \cdot_i)$, and let $D = (B, F, \star)$. Since $C_0 \simeq C_1$, there must exist morphisms $(g_0, h_0) : C_0 \to C_1$ and $(g_1, h_1) : C_1 \to C_0$ such that $(g_1 \circ g_0, \mathrm{id}_{E_0}) : C_0 \to C_0$ and $(g_0 \circ g_1, \mathrm{id}_{E_1}) : C_1 \to C_1$ are both morphisms.

Let $C_i \otimes D = (A_i \times B, \hom(C_i, D^*), \diamond_i)$. Consider the morphisms $(g_i', h_i') : C_i \otimes D \to C_{1-i} \otimes D$, where $g_i' : A_i \times B \to A_{1-i} \times B$ is given by $g_i'(a, b) = (g_i(a), b)$ and $h_i' : \hom(C_{1-i}, D^*) \to \hom(C_I, D^*)$ is given by $h_i'(g, h) = (g, h) \circ (g_i, h_i)$.

To see that these are morphisms, observe that for any $(a, b) \in A_i \times B$ and $(g, h) : C_{1-i} \to D^*$, we have

$$\begin{aligned} g_i'(a, b) \diamond_{1-i} (g, h) &= (g_i(a), b) \diamond_{1-i} (g, h) \\ &= b \star g(g_i(a)) \\ &= b \star (g \circ g_i)(a)) \\ &= (a, b) \diamond_i (g \circ g_i, h_i \circ h) \\ &= (a, b) \diamond_i h_i'(g, h) . \end{aligned}$$

Finally, we need to show that (g_0, h_0) and (g_1, h_1) compose to something homotopic to the identity in both orders. This is equivalent to saying that $(g_0' \circ g_1, \mathrm{id}_{\hom(C_1, D^*)})$ and $(g_1' \circ g_0, \mathrm{id}_{\hom(C_0, D^*)})$ are both morphisms. Indeed, for all $(a, b) \in A_i \times B$ and $(g, h) : C_i \to D^*$, since $(g_{1-i}' \circ g_i, \mathrm{id}_{E_i})$ is a morphism, we have

$$\begin{aligned} g_{1-i}'(g_i'(a, b)) \diamond_i (g, h) &= (g_{1-i}'(g_i'(a)), b) \diamond_i (g, h) \\ &= g_{1-i}'(g_i'(a)) \cdot_i h(b) \\ &= a \cdot_i h(b) \\ &= (a, b) \diamond_i (g, h). \end{aligned}$$

□

2.3. Distributivity

Claim: $\otimes$ distributes over $\oplus$, so for all Cartesian frames C_0, C_1, and D, $(C_0 \oplus C_1) \otimes D \cong (C_0 \otimes D) \oplus (C_1 \otimes D)$.

Proof: Since $\oplus$ is the categorical coproduct, there exist morphisms $\iota_0 : C_0 \to C_0 \oplus C_1$ and $\iota_1 : C_1 \to C_0 \oplus C_1$ such that for any morphisms $\phi_0 : C_0 \to D^*$ and $\phi_1 : C_1 \to D^*$, there exists a unique morphism $\phi : C_0 \otimes C_1 \to D^*$ such that $\phi_i = \phi \circ \iota_i$.

Let $C_i = (A_i, E_i, \cdot_i)$, and let $D = (B, F, \star)$. Consider the isomorphism $(g, h) : (C_0 \otimes D) \oplus (C_1 \otimes D) \to (C_0 \oplus C_1) \otimes D$, where $g : (A_0 \times B) \sqcup (A_1 \times B) \to (A_0 \sqcup A_1) \times B$ is the natural bijection that sends (a, b) to (a, b), and $h : \hom(C_0 \oplus C_1, D^*) \to \hom(C_0, D^*) \times \hom(C_1, D^*)$ is given by $h(\phi) = (\phi \circ \iota_0, \phi \circ \iota_1)$.

Clearly, g is an bijection. h is also a bijection, since it is inverse to the function that sends (ϕ_0, ϕ_1) to the unique ϕ as above. Thus, all that remains to show is that (g, h) is a morphism.

Let $\diamond = \text{Eval}((C_0 \otimes D) \oplus (C_1 \otimes D))$ and let $\bullet = \text{Eval}((C_0 \oplus C_1) \otimes D)$. Given $(a, b) \in (A_0 \times B) \sqcup (A_1 \times B)$ and $(g', h') \in \hom(C_0 \oplus C_1, D^*)$, without loss of generality, assume that $a \in A_0$. Let $(g_0', h_0') = (g', h') \circ \iota_0$. Observe that since the function on agents in ι_0 is the inclusion of A_0 into $A_0 \sqcup A_1$, we have that g_0' is g' restricted to A_0. Thus, we have

$$\begin{aligned} g(a, b) \bullet (g', h') &= (a, b) \bullet (g', h') \\ &= b \star g'(a) \\ &= b \star g_0'(a) \\ &= (a, b) \diamond (g_0', h_0') \\ &= (a, b) \diamond h(g', h'). \end{aligned}$$

□

2.4. Tensor is for Disjoint Agents

It doesn't really make sense to talk about $C \otimes D$ when C and D's agents are the same agent, or otherwise overlap. This is because $C \otimes D$'s agent can make choices for both C and D, and if C and D overlap, $C \otimes D$'s agent could make choices for the intersection in two contradictory ways.

If you try to take the tensor of two frames whose agents overlap, you get a frame with an agent but no possible worlds.

Claim: If $\text{Ensure}(C) \cap \text{Prevent}(D)$ is nonempty, then $C \otimes D \simeq \top$.

Proof: Let $C = (A, E, \cdot)$, and let $D = (B, F, \star)$. Consider some $S \in \text{Ensure}(C) \cap \text{Prevent}(D)$. There is some $a \in A$ such that $a \cdot e \in S$ for all $e \in E$, and some $b \in B$ such that $b \star f \notin S$ for all $f \in F$. First, observe that $\text{Agent}(C \otimes D)$ is nonempty, since it contains (a, b). Next, observe that $\text{Env}(C \otimes D)$ is empty, since if there were a morphism $(g, h) : C \to D^*$, it would need to satisfy $b \star g(a) = a \cdot h(b)$, which is impossible since the

left hand side is not in S, while the right hand side is in S. Thus, C ⊗ D has empty environment and nonempty agent, so C ⊗ D ≃ ⊤. □

Tensoring an agent with itself lets you play "both" agents, which has the neat consequence that if the agent has any control, you can have the agent make two different choices that put you in two different possible worlds, which is a contradiction. The result is that the agent has no possible worlds.

Corollary: If Ctrl(C) is nonempty, then C ⊗ C ≃ ⊤.

Proof: Trivial. □

3. Tensor is Relative to a Coarse World Model

Recall that for any function $p : W \to V$, the functor $p^\circ : \mathrm{Chu}(W) \to \mathrm{Chu}(V)$ preserves sums and products, meaning that for any Cartesian frames C and D over W, $p^\circ(C \oplus D) = p^\circ(C) \oplus p^\circ(D)$ and $p^\circ(C\&D) = p^\circ(C)\&p^\circ(D)$. However, the same is not true for ⊗. To see this, let's go back to the voting example above.

Let's assume that Jack, Kate, and Luke have a party if and only if a majority vote in favor, and let $V = \{Y, N\}$ be the two-element world that only tracks whether or not they have a party. Let $p : W \to V$ be the function such that $p(\varepsilon) = p(J) = p(K) = p(L) = N$ and $p(JK) = p(JL) = p(KL) = p(JKL) = Y$. Then,

$$p^\circ(C_J) \cong p^\circ(C_K) \cong \begin{pmatrix} N & Y & Y & Y \\ N & N & N & Y \end{pmatrix} \simeq \begin{pmatrix} N & Y & Y \\ N & N & Y \end{pmatrix},$$

and

$$p^\circ(C_J \otimes C_K) \cong p^\circ(C_L^*) \cong \begin{pmatrix} Y & Y \\ N & Y \\ N & Y \\ N & N \end{pmatrix} \simeq \begin{pmatrix} Y & Y \\ N & Y \\ N & N \end{pmatrix},$$

but

$$\begin{matrix} N & Y & Y & Y \\ (N & N & N & Y) \end{matrix} \otimes \begin{matrix} N & Y & Y & Y \\ (N & N & N & Y) \end{matrix} \not\simeq \begin{pmatrix} Y & Y \\ N & Y \\ N & Y \\ N & N \end{pmatrix}.$$

We can see that $p^\circ(C_J \otimes C_K)$ is not equivalent to $p^\circ(C_J) \otimes p^\circ(C_K)$ by observing that the latter has a constant N environment while the former doesn't.

Let $p^\circ(C_J) \cong p^\circ(C_K) \cong (A, E, \cdot)$, and let $e_N \in E$ denote the environment such that $a \cdot e_N = N$ for both $a \in A$. (In the matrix representation above, this is the first column.) Observe that there exists a morphism $(g, h) : (A, E, \cdot) \to (A, E, \cdot)^*$, where g and h are both the constant e_N function. This is a morphism because for all $a_0, a_1 \in A$, $a_0 \cdot h(a_1) = a_1 \cdot g(a_0) = N$. This gives an environment in $p^\circ(C_J) \otimes p^\circ(C_K)$, all of whose entries must be N. $p^\circ(C_J \otimes C_K)$ has no such environment, so $p^\circ(C_J \otimes C_K)$ cannot be isomorphic to $p^\circ(C_J) \otimes p^\circ(C_K)$, or even biextensionally equivalent. Indeed:

$$p^\circ(C_J) \otimes p^\circ(C_K) \simeq \begin{pmatrix} N & Y & Y & Y & Y & Y \\ N & N & N & Y & Y & Y \\ N & N & Y & N & Y & Y \\ N & N & N & N & N & Y \end{pmatrix}.$$

To see what is going on here, consider another example where Jack and Kate and Luke vote on whether to have a party, but whether or not the party happens is not just a function of the majority's vote. Instead, after the three people cast their votes, a coin is flipped:

- If heads, the votes are tallied and majority wins as normal.
- If tails, one of the three voters is selected at random to be dictator, and the party happens if and only if they voted in favor.

Let us work up to biextensional collapse. Let D_J be the Cartesian frame over V representing Jack's perspective. We have

$$D_J \simeq \begin{pmatrix} N & Y & Y \\ N & N & Y \end{pmatrix},$$

where the top row represents voting for the party, and the bottom row represents voting against.

The first column represents environments where the party does not happen *and* Jack's vote didn't matter—either the coin came up heads and the others both voted against, or Kate or Luke became dictator and voted against. The third column similarly represents outcomes where the party happens regardless of how Jack votes. The second column represents all environments in which Jack's vote matters, so either he is dictator, or Kate and Luke's votes were split.

Similarly, let D_K be the Cartesian frame over V representing Kate's perspective,

$$D_K \simeq \begin{pmatrix} N & Y & Y \\ N & N & Y \end{pmatrix}.$$

Then,

$$D_J \otimes D_K \simeq \begin{pmatrix} N & Y & Y & Y & Y & Y \\ N & N & N & Y & Y & Y \\ N & N & Y & N & Y & Y \\ N & N & N & N & N & Y \end{pmatrix}.$$

The rows represent, in order: both voting in favor; Jack voting in favor but Kate voting against; Kate voting in favor but Jack voting against; and both voting against.

The columns represent, in order: Luke is dictator and votes against; majority rules and Luke votes against; Kate is dictator; Jack is dictator; majority rules and Luke votes in favor; and Luke is dictator and votes in favor.

Here, $D_J \otimes D_K$ looks more like what we would expect Jack and Kate working together on a team to look like. However, up to biextensional equivalence, D_J and D_K are the same as $p^\circ(C_J)$ and $p^\circ(C_K)$.

When we forget the actual votes and only look at whether the party happens, then up to biextensional collapse, the Cartesian frame representing Jack's perspective no longer has any way to distinguish between the simple majority rule vote and the complicated voting system with coins and dictators.

In general, just looking at two Cartesian frames does not tell you all of the information about the relationships between the people we might be using the frames to model.

The Cartesian frames over V representing Jack and Kate's perspectives do not have any information that distinguishes between the two vote counting schemes.

When taking a tensor, we automatically include all of the possible ways the two agents can embed in each other's environments, even if a given embedding doesn't make sense in a given interpretation.

4. Par

Our next multiplicative operation is $⅋$, which is pronounced "par."

Definition: Let $C = (A, E, \cdot)$ and $D = (B, F, \star)$ be Cartesian frames over W.

$C⅋D = (\hom(C^*, D), E \times F, \diamond)$, where $(g, h) \diamond (e, f) = g(e) \star f = h(f) \cdot e$.

Claim: $⅋$ is De Morgan dual to $\otimes$, so $C⅋D = (C^* \otimes D^*)^*$.

Proof: Trivial. □

$⅋$ has much less of an intuitive interpretation than $\otimes$. One reason for this is that in order to par two agents together, they have to be large enough that each other's environments embed within them. If C and D are not large enough, we will have that $C⅋D \simeq 0$. (I am being informal with the word "large" here.)

One way that C and D can fail to be large enough is if $\text{Ensure}(C^*) \cap \text{Prevent}(D^*)$ is nonempty, which is dual to the above result about tensor being for disjoint agents. It is actually pretty difficult for C and D to be large enough. If there is any fact about the world that is determined outside of both agents, $C⅋D$ will be trivial.

We had a dual restriction for $\otimes$, but it didn't get in the way nearly as often: simple intuitive examples tend to be about small agents interacting with a large environment, so it is easy to imagine two agents that are disjoint. It is much harder to imagine simple examples of two agents that cover, which (informally) is what you would have to have for $⅋$ to be nontrivial.

I expect to not use $⅋$ very often, but I am including it here for completeness.

Claim: ⅋ is commutative and associative, and $\bot$ is the identity of ⅋ (up to isomorphism).

Proof: Trivial from the fact that ⅋ is De Morgan dual to $\otimes$ and $1^* \cong \bot$. □

Claim: If $C_0 \simeq C_1$ and $D_0 \simeq D_1$, then $C_0 ⅋ D_0 \simeq C_1 ⅋ D_1$.

Proof: Trivial from the fact that ⅋ is De Morgan dual to $\otimes$, and $\simeq$ is preserved by $-^*$. □

Claim: ⅋ distributes over &, so for all Cartesian frames C_0, C_1, and D, we have $(C_0 \& C_1) ⅋ D \cong (C_0 ⅋ D) \& (C_1 ⅋ D)$.

Proof: Trivial from the fact that ⅋ is De Morgan dual to $\otimes$, and & is De Morgan dual to $\oplus$. □

5. Lollipop

We have one more operation to introduce, $\multimap$ (pronounced "lollipop"), which is a Cartesian frame that can be thought of as representing the collection of morphisms between two Cartesian frames.

Definition: Given two Cartesian frames over W, $C = (A, E, \cdot)$ and $D = (B, F, \star)$, we let $C \multimap D$ denote the Cartesian frame $C \multimap D = (\hom(C, D), A \times F, \diamond)$, where $\diamond$ is given by $(g, h) \diamond (a, f) = g(a) \star f = a \cdot h(f)$.

One way to interpret $C \multimap D$ is as "D with a C-shaped hole in it." Indeed, let us think about Agent($C \multimap D$). and Env($C \multimap D$) separately.

Agent($C \multimap D$) $= \hom(C, D)$ is the collection of morphisms from C to D. Morphisms from C to D are exactly interfaces through which the agent of C can interact with the environment of D. We can also think of this as the collection of interfaces that allow the agent of C to fill the role of the agent of D. This makes sense. The collection of

ways that a "D with a C-shaped hole in it" can be is exactly the collection of interfaces that allow us to get a possible agent of D from a possible agent of C.

Similarly, $Env(C \multimap D) = A \times F$ makes sense as the environment of a "D with a C-shaped hole in it." The environment needs to supply an environment for D, and also fill in the hole with an agent for C.

Previously, C's agent might have been part of D's agent; in $C \multimap D$, however, this part of D gets moved into the environment.

Imagine a football team D with one team member, C, removed—the team with a football-player-shaped hole in it. Its environment, naturally, is pairs of "the kind of environment you get for a football team" and "the removed teammate".

Lollipop can be easily constructed from our other operations.

Claim: $C \multimap D \cong C^* \parr D \cong (C \otimes D^*)^*$.

Proof: Trivial. □

Lollipop is well-defined up to biextensional equivalence.

Claim: If $C_0 \simeq C_1$ and $D_0 \simeq D_1$, then $C_0 \multimap D_0 \simeq C_1 \multimap D_1$.

Proof: Trivial. □

Lollipop also has some identity-like properties.

Claim: For all Cartesian Frames C, $C \cong 1 \multimap C$ and $C^* \cong C \multimap \bot$.

Proof: $1 \multimap C \cong (1 \otimes C^*)^* \cong C^{**} \cong C$ and $C \multimap \bot \cong (C \otimes 1)^* \cong C^*$. □

This last result is especially interesting because we can actually think of $C \multimap \bot$ as an alternative definition for C^*.

In "Tensor is Relative to a Coarse World Model" above, we noted that two agents working together might sometimes have strictly fewer possible environments than show up in the tensor. In the next post, we will introduce the concept of a *sub-tensor*,

which allows us to represent teams that have fewer possible environments than the tensor. Similarly, *sub-sum* will be sum with spurious possible environments removed.

Sub-Sums and Sub-Tensors

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the eighth post in the Cartesian frames sequence. Here, we define new versions of the sum and tensor of Cartesian frames that can delete spurious possible environments from the frame.

1. Motivating Examples

The sum $C \oplus D$ of C and D is supposed to represent an agent that can do anything either C or D can do, while the tensor, $C \otimes D$, is supposed to represent an agent that can do anything C and D can do working freely together on a team. However, sometimes these operations produce Cartesian frames with more environments than we would expect.

Consider two players, Alice and Bob, in a prisoner's dilemma. We will let $W = \{0, 1, 2, 3\}$ be the space of utilities for Alice. The Cartesian frame for Alice looks like

$$C_0 = \begin{pmatrix} 2 & 0 \\ 3 & 1 \end{pmatrix},$$

where the top row represents Alice cooperating, and the left column represents Bob cooperating.

Let $C_1 = \begin{pmatrix} 2 & 0 \end{pmatrix}$ represent Alice committed to cooperating, and let $C_2 = \begin{pmatrix} 3 & 1 \end{pmatrix}$ represent Alice committed to defecting. Since the real Alice can either cooperate or defect, one might expect that Alice (C_0) would equal the sum ($C_1 \oplus C_2$) of Alice cooperating with Alice defecting. However,

$$C_1 \oplus C_2 = \begin{pmatrix} 2 & 0 & 2 & 0 \\ 3 & 1 & 1 & 3 \end{pmatrix}.$$

The last two columns are spurious environments that represent Bob copying Alice's move, and Bob doing the opposite of Alice's move. However, since Bob cannot see Alice's move, Bob should not be able to implement these policies: Bob can only choose to cooperate or defect.

Next, consider a unilateralist's curse game where two players each have access to a button that destroys the Earth. If either player pushes the button, the Earth is destroyed. Otherwise, the Earth is not destroyed. $W = \{0, 1\}$, where 0 represents the world being destroyed and 1 represents the world not being destroyed.

Here, both players have the Cartesian frame

$$D_1 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

where the first row represents pressing the button, and the first column represents the other player pressing the button.

The two players together can be expressed with the Cartesian frame

$$D_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

where the rows in order represent: both players pressing the button; the first player pressing the button; the second player pressing the button; and neither player pressing the button.

One might expect that $D_1 \otimes D_1$ would be D_2, but in fact,

$$D_1 \otimes D_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

The second possible environment, in which the earth is just destroyed regardless of what the players do, is spurious.

In both of the above examples, the spurious environments are only spurious because of our interpretation. In the prisoner's dilemma case, $C_1 \oplus C_2$ would be correct if Alice and Bob were playing a modified dilemma where Bob can see Alice's choice. In the unilateralist's curse example, $D_1 \otimes D_1$ would be correct if there were three people playing the game. The problem is that the $\oplus$ and $\otimes$ operations do not see our interpretation, and so include all possible environments.

2. Deleting Spurious Environments

We introduce two new concepts, called *sub-sum* and *sub-tensor*, which represents sum and tensor with some (but not too many) spurious environments removed.

Definition: Let $C = (A, E, \cdot)$, and let $D = (B, F, \star)$. A sub-sum of C and D is a Cartesian frame of the form $(A \sqcup B, X, \diamond)$, where $X \subseteq \text{Env}(C \oplus D)$ and $\diamond$ is Eval$(C \oplus D)$ restricted to $(A \sqcup B) \times X$, such that $C \simeq (A, X, \diamond_C)$ and $D \simeq (B, X, \diamond_D)$, where $\diamond_C$ is $\diamond$ restricted to $A \times X$ and $\diamond_D$ is $\diamond$ restricted to $B \times X$. Let $C \boxplus D$ denote the set of all sub-sums of C and D.

Definition: Let $C = (A, E, \cdot)$, and let $D = (B, F, \star)$. A sub-tensor of C and D is a Cartesian frame of the form $(A \times B, X, \bullet)$, where $X \subseteq \text{Env}(C \otimes D)$ and $\bullet$ is Eval$(C \otimes D)$ restricted to $(A \times B) \times X$, such that $C \simeq (A, B \times X, \bullet_C)$ and $D \simeq (B, A \times X, \bullet_D)$, where $\bullet_C$ and $\bullet_D$ are given by $a \bullet_C (b, x) = (a, b) \bullet x$ and $b \bullet_D (a, x) = (a, b) \bullet x$. Let $C \boxtimes D$ denote the set of all sub-tensors of C and D.

Thus, we define $C \boxplus D$ and $C \boxtimes D$ to be sets of Cartesian frames that can be obtained by deleting columns from from $C \oplus D$ and $C \otimes D$, respectively, but we have an extra restriction to ensure that we do not delete too many columns.

We will discuss later how to interpret the extra restriction, but first let us go back to our above examples.

If $C_1 = \begin{pmatrix} 2 & 0 \end{pmatrix}$ and $C_2 = \begin{pmatrix} 3 & 1 \end{pmatrix}$, then $C_1 \boxplus C_2$ has 7 elements:

$$\begin{pmatrix} 2 & 0 & 2 & 0 \\ 3 & 1 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 2 & 0 & 2 \\ 3 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 0 & 0 \\ 3 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 2 & 2 & 0 \\ 3 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 0 & 2 & 0 \\ 1 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 3 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 1 & 3 \end{pmatrix}$$

The 9 Cartesian frames that can be obtained by deleting columns from $C \oplus D$ that are not in $C_1 \boxplus C_2$ are:

$$\begin{pmatrix} 2 & 2 \\ 3 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 3 & 3 \end{pmatrix}, \begin{pmatrix} 0 & 2 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \end{pmatrix}, ( ).$$

The Cartesian frames above in $C_1 \boxplus C_2$ are exactly those with all four entries, 0, 1, 2, and 3. This is because the extra restriction to be in $C_1 \boxplus C_2$ is exactly that if you delete the bottom row, you get an object biextensionally equivalent to $\begin{pmatrix} 2 & 0 \end{pmatrix}$, and if you delete the top row, you get an object biextensionally equivalent to $\begin{pmatrix} 3 & 1 \end{pmatrix}$.

Similarly, from the unilateralist's curse example,

$$D_1 \boxtimes D_1 = \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{pmatrix} \right\}.$$

It is easy to see that there are no other Cartesian frames in $D_1 \boxtimes D_1$, since there are only four subsets of the two element environment of $D_1 \otimes D_1$, and the Cartesian frames corresponding to the other two subsets do not have 1 in their image, so we cannot build anything biextensionally equivalent to D_1 out of them.

Conversely, let $C = (A, E, \cdot)$ and $D = (B, F, \star)$ both be D_1, and notice that if

$$(A \times B, X, \bullet) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

then $(A, B \times X, \bullet_C)$ and $(B, A \times X, \bullet_D)$ are both two-by-two matrices with a single 1 entry and three 0 entries, and so must be isomorphic to D_1. Similarly, if

$$(A \times B, X, \bullet) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{pmatrix},$$

then $(A, B \times X, \bullet_C)$ and $(B, A \times X, \bullet_D)$ are both two-by-four matrices with a single 1 entry and seven 0 entries, and so must by biextensionally equivalent to D_1.

3. Properties of Sub-Sums and Sub-Tensors

3.1. Sub-Sums and Sub-Tensors Are Commutative

Claim: For any Cartesian frames C_0 and C_1, there is a bijection between $C_0 \boxplus C_1$ and $C_1 \boxplus C_0$ that preserves Cartesian frames up to isomorphism. Similarly, there is a bijection between $C_0 \boxtimes C_1$ and $C_1 \boxtimes C_0$ that preserves Cartesian frames up to isomorphism.

Proof: Trivial. □

3.2. Tensors Need Not Be Sub-Tensors

For any Cartesian frames C and D with nonempty environments, we have that $C \oplus D \in C \boxplus D$. However, sometimes $C \otimes D \notin C \boxtimes D$. Indeed, sometimes $C \boxtimes D = \{\}$.

For example, if C and D both have nonempty image, but there are no morphisms from C to D^*, then $C \otimes D$ has no environments, and it is easy to see that $C \boxtimes D$ must be empty.

3.3. Sub-Sums and Sub-Tensors Are Superagents

Claim: For any Cartesian frames C_0 and C_1, and for any $D_0 \in C_0 \boxplus C_1$, we have $C_0 \triangleleft D_0$ and $C_1 \triangleleft D_0$. Similarly, for any $D_1 \in C_0 \boxtimes C_1$, we have $C_0 \triangleleft D_1$ and $C_1 \triangleleft D_1$.

Proof: Let $C_i = (A_i, E_i, \cdot_i)$ and $D_i = (B_i, F_i, \star_i)$. First, we show $C_0 \triangleleft D_0$ using the currying definition of subagent. Observe $B_0 = A_0 \sqcup A_1$. Consider the Cartesian frame $(A_0, \{e\}, \diamond)$ over B_0, where $\diamond$ is given by $a \diamond e = a$. Observe that the definition of sub-sum says that $C_0 \simeq D_0^{\circ}(A_0, \{e\}, \diamond)$, so $C_0 \triangleleft D_0$. Therefore, $C_0 \triangleleft D_0$, and by commutativity, we also have $C_1 \triangleleft D_0$.

Similarly, we show $C_0 \triangleleft D_1$ using the currying definition of subagent. Observe that $B_1 = A_0 \times A_1$. Consider the Cartesian frame $(A_0, A_1, \bullet)$ over B_1, where $\bullet$ is given by $a_0 \bullet a_1 = (a_0, a_1)$. Observe that the definition of sub-tensor says that $C_0 \simeq D_1^{\circ}(A_0, A_1, \bullet)$. Therefore, $C_0 \triangleleft D_1$, and by commutativity, we also have $C_1 \triangleleft D_1$. □

Observe that in the above proof, the Cartesian frame over B_0 we constructed to show that sub-sums are superagents had a singleton environment, and the Cartesian frame over B_1 we constructed to show that sub-tensors are superagents had a surjective evaluation function. The relevance of this observation will become clear later.

3.4. Biextensional Equivalence

As we do with most of our definitions, we will now show that sub-sums and sub-tensors are well-defined up to biextensional equivalence.

Claim: Given two Cartesian frames over W, C_0 and C_1, and given any $D \in C_0 \boxplus C_1$, we have that for all $C_0' \simeq C_0$ and $C_1' \simeq C_1$, there exists a $D' \simeq D$, with $D' \in C_0' \boxplus C_1'$.

Proof: Let $C_0 = (A_0, E_0, \cdot_0)$, let $C_1 = (A_1, E_1, \cdot_1)$, and let $D = (A_0 \sqcup A_1, X, \star)$ be an element of $C_0 \boxplus C_1$, so $X \subseteq E_0 \times E_1$, and $a \star (e_0, e_1) = a \cdot_0 e_0$ if $a \in A_0$, and $a \star (e_0, e_1) = a \cdot_1 e_1$ if $a \in A_1$.

The fact that $D \in C_0 \boxplus C_1$ tells us that for $i \in \{0, 1\}$, $D_i \simeq C_i$, where $D_i = (A_i, X, \star_i)$ with $\star_i$ given by $a \star_0 (e_0, e_1) = a \cdot_0 e_0$ and $a \star_1 (e_0, e_1) = a \cdot_0 e_1$.

For $i \in \{0, 1\}$, let $C_i' = (A_i, E_i', \cdot_i')$ satisfy $C_i' \simeq C_i$. Thus, there exist morphisms $(g_i, h_i) : C_i \to C_i'$, and $(g_i', h_i') : C_i' \to C_i$, such that $(g_0, h_0) \circ (g_0', h_0')$, $(g_0', h_0') \circ (g_0, h_0)$, $(g_1, h_1) \circ (g_1', h_1')$, and $(g_1', h_1') \circ (g_1, h_1)$ are all homotopic to the identity.

We define a function $f : E_0 \times E_1 \to E_0' \times E_1'$ by $f(e_0, e_1) = (h_0(e_0), h_1(e_1))$. Then, we define $X' \subseteq E_0' \times E_1'$ by $X' = \{f(x) \mid x \in X\}$, and let $D' = (A_0' \sqcup A_1', X', \star')$, where $a \star' (e_0, e_1) = a \cdot_0' e_0$ if $a \in A_0'$, and $a \star' (e_0, e_1) = a \cdot_1' e_1$ if $a \in A_1'$. We need to show $D' \in C_0 \boxplus C_1$, and that $D' \simeq D$.

To show that $D' \simeq D$, we will construct a pair of morphisms $(j, k) : D \to D'$ and $(j', k') : D' \to D$ that compose to something homotopic to the identity in both orders. We define $j : A_0 \sqcup A_1 \to A_0' \sqcup A_1'$ by $j(a) = g_0(a)$ if $a \in A_0$, and $j(a) = g_1(a)$ if $a \in A_1$. We similarly define $j' : A_0' \sqcup A_1' \to A_0 \sqcup A_1$ by $j'(a) = g_0'(a)$ if $a \in A_0'$, and $j'(a) = g_1'(a)$ if $a \in A_1'$. We define $k' : X \to X'$ by $k'(x) = f(x)$, which is clearly a function into X', by the definition of X'. Further, k' is surjective, and thus has a right inverse. We choose $k : X' \to X$ to be any right inverse to k', so $f(k(x)) = x$ for all $x \in X'$.

To see that (j', k') is a morphism, observe that for $a \in A_0' \sqcup A_1'$, and $(e_0, e_1) \in X$, if $a \in A_i'$, then

$$\begin{aligned} j'(a) \star (e_0, e_1) &= g_i'(a) \cdot_i e_i \\ &= a \cdot_i' h_i(e_i) \\ &= a \star' (h_0(e_0), h_1(e_1)) \\ &= a \star' k'(e_0, e_1)). \end{aligned}$$

To see that (j, k) is a morphism, consider an arbitrary $a \in A_0 \sqcup A_1$ and $(e_0', e_1') \in X'$, and let $(e_0, e_1) = k(e_0', e_1')$. Then, if $a \in A_i$, we have

$$\begin{aligned} j(a) \star' (e_0', e_1') &= j(a) \star' f(e_0, e_1) \\ &= j(a) \star' (h_0(e_0), h_1(e_1)) \\ &= g_i(a) \cdot_i' h_i(e_i) \\ &= g_i'(g_i(a)) \cdot_i e_i \\ &= a \cdot_i e_i \\ &= a \star (e_0, e_1) \\ &= a \star k(e_0', e_1'). \end{aligned}$$

To see that $(j', k') \circ (j, k)$ is homotopic to the identity on D, observe that for all $a \in A_0 \sqcup A_1$ and $(e_0, e_1) \in X$, we have that if $a \in A_i$,

$$\begin{aligned} j'(j(a)) \star (e_0, e_1) &= g_i'(g_i(a)) \cdot_i e_i \\ &= a \cdot_i e_i \\ &= a \star (e_0, e_1). \end{aligned}$$

Similarly, to see that $(j, k) \circ (j', k')$ is homotopic to the identity on D', observe that for all $a \in A_0 \sqcup A_1$ and $(e_0, e_1) \in X'$, we have that if $a \in A_i$,

$$\begin{aligned} j(j'(a)) \star' (e_0, e_1) &= g_i(g_i'(a)) \cdot_i' e_i \\ &= a \cdot_i' e_i \\ &= a \star' (e_0, e_1). \end{aligned}$$

Thus, $D' \simeq D$.

To see $D' \in C_0 \boxplus C_1$, we need to show that $D_i \simeq C_i$, where $D_i = (A_i, X', \star_i')$ with $\star_i'$ given by $a \star_0' (e_0, e_1) = a \cdot_0' e_0$ and $a \star_1' (e_0, e_1) = a \cdot_1' e_1$. It suffices to show that $D_i \simeq D_i$, since $D_i \simeq C_i \simeq C_i$.

For $i \in \{0, 1\}$, we construct morphisms $(j_i, k_i) : D_i \to D_i$, and $(j_i, k_i) : D_i \to D_i$. We define $j_i = g_i$, $j_i' = g_i'$, $k_i = k$, and $k_i' = k'$.

To see that (j_i, k_i) is a morphism, observe that for all $a \in A_i$ and $x \in X'$, we have

$$\begin{aligned} j_i(a) \star_i' x &= g_i(a) \star_i' x \\ &= j(a) \star' x \\ &= a \star k(x) \\ &= a \star_i k_i(x), \end{aligned}$$

and to see that (j_i', k_i') is a morphism, observe that for all $a \in A_i$, and $x \in X$, we have

$$\begin{aligned} j_i'(a) \star_i' x &= g_i'(a) \star_i' x \\ &= j'(a) \star x \\ &= a \star' k'(x) \\ &= a \star_i' k_i'(x). \end{aligned}$$

To see $(j_i', k_i') \circ (j_i, k_i)$ is homotopic to the identity on D_i, observe that for all $a \in A_i$ and $x \in X$, we have

$$\begin{aligned} j_i'(j_i(a)) \star_i' x &= g_i'(g_i(a)) \star_i' x \\ &= j'(j(a)) \star x \\ &= a \star x \\ &= a \star_i x, \end{aligned}$$

and similarly, to show that $(j_i, k_i) \circ (j_i', k_i')$ is homotopic to the identity on D_i', observe that for all $a \in A_i'$ and $x \in X'$, we have

$$\begin{aligned} j_i'(j_i(a)) \star_i' x &= g_i'(g_i(a)) \star_i' x \\ &= j(j'(a)) \star' x \\ &= a \star' x \\ &= a \star_i' x. \end{aligned}$$

Thus, we have that $D_i' \simeq D_i$, completing the proof. □

We have a similar result for sub-tensors, whose proof directly mirrors the proof for sub-sums:

Claim: Given two Cartesian frames over W, C_0 and C_1, and any $D \in C_0 \boxtimes C_1$, we have that for all $C_0' \simeq C_0$ and $C_1' \simeq C_1$, there exists a $D' \simeq D$, with $D' \in C_0' \boxtimes C_1'$.

Proof: Let $C_0 = (A_0, E_0, \cdot_0)$, let $C_1 = (A_1, E_1, \cdot_1)$, and let $D = (A_0 \times A_1, X, \star)$ be an element of $C_0 \boxtimes C_1$, so $X \subseteq \hom(C_0, C_1^*)$, and

$$\begin{aligned} (a, b) \star (g, h) &= a \cdot_0 h(b) \\ &= b \cdot_1 g(a). \end{aligned}$$

The fact that $D \in C_0 \boxtimes C_1$ tells us that for $i \in \{0, 1\}$, $D_i \simeq C_i$, where $D_i = (A_i, A_{1-i} \times X, \star_i)$ with $\star_i$ given by

$$\begin{aligned} a \star_0 (b, (g, h)) &= b \star_1 (a, (g, h)) \\ &= (a, b) \star (g, h). \end{aligned}$$

For $i \in \{0,1\}$, let $C_i' = (A_i, E_i, \cdot_i')$ satisfy $C_i' \simeq C_i$. Thus, there exist morphisms $(g_i, h_i) : C_i \to C_i'$, and $(g_i', h_i') : C_i' \to C_i$, such that $(g_0, h_0) \circ (g_0', h_0')$, $(g_0', h_0') \circ (g_0, h_0)$, $(g_1, h_1) \circ (g_1', h_1')$, and $(g_1', h_1') \circ (g_1, h_1)$ are all homotopic to the identity.

We define a function $f : \hom(C_0, {C_1}^*) \to \hom(C_0', {C_1'}^*)$ by $f(g,h) = (h_1, g_1) \circ (g,h) \circ (g_0', h_0')$. This function is well-defined, since $(h_1, g_1) = (g_1, h_1)^* \in \hom({C_1}^*, {C_1'}^*)$ and $(h_0', g_0') \in \hom(C_0', C_0)$.

Then, we define $X' \subseteq \hom(C_0', {C_1'}^*)$ by $X' = \{f(g,h) \mid (g,h) \in X\}$, and let $D' = (A_0 \times A_1, X', \star')$, where

$$\begin{aligned}(a, b) \star' (g, h) &= a \cdot_0' h(b) \\ &= b \cdot_1' g(a).\end{aligned}$$

We need to show that $D' \in C_0' \boxtimes C_1'$, and that $D' \simeq D$.

To show that $D' \simeq D$, we will construct a pair of morphisms $(j,k) : D \to D'$ and $(j',k') : D' \to D$ that compose to something homotopic to the identity in both orders. We define $j : A_0 \times A_1 \to A_0' \times A_1'$ by $j(a,b) = (g_0(a), g_1(b))$, and we similarly define $j' : A_0' \times A_1' \to A_0 \times A_1$ by $j'(a,b) = (g_0'(a), g_1'(b))$. We define $k' : X \to X'$ by $k'(x) = f(x)$, which is clearly a function into X', by the definition of X'. Further, k' is surjective, and thus has a right inverse. We choose $k : X' \to X$ to be any right inverse to k', so $f(k(x)) = x$ for all $x \in X'$.

To see (j', k') is a morphism, observe that for $(a,b) \in A_0' \times A_1'$, and $(g,h) \in X$, we have

$$\begin{aligned}j'(a, b) \star (g, h) &= (g_0'(a), g_1'(b)) \star (g, h) \\ &= g_1'(b) \cdot_1 g(g_0'(a)) \\ &= b \cdot_1' h_1(g(g_0'(a))) \\ &= (a, b) \star' (h_1 \circ g \circ g_0', h_0 \circ h \circ g_1') \\ &= (a, b) \star' k'(g, h).\end{aligned}$$

To see that (j,k) is a morphism, consider an arbitrary $(a,b) \in A_0 \times A_1$ and $(g',h') \in X'$, and let $(g,h) = k(g',h')$. Then, we have:

$$
\begin{aligned}
j(a,b) \star' (g',h') &= (g_0(a), g_1(b)) \star' f(g,h) \\
&= (g_0(a), g_1(b)) \star' (h_1', h_1') \circ (g,h) \circ (g_0', h_0') \\
&= g_1(b) \cdot_1' h_1'(g(g_0'(g_0(a)))) \\
&= g_1'(g_1(b)) \cdot_1' g(g_0'(g_0(a))) \\
&= b \cdot_1' g(g_0'(g_0(a))) \\
&= g_0'(g_0(a)) \cdot_0' h(b) \\
&= a \cdot_0' h(b) \\
&= (a,b) \star (g,h) \\
&= (a,b) \star k(g',h').
\end{aligned}
$$

To see that $(j',k') \circ (j,k)$ is homotopic to the identity on D, observe that for all $(a,b) \in A_0 \times A_1$ and $(g,h) \in X$, we have:

$$
\begin{aligned}
j'(j(a,b)) \star (g,h) &= (g_0'(g_0(a)), g_1'(g_1(b))) \star (g,h) \\
&= g_1'(g_1(b)) \cdot_1' g(g_0'(g_0(a))) \\
&= b \cdot_1' g(g_0'(g_0(a))) \\
&= g_0'(g_0(a)) \cdot_0' h(b) \\
&= a \cdot_0' h(b) = (a,b) \star (g,h).
\end{aligned}
$$

Similarly, to see that $(j,k) \circ (j',k')$ is homotopic to the identity on D', observe that for all $(a,b) \in A_0' \times A_1'$ and $(g,h) \in X$, we have:

$$
\begin{aligned}
j(j'(a,b)) \star' (g,h) &= (g_0(g_0'(a)), g_1(g_1'(b))) \star' (g,h) \\
&= g_1(g_1'(b)) \cdot_1' g(g_0(g_0'(a))) \\
&= b \cdot_1' g(g_0(g_0'(a))) \\
&= g_0(g_0'(a)) \cdot_0' h(b) \\
&= a \cdot_0' h(b) \\
&= (a,b) \star' (g,h).
\end{aligned}
$$

Thus, $D' \simeq D$.

To see $D' \in C_0 \boxtimes C_1$, we need to show that $D_i \simeq C_i$, where $D_i = (A_i, A_{1-i} \times X', \star_i)$ with $\star_i$ given by

$$\begin{aligned} a \star_0 (b, (g, h)) &= b \star_1 (a, (g, h)) \\ &= (a, b) \star' (g, h). \end{aligned}$$

It suffices to show that $D_i \simeq D_i$, since $D_i \simeq C_i \simeq C_i$.

For $i \in \{0, 1\}$, we construct morphisms $(j_i, k_i) : D_i \to D_i$, and $(j_i, k_i) : D_i \to D_i$. We define $j_i = g_i$ and $j_i = g_i$.

We define $k_i : (A_{1-i} \times X') \to (A_{1-i} \times X)$ by $k_i(a, x) = (g_{1-i}(a), k(x))$, and similarly define

$k_i : (A_{1-i} \times X) \to (A_{1-i} \times X')$ by $k_i(a, x) = (g_{1-i}(a), k'(x))$.

To see that (j_0, k_0) is a morphism, observe that for all $a \in A_0$ and $(a_1, (g, h)) \in A_1 \times X'$, we have:

$$\begin{aligned} a \star_0 k_0(b, (g, h)) &= a \star_0 (g_1(b), k(g, h)) \\ &= (a, g_1(b)) \star k(g, h) \\ &= j(a, g_1(b)) \star' (g, h) \\ &= (g_0(a), g_1(g_1(b))) \star' (g, h) \\ &= g_1(g_1(b)) \cdot_1 g(g_0(a)) \\ &= b \cdot_1 g(g_0(a)) \\ &= (g_0(a), b) \star' (g, h) \\ &= j_0(a) \star_0 (b, (g, h)). \end{aligned}$$

To see that (j_1, k_1), (j_0, k_0), and (j_1, k_1) are morphisms is similar.

To see $(j_0, k_0) \circ (j_0, k_0)$ is homotopic to the identity on D_0, observe that for all $a \in A_0$ and $(b, (g, h)) \in A_1 \times X$, we have

$$
\begin{aligned}
j_i'(j_i(a)) \star_i' x &= (g_i'(g_i(a)), b) \star (g, h) \\
&= g_i'(g_i(a)) \cdot_0 h(b) \\
&= a \cdot_0 h(b) \\
&= (a, b) \star (g, h) \\
&= a \star_0 (b, (g, h)),
\end{aligned}
$$

and seeing that $(j_1', k_1') \circ (j_1, k_1)$, $(j_0, k_0) \circ (j_0', k_0')$, and $(j_1', k_1') \circ (j_1', k_1')$ are homotopic to the identity is similar.

Thus, we have that $D_i' \simeq D_i$, completing the proof. □

In our next post, we will use sub-sum and sub-tensor to define *additive subagents*, which are like agents that have committed to restrict their class of options; and *multiplicative subagents*, which are like agents that are contained inside other agents. We will also introduce the concept of *sub-environments*.

Additive and Multiplicative Subagents

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the ninth post in the Cartesian frames sequence. Here, we refine our notion of subagent into additive and multiplicative subagents. As usual, we will give many equivalent definitions.

The additive subagent relation can be thought of as representing the relationship between an agent that has made a commitment, and the same agent before making that commitment. The multiplicative subagent relation can be thought of as representing the relationship between a football player and a football team.

Another way to think about the distinction is that additive subagents have fewer options, while multiplicative subagents have less refined options.

We will introduce these concepts with a definition using sub-sums and sub-tensors.

1. Definitions of Additive and Multiplicative Subagent

1.1. Sub-Sum and Sub-Tensor Definitions

Definition: C is an additive subagent of D, written $C \triangleleft_+ D$, if there exists a C' and a $D' \simeq D$ with $D' \in C \boxplus C'$.

Definition: C is a multiplicative subagent of D, written $C \triangleleft_\times D$, if there exists a C' and $D' \simeq D$ with $D' \in C \boxtimes C'$.

These definitions are nice because they motivate the names "additive" and "multiplicative." Another benefit of these definitions is that they draw attention to the Cartesian frames given by C'. This feature is emphasized more in the below (clearly equivalent) definition.

1.2. Brother and Sister Definitions

Definition: C' is called a brother to C in D if $D \simeq D'$ for some $D' \in C \boxplus C'$. Similarly, C' is called a sister to C in D if $D \simeq D'$ for some $D' \in C \boxtimes C'$.

E.g., one "sister" of a football player will be the entire rest of the football team. One "brother" of a person that precommitted to carry an umbrella will be the

counterfactual version of themselves that instead precommitted to *not* carry an umbrella.

This allows us to trivially restate the above definitions as:

Definition: We say $C \triangleleft_+ D$ if C has a brother in D and $C \triangleleft_\times D$ if C has a sister in D.

Claim: This definition is equivalent to the ones above.

Proof: Trivial. □

1.3. Committing and Externalizing Definitions

Next, we will give the committing definition of additive subagent and an externalizing definition of multiplicative subagent. These definitions are often the easiest to work with directly in examples.

We call the following definition the "committing" definition because we are viewing C as the result of D making a commitment (up to biextensional equivalence).

Definition: Given Cartesian frames C and D over W, we say $C \triangleleft_+ D$ if there exist three sets X, Y, and Z, with $X \subseteq Y$, and a function $f : Y \times Z \to W$ such that $C \simeq (X, Z, \diamond)$ and $D \simeq (Y, Z, \bullet)$, where $\diamond$ and $\bullet$ are given by $x \diamond z = f(x, z)$ and $y \bullet z = f(y, z)$.

Claim: This definition is equivalent to the sub-sum and brother definitions of $\triangleleft_+$.

Proof: First, assume that C has a brother in D. Let $C = (A, E, \cdot)$, and let $D = (B, F, \star)$. Let $C' = (A', E', \cdot')$ be brother to C in D. Let $D' = (B', F', \star')$ be such that $D' \simeq D$ and $D' \in C \boxplus C'$. Then, if we let $X = A$, let $Y = B' = A \sqcup A'$, let $Z = F'$, and let $f(y, z) = y \star' z$, we get $D \simeq D' = (Y, Z, \bullet)$, where $y \bullet z = f(y, z)$, and by the definition of sub-sum, $C \simeq (X, Z, \diamond)$, where $x \diamond z = f(x, z)$.

Conversely, let X, Y, and Z be arbitrary sets with $X \subseteq Y$, and let $f : Y \times Z \to W$. Let $C \simeq C_0 = (X, Z, \diamond_0)$, and let $D \simeq D' = (Y, Z, \bullet)$, where $x \diamond_0 z = f(x, z)$ and $y \bullet z = f(y, z)$. We want to show that C has a brother in D. It suffices to show that C_0 has a brother in

D, since sub-sum is well-defined up to biextensional equivalence. Indeed, we will show that $C_1 = (Y \setminus X, Z, \diamond_1)$ is brother to C_0 in D, where $\diamond_1$ is given by $x \diamond_1 z = f(x, z)$.

Observe that $C_0 \oplus C_1 = (Y, Z \times Z, \bullet')$, where $\bullet'$ is given by

$$\begin{aligned} y \bullet' (z_0, z_1) &= y \diamond_0 z_0 \\ &= y \bullet z_0 \end{aligned}$$

if $y \in X$, and is given by

$$\begin{aligned} y \bullet' (z_0, z_1) &= y \diamond_1 z_1 \\ &= y \bullet z_1 \end{aligned}$$

otherwise. Consider the diagonal subset $S \subseteq Z \times Z$ given by $S = \{(z, z) \mid z \in Z\}$. Observe that the map $z \mapsto (g_z, h_z)$ is a bijection from Z to S. Observe that if we restrict $\bullet'$ to $Y \times S$, we get $\bullet'' : Y \times S \to W$ given by $y \bullet'' (z, z) = y \bullet z$. Thus $(Y, S, \bullet'') \cong (Y, Z, \bullet)$, with the isomorphism coming from the identity on Y, and the bijection between S and Z.

If we further restrict $\bullet''$ to $X \times S$ or $(Y \setminus X) \times S$, we get $\bullet_0$ and $\bullet_1$ respectively, given by $x \bullet_0 = x \diamond_0 z$ and $x \bullet_1 (z, z) = x \diamond_1 z$. Thus $(X, S, \bullet_0) \cong (X, Z, \diamond_0)$ and $(Y \setminus X, S, \bullet_1) \cong (Y \setminus X, Z, \diamond_1)$, with the isomorphisms coming from the identities on Y and $X \setminus Y$, and the bijection between S and Z.

Thus $(Y, S, \bullet'') \in C_0 \boxplus C_1$, and $(Y, S, \bullet'') \cong D' \simeq D$, so C_1 is brother to C_0 in D, so C has a brother in D. □

Next, we have the externalizing definition of multiplicative subagent. Here, we are viewing C as the result of D sending some of its decisions into the environment (up to biextensional equivalence).

Definition: Given Cartesian frames C and D over W, we say $C \triangleleft_\times D$ if there exist three sets X, Y, and Z, and a function $f : X \times Y \times Z \to W$ such that $C \simeq (X, Y \times Z, \diamond)$ and

$D \simeq (X \times Y, Z, \bullet)$, where $\diamond$ and $\bullet$ are given by $x \diamond (y,z) = f(x,y,z)$ and $(x,y) \bullet z = f(x,y,z)$.

Claim: This definition is equivalent to the sub-tensor and sister definitions of $\triangleleft_{\times}$.

Proof: First, assume that C has a sister in D. Let $C = (A, E, \cdot)$, and let $D = (B, F, \star)$. Let $C' = (A', E', \cdot')$ be sister to C in D. Let $D' = (B', F', \star')$ be such that $D' \simeq D$ and $D' \in C \boxtimes C'$. Then, if we let $X = A$, let $Y = A'$, let $Z = F' \subseteq \hom(C, C'^{*})$, and let

$$\begin{aligned} f(x, y, (g, h)) &= x \cdot h(y) \\ &= y \cdot' g(x), \end{aligned}$$

we get $D \simeq D' = (X \times Y, Z, \bullet)$, where $(x,y) \bullet z = f(x,y,z)$, and by the definition of sub-tensor, $C \simeq (X, Y \times Z, \diamond)$, where $x \diamond (y,z) = f(x,y,z)$.

Conversely, let X, Y, and Z be arbitrary sets, and let $f : X \times Y \times Z \to W$. Let $C \simeq C_0 = (X, Y \times Z, \diamond_0)$, and let $D \simeq D' = (X \times Y, Z, \bullet)$, where $x \diamond_0 (y,z) = (x,y) \bullet z = f(x,y,z)$. We will assume for now that at least one of X and Y is nonempty, as the case where both are empty is degenerate.

We want to show that C_0 has a sister in D. It suffices to show that C_0 has a sister in D, since sub-tensor is well-defined up to biextensional equivalence. Indeed, we will show that $C_1 = (Y, X \times Z, \diamond_1)$ is sister to C_0 in D, where $\diamond_1$ is given by $y \diamond_1 (x,z) = f(x,y,z)$.

Observe that $C_0 \otimes C_1 = (X \times Y, \hom(C_0, C_1^{*}), \bullet')$, where $\bullet'$ is given by

$$\begin{aligned} (x, y) \bullet' (g, h) &= x \diamond_0 h(y) \\ &= y \star_1 g(x). \end{aligned}$$

For every $z \in Z$, there is a morphism $(g_z, h_z) : C_0 \to C_1^{*}$, where $g_z : X \to X \times Z$ is given by $g_z(x) = (x,z)$, and $g_z : Y \to Y \times Z$ is given by $g_z(x) = (x,z)$. This is clearly a morphism. Consider the subset $S \subseteq \hom(C_0, C_1^{*})$ given by $S = \{(g_z, h_z) \mid z \in Z\}$.

Observe that the map $z \mapsto (g_z, h_z)$ is a bijection from Z to S. (We need that at least one of X and Y is nonempty here for injectivity.)

If we restrict $\bullet'$ to $(X \times Y) \times S$, we get $\bullet'' : (X \times Y) \times S \to W$ given by $y \bullet'' (g_z, h_z) = y \bullet z$. Thus, $(X \times Y, S, \bullet'') \cong (X \times Y, Z, \bullet)$, with the isomorphism coming from the identity on $X \times Y$, and the bijection between S and Z.

To show that $(X \times Y, S, \bullet'') \in C_0 \boxtimes C_1$, we need to show that $C_0 \simeq (X, Y \times S, \bullet_0)$ and $C_1 \simeq (Y, X \times S, \bullet_1)$, where $\bullet_0$ and $\bullet_1$ are given by

$$\begin{aligned} x \bullet_0 (y, (g_h, z_h)) &= y \bullet_1 (x, (g_h, z_h)) \\ &= (x, y) \bullet'' (g_z, h_z). \end{aligned}$$

Indeed, $x \bullet_0 (y, (g_h, z_h)) = x \diamond_0 (y, z)$ and $y \bullet_1 (x, (g_h, z_h)) = y \diamond_1 (x, z)$, so $(X, Y \times S, \bullet_0) \cong (X, Y \times Z, \diamond_0) = C_0$ and $(Y, X \times S, \bullet_1) \cong (Y, X \times Z, \diamond_1) = C_1$, with the isomorphisms coming from the identities on X and Y, and the bijection between S and Z.

Thus $(X \times Y, S, \bullet'') \in C_0 \boxplus C_1$, and $(Y, S, \bullet'') \cong D' \simeq D$, so C_1 is sister to C_0 in D, so C has a sister in D.

Finally, in the case where X and Y are both empty, C ≅null, and either D ≃ null or D ≃ 0, depending on whether Z is empty. It is easy to verify that null ⊠ null = {0, null}, since null ⊗ null ≅0, taking the two subsets of the singleton environment in 0 yields 0 and null as candidate sub-tensors, and both are valid sub-tensors, since either way, the conditions reduce to null ≃ null. □

Next, we have some definitions that more directly relate to our original definitions of subagent.

1.4. Currying Definitions

Definition: We say $C \triangleleft_+ D$ if there exists a Cartesian frame M over Agent(D) with $|\text{Env}(M)| = 1$, such that $C \simeq D^\circ(M)$.

Claim: This definition is equivalent to all of the above definitions of $\triangleleft_+$.

Proof: We show equivalence to the committing definition.

First, assume that there exist three sets X, Y, and Z, with $X \subseteq Y$, and a function $p : Y \times Z \to W$ such that $C \simeq (X, Z, \diamond)$ and $D \simeq (Y, Z, \bullet)$, where $\diamond$ and $\bullet$ are given by $x \diamond z = p(x, z)$ and $y \bullet z = p(y, z)$.

Let $D = (B, F, \star)$, and let $(g_0, h_0) : D \to (Y, Z, \bullet)$ and $(g_1, h_1) : (Y, Z, \bullet) \to D$ compose to something homotopic to the identity in both orders.

We define M, a Cartesian frame over B, by $M = (X, \{e\}, \cdot)$, where $\cdot$ is given $x \cdot e = g_1(x)$. Observe that $D^\circ(M) = (X, \{e\} \times F, \star')$, where $\star'$ is given by

$$\begin{aligned} x \star' (e, f) &= (x \cdot e) \star f \\ &= g_1(x) \star f \\ &= x \bullet h_1(f). \end{aligned}$$

To show that $(X, Z, \diamond) \simeq D^\circ(M)$, we construct morphisms $(g_2, h_2) : (X, Z, \diamond) \to D^\circ(M)$ and $(g_3, h_3) : D^\circ(M) \to (X, Z, \diamond)$ that compose to something homotopic to the identity in both orders. Let g_2 and g_3 both be the identity on X. Let $h_2 : \{e\} \times F \to Z$ be given by $h_2(e, f) = h_1(f)$, and let $h_3 : Z \to \{e\} \times F$ be given by $h_3(z) = (e, h_0(z))$.

We know (g_2, h_2) is a morphism, since for all $x \in X$ and $(e, f) \in \{e\} \times F$, we have

$$\begin{aligned} g_2(x) \star' (e, f) &= x \star' (e, f) \\ &= x \bullet h_1(f) \\ &= x \diamond h_1(f) \\ &= x \diamond (h_2(e, f)). \end{aligned}$$

We also have that (g_3, h_3) is a morphism, since for all $x \in X$ and $z \in Z$, we have

$$\begin{aligned} g_3(x) \diamond z &= x \diamond z \\ &= x \bullet z \\ &= x \bullet h_1(h_0(z)) \\ &= x \star' (e, h_0(z)) \\ &= x \star' h_3(z). \end{aligned}$$

Observe that (g_2, h_2) and (g_3, h_3) clearly compose to something homotopic to the identity in both orders, since $g_2 \circ g_3$ and $g_3 \circ g_2$ are the identity on X.

Thus, $C \simeq (X, Z, \diamond) \simeq D^\circ(M)$, and $|\text{Env}(M)| = 1$.

Conversely, assume $C \simeq D^\circ(M)$, with $|\text{Env}(M)| = 1$. We define $Y = \text{Agent}(D)$ and $Z = \text{Env}(D)$. We define $f : Y \times Z \to W$ by $f(y, z) = y \bullet z$, where $\bullet = \text{Eval}(D)$.

Let $X \subseteq Y$ be given by $X = \text{Image}(M)$. Since $|\text{Env}(M)| = 1$, we have $M \simeq \perp_X$. Thus, $C \simeq D^\circ(M) \simeq D^\circ(\perp_X)$. Unpacking the definition of $D^\circ(\perp_X)$, we get $D^\circ(\perp_X) = (X, \{e\} \times Z, \cdot)$, where $\cdot$ is given by $x \cdot (e, z) = f(x, z)$, which is isomorphic to $(X, Z, \diamond)$, where $\diamond$ is given by $x \diamond z = f(x, z)$. Thus $C \simeq (X, Z, \diamond)$ and $D = (Y, Z, \bullet)$, as in the committing definition. □

Definition: We say $C \triangleleft_\times D$ if there exists a Cartesian frame M over Agent(D) with $\text{Image}(M) = \text{Agent}(D)$, such that $C \simeq D^\circ(M)$.

Claim: This definition is equivalent to all of the above definitions of $\triangleleft_\times$.

Proof: We show equivalence to the externalizing definition.

First, assume there exist three sets X, Y, and Z, and a function $p : X \times Y \times Z \to W$ such that $C \simeq (X, Y \times Z, \diamond)$ and $D \simeq (X \times Y, Z, \bullet)$, where $\diamond$ and $\bullet$ are given by $x \diamond (y, z) = (x, y) \bullet z = p(x, y, z)$.

Let $D = (B, F, \star)$, and let $(g_0, h_0) : D \to (X \times Y, Z, \bullet)$ and $(g_1, h_1) : (X \times Y, Z, \bullet) \to D$ compose to something homotopic to the identity in both orders.

We define $B' = B \sqcup \{a\}$, and we define M, a Cartesian frame over B, by $M = (X, Y \times B', \cdot)$, where $\cdot$ is given by $x \cdot (y, b) = b$ if $b \in B$ and $g_0(b) = (x, y)$, and $x \cdot (y, b) = g_1(x, y)$ otherwise. Clearly, Image(M) = B, since for any $b \in B$, if we let $(x, y) = g_0(b)$, we have $x \cdot (y, b) = b$.

Observe that for all $x \in X$, $y \in Y$, $b \in B'$ and $f \in F$, if $b \in B$ and $g_0(b) = (x, y)$, then

$$\begin{aligned} (x \cdot (y, b)) \star f &= b \star f \\ &= g_1(g_0(b)) \star f \\ &= g_1(x, y) \star f, \end{aligned}$$

and on the other hand, if $b = a$ or $g_0(b) \neq (x, y)$, we also have $(x \cdot (y, b)) \star f = g_1(x, y) \star f$.

Thus, we have that $D^\circ(M) = (X, Y \times B' \times F, \star')$, where $\star'$ is given by

$$\begin{aligned} x \star' (y, b, f) &= (x \cdot (y, b)) \star f \\ &= g_1(x, y) \star f \\ &= (x, y) \bullet h_1(f). \end{aligned}$$

To show that $(X, Y \times Z, \diamond) \simeq D^\circ(M)$, we construct morphisms $(g_2, h_2) : (X, Y \times Z, \diamond) \to D^\circ(M)$ and $(g_3, h_3) : D^\circ(M) \to (X, Y \times Z, \diamond)$ that compose to something homotopic to the identity in both orders. Let g_2 and g_3 both be the identity on X. Let $h_2 : Y \times B' \times F \to Y \times Z$ be given by $h_2(y, b, f) = (y, h_1(f))$, and let $h_3 : Y \times Z \to Y \times B' \times F$ be given by $h_3(y, z) = (y, a, h_0(z))$.

We know (g_2, h_2) is a morphism, since for all $x \in X$ and $(y, b, f) \in Y \times B' \times F$,

$$
\begin{aligned}
g_2(x) \star' (y, b, f) &= x \star' (y, b, f) \\
&= (x, y) \bullet h_1(f) \\
&= p(x, y, h_1(f)) \\
&= x \diamond (y, h_1(f)) \\
&= x \diamond (h_2(y, b, f)).
\end{aligned}
$$

We also have that (g_3, h_3) is a morphism, since for all $x \in X$ and $(y, z) \in Y \times Z$, we have

$$
\begin{aligned}
g_3(x) \diamond (y, z) &= x \diamond z \\
&= x \diamond (y, z) \\
&= p(x, y, z) \\
&= (x, y) \bullet z \\
&= (x, y) \bullet h_1(h_0(z)) \\
&= (x, y) \star' (y, a, h_0(z)) \\
&= x \star' h_3(y, z).
\end{aligned}
$$

Observe that (g_2, h_2) and (g_3, h_3) clearly compose to something homotopic to the identity in both orders, since $g_2 \circ g_3$ and $g_3 \circ g_2$ are the identity on X.

Thus, $C \simeq (X, Z, \diamond) \simeq D^\circ(M)$, where Image(M) = Agent(D).

Conversely, assume $C \simeq D^\circ(M)$, with Image(M) = Agent(D). Let X = Agent(M), let Y = Env(M), and let Z = Env(D). Let $f : X \times Y \times Z \to W$ be given by $f(x, y, z) = (x \cdot y) \star z$, where $\cdot$ = Eval(M) and $\star$ = Eval(D).

Thus $C \simeq D^\circ(M) \cong (X, Y \times Z, \diamond)$, where $\diamond$ is given by $x \diamond (y, z) = (x \cdot y) \star z = f(x, y, z)$. All that remains to show is that $D \simeq (X \times Y, Z, \bullet)$, where $(x, y) \bullet z = f(x, y, z)$. Let $D = (B, Z, \star)$.

We construct morphisms $(g_0, h_0) : D \to (X \times Y, Z, \bullet)$ and $(g_1, h_1) : D \to (X \times Y, Z, \bullet)$ that compose to something homotopic to the identity in both orders. Let h_0 and h_1 be the identity on Z. Let $g_1 : X \times Y \to B$ be given by $g_1(x, y) = x \cdot y$. Since g_1 is surjective, it has a right inverse. Let $g_0 : B \to X \times Y$ be any choice of right inverse of g_1, so $g_1(g_0(b)) = b$ for all $b \in B$.

We know (g_1, h_1) is a morphism, since for all $(x, y) \in X \times Y$ and $z \in Z$,

$$\begin{aligned} g_1(x, y) \star z &= (x \cdot y) \star z \\ &= f(x, y, z) \\ &= (x, y) \bullet z \\ &= (x, y) \bullet h_1(z). \end{aligned}$$

To see that (g_0, h_0) is a morphism, given $b \in B$ and $z \in Z$, let $(x, y) = g_0(b)$, and observe

$$\begin{aligned} g_0(b) \bullet z &= (x, y) \bullet z \\ &= f(x, y, z) \\ &= (x \cdot y) \star z \\ &= g_1(x, y) \star z \\ &= g_1(g_0(b)) \star z \\ &= b \star h_0(z). \end{aligned}$$

(g_0, h_0) and (g_1, h_1) clearly compose to something homotopic to the identity in both orders, since $h_0 \circ h_1$ and $h_1 \circ h_0$ are the identity on Z. Thus $D \simeq (X \times Y, Z, \bullet)$, completing the proof. □

Consider two Cartesian frames C and D, and let M be a frame whose possible agents are Agent(C) and whose possible worlds are Agent(D). When C is a subagent of D, (up to biextensional equivalence) there exists a function from Agent(C), paired with Env(M), to Agent(D).

Just as we did in "Subagents of Cartesian Frames" §1.2 (Currying Definition), we can think of this function as a (possibly) nondeterministic function from Agent(C) to Agent(D), where Env(M) represents the nondeterminism. In the case of additive subagents, Env(M) is a singleton, meaning that the function from Agent(C) to Agent(D) is actually deterministic. In the case of multiplicative subagents, the (possibly) nondeterministic function is surjective.

Recall that in "Sub-Sums and Sub-Tensors" §3.3 (Sub-Sums and Sub-Tensors Are Superagents), we constructed a frame with a singleton environment to prove that sub-sums are superagents, and we constructed a frame with a surjective evaluation function to prove that sub-tensors are superagents. The currying definitions of $\triangleleft_+$ and $\triangleleft_\times$ show why this is the case.

1.5. Categorical Definitions

We also have definitions based on the categorical definition of subagent. The categorical definition of additive subagent is almost just swapping the quantifiers from our original categorical definition of subagent. However, we will also have to weaken the definition slightly in order to only require the morphisms to be homotopic.

Definition: We say $C \triangleleft_+ D$ if there exists a single morphism $\phi_0 : C \to D$ such that for every morphism $\phi : C \to \bot$ there exists a morphism $\phi_1 : D \to \bot$ such that ϕ is homotopic to $\phi_1 \circ \phi_0$.

Claim: This definition is equivalent to all the above definitions of $\triangleleft_+$.

Proof: We show equivalence to the committing definition.

First, let $C = (A, E, \cdot)$ and $D = (B, F, \bullet)$ be Cartesian frames over W, and let $(g_0, h_0) : C \to D$ be such that for all $(g, h) : C \to \bot$, there exists a $(g', h') : D \to \bot$ such that (g, h) is homotopic to $(g', h') \circ (g_0, h_0)$. Let $\bot = (W, \{i\}, \star)$.

Let $Y = B$, let $Z = F$, and let $X = \{g_0(a) \mid a \in A\}$. Let $f : Y \times Z \to W$ be given by $f(y, z) = y \bullet z$. We already have $D = (Y, Z, \bullet)$, and our goal is to show that $C \simeq (X, Z, \diamond)$, where $\diamond$ is given by $x \diamond z = f(x, z)$.

We construct $(g_1, h_1) : C \to (X, Z, \diamond)$ and $(g_2, h_2) : (X, Z, \diamond) \to C$ that compose to something homotopic to the identity in both orders.

We define $g_1 : A \to X$ by $g_1(a) = g_0(a)$. g_1 is surjective, and so has a right inverse. We let $g_2 : X \to A$ be any right inverse to g_1, so $g_1(g_2(x)) = x$ for all $x \in X$. We let $h_1 : Z \to E$ be given by $h_1(z) = h_0(z)$.

Defining $h_2 : E \to Z$ will be a bit more complicated. Given an $e \in E$, let (g_e, h_e) be the morphism from C to $\perp$, given by $h_e(i) = e$ and $g_e(a) = a \cdot e$. Let $(g'_e, h'_e) : D \to \perp$ be such that (g'_e, h'_e) is homotopic to $(g'_e, h'_e) \circ (g_0, h_0)$. We define h_2 by $h_2(e) = h'_e(i)$.

We trivially have that (g_1, h_1) is a morphism, since for all $a \in A$ and $z \in Z$,

$$\begin{aligned} g_1(a) \diamond z &= g_0(a) \bullet z \\ &= a \cdot h_0(z) \\ &= a \cdot h_1(z). \end{aligned}$$

To see that (g_2, h_2) is a morphism, consider $x \in X$ and $e \in E$, and define (g_e, h_e) and (g'_e, h'_e) as above. Then,

$$\begin{aligned} x \diamond h_2(e) &= g_1(g_2(x)) \diamond h'_e(i) \\ &= g'_e(g_0(g_2(x))) \star i \\ &= g_2(x) \cdot h_e(i) \\ &= g_2(x) \cdot e. \end{aligned}$$

We trivially have that $(g_1, h_1) \circ (g_2, h_2)$ is homotopic to the identity, since $g_1 \circ g_2$ is the identity on X. To see that $(g_2, h_2) \circ (g_1, h_1)$ is homotopic to the identity on C, observe that for all $a \in A$ and $e \in E$, defining (g_e, h_e) and (g'_e, h'_e) as above,

$$
\begin{aligned}
g_2(g_1(a)) \cdot e &= g_1(a) \diamond h_2(e) \\
&= g_0(a) \diamond h'_e(i) \\
&= g'_e(g_0(a)) \star i \\
&= a \star h_e(i) \\
&= a \cdot e .
\end{aligned}
$$

Thus $C \simeq (X, Z, \diamond)$, and $C \lhd_+ D$ according to the committing definition.

Conversely, let X, Y, and Z be arbitrary sets with $X \subseteq Y$, let $f : Y \times Z \to W$, and let $C \simeq (X, Z, \diamond)$ and $D \simeq (Y, Z, \bullet)$, where $\diamond$ and $\bullet$ are given by $x \diamond z = f(x, z)$ and $y \bullet z = f(y, z)$.

Let $(g_1, h_1) : C \to (X, Z, \diamond)$ and $(g_2, h_2) : (X, Z, \diamond) \to C$ compose to something homotopic to the identity in both orders, and let $(g_3, h_3) : D \to (Y, Z, \bullet)$ and $(g_4, h_4) : (Y, Z, \bullet) \to D$ compose to something homotopic to the identity in both orders. Let $(g_0, h_0) : (X, Z, \diamond) \to (Y, Z, \bullet)$ be given by g_0 is the embedding of X in Y and h_0 is the identity on Z. (g_0, h_0) is clearly a morphism.

We let $\phi : C \to D = (g_4, h_4) \circ (g_0, h_0) \circ (g_1, h_1)$.

Given a $(g, h) : C \to \perp$, our goal is to construct a $(g', h') : D \to \perp$ such that (g, h) is homotopic to $(g', h') \circ \phi$.

Let $\perp = (W, \{i\}, \star)$, let $C = (A, E, \cdot_0)$, and let $D = (B, F, \cdot_1)$. Let $h' : \{i\} \to F$ be given by $h' = h_3 \circ h_2 \circ h$. Let $g' : B \to W$ be given by $g'(b) = b \cdot_1 h'(i)$. This is clearly a morphism, since for all $b \in B$ and $i \in \{i\}$,

$$
\begin{aligned}
g'(b) \star i &= g'(b) \\
&= b \cdot h'(i) .
\end{aligned}
$$

To see that (g, h) is homotopic to $(g', h') \circ (g_4, h_4) \circ (g_0, h_0) \circ (g_1, h_1)$, we just need to check that $(g, h_1 \circ h_0 \circ h_4 \circ h') : C \to \bot$ is a morphism. Or, equivalently, that $(g, h_1 \circ h_4 \circ h_3 \circ h_2 \circ h) : C \to \bot$, since h_0 is the identity, and $h' = h_3 \circ h_2 \circ h$.

Indeed, for all $a \in A$ and $i \in \{i\}$,

$$\begin{aligned} g(a) \star i &= a \cdot_0 h(a) \\ &= a \cdot_0 h_1(h_2(h(a))) \\ &= g_1(a) \diamond h_2(h(a)) \\ &= g_1(a) \bullet h_2(h(a)) \\ &= g_1(a) \bullet h_4(h_3(h_2(h(a)))) \\ &= g_1(a) \diamond h_4(h_3(h_2(h(a)))) \\ &= a \cdot_0 h_1(h_4(h_3(h_2(h(a))))). \end{aligned}$$

Thus (g, h) is homotopic to $(g', h') \circ \phi$, completing the proof. □

Definition: We say $C \triangleleft_\times D$ if for every morphism $\phi : C \to \bot$, there exist morphisms $\phi_0 : C \to D$ and $\phi_1 : D \to \bot$ such that $\phi = \phi_1 \circ \phi_0$, and for every morphism $\psi : 1 \to D$, there exist morphisms $\psi_0 : 1 \to C$ and $\psi_1 : C \to D$ such that $\psi = \psi_1 \circ \psi_0$.

Before showing that this definition is equivalent to all of the above definitions, we will give one final definition of multiplicative subagent.

1.6. Sub-Environment Definition

First, we define the concept of a sub-environment, which is dual to the concept of a sub-agent.

Definition: We say C is a sub-environment of D, written $C \triangleleft^* D$, if $D^* \triangleleft C^*$.

We can similarly define additive and multiplicative sub-environments.

Definition: We say C is an additive sub-environment of D, written $C \overset{*}{\triangleleft_+} D$, if $D^* \triangleleft_+ C^*$.

We say C is an multiplicative sub-environment of D, written $C \overset{*}{\triangleleft_\times} D$, if $D^* \triangleleft_\times C^*$.

This definition of a multiplicative sub-environment is redundant, because the set of frames with multiplicative sub-agents is exactly the set of frames with multiplicative sub-environments, as shown below:

Claim: $C \triangleleft_\times D$ if and only if $C \overset{*}{\triangleleft}_\times D$.

Proof: We prove this using the externalizing definition of $\triangleleft_\times$.

If $C \triangleleft_\times D$, then for some X, Y, Z, and $f : X \times Y \times Z \to W$, we have $C \simeq (X, Y \times Z, \diamond)$ and $D \simeq (X \times Y, Z, \bullet)$, where $\diamond$ and $\bullet$ are given by $x \diamond (y, z) = f(x, y, z)$ and $(x, y) \bullet z = f(x, y, z)$.

Observe that $D^* \simeq (Z, Y \times X, \cdot)$ and $C^* \simeq (Z \times Y, X, \star)$, where $\cdot$ and $\star$ are given by $z \cdot (y, x) = f(x, y, z)$ and $(z, y) \star x = f(x, y, z)$. Taking $X' = Z$, $Y' = Y$, $Z' = X$, and $f'(x, y, z) = f(z, y, x)$, this is exactly the externalizing definition of $D^* \triangleleft_\times C^*$, so $C \overset{*}{\triangleleft}_\times D$.

Conversely, if $C \overset{*}{\triangleleft}_\times D$, then $D^* \triangleleft_\times C^*$, so $C \cong \{C^*\}^* \triangleleft_\times \{D^*\}^* \cong D$. □

We now give the sub-environment definition of multiplicative subagent:

Definition: We say $C \triangleleft_\times D$ if $C \triangleleft D$ and $C \triangleleft^* D$. Equivalently, we say $C \triangleleft_\times D$ if $C \triangleleft D$ and $D^* \triangleleft C^*$.

Claim: This definition is equivalent to the categorical definition of $\triangleleft_\times$.

Proof: The condition that for every morphism $\phi : C \to \bot$, there exist morphisms $\phi_0 : C \to D$ and $\phi_1 : D \to \bot$ such that $\phi = \phi_1 \circ \phi_0$, is exactly the categorical definition of $C \triangleleft D$.

The condition that for every morphism $\psi : 1 \to D$, there exist morphisms $\psi_0 : 1 \to C$ and $\psi_1 : C \to D$ such that $\psi = \psi_1 \circ \psi_0$, is equivalent to saying that for every morphism $\psi^* : D^* \to \bot$, there exist morphisms $\overset{*}{\psi}_0 : C^* \to \bot$ and $\overset{*}{\psi}_1 : D^* \to C^*$ such that $\psi^* = \overset{*}{\psi}_1 \circ \overset{*}{\psi}_0$. This is the categorical definition of $D^* \triangleleft C^*$. □

Claim: The categorical and sub-environment definitions of $\triangleleft_\times$ are equivalent to the other four definitions of multiplicative subagent above: sub-tensor, sister, externalizing, and currying.

Proof: We show equivalence between the externalizing and sub-environment definitions. First, assume that $C = (A, E, \cdot)$ and $D = (B, F, \star)$ are Cartesian frames over W with $C \triangleleft D$ and $C \triangleleft^* D$.

We define $X = A$, $Z = F$, and $Y = \hom(C, D)$. We define $p : X \times Y \times Z \to W$ by

$$\begin{aligned} p(a, (g, h), f) &= g(a) \star f \\ &= a \cdot h(f). \end{aligned}$$

We want to show that $C \simeq (X, Y \times Z, \diamond)$, and $D \simeq (X \times Y, Z, \bullet)$, where $\diamond$ and $\bullet$ are given by $x \diamond (y, z) = (x, y) \bullet z = p(x, y, z)$.

To see $C \simeq (X, Y \times Z, \diamond)$, we construct $(g_0, h_0) : C \to (X, Y \times Z, \diamond)$ and $(g_1, h_1) : (X, Y \times Z, \diamond) \to C$ that compose to something homotopic to the identity in both orders. Let g_0 and g_1 be the identity on X and let $h_0 : Y \times Z \to E$ be defined by $h_0((g, h), f) = h(f)$. By the covering definition of subagent, h_0 is surjective, and so has a right inverse. Let $h_1 : E \to Y \times Z$ be any right inverse of h_0, so $h_0(h_1(e)) = e$ for all $e \in E$.

We know (g_0, h_0) is a morphism, because for all $a \in A$ and $((g, h), f) \in Y \times Z$,

$$\begin{aligned} g_0(a) \diamond ((g, h), f) &= a \diamond ((g, h), f) \\ &= p(a, (g, h), f) \\ &= a \cdot h(f) \\ &= a \cdot h_0((g, h), f). \end{aligned}$$

We know (g_1, h_1) is a morphism, since for $x \in X$ and $e \in E$, if $((g, h), f) = h_1(e)$,

$$
\begin{aligned}
g_1(x) \cdot e &= x \cdot h_0((g, h), f) \\
&= x \cdot h(f) \\
&= p(x, (g, h), f) \\
&= x \diamond ((g, h), f) \\
&= x \diamond h_1(e).
\end{aligned}
$$

(g_0, h_0) and (g_1, h_1) clearly compose to something homotopic to the identity in both orders, since $g_0 \circ g_1$ and $g_1 \circ g_0$ are the identity on X.

To see $D \simeq (X \times Y, Z, \bullet)$, we construct $(g_2, h_2) : D \to (X \times Y, Z, \bullet)$ and $(g_3, h_3) : (X \times Y, Z, \bullet) \to D$ that compose to something homotopic to the identity in both orders. Let h_2 and h_3 be the identity on Z and let $g_3 : X \times Y \to B$ be defined by $g_3(a, (g, h)) = g(a)$. By the covering definition of subagent and the fact that $D^* \triangleleft C^*$, g_3 is surjective, and so has a right inverse. Let $g_2 : B \to X \times Y$ be any right inverse of g_3, so $g_3(g_2(b)) = b$ for all $b \in B$.

We know (g_3, h_3) is a morphism, because for all $f \in F$ and $(a, (g, h)) \in X \times Y$,

$$
\begin{aligned}
g_3(a, (g, h)) \star f &= g(a) \star f \\
&= p(a, (g, h), f) \\
&= (a, (g, h)) \bullet f \\
&= (a, (g, h)) \bullet h_3(f).
\end{aligned}
$$

We know (g_2, h_2) is a morphism, since for $z \in Z$ and $b \in B$, if $(a, (g, h)) = g_2(b)$,

$$
\begin{aligned}
g_2(b) \bullet z &= (a, (g, h)) \bullet z \\
&= p(a, (g, h), z) \\
&= g(a) \star z \\
&= g_3(a, (g, h)) \star z \\
&= b \star h_2(z).
\end{aligned}
$$

Observe that (g_2, h_2) and (g_3, h_3) clearly compose to something homotopic to the identity in both orders, since $h_2 \circ h_3$ and $h_3 \circ h_2$ are the identity on Z.

Thus, $C \simeq (X, Y \times Z, \diamond)$, and $D \simeq (X \times Y, Z, \bullet)$.

Conversely, if $C \triangleleft_\times D$ according to the externalizing definition, then we also have $D^* \triangleleft_\times C^*$. However, by the currying definitions of multiplicative subagent and of subagent, multiplicative subagent is stronger than subagent, so $C \triangleleft D$ and $D^* \triangleleft C^*$. □

2. Basic Properties

Now that we have enough definitions of additive and multiplicative subagent, we can cover some basic properties.

First: Additive and multiplicative subagents are subagents.

Claim: If $C \triangleleft_+ D$, then $C \triangleleft D$. Similarly, if $C \triangleleft_\times D$, then $C \triangleleft D$.

Proof: Clear from the currying definitions. □

Additive and multiplicative subagent are also well-defined up to biextensional equivalence.

Claim: If $C \triangleleft_+ D$, $C' \simeq C$, and $D' \simeq D$, then $C' \triangleleft_+ D'$. Similarly, if $C \triangleleft_\times D$, $C' \simeq C$, and $D' \simeq D$, then $C' \triangleleft_\times D'$.

Proof: Clear from the committing and externalizing definitions. □

Claim: Both $\triangleleft_+$ and $\triangleleft_\times$ are reflexive and transitive.

Proof: Reflexivity is clear from the categorical definitions. Transitivity of $\triangleleft_\times$ is clear from the transitivity of $\triangleleft$ and the sub-environment definition. Transitivity of $\triangleleft_+$ can be seen using the categorical definition, by composing the morphisms and using the fact that being homotopic is preserved by composition. □

3. Decomposition Theorems

We have two decomposition theorems involving additive and multiplicative subagents.

3.1. First Decomposition Theorem

Theorem: $C_0 \triangleleft C_1$ if and only if there exists a C_2 such that $C_0 \triangleleft_\times C_2 \triangleleft_+ C_1$.

Proof: We will use the currying definitions of subagent and multiplicative subagent, and the committing definition of additive subagent. Let $C_0 = (A_0, E_0, \cdot_0)$ and $C_1 = (A_1, E_1, \cdot_1)$. If $C_0 \triangleleft C_1$, there exists some Cartesian frame D over A_1 such that $C_0 = \overset{\circ}{C_1}(D)$.

Let $C_2 = (\text{Image}(D), E_1, \cdot_2)$, where $\cdot_2$ is given by $a \cdot_2 e = a \cdot_1 e$. C_2 is created by deleting some possible agents from C_1, so by the committing definition of additive subagent $C_2 \triangleleft_+ C_1$.

Also, if we let D' be the Cartesian frame over Image(D) which is identical to D, but on a restricted codomain, then we clearly have that $\overset{\circ}{C_1}(D) \cong \overset{\circ}{C_2}(D')$. Thus $C_0 \simeq \overset{\circ}{C_2}(D')$ and $\text{Image}(D') = \text{Agent}(C_2)$, so $C_0 \triangleleft_\times C_2$.

The converse is trivial, since subagent is weaker than additive and multiplicative subagent and is transitive. □

Imagine that that a group of kids, Alice, Bob, Carol, etc., is deciding whether to start a game of baseball or football against another group. If they choose baseball, they form a team represented by the frame C_B, while if they choose football, they form a team represented by the frame C_F. We can model this by imagining that C_0 is the group's initial state, and C_B and C_F are precommitment-style subagents of C_0.

Suppose the group chooses football. C_F's choices are a function of Alice-the-football-player's choices, Bob-the-football-player's choices, etc. (Importantly, Alice here has different options and a different environment than if the original group had chosen baseball. So we will need to represent Alice-the-football-player, C_{AF}, with a different frame than Alice-the-baseball-player, C_{AB}; and likewise for Bob and the other team members.)

It is easy to see in this case that the relationship between Alice-the-football-player's frame (C_{AF}) and the entire group's initial frame (C_0) can be decomposed into the additive relationship between C_0 and C_F and the multiplicative relationship between C_F and C_{AF}, in that order.

The first decomposition theorem tells us that *every* subagent relation, even ones that don't seem to involve a combination of "making a commitment" and "being a team," can be decomposed into a combination of those two things. I've provided an example above where this factorization feels natural, but other cases may be less natural.

Using the framing from our discussion of the currying definitions: this decomposition is always possible because we can always decompose a possibly-nondeterministic function f into (1) a possibly-nondeterministic surjective function onto f's image, and (2) a deterministic function embedding f's image in f's codomain.

3.2. Second Decomposition Theorem

Theorem: There exists a morphism from C_0 to C_1 if and only if there exists a C_2 such that $C_0 \overset{*}{\triangleleft_+} C_2 \triangleleft_+ C_1$.

Proof: First, let $C_0 = (A, E, \cdot)$, let $C_1 = (B, F, \star)$, and let $(g, h) : C_0 \to C_1$. We let $C_2 = (A, F, \diamond)$, where $a \diamond f = g(a) \star f = a \cdot h(f)$.

First, we show $C_2 \triangleleft_+ C_1$, To do this, we let $B' \subseteq B$ be the image of g, and let $C_2' = (B', F, \star')$, where $\star'$ is given by $b \star' f = b \star f$. By the committing definition of additive subagent, it suffices to show that $C_2' \simeq C_2$.

We define $(g_0, h_0) : C_2 \to C_2'$ and $(g_1, h_1) : C_2' \to C_2$ as follows. We let h_0 and h_1 be the identity on F. We let $g_0 : A \to B'$ be given by $g_0(a) = g(a)$. Observe that g_0 is surjective, and thus has a right inverse. Let g_1 be any right inverse to g_0, so $g_0(g_1(b)) = b$ for all $b \in B'$.

We know (g_0, h_0) is a morphism, since for all $a \in A$ and $f \in F$, we have

$$\begin{aligned} g_0(a) \star' f &= g(a) \star f \\ &= a \diamond f \\ &= a \diamond h_0(f). \end{aligned}$$

Similarly, we know (g_1, h_1) is a morphism, since for all $b \in B'$ and $f \in F$, we have

$$\begin{aligned} g_1(b) \diamond f &= g_1(b) \diamond h_0(f) \\ &= g_0(g_1(b)) \star' f \\ &= b \star' f \\ &= b \star' h_1(f). \end{aligned}$$

Clearly, $(g_0, h_0) \circ (g_1, h_1)$ and $(g_1, h_1) \circ (g_0, h_0)$ are homotopic to the identity, since $h_0 \circ h_1$ and $h_1 \circ h_0$ are the identity on F. Thus, $C_2' \simeq C_2$.

The fact that $C_0 \overset{*}{\triangleleft_+} C_2$, or equivalently $C_2' \overset{*}{\triangleleft_+} C_0$, is symmetric, since the relationship between C_2' and C_0 is the same as the relationship between C_2 and C_1.

Conversely, if $C_2 \triangleleft_+ C_1$, there is a morphism from C_2 to C_1 by the categorical definition of additive subagent. Similarly, if $C_0 \overset{*}{\triangleleft_+} C_2$, then $C_2' \overset{*}{\triangleleft_+} C_0$, so there is a morphism from C_2' to C_0, and thus a morphism from C_0 to C_2. These compose to a morphism from C_0 to C_1. □

When we introduced morphisms and described them as "interfaces," we noted that every morphism $(g, h) : C_0 \to C_1$ implies the existence of an intermediate frame C_2 that represents Agent(C_0) interacting with Env(C_1). The second decomposition theorem formalizes this claim, and also notes that this intermediate frame is a super-environment of C_0 and a subagent of C_1.

In our next post, we will provide several methods for constructing additive and multiplicative subagents: "Committing, Assuming, Externalizing, and Internalizing."

Committing, Assuming, Externalizing, and Internalizing

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the tenth post in the Cartesian frames sequence.

Here, we define a bunch of ways to construct new (additive/multiplicative) (sub/super)-(agents/environments) from a given Cartesian frame. Throughout this post, we will start with a single Cartesian frame over W, $C = (A, E, \cdot)$.

We will start by defining operations from taking subsets and partitions of A and E. We will then define similar operations from taking subsets and partitions of W.

1. Definitions from Agents and Environments

1.1. Committing

Definition: Given a subset $B \subseteq A$, let $\text{Commit}^B(C)$ denote the Cartesian frame $(B, E, \star)$, with $\star$ given by $b \star e = b \cdot e$. Let $\text{Commit}^{\setminus B}(C)$ denote the Cartesian frame $(A \setminus B, E, \diamond)$, with $\diamond$ given by $a \diamond e = a \cdot e$.

$\text{Commit}^B(C)$ represents the perspective you get when the agent of C makes a commitment to choose an element of B, while $\text{Commit}^{\setminus B}(C)$ represents the perspective you get when the agent of C makes a commitment to choose an element outside of B.

Claim: For all $B \subseteq A$, $\text{Commit}^B(C) \triangleleft_+ C$ and $\text{Commit}^{\setminus B}(C) \triangleleft_+ C$. Further, $\text{Commit}^{\setminus B}(C)$ and $\text{Commit}^B(C)$ are brothers in C.

Proof: That $\text{Commit}^B(C) \triangleleft_+ C$ and $\text{Commit}^{\setminus B}(C) \triangleleft_+ C$ is trivial from the committing definition of additive subagent.

Observe that $\text{Commit}^{B}(C) \oplus \text{Commit}^{\setminus B}(C) = (A, E \times E, \bullet)$, where $\bullet$ is given by $a \bullet (e, f) = a \cdot e$ if $a \in B$, and $a \bullet (e, f) = a \cdot f$ if $a \notin B$. Let $D \subset E \times E$ be the diagonal, $\{(e, e) \mid e \in E\}$. We clearly have that $(A, D, \bullet')$ is in $\text{Commit}^{B}(C) \boxplus \text{Commit}^{\setminus B}(C)$, where $\bullet'$ is the restriction of $\bullet$ to $A \times D$, and that $(A, D, \bullet') \cong C$; so $\text{Commit}^{\setminus B}(C) \triangleleft_{+} C$ and $\text{Commit}^{B}(C) \triangleleft_{+} C$ are brothers in C. □

Claim: $\text{Commit}^{\setminus B}(C) \cong \text{Commit}^{A \setminus B}(C)$

Proof: Trivial. □

1.2. Assuming

Assuming is the dual operation to committing.

Definition: Given a subset $F \subseteq E$, let $\text{Assume}^{F}(C)$ denote the Cartesian frame $(A, F, \star)$, with $\star$ given by $a \star f = a \cdot f$. Let $\text{Assume}^{\setminus F}(C)$ denote the Cartesian frame $(A, E \setminus F, \diamond)$, with $\diamond$ given by $a \diamond e = a \cdot e$.

$\text{Assume}^{F}(C)$ represents the perspective you get when you assume the environment is chosen from F, while $\text{Assume}^{\setminus F}(C)$ represents the perspective you get when you assume the environment is chosen from outside of F.

In "Introduction to Cartesian Frames" §3.2 (Examples of Controllables), I noted the counter-intuitive result that agents have no control in worlds where a meteor (or other event) could have prevented their existence:

$$C_0 = \begin{array}{c} \\ u \\ n \\ u \leftrightarrow r \\ u \leftrightarrow s \end{array} \begin{array}{c} \begin{array}{ccc} r & s & m \end{array} \\ \begin{pmatrix} ur & us & m \\ nr & ns & m \\ ur & ns & m \\ nr & us & m \end{pmatrix} \end{array}.$$

Here, we see that we can use $\text{Assume}^{\setminus F}(C)$ to recover the more intuitive idea of "control." The subagent modified by the assumption "there's no meteor" can have controllables, even though the original agent has no controllables:

$$\text{Assume}^{\setminus\{m\}}(C_0) = \begin{array}{c} \\ u \\ n \\ u \leftrightarrow r \\ u \leftrightarrow s \end{array} \begin{array}{c} \begin{array}{cc} r & s \end{array} \\ \left(\begin{array}{cc} ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{array}\right) \end{array}.$$

Claim: For all $F \subseteq E$, $(\text{Assume}^{F}(C))^* \cong \text{Commit}^{F}(C^*)$ and $(\text{Assume}^{\setminus F}(C))^* \cong \text{Commit}^{\setminus F}(C^*)$. Similarly, for all $B \subseteq A$, $(\text{Commit}^{B}(C))^* \cong \text{Assume}^{B}(C^*)$ and $(\text{Commit}^{\setminus B}(C))^* \cong \text{Assume}^{\setminus B}(C^*)$.

Proof: Trivial. □

Claim: For all $F \subseteq E$, $C \overset{*}{\triangleleft_+} \text{Assume}^{F}(C)$ and $C \overset{*}{\triangleleft_+} \text{Assume}^{\setminus F}(C)$.

Proof: Trivial. □

1.3. Externalizing

Note that for the following definitions, when we say "X is a partition of Y," we mean that X is a set of nonempty subsets of Y, such that for each $y \in Y$, there exists a unique $x \in X$ such that $y \in x$.

Definition: Given a partition X of Y, let Y/X denote the set of all functions q from X to Y such that $q(x) \in x$ for all $x \in X$.

Definition: Given a partition B of A, let $\text{External}^{B}(C)$ denote the Cartesian Frame $(A/B, B \times E, \star)$, where $\star$ is given by $q \star (b, e) = q(b) \cdot e$. Let $\text{External}^{/B}(C)$ denote the Cartesian Frame $(B, A/B \times E, \diamond)$, where $\diamond$ is given by $b \diamond (q, e) = q(b) \cdot e$.

We say "externalizing B" for External^B and "externalizing mod B" for $\text{External}^{/B}$.

$\text{External}^B(C)$ can be thought of the result of the agent of C first factoring its choice into choosing an equivalence class in B, and choosing an element of each equivalence class, and then externalizing the part of itself that chooses an equivalence class. I.e., we are drawing a new Cartesian frame which treats the choice of equivalence class as part of the environment, rather than part of the agent.

Similarly, $\text{External}^{/B}(C)$ can be thought of the result of the agent of C factoring as above, then externalizing the part of itself that chooses an element of each equivalence class.

Claim: For all partitions B of A, $\text{External}^B(C) \triangleleft_\times C$ and $\text{External}^{/B}(C) \triangleleft_\times C$. Further, $\text{External}^B(C)$ and $\text{External}^{/B}(C)$ are sisters in C.

Proof: Let $\text{External}^B(C) = (A/B, B \times E, \star)$ and $\text{External}^{/B}(C) = (B, A/B \times E, \diamond)$.

First, observe that for every $e \in E$, there exists a morphism $(g_e, h_e) : \text{External}^B(C) \to (\text{External}^{/B}(C))^*$, with $g_e :\to A/B \times E$ given by $g_e(q) = (q, e)$, and $h_e : B \to B \times E$ given by $h_e(b) = (b, e)$. To see that this is a morphism, observe that for all $q \in A/B$ and $b \in B$,

$$\begin{aligned} g_e(q) \diamond b &= q(b) \cdot e \\ &= q \star h_e(b). \end{aligned}$$

Let $E' \subseteq \hom(\text{External}^B(C), (\text{External}^{/B}(C))^*)$ be given by $E' = \{(g_e, h_e) \mid e \in E\}$, and let $D = (A/B \times B, E', \bullet)$, where

$$\begin{aligned} (q, b) \bullet (g_e, h_e) &= g_e(q) \diamond b \\ &= q \star h_e(b) \\ &= q(b) \cdot e. \end{aligned}$$

Our goal is to show that $D \in \text{External}^B(C) \boxtimes \text{External}^{/B}(C)$, and that $D \simeq C$.

To see $D \simeq C$, we define $(g_0, h_0) : D \to C$ and $(g_1, h_1) : C \to D$ as follows.

First, $g_0 : A/p \times B \to A$ is given by $g_0(q, b) = q(b)$. We first need to confirm that g_0 is surjective. Given any $a \in A$, we can let $b \in B$ be the set with $a \in b$ and construct a function $q \in A/B$ by saying $q(b) = a$, and for each $b' \neq b$, choosing an $a' \in b'$, and saying $q(b') = a'$. Observing that $g_0(q, b) = a$, we have that g_0 is surjective and thus has a right inverse.

We choose $g_1 : A \to A/B \times B$ to be any right inverse to g_0. Similarly, we define $h_0 : E \to E'$ by $h_0(e) = (g_e, h_e)$, which is clearly surjective, and define $h_1 : E' \to E$ to be any right inverse to h_0. (Indeed, h_0 is bijective as long as A is nonempty.)

Then, for all $(q, b) \in A/B \times B$ and $e \in E$, we have

$$\begin{aligned} g_0(q, b) \cdot e &= q(b) \cdot e \\ &= (q, b) \bullet (g_e, h_e) \\ &= q, b \bullet h_0(e), \end{aligned}$$

so (g_0, h_0) is a morphism. This also gives us that for all $a \in A$ and $e' \in E'$ we have

$$\begin{aligned} g_1(a) \bullet e' &= g_1(a) \bullet h_0(h_1(e')) \\ &= g_0(g_1(a)) \cdot h_1(e) \\ &= a \cdot h_1(e), \end{aligned}$$

so (g_1, h_1) is a morphism. We know $(g_0, h_0) \circ (g_1, h_1)$ is homotopic to the identity on C, since $g_0 \circ g_1$ is the identity on A, and we know that $(g_1, h_1) \circ (g_0, h_0)$ is homotopic to the identity on D, since $h_0 \circ h_1$ is the identity on E'. Thus, $D \simeq C$.

To show that $D \in \text{External}^B(C) \boxtimes \text{External}^{/B}(C)$, it suffices to show that

$$\begin{aligned} \text{External}^B(C) &= (A/B, B \times E, \star) \\ &\simeq (A/B, B \times E', \star'), \end{aligned}$$

and

$$\begin{aligned}\text{External}^{/B}(C) &= (B, A/B \times E, \diamond) \\ &\simeq (B, A/B \times E', \diamond'),\end{aligned}$$

where $\star'$ and $\diamond'$ are given by

$$\begin{aligned}q \star' (b, (g_e, h_e)) &= b \diamond' (q, (g_e, h_e)) \\ &= q \star h_e(b) \\ &= b \diamond g_e(q) \\ &= q(b) \cdot e.\end{aligned}$$

Indeed, we show that if A is nonempty, $(A/B, B \times E, \star) \cong (A/B, B \times E', \star')$, and $(B, A/B \times E, \diamond) \cong (B, A/B \times E', \diamond')$.

If A is nonempty, then the function from E to E' given by $e \mapsto (g_e, h_e)$ is a bijection, since it is clearly surjective, and is injective because e is uniquely defined by $g_e(a) = (a, e)$. This gives a bijection between $B \times E$ and $B \times E'$, and we have that for all $q \in A/B$, $b \in B$, and $e \in E$,

$$\begin{aligned}q \star' (b, (g_e, h_e)) &= q(b) \cdot e \\ &= q \star (b, e).\end{aligned}$$

Similarly, we have a bijection between $A/B \times E$ and $A/B \times E'$, and for all $q \in A/B$, $b \in B$, and $e \in E$,

$$\begin{aligned}b \diamond' (q, (g_e, h_e)) &= q(b) \cdot e \\ &= b \diamond (q, e).\end{aligned}$$

If A is empty, then B is empty, and A/p is a singleton empty function, so $(A/B, B \times E, \star) \simeq (A/B, B \times E', \star') \simeq \top$, and we either have $(B, A/B \times E, \diamond) \simeq (B, A/B \times E', \diamond') \simeq 0$ or $(B, A/B \times E, \diamond) \simeq (B, A/B \times E', \diamond') \simeq \text{null}$, depending on whether or not E is empty.

Thus, $D \in \text{External}^{B}(C) \boxtimes \text{External}^{/B}(C)$, so $\text{External}^{B}(C)$ and $\text{External}^{/B}(C)$ are sisters in C. □

1.4. Internalizing

Definition: Given a partition F of E, let $\text{Internal}^{F}(C)$ denote the Cartesian Frame $(F \times A, E/F, \star)$, where $\star$ is given by $(f, a) \star q = a \cdot q(f)$. Let $\text{Internal}^{/F}(C)$ denote the Cartesian Frame $(E/F \times A, F, \diamond)$, where $\diamond$ is given by $(q, a) \diamond f = a \cdot q(f)$.

We say "internalizing p" for Internal^{p} and "internalizing mod p" for $\text{Internal}^{/p}$.

Claim: For all partitions F of E, $(\text{Internal}^{F}(C))^* \cong \text{External}^{F}(C^*)$ and $(\text{Internal}^{/F}(C))^* \cong \text{External}^{/F}(C^*)$. Similarly, for all partitions B of A, $(\text{External}^{B}(C))^* \cong \text{Internal}^{B}(C^*)$ and $(\text{External}^{/B}(C))^* \cong \text{Internal}^{/B}(C^*)$.

Proof: Trivial. □

Claim: For all partitions F of E, $C \triangleleft_{\times} \text{Internal}^{F}(C)$ and $C \triangleleft_{\times} \text{Internal}^{/F}(C)$.

Proof: This follows from the fact that $(\text{Internal}^{F}(C))^* \cong \text{External}^{F}(C^*) \triangleleft_{\times} C^*$ and $(\text{Internal}^{/F}(C))^* \cong \text{External}^{/F}(C^*) \triangleleft_{\times} C^*$, and the fact that multiplicative subagent is equivalent to multiplicative sub-environment. □

2. Definitions from Worlds

The above definitions are dependent on subsets and partitions of a given A and E, and thus do not represent a single operation that can be applied to an arbitrary Cartesian frame over W. We will now use the above eight definitions to define another eight operations that instead use subsets and partitions of W.

Once we have the following definitions in hand, our future references to committing, assuming, externalizing, and internalizing will use the definitions from worlds unless

noted otherwise.

2.1. Committing

Definition: Given a set $S \subseteq W$, we define $\text{Commit}_S(C) = \text{Commit}^B(C)$ and $\text{Commit}_{\setminus S}(C) = \text{Commit}^{\setminus B}(C)$, where $B \subseteq A$ is given by $B = \{a \in A \mid \forall e \in E, a \cdot e \in S\}$.

Claim: For all $S \subseteq W$, $\text{Commit}_S(C) \lhd_+ C$ and $\text{Commit}_{\setminus S}(C) \lhd_+ C$. Further, $\text{Commit}_{\setminus S}(C)$ and $\text{Commit}_S(C)$ are brothers in C.

Proof: Trivial. □

Unlike before, it is not necessarily the case that $\text{Commit}_{\setminus S}(C) \cong \text{Commit}_{W \setminus S}(C)$. This is because there might be rows that contains both elements of S and elements of $W \setminus S$.

2.2. Assuming

Definition: Given $S \subseteq W$, we define $\text{Assume}_S(C) = \text{Assume}^F(C)$ and $\text{Assume}_{\setminus S}(C) = \text{Assume}^{\setminus F}(C)$, where $F \subseteq E$ is given by $F = \{e \in E \mid \forall a \in A, a \cdot e \in S\}$.

Claim: For all $S \subseteq W$, $(\text{Assume}_S(C))^* \cong \text{Commit}_S(C^*)$ and $(\text{Assume}_{\setminus S}(C))^* \cong \text{Commit}_{\setminus S}(C^*)$, $(\text{Commit}_S(C))^* \cong \text{Assume}_S(C^*)$ and $(\text{Commit}_{\setminus S}(C))^* \cong \text{Assume}_{\setminus S}(C^*)$.

Proof: Trivial. □

Claim: For all $S \subseteq W$, $C \overset{*}{\lhd_+} \text{Assume}_S(C)$ and $C \overset{*}{\lhd_+} \text{Assume}_{\setminus S}(C)$.

Proof: Trivial. □

2.3. Externalizing

Definition: Given a partition V of W, let $v : W \to V$ send each element $w \in W$ to the part that contains it. We define $\text{External}_V(C) = \text{External}^B(C)$ and $\text{External}_{/V}(C) = \text{External}^{/B}(C)$, where $B = \{\{a' \in A \mid \forall e \in E,\ v(a' \cdot e) = v(a \cdot e)\} \mid a \in A\}$.

Claim: For all partitions V of W, $\text{External}_V(C) \triangleleft_\times C$ and $\text{External}_{/V}(C) \triangleleft_\times C$. Further, $\text{External}_V(C)$ and $\text{External}_{/V}(C)$ are sisters in C.

Proof: Trivial. □

2.4. Internalizing

Definition: Given a partition V of W, let $v : W \to V$ send each element $w \in W$ to the part that contains it. We define $\text{Internal}_V(C) = \text{Internal}^F(C)$ and $\text{Internal}_{/V}(C) = \text{Internal}^{/F}(C)$, where $F = \{\{e' \in E \mid \forall a \in a,\ v(a \cdot e') = v(a \cdot e)\} \mid e \in E\}$.

Claim: For all partitions V of W, $C \triangleleft_\times \text{Internal}_V(C)$ and $C \triangleleft_\times \text{Internal}_{/V}(C)$.

Proof: Trivial. □

Claim: For all partitions V of W, $(\text{Internal}_V(C))^* \cong \text{External}_V(C^*)$, $(\text{Internal}_{/V}(C))^* \cong \text{External}_{/V}(C^*)$, $(\text{External}_V(C))^* \cong \text{Internal}_V(C^*)$, and $(\text{External}_{/V}(C))^* \cong \text{Internal}_{/V}(C^*)$.

Proof: Trivial. □

3. Basic Properties

3.1. Biextensional Equivalence

Committing and assuming are well-defined up to biextensional equivalence.

Claim: If $C_0 \simeq C_1$ are Cartesian frames over W, then for any subset $S \subseteq W$, $\text{Commit}_S(C_0) \simeq \text{Commit}_S(C_1)$, $\text{Commit}_{\setminus S}(C_0) \simeq \text{Commit}_{\setminus S}(C_1)$, $\text{Assume}_S(C_0) \simeq \text{Assume}_S(C_1)$, and $\text{Assume}_{\setminus S}(C_0) \simeq \text{Assume}_{\setminus S}(C_1)$.

Proof: Let $C_i = (A_i, E_i, \cdot_i)$, and let $(g_0, h_0) : C_0 \to C_1$ and $(g_1, h_1) : C_1 \to C_0$ compose to something homotopic to the identity in both orders. Let $B_i \subset A_i$ be defined by $B_i = \{a \in A_i \mid \forall e \in E_i,\ a \cdot_i e \in S\}$.

Observe that if $b \in B_0$, then for all $e \in E_1$, $g_0(b) \cdot_1 e = b \cdot_0 h_0(e) \in S$, so $g_0(b) \in B_1$.

Similarly, if $b \in B_1$, then $g_1(b) \in B_0$. Thus, if we let $g_i' : B_i \to B_{1-i}$ be given by $g_i'(b) = g_i(b)$, we get morphisms $(g_i', h_i') : \text{Commit}_S(C_i) \to \text{Commit}_S(C_{1-i})$, which are clearly morphisms, since they are restrictions of our original morphisms (g_i, h_i).

Since (g_0, h_0) and (g_1, h_1) compose to something homotopic to the identity in both orders, $(\text{id}_{A_i}, h_{1-i} \circ h_i) : C_i \to C_i$ is a morphism, so $(\text{id}_{B_i}, h_{1-i} \circ h_i) : \text{Commit}_S(C_i) \to \text{Commit}_S(C_i)$ is a morphism, so (g_0', h_0') and (g_1', h_1') compose to something homotopic to the identity in both orders. Thus $\text{Commit}_S(C_0) \simeq \text{Commit}_S(C_1)$.

Similarly, if $b \in A_0 \setminus B_0$, then there exists an $e \in E_0$ such that $b \cdot_0 e \notin S$. But then

$$\begin{aligned} g_0(b) \cdot_1 h_1(e) &= g_1(g_0(b)) \cdot_0 e \\ &= b \cdot_0 e \end{aligned}$$

$\notin S$, so $g_0(b) \in A_1 \setminus B_1$. Similarly, if $b \in A_1 \setminus B_1$, then $g_1(b) \in A_0 \setminus B_0$. Thus, if we let $g_i'' : A_i \setminus B_i \to A_{1-i} \setminus B_{1-i}$ be given by $g_i''(b) = g_i(b)$, we get morphisms $(g_i'', h_i'') : \text{Commit}_{\setminus S}(C_i) \to \text{Commit}_{\setminus S}(C_{1-i})$, which (similarly to above) compose to something homotopic to the identity in both orders. Thus, $\text{Commit}_{\setminus S}(C_0) \simeq \text{Commit}_{\setminus S}(C_1)$.

We know that $\text{Assume}_S(C_0) \simeq \text{Assume}_S(C_1)$ and $\text{Assume}_{\setminus S}(C_0) \simeq \text{Assume}_{\setminus S}(C_1)$, because

$$(\text{Assume}_S(C_0))^* \cong \text{Commit}_S(C_0^*) \simeq \text{Commit}_S(C_1^*) \cong \text{Assume}_S(C_1)$$

and

$$(\text{Assume}_{\setminus S}(C_0))^* \cong \text{Commit}_{\setminus S}(C_0^*) \simeq \text{Commit}_{\setminus S}(C_1^*) \cong \text{Assume}_{\setminus S}(C_1).$$

□

Externalizing and internalizing are also well-defined up to biextensional equivalence.

Claim: If $C_0 \simeq C_1$ are Cartesian frames over W, then for all partitions V of W, $\text{External}_V(C_0) \simeq \text{External}_V(C_1)$, $\text{External}_{/V}(C_0) \simeq \text{External}_{/V}(C_1)$, $\text{Internal}_V(C_0) \simeq \text{Internal}_V(C_1)$, and $\text{Internal}_{/V}(C_0) \simeq \text{Internal}_{/V}(C_1)$.

Proof: Let $C_i = (A_i, E_i, \cdot_i)$, and let $(g_0, h_0) : C_0 \to C_1$ and $(g_1, h_1) : C_1 \to C_0$ compose to something homotopic to the identity in both orders. Let V be a partition of W, and let $v : W \to V$ send each element $w \in W$ to the part that contains it. Let B_i be the partition of A_i defined by $B_i = \{\{a' \in A_i \mid \forall e \in E_i,\ v(a' \cdot e) = v(a \cdot e)\} \mid a \in A_i\}$.

Let $\beta_i : A_i \to B_i$, send each element of A_i to its part in B_i, so $\beta_i(a) = \{a' \in A_i \mid \forall e \in E_i,\ v(a' \cdot e) = v(a \cdot e)\}$. Since β_i is surjective, it has a right

inverse. Let $\alpha_i : B_i \to A_i$ be any choice of right inverse to β_i. This gives a pair of functions $\iota_i : B_i \to B_{1-i}$ given by $\iota_i = \beta_{1-i} \circ g_i \circ \alpha_i$.

We start by showing that ι_0 and ι_1 are inverses, and thus bijections between B_0 and B_1. We do this by showing that $\beta_i \circ g_{1-i} \circ g_i = \beta_i$, and that $\iota_i \circ \beta_i = \beta_{1-i} \circ g_i$, and thus we will have

$$\begin{aligned} \iota_{1-i} \circ \iota_i &= \iota_{1-i} \circ \beta_{1-i} \circ g_i \circ \alpha_i \\ &= \beta_i \circ g_{1-i} \circ g_i \circ \alpha_i \\ &= \beta_i \circ \alpha_i , \end{aligned}$$

which is the identity on B_i.

To see that $\beta_i \circ g_{1-i} \circ g_i = \beta_i$, observe that for all $a \in A_i$, we have that for all $e \in E_i$, $v(a \cdot_i e) = v(g_{1-i}(g_i(a)) \cdot_i e)$, so, $\beta_i(a) = \beta_i(g_{1-i}(g_i(a)))$. Thus, $\beta_i = \beta_i \circ g_{1-i} \circ g_i$.

To see that $\iota_i \circ \beta_i = \beta_{1-i} \circ g_i$, first observe that for all $a \in A_i$, we have that $\beta_i(\alpha_i(\beta_i(a))) = \beta_i(a)$, and thus, for all $e \in E_{1-i}$,

$$\begin{aligned} v(g_i(a) \cdot_{1-i} e) &= v(a \cdot_i h_i(e)) \\ &= v(\alpha_i(\beta_i(a)) \cdot_i h_i(e)) \\ &= v(g_i(\alpha_i(\beta_i(a))) \cdot_{1-i} e) . \end{aligned}$$

Thus, $\beta_{1-i}(g_i(a)) = \beta_{1-i}(g_i(\alpha_i(\beta_i(a))))$. Thus, we have

$$\begin{aligned} \beta_{1-i} \circ g_i &= \beta_{1-i} \circ g_i \circ \alpha_i \circ \beta_i \\ &= \iota_i \circ \beta_i . \end{aligned}$$

This also gives us functions $f_i : A_i/B_i \to A_{1-i}/B_{1-i}$, by $f_i(q) = g_i \circ q \circ \iota_{1-i}$. To ehow that these functions are well-defined, we need to show that for all $q \in A_i/B_i$, $f_i(q)$ is in fact in A_{1-i}/B_{1-i}, by showing that for all $b \in B_{1-i}$, $g_i(q(\iota_{1-i}(b))) \in b$, or equivalently that $\beta_{1-i} \circ g_i \circ q \circ \iota_{1-i}$ is the identity on B_{1-i}. Since $q \in A_i/B_i$, we already have that $\beta_i \circ q$ is the identity of B_i. Thus, we have that

$$\begin{aligned}\beta_{1-i} \circ g_i \circ q \circ \iota_{1-i} &= \iota_i \circ \beta_i \circ q \circ \iota_{1-i} \\ &= \iota_i \circ \iota_{1-i}\end{aligned}$$

is the identity on B_{1-i}.

We are now ready to demonstrate that $\mathrm{External}_V(C_0) \simeq \mathrm{External}_V(C_1)$.

Let $\mathrm{External}_V(C_i) = (A_i/B_i, B_i \times E_i, \star_i)$, and define

$$(g_i', h_i') : (A_i/B_i, B_i \times E_i, \star_i) \to (A_{1-i}/B_{1-i}, B_{1-i} \times E_{1-i}, \star_{1-i})$$

by $g_i' = f_i$, while $h_i' : B_{1-i} \times E_{1-i} \to B_i \times E_i$ is given by $h_i'(b, e) = (\iota_{1-i}(b), h_i(e))$.

To see that (g_i', h_i') is a morphism, observe that for all $q \in A_i/B_i$, and $(b, e) \in B_{1-i} \times E_{1-i}$, we have

$$\begin{aligned}g_i'(q) \star_{1-i} (b, e) &= f_i(q) \star_{1-i} (b, e) \\ &= f_i(q)(b) \cdot_{1-i} e \\ &= g_i(q(\iota_{1-i}(b))) \cdot_{1-i} e \\ &= q(\iota_{1-i}(b)) \cdot_i h_i(e) \\ &= q \star_i (\iota_{1-i}(b), h_i(e)) \\ &= q \star_i h_i'(b, e).\end{aligned}$$

To see that $(g_{1-i}', h_{1-i}') \circ (g_i', h_i')$ is homotopic to the identity, we show that

$$(\mathrm{id}_{A_i/B_i}, h_i' \circ h_{1-i}') : (A_i/B_i, B_i \times E_i, \star_i) \to (A_i/B_i, B_i \times E_i, \star_i)$$

is a morphism. Indeed, for all $q \in A_i/B_i$ and $(b, e) \in B_i \times E_i$,

$$
\begin{aligned}
q \star_i h_i'(h_{1-i}'(b, e)) &= q \star_i (b, h_i(h_{1-i}(e))) \\
&= q(b) \cdot_i h_i(h_{1-i}(e)) \\
&= q(b) \cdot_i e = q \star_i (b, e).
\end{aligned}
$$

Thus, $\mathrm{External}_V(C_0) \simeq \mathrm{External}_V(C_1)$

Similarly, let $\mathrm{External}_{/V}(C_i) = (B_i, A_i/B_i \times E_i, \diamond_i)$, and define

$$
(g_i'', h_i'') : (B_i, A_i/B_i \times E_i, \diamond_i) \to (B_{1-i}, A_{1-i}/B_{1-i} \times E_{1-i}, \diamond_{1-i})
$$

by $g_i'' = \iota_i$, and $h_i'' : A_{1-i}/B_{1-i} \times E_{1-i} \to A_i/B_i \times E_i$ is given by $h_i''(q, e) = (f_{1-i}(q), h_i(e))$.

To see that (g_i'', h_i'') is a morphism, observe that for all $q \in B_i$, and $(q, e) \in A_{1-i}/B_{1-i} \times E_{1-i}$, we have

$$
\begin{aligned}
g_i''(b) \diamond_{1-i} (q, e) &= \iota_i(b) \diamond_{1-i} (q, e) \\
&= q(\iota_i(b)) \cdot_{1-i} e \\
&= q(\iota_i(b)) \cdot_{1-i} h_{1-i}(h_i(e)) \\
&= g_{1-i}(q(\iota_i(b))) \cdot_i h_i(e) \\
&= f_{1-i}(q)(b) \cdot_i h_i(e) \\
&= b \diamond_i (f_{1-i}(q), h_i(e)) \\
&= b \diamond_i h_i''(q, e).
\end{aligned}
$$

Clearly, $(g_{1-i}'', h_{1-i}'') \circ (g_i'', h_i'')$ is homotopic to the identity, since $g_{1-i}'' \circ g_i''$ is the identity on B_i. Thus, $\mathrm{External}_{/V}(C_0) \simeq \mathrm{External}_{/V}(C_1)$.

We know that $\mathrm{Internal}_V(C_0) \simeq \mathrm{Internal}_V(C_1)$ and $\mathrm{Internal}_{/V}(C_0) \simeq \mathrm{Internal}_{/V}(C_1)$, because

$$
\begin{aligned}
(\,\mathrm{Internal}_{\mathsf{V}}\,(\,\mathsf{C}_0\,)\,)^* &\cong \mathrm{External}_{\mathsf{V}}\,(\,\mathsf{C}_0^*\,) \\
&\simeq \mathrm{External}_{\mathsf{V}}\,(\,\mathsf{C}_1^*\,) \\
&\cong \mathrm{Internal}_{\mathsf{V}}\,(\,\mathsf{C}_1\,)
\end{aligned}
$$

and

$$
\begin{aligned}
(\,\mathrm{Internal}_{/\mathsf{V}}\,(\,\mathsf{C}_0\,)\,)^* &\cong \mathrm{External}_{/\mathsf{V}}\,(\,\mathsf{C}_0^*\,) \\
&\simeq \mathrm{External}_{/\mathsf{V}}\,(\,\mathsf{C}_1^*\,) \\
&\cong \mathrm{Internal}_{/\mathsf{V}}\,(\,\mathsf{C}_1\,)\,.
\end{aligned}
$$

□

3.2. Committing and Assuming Can Be Defined Using Lollipop and Tensor

Claim: $\mathrm{Commit}_S(C) \cong 1_S \multimap C$ and $\mathrm{Assume}_S(C) \cong 1_S \otimes C$.

Proof: Let $C = (A, E, \cdot)$, and let $1_S = (\{a\}, S, \diamond)$.

Let $\mathrm{Commit}_S(C) = (B, E, \star)$, where $B = \{b \in A \mid \forall e \in e,\ b \cdot e \in S\}$, and $b \star e = b \cdot e$.

Let $1_S \multimap C = (\hom(1_S, C), \{a\} \times E, \bullet)$, where

$$
\begin{aligned}
(\,g\,,\,h\,) \bullet (\,a\,,\,e\,) &= g\,(\,a\,) \cdot e \\
&= a \diamond h\,(\,e\,) \\
&= h\,(\,e\,)\,.
\end{aligned}
$$

We construct an isomorphism $(g_0, h_0) : (1_S \multimap C) \to \mathrm{Commit}_S(C)$, by defining

$g_0 : \hom(1_S, C) \to B$ by $g_0(g, h) = g(a)$, and by defining $h_0 : E \to \{a\} \times E$ by $h_0(e) = (a, e)$

.

First, we need to show that g_0 is a well-defined function into B. Observe that for all $(g, h) \in \hom(1_S, C)$, and for all $e \in E$,

$$\begin{aligned} g_0(g, h) \cdot e &= g(a) \cdot e \\ &= h(e) \end{aligned}$$

$\in S$, and so $g_0(g, h) \in B$.

Next, we show that (g_0, h_0) is a morphism, by showing that for all $(g, h) \in \hom(1_S, C)$ and $e \in E$,

$$\begin{aligned} g_0(g, h) \star e &= g(a) \star e \\ &= g(a) \cdot e \\ &= a \diamond h(e) \\ &= (g, h) \bullet (a, e) \\ &= (g, h) \bullet h_0(e). \end{aligned}$$

Finally, to show that (g_0, h_0), we need to show that g_0 and h_0 are bijections. Clearly, h_0 is a bijection. To see that g_0 is injective, observe that if $g_0(g, h) = g_0(g', h')$, then $g(a) = g'(a)$, so $g = g'$. Further, for all $e \in E$,

$$\begin{aligned} h(e) &= a \diamond h(e) \\ &= g(a) \cdot e \\ &= g'(a) \cdot e \\ &= a \diamond h'(e) \\ &= h'(e), \end{aligned}$$

so $h = h'$. Thus g_0 is injective. To see that g_0 is surjective, observe that for all $b \in B$, there exists a morphism $(g_b, h_b) : 1_S \to C$, given by $g_b(a) = b$, and $h_b(e) = b \star e$. This is a morphism because, for all $a \in \{a\}$ and $e \in E$,

$$
\begin{aligned}
g_b(a) \star e &= b \star e \\
&= h_b(e) \\
&= a \diamond h_b(e).
\end{aligned}
$$

Since

$$
\begin{aligned}
g_0(g_b, h_b) &= g_b(a) \\
&= b,
\end{aligned}
$$

we have that g_0 is surjective, and thus (g_0, h_0) is an isomorphism between $1_S \multimap C$ and $\text{Commit}_S(C)$.

To see that $\text{Assume}_S(C) \cong 1_S \otimes C$, observe that

$$
\begin{aligned}
\text{Assume}_S(C) &\cong \\
&\quad \text{Commit}_S(C^*)^* \\
&\cong \\
&\quad (1_S \to C^*)^* \\
&\cong \\
&\quad (1_S^* \parr C^*)^* \\
&\cong \\
&\quad 1_S \otimes C.
\end{aligned}
$$

□

Recall that we can think of 1_S as a powerless agent that has been promised S. $1_S \otimes C$, then, is a team consisting of Agent(C) alongside an agent that has been promised S.

In order for these two to form a team, the promise S must still hold for the team as a whole; and since Agent(1_S) is powerless, the resultant team is exactly Agent(C) joined with the promise, i.e., $\text{Assume}_S(C)$.

$\text{Commit}_S(C) \cong 1_S \multimap C$ is less intuitive. $1_S \multimap C$ is "C with a hole in it shaped like a promise that S happens." In effect, an agent-and-hole can only "fit" such a promise into itself by being the kind of agent-and-hole that always guarantees S will happen.

It will sometimes be helpful to think about assuming and committing in terms of 1_S, as this highlights the close relationship between these operations and the other objects and operations we've been working with.[1]

4. Idempotence

We will show that all eight of the new definition from worlds are idempotent (up to isomorphism). We will do this by in each case describing the subset of Cartesian frames over W that each operation projects onto, and showing that the operation is indeed fixed on that subset.

4.1. Committing and Assuming

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and subset S of W, $\text{Commit}_S(C) \triangleleft \perp_S$ and $\text{Assume}_S(C) \triangleleft \perp_S$.

Proof: Trivial. □

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and subset S of W, with $C \triangleleft \perp_S$, $\text{Commit}_S(C) \cong \text{Assume}_S(C) \cong C$.

Proof: Trivial. □

Corollary: For any subset S of W, Commit_S and Assume_S are idempotent.

Proof: Trivial. □

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and subset S of W, if $\text{Commit}_{\setminus S} = (A', E', \cdot')$, then for all $a' \in A'$, there exists an $e' \in E'$ such that $a' \cdot' e' \notin S$.

Proof: Trivial. □

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and subset S of W, if for all $a \in A$, there exists an $e \in E$ such that $a \cdot e \notin S$, then $\text{Commit}_{\setminus S}(C) \cong C$.

Proof: Trivial. □

Corollary: For any subset S of W, $\text{Commit}_{\setminus S}$ is idempotent.

Proof: Trivial. □

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and subset S of W, if $\text{Assume}_{\setminus S}(C) = (A', E', \cdot')$, then for all $e' \in E'$, there exists an $a' \in A'$ such that $a' \cdot' e' \notin S$.

Proof: Trivial. □

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and subset S of W, if for all $e \in E$, there exists an $a \in A$ such that $a \cdot e \notin S$, then $\text{Assume}_{\setminus S}(C) \cong C$.

Proof: Trivial. □

Corollary: For any subset S of W, $\text{Assume}_{\setminus S}(C)$ is idempotent.

Proof: Trivial. □

4.2. Externalizing

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and partition V of W, let $v : W \to V$ send each element of W to its part in V. If $\text{External}_V(C) = (A', E', \cdot')$, then A' is nonempty and for all $a_0', a_1' \in A'$ and $e' \in E'$, we have $v(a_0' \cdot' e') = v(a_1' \cdot' e')$.

Proof: Let B be defined, as in the definition of External_V, as $B = \{\{a' \in A \mid \forall e \in E,\ v(a' \cdot e) = v(a \cdot e)\} \mid a \in A\}$. A' is A/B, the set of functions from B to A that sends each part in B to an element of that part, and $E' = B \times E$. A' is clearly

nonempty. Consider an arbitrary $a_0', a_1' \in A'$ and $(e, b) \in E'$. Since $a_0'(b), a_0'(b) \in b$ are in the same part, we have that $a_0'(b) \cdot e = a_0'(b) \cdot e$, and thus $v(a_0' \cdot' (b, e)) = v(a_1' \cdot' (b, e))$. □

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and partition V of W, let $v : W \to V$ send each element of W to its part in V. If A is nonempty and for all $a_0, a_1 \in A$ and $e \in E$, we have $v(a_0 \cdot e) = v(a_1 \cdot e)$, then $\text{External}_V(C) \cong C$.

Proof: Let B be defined, as in the definition of External_V, as $B = \{\{a' \in A \mid \forall e \in E,\ v(a' \cdot e) = v(a \cdot e)\} \mid a \in A\}$. If A is nonempty and for all $a_0, a_1 \in A$ and $e \in E$, we have $v(a_0 \cdot e) = v(a_1 \cdot e)$, then B has only one part, $B = \{A\}$. Thus, $\text{External}_V(C) = (A/\{A\}, \{A\} \times E, \star)$, where $\star$ is given by $q \star (A, e) = q(A) \cdot e$.

Let $(g, h) : \text{External}_V(C) \to C$ be given by $g(q) = q(A)$, and $h(e) = (A, e)$. This is trivially a morphism, and both g and h are trivially bijections, so $\text{External}_V(C) \cong C$.□

Corollary: For any partition V of W, External_V is idempotent.

Proof: Trivial. □

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and partition V of W, let $v : W \to V$ send each element of W to its part in V. If $\text{External}_{/V} = (A', E', \cdot')$, then for all $a_0' \neq a_1' \in A'$ there exists an $e' \in E'$, such that $v(a_0' \cdot' e') \neq v(a_1' \cdot' e')$.

Proof: Let B be defined once again as $B = \{\{a' \in A \mid \forall e \in E,\ v(a' \cdot e) = v(a \cdot e)\} \mid a \in A\}$. $A' = B$ and $E' = A/B \times E$. Since A/B is clearly nonempty, fix any $q \in A/B$. Observe that if $a_0' \neq a_1'$, then $q(a_0')$ and $q(a_1')$ are in

different parts in B, so there exists an $e \in E$ such that $v(q(a_0') \cdot e) \neq v(q(a_1') \cdot e)$. Thus $v(a_0' \cdot' (q, e)) \neq v(a_1' \cdot' (q, e))$. □

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and partition V of W, let $v : W \to V$ send each element of W to its part in V. If for all $a_0 \neq a_1 \in A$ there exists an $e \in E$, such that $v(a_0 \cdot e) \neq v(a_1 \cdot e)$, then $\text{External}_V(C) \cong C$.

Proof: Again, let B be defined again as $B = \{\{a' \in A \mid \forall e \in E,\ v(a' \cdot e) = v(a \cdot e)\} \mid a \in A\}$. If for all $a_0 \neq a_1 \in A$ there exists an $e \in E$ such that $v(a_0 \cdot e) \neq v(a_1 \cdot e)$, then every element of B is a singleton. Thus $A/B = \{q\}$ is a singleton, and q is a bijection.

$\text{External}_{/V}(C) = (B, \{q\} \times E, \star)$, where $\star$ is given by $b \star (q, e) = q(b) \cdot e$. Let $(g, h) : \text{External}_{/V}(C) \to C$ be given by $g(b) = q(b)$, and $h(e) = (q, e)$. This is trivially a morphism and both g and h are trivially bijections, so $\text{External}_{/V}(C) \cong C$. □

Corollary: For any partition V of W, $\text{External}_{/V}(C)$ is idempotent.

Proof: Trivial. □

4.3. Internalizing

Using duality, we also get all of the following analogous results for internalizing.

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and partition V of W, let $v : W \to V$ send each element of W to its part in V. If $\text{Internal}_V(C) = (A', E', \cdot')$, then E' is nonempty and for all $e_0', e_1' \in E'$ and $a' \in A'$, we have $v(a' \cdot' e_0') = v(a' \cdot' e_1')$.

Proof: Trivial. □

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and partition V of W, let $v : W \to V$ send each element of W to its part in V. If for all $e_0, e_1 \in E$ and $a \in A$ we have $v(a \cdot e_0) = v(a \cdot e_1)$, then $\text{Internal}_V(C) \cong C$.

Proof: Trivial. □

Corollary: For any partition V of W, $\text{Internal}_V(C)$ is idempotent.

Proof: Trivial. □

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and partition V of W, let $v : W \to V$ send each element of W to its part in V. If $\text{Internal}_{/V}(C) = (A', E', \cdot')$, then for all $e_0' \neq e_1' \in E'$ there exists an $a' \in A'$, such that $v(a' \cdot' e_0') \neq v(a' \cdot' e_0')$.

Proof: Trivial. □

Claim: For any Cartesian frame $C = (A, E, \cdot)$ over W and partition V of W, let $v : W \to V$ send each element of W to its part in V. If for all $e_0 \neq e_1 \in A$ there exists an $a \in A$, such that $v(a \cdot e_0) \neq v(a \cdot e_1)$, then $\text{Internal}_V(C) \cong C$.

Proof: Trivial. □

Corollary: For any partition V of W, $\text{Internal}_{/V}(C)$ is idempotent.

Proof: Trivial. □

Our new assuming, internalizing, and externalizing operations will also provide a new lens for us to better understand observables. We turn to this in our next post, "Eight Definitions of Observability."

Footnotes

1. This section is a good distillation of 1_S as it relates to multiplicative operations. The additive role of 1_S is quite different from this, and quite varied. There isn't a single interpretation for 1_S in additive contexts, beyond the basic interpretation we provided in "Biextensional Equivalence" that 1_S is "a powerless, all-knowing agent... plus a promise from the environment that the world will be in S." ↩

Eight Definitions of Observability

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the eleventh post in the Cartesian frames sequence. Here, we compare eight equivalent definitions of observables, which emphasize different philosophical interpretations.

Throughout this post, we let $C = (A, E, \cdot)$ be a Cartesian frame over a nonempty set W, we let $V = \{S_1, \dots, S_n\}$ be a finite partition of W, and we let $v : W \to V$ send each element of W to its part in V.

The condition that V is finite is an important one. Many of the definitions below can be extended to infinite partitions, and the theory of observability for infinite partitions is probably nice, but we are not discussing it here. The condition that W is nonempty is just ruling out some degenerate cases

1. Definition from Subsets

The definitions in this post will talk about when a finite partition V of W is observable in C. This will make some of the definitions more elegant, and it is easy to translate back and forth between the new definitions of the observability of a finite partition and the old definitions of the observability of a subset.

Definition: We say C's agent can observe a finite partition V of W if for all parts $S_i \in V$, $S_i \in \text{Obs}(C)$. We let $\text{Obs}'(C)$ denote the set of all finite partitions of W that are observable in C.

Claim: For any nonempty strict subset $S \subset W$, C's agent can observe S if and only if C'sagent can observe $\{S, (W \setminus S)\}$.

Proof: If C's agent can observe $\{S, (W \setminus S)\}$, then clearly C's agent can observe S. If C'sagent can observe S, then since observability is closed under complements, C's agent can observe $W \setminus S$, and so can observe $\{S, (W \setminus S)\}$. □

1.1. Example

In "Introduction to Cartesian Frames," we gave the example of an agent that can choose between unconditionally carrying an umbrella, unconditionally carrying no umbrella, carrying an umbrella iff it's raining, and carrying an umbrella iff it's sunny:

$$C_0 = \begin{array}{c} \\ u \\ n \\ u \leftrightarrow r \\ u \leftrightarrow s \end{array} \begin{array}{c} \begin{array}{cc} r & s \end{array} \\ \begin{pmatrix} ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{pmatrix} \end{array}$$

Here, $\text{Obs}(C_0) = \{\{\}, \{ur, nr\}, \{us, ns\}, W\}$, so the partition $V = \{R, S\}$ is observable in C_0, where $R = \{ur, nr\}$ and $S = \{us, ns\}$.

As we go through the definitions in this post, we will repeatedly return to C_0 and show how to understand C_0's observables in terms of our new definitions.

Before presenting fundamentally new definitions, we will modify our two old definitions to be about finite partitions instead of subsets.

2. Conditional Policies Definition

Definition: We say that C's agent can observe a finite partition V of W if for all functions $f : V \to A$, there exists an element $a_f \in A$ such that for all $e \in E$, $f(v(a_f \cdot e)) \cdot e = a_f \cdot e$.

Claim: This definition is equivalent to the definition from subsets.

Proof: We work by induction on the number of parts in V. Since W is nonempty, V has at least one part. If $V = \{W\}$ has one part, we clearly have that C's agent can observe V under the definition from subsets. For the conditional policies definition, we also have that C's agent can observe V, since we can take $a_f = f(W)$, and thus, for all $e \in E$,

$$\begin{aligned} f(v(a_f \cdot e)) \cdot e &= f(W) \cdot e \\ &= a_f \cdot e. \end{aligned}$$

If $V = \{S_1, \ldots, S_n\}$ has n parts, consider the partition $V' = \{S_1 \cup S_2, S_3, \ldots, S_n\}$ which unions together the first two parts S_1 and S_2 of V. Let $v' : W \to V'$ send each element of W to its part in V'.

First, assume that C's agent can observe V according to the definition from subsets. Then, since observability of subsets is closed under unions, C's agent can also observe V' under the definition from subsets, and thus also under the conditional policies definition.

Given a function $f : V \to A$, let $f' : V' \to A$ be given by $f'(S_1 \cup S_2) = f(S_2)$, and $f'(S_i) = f(S_i)$ on all other inputs. Since C's agent can observe V' under the conditional policies definition, we can let $a_{f'}$ be such that for all $e \in E$, $f'(v'(a_{f'} \cdot e)) \cdot e = a_{f'} \cdot e$.

Choose an $a_f \in A$ such that $a_f \in \text{if}(S_1, f(S_1), a_f')$, which we can do because S_1 is observable in C. Observe that for all $e \in E$, we have that if $a_f \cdot e \in S_1$, then

$$\begin{aligned} f(v(a_f \cdot e)) \cdot e &= f(S_1) \cdot e \\ &= a_f \cdot e, \end{aligned}$$

if $a_f \cdot e \in S_2$, we have $a_f \cdot e = a_{f'} \cdot e$, and thus

$$\begin{aligned} f(v(a_f \cdot e)) \cdot e &= f(S_2) \cdot e \\ &= f'(S_1 \cup S_2) \cdot e \\ &= f'(v'(a_{f'} \cdot e)) \cdot e \\ &= a_{f'} \cdot e \\ &= a_f \cdot e, \end{aligned}$$

and finally if $a_f \cdot e \in S_i$ for some $i \neq 1, 2$, we still have have $a_f \cdot e = a_{f'} \cdot e$, and thus

$$\begin{aligned} f(v(a_f \cdot e)) \cdot e &= f(S_i) \cdot e \\ &= f'(S_i) \cdot e \\ &= f'(v'(a_{f'} \cdot e)) \cdot e \\ &= a_{f'} \cdot e \\ &= a_f \cdot e. \end{aligned}$$

Thus, C's agent can observe V according to the conditional policies definition.

Conversely, if C's agent can observe V according to the conditional policies definition, then to show that C's agent can observe V according to the definition from subsets, it suffices to show that the agent can observe S_i for all $S_i \in V$. Thus, we need to show that for any $a_0, a_1 \in A$, there exists an $a_2 \in A$ with $a_2 \in \text{if}(S_i, a_0, a_1)$.

Indeed, if we let $f : V \to A$ send S_i to a_0, and send all other inputs to a_1, then we can take an a_f such that for all $e \in E$, $f(v(a_f \cdot e)) \cdot e = a_f \cdot e$. But then, if $a_f \cdot e \in S_i$, then

$$\begin{aligned} a_f \cdot e &= f(v(a_f \cdot e)) \cdot e \\ &= f(S_1) \cdot e \\ &= a_0 \cdot e, \end{aligned}$$

and otherwise,

$$\begin{aligned} a_f \cdot e &= f(v(a_f \cdot e)) \cdot e \\ &= a_1 \cdot e. \end{aligned}$$

Thus, C's agent can observe V according to the definition from subsets. □

2.1. Example

Let $C_0 = (A, E, \cdot)$ be defined as in the §1.1 example, with $R = \{ur, nr\}$, $S = \{us, ns\}$, and $V = \{R, S\}$.

$A = \{u, n, u \leftrightarrow r, u \leftrightarrow s\}$ is a four-element set, and $V = \{R, S\}$ is a two-element set, so there are sixteen functions $f : V \to A$. For each function, there is a possible agent $a_f \in A$ that satisfies $f(v(a_f \cdot e)) \cdot e = a_f \cdot e$ for all $e \in E$. We can illustrate the sixteen functions and the corresponding $a_f \in A$ in a sixteen-row table:

$f(R)$	$f(S)$	a_f
u	u	u
u	n	$u \leftrightarrow r$
u	$u \leftrightarrow r$	$u \leftrightarrow r$
u	$u \leftrightarrow s$	u
n	u	$u \leftrightarrow s$
n	n	n
n	$u \leftrightarrow r$	n
n	$u \leftrightarrow s$	$u \leftrightarrow s$
$u \leftrightarrow r$	u	u
$u \leftrightarrow r$	n	$u \leftrightarrow r$
$u \leftrightarrow r$	$u \leftrightarrow r$	$u \leftrightarrow r$
$u \leftrightarrow r$	$u \leftrightarrow s$	u
$u \leftrightarrow s$	u	$u \leftrightarrow s$
$u \leftrightarrow s$	n	n
$u \leftrightarrow s$	$u \leftrightarrow r$	n
$u \leftrightarrow s$	$u \leftrightarrow s$	$u \leftrightarrow s$

Since there is an $a_f \in A$ for each function, C_0's agent can observe V according to the conditional policies definition.

3. Additive Definitions

Next, we give an additive definition of observables. This is a version of our categorical definition of observables from "Controllables and Observables, Revisited," modified to be about finite partitions.

Definition: We say C's agent can observe a finite partition $V = \{S_1, \ldots, S_n\}$ of W if there exist $C_1, \cdots C_n$, Cartesian frames over W, with $C_i \triangleleft \perp_{S_i}$ such that $C \simeq C_1 \& \ldots \& C_n$.

This can also be strengthened to a constructive version of the additive definition, which we will call the assuming definition.

Definition: We say C's agent can observe a finite partition $V = \{S_1, \dots, S_n\}$ of W if $C \simeq \text{Assume}_{S_1}(C) \& \dots \& \text{Assume}_{S_n}(C)$.

Claim: These definitions are equivalent to each other and the definitions above.

Proof: We assume that $n \geq 2$, and that A is nonempty. The case where $n = 1$ and the case where $A = \{\}$ are trivial.

If C's agent can observe V according to the assuming definition of observables, then it can also clearly observe V according to the additive definition, since $\text{Assume}_{S_1}(C) \triangleleft \perp_{S_1}$.

Next, assume that C's agent can observe V according to the additive definition. We will show that C's agent can observe S_1. Consider the pair of Cartesian frames C_1 and $C_2 \& \dots \& C_n$. Observe that $C_1 \triangleleft \perp_{S_1}$ and that $C_2 \& \dots \& C_n \triangleleft \perp_{W \setminus S_1}$, and that $C \simeq C_1 \& (C_2 \& \dots \& C_n)$. Thus, S_1 is observable in C. Symmetrically, S_i is observable in C for all $i = 1, \dots n$, and thus V is observable in C according to the definition from subsets.

Finally, assume that C's agent can observe V according to the conditional policies definition (and also the definition from subsets). We will show that $C \simeq C_1 \& \dots \& C_n$, where $C_i = \text{Assume}_{S_i}(C)$.

We have $C_1 \& \dots \& C_n = (A^n, E_1 \sqcup \dots \sqcup E_n, \star)$, where $C_i = (A, E_i, \cdot_i)$, and $\star$ is given by $(a_1, \dots, a_n) \star e = a_i \cdot e$, where $e \in E_i$.

First observe that for every $e \in E$, there is a unique $i \in \{1, \dots, n\}$ such that $e \in E_i$. This is because there exists an $a_0 \in A$, and from the definition from subsets, C's agent can observe each S_i, and so given an $e \in E$, if $a_0 \cdot e \in S_i$, it must be the case that for all $a \in A$, $a \cdot e \in S_i$. Thus, we have that that $E = E_1 \sqcup \dots \sqcup E_n$.

We construct $(g_0, h_0) : (A^n, E, \star) \to C$ and $(g_1, h_1) : C \to (A^n, E, \star)$ which compose to something homotopic to the identity in each order. Let $g_1 : A \to A^n$ be the diagonal, given by $g_1(a) = (a, \dots, a)$. Let h_0 and h_1 be the identity on E. Let g_0 be given by $g_0(a_1, \dots, a_n) = a_f$, where $f : V \to A$ is given by $f(S_i) = a_i$, and a_f satisfies $f(v(a_f \cdot e)) \cdot e = a_f \cdot e$ for all $e \in E$, which is possible by the conditional policies definition.

To see that (g_1, h_1) is a morphism, observe that for all $a \in A$ and $e \in E$,

$$\begin{aligned} g_1(a) \star e &= (a, \dots, a) \star e \\ &= a \cdot e \\ &= a \cdot h_1(e). \end{aligned}$$

To see that (g_0, h_0) is a morphism, observe that for all $(a_1, \dots, a_n) \in A$, and $e \in E$, if we let $f : V \to A$ be given by $f(S_i) = a_i$, we have

$$\begin{aligned} g_0(a_1, \dots, a_n) \cdot e &= f(v(g_0(a_1, \dots, a_n) \cdot e)) \cdot e \\ &= f(S_i) \cdot e \\ &= a_i \cdot e \\ &= (a_1, \dots, a_n) \star e \\ &= (a_1, \dots, a_n) \star h_0(e), \end{aligned}$$

where i is such that $e \in E_i$. The fact that (g_0, h_0) and (g_1, h_1) compose to something homotopic to the identity in both orders follows from the fact that $h_0 \circ h_1$ and $h_1 \circ h_0$ are the identity on E. Thus, $C \simeq \text{Assume}_{S_1}(C) \& \dots \& \text{Assume}_{S_n}(C)$, and so V is observable in C according to the assuming definition. □

3.1. Example

Let C_0 be defined as in the previous examples, with $R = \{ur, nr\}$ and $S = \{us, ns\}$. By the assuming definition, there exist two frames

$$C_1 = \text{Assume}_R(C_0) = \begin{array}{c} \\ u \\ n \end{array} \begin{array}{c} r \\ \begin{pmatrix} ur \\ nr \end{pmatrix} \end{array}$$

and

$$C_2 = \text{Assume}_S(C_0) = \begin{array}{c} \\ u \\ n \end{array} \begin{array}{c} s \\ \begin{pmatrix} us \\ ns \end{pmatrix} \end{array}$$

such that $C_0 \simeq C_1 \& C_2$.

This example both illustrates the idea behind the additive definitions, and shows the construction used in the assuming definition. This is also the same example we provided to illustrate products of Cartesian frames in "Additive Operations on Cartesian Frames."

Another way of thinking about the additive definition of observables: Recall "Committing, Assuming, Externalizing, and Internalizing" §3.2 (Committing and Assuming Can Be Defined Using Lollipop and Tensor), where we saw that $\text{Assumes}_S(C) \cong 1_S \otimes C$.

This means that (up to isomorphism) we can restate $C_0 \simeq C_1 \& C_2$ as $C_0 \simeq (1_R \otimes C_0) \,\&\, (1_S \otimes C_0)$, i.e.,

$$\begin{pmatrix} ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{pmatrix} \simeq \begin{pmatrix} ur & nr \end{pmatrix} \otimes \begin{pmatrix} ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{pmatrix} \,\&\, \begin{pmatrix} us & ns \end{pmatrix} \otimes \begin{pmatrix} ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{pmatrix}.$$

This (equivalent) framing makes it easier to keep track of what "assuming" is doing categorically, so that we can see what interfaces between frames we are relying on when we say that something is "observable" using an additive definition.

4. Multiplicative Definitions

Our multiplicative definitions will depend on a notion of agents being powerless outside of a subset.

4.1. Powerless Outside of a Subset

Definition: Given a subset S of W, we say that C's agent is powerless outside S if for all $e \in E$, and all $a_0, a_1 \in A$, if $a_0 \cdot e \notin S$, then $a_0 \cdot e = a_1 \cdot e$.

To say that C's agent is powerless outside S is to say that the if the world is at all dependent on C's agent, then the world must be in S.

Here are some lemmas about being powerless outside of a subset, which we will use later.

Lemma: If C's agent is powerless outside S and $T \supseteq S$, then C's agent is powerless outside T.

Proof: Trivial. □

Lemma: If C and D's agents are both powerless outside S, then $C \otimes D$'s agent is powerless outside S.

Proof: Let $D = (B, F, \star)$, and let $C \otimes D = (A \times B, \hom(C, D^*), \diamond)$. Consider some $(a_0, b_0), (a_1, b_1) \in A \times B$ and $(g, h) \in \hom(C, D^*)$. We will use the fact that if $a_0 \cdot h(b_0) \notin S$ then $a_0 \cdot h(b_0) = a_1 \cdot h(b_0)$, and the fact that if $b_0 \star g(a_1) \notin S$ then $b_0 \star g(a_1) = b_1 \star g(a_1)$.

Observe that if $(a_0, b_0) \diamond (g, h) \notin S$, then

$$
\begin{aligned}
(a_0, b_0) \diamond (g, h) &= a_0 \cdot h(b_0) \\
&= a_1 \cdot h(b_0) \\
&= b_0 \star g(a_1) \\
&= b_1 \star g(a_1) \\
&= (a_1, b_1) \star (g, h).
\end{aligned}
$$

□

Now, we are ready for our first truly new definition of the observability of a finite partition.

4.2. Multiplicative Definitions of Observables

Definition: We say that C's agent can observe a finite partition $V = \{S_1, \ldots, S_n\}$ of W if $C \simeq C_1 \otimes \cdots \otimes C_n$, where each C_i's agent is powerless outside S_i.

Again, we also have a constructive version of this definition:

Definition: We say that C's agent can observe a finite partition $V = \{S_1, \ldots, S_n\}$ of W if $C \simeq C_1 \otimes \cdots \otimes C_n$, where $C_i = \text{Assume}_{S_i}(C)\&1_{T_i}$, where $T_i = (W \setminus S_i) \cap \text{Image}(C)$.

Claim: These definitions are equivalent to each other and equivalent to the definitions above.

Proof: First, observe that if C's agent can observe V according to the constructive version of the multiplicative definition, it can also observe V according to the nonconstructive version of the multiplicative definition, since the agent of $\text{Assume}_{S_i}(C)\&1_{T_i}$ is clearly powerless outside S_i.

Next, we show that if C's agent can observe V according to the nonconstructive multiplicative definition, it can also observe V according to the definition from subsets. Let $C \simeq D = C_1 \otimes \cdots \otimes C_n$, where each C_i's agent is powerless outside S_i. It suffices to show that D's agent can observe V, since the definition from subsets is equivalent to the additive definition, and thus closed under biextensional equivalence. Thus, it suffices to show that D's agent can observe S_i for all $i = 1, \ldots, n$. We will show that D's agent can observe S_1, and the rest will follows by symmetry.

Let $C_1 = (A_1, E_1, \cdot_1)$, and let $D_1 = (B_1, F_1, \star_1) = C_2 \otimes \cdots \otimes C_n$. We start by showing that D_1's agent is powerless outside $W \setminus S_1$. We have that the agents of $C_2, \ldots, C_n$ are all powerless outside $W \setminus S_1$, since being powerless outside something is closed under supersets. Thus we have that D_1's agent is powerless outside $W \setminus S_1$, since being powerless outside $W \setminus S_1$ is closed under tensor.

Thus, we have $D = (A_1 \times B_1, \hom(C, D^*), \diamond) = C_1 \otimes D_1$, with C_1's agent powerless outside S_1 and D_1's agent powerless outside $W \setminus S_1$. Given an arbitrary $(a_1, b_1), (a_2, b_2) \in A_1 \times B_1$, we will show that $(a_1, b_2) \in \text{if}(S_1, (a_1, b_1), (a_2, b_2))$, and thus show that D 's agent can observe S_1.

It suffices to show that for all $(g, h) : C \to D^*$, if $(a_1, b_2) \diamond (g, h) \in S_1$, then $(a_1, b_2) \diamond (g, h) = (a_1, b_1) \diamond (g, h)$, and if $(a_1, b_2) \diamond (g, h) \notin S_1$, then $(a_1, b_2) \diamond (g, h) = (a_2, b_2) \diamond (g, h)$. Indeed, if $(a_1, b_2) \diamond (g, h) \in S_1$, then, since D_1's agent is powerless outside $W \setminus S_1$, we have

$$
\begin{aligned}
(a_1, b_2) \diamond (g, h) &= b_2 \star_1 g(a_1) \\
&= b_1 \star_1 g(a_1) \\
&= (a_1, b_1) \diamond (g, h).
\end{aligned}
$$

Similarly, if $(a_1, b_2) \diamond (g, h) \notin S_1$, then, since C_1's agent is powerless outside S, we have

$$\begin{aligned}(a_1, b_2) \diamond (g, h) &= a_1 \cdot_1 h(b_2) \\ &= a_2 \cdot_1 h(b_2) \\ &= (a_2, b_2) \diamond (g, h).\end{aligned}$$

Thus, D's agent can observe S_1, so C's agent can observe V according to the definition from subsets.

Finally, we assume that C's agent can observe V according to the assuming definition, and show that C's agent can observe V according to the constructive version of the multiplicative definition.

We work by induction on n, the number of parts. The case where n = 1 is trivial. Let $C \simeq \text{Assume}_{S_1}(C) \& \ldots \& \text{Assume}_{S_n}(C)$. Thus, we also have that $C \simeq \text{Assume}_{S_1 \cup S_2}(C) \& \text{Assume}_{S_3}(C) \& \ldots \& \text{Assume}_{S_n}(C)$, and so by induction, we have that $C \simeq (\text{Assume}_{S_1 \cup S_2}(C) \& 1_{T_1 \cap T_2}) \otimes C_3 \otimes \cdots \otimes C_n$, where C_i and T_i are as in the constructive multiplicative definition. Thus, it suffices to show that

$$\begin{aligned}\text{Assume}_{S_1 \cup S_2}(C) \& 1_{T_1 \cap T_2} &\simeq C_1 \otimes C_2 \\ &= (\text{Assume}_{S_1}(C) \& 1_{T_1}) \otimes (\text{Assume}_{S_2}(C) \& 1_{T_2}).\end{aligned}$$

First, observe that we have $C \simeq D_1 \& D_2 \& D_3$, where $D_1 = \text{Assume}_{S_1}(C)$, $D_2 = \text{Assume}_{S_2}(C)$, and $D_3 = \text{Assume}_{S_3}(C) \& \ldots \& \text{Assume}_{S_n}(C)$. Let $D_i = (B_i, F_i, \star_i)$. Let $R_i = \text{Image}(D_i)$.

Observe that $T_1 = R_2 \cup R_3$, $T_2 = R_1 \cup R_3$, and $T_1 \cup T_2 = R_3$, and observe that $\text{Assume}_{S_1 \cup S_2}(C) \simeq D_1 \& D_2$. Thus it suffices to show that $(D_1 \& 1_{R_2 \cup R_3}) \otimes (D_2 \& 1_{R_1 \cup R_3}) \simeq D_1 \& D_2 \& 1_{R_3}$.

Let $D_1 \& 1_{R_2 \cup R_3} = (B_1, F_1 \sqcup R_2 \sqcup R_3, \bullet_1)$, let $D_2 \& 1_{R_1 \cup R_3} = (B_2, F_2 \sqcup R_1 \sqcup R_3, \bullet_2)$, and let $D_1 \& D_2 \& 1_{R_3} = (B_1 \times B_2, F_1 \sqcup F_2 \sqcup R_3, \bullet_3)$ where $\bullet_1$, $\bullet_2$, and $\bullet_3$ are all given by $b \bullet_i f = b \star_1 f$ if $f \in F_1$, $b \bullet_i f = b \star_2 f$ if $f \in F_2$, and $b \bullet_i f = f$ otherwise.

Let $H = \hom(B_1 \& 1_{R_2 \cup R_3}, (D_2 \& 1_{R_1 \cup R_3})^*)$. Let $(D_1 \& 1_{R_2 \cup R_3}) \otimes (D_2 \& 1_{R_1 \cup R_3}) = (B_1 \times B_2, H, \bullet_4))$, where

$$\begin{aligned}(b_1, b_2) \bullet_4 (g, h) &= b_1 \bullet_1 h(b_2) \\ &= b_2 \bullet_2 g(b_1).\end{aligned}$$

Observe that for any $f_1 \in F_1$, there is a $(g_{f_1}, h_{f_1}) \in H$, given by $g_{f_1}(b_1) = b_1 \cdot_1 f_1$ and $h_{f_1}(b_2) = f_1$. This is clearly a morphism, since

$$\begin{aligned}b_1 \bullet_1 h_{f_1}(b_2) &= b_1 \bullet_1 f_1 \\ &= b_1 \cdot_1 f_1 \\ &= g_{f_1}(b_1) \\ &= b_2 \bullet_2 g_{f_1}(b_1).\end{aligned}$$

Similarly, for any $f_2 \in F_2$, there is a morphism $(g_{f_2}, h_{f_2}) \in H$ given by $g_{f_2}(b_1) = f_2$ and $h_{f_2}(b_2) = b_2 \cdot_2 f_2$. Finally, for any $r \in R_3$, there is a morphism $(g_r, h_r) \in H$, given by $g_r(b_1) = h_r(b_2) = r$, which is also clearly a morphism.

We show that these are in fact all of the morphisms in H. Indeed, let (g, h) be a morphism in H, let b_1 be an element of B_1, and let b_2 be an element of b_2. Let

$$\begin{aligned}r &= b_2 \bullet_2 g(b_1) \\ &= b_1 \bullet_1 h(b_2).\end{aligned}$$

If $r \in R_3$, then $g(b_1) = h(b_2) = r$, so given any $b_1 \in B_1$,

$$\begin{aligned} b_2' \bullet_2 g(b_1') &= b_1' \bullet_1 h(b_2') \\ &= r \end{aligned}$$

$\in R_3$, and so

$$\begin{aligned} g(b_1') &= b_2' \bullet_2 g(b_1') \\ &= r. \end{aligned}$$

Similarly, for any $b_2' \in B_2$, $h(b_2') = r$ and so $(g, h) = (g_r, h_r)$.

If $r \in R_1$, then $g(b_1') = r$, and $h(b_2') \in F_1$. Let $f_1 = h(b_2')$. Given any $b_1' \in B_1$,

$$\begin{aligned} b_2' \bullet_2 g(b_1') &= b_1' \bullet_1 h(b_2') \\ &= b_1' \bullet_1 f_1 \end{aligned}$$

$\in R_1$, so

$$\begin{aligned} g(b_1') &= b_2' \bullet_2 g(b_1') \\ &= b_1' \bullet_1 f_1 \\ &= b_1' \cdot_1 f_1. \end{aligned}$$

Given any $b_2' \in B_2$,

$$\begin{aligned} b_1' \bullet_1 h(b_2') &= b_2' \bullet_2 g(b_1') \\ &= b_2' \bullet_2 r \\ &= r \end{aligned}$$

$\in R_1$, and so

$$\begin{aligned} h(b_2') &= b_1' \bullet_1 h(b_2') \\ &= r. \end{aligned}$$

Thus,

$$(g, h) = ($$

.

Finally, if $r \in R_2$, we similarly have $(g, h) = (g_{f_2}, h_{f_2})$, where $f_2 = g(b_1) \in F_2$.

We construct a pair of morphisms

$$(g_0, h_0) : (B_1 \times B_2, H, \bullet_4) \to (B_1 \times B_2, F_1 \sqcup F_2 \sqcup R_3, \bullet_3)$$

and

$$(g_1, h_1) : (B_1 \times B_2, F_1 \sqcup F_2 \sqcup R_3, \bullet_3) \to (B_1 \times B_2, H, \bullet_4),$$

by letting g_0 and g_1 be the identity on $B_1 \times B_2$, letting $h_0 : F_1 \sqcup F_2 \sqcup R_3 \to H$ be given by $h_0(f) = (g_f, h_f)$ as above. Since we have shown that h_0 is surjective, we let h_1 be any right inverse to h_0. It is easy to show that both of these are morphisms by the

construction of (g_f, h_f), and they compose to something homotopic to the identity in both orders since $g_0 \circ g_1$ and $g_1 \circ g_0$ are the identity of $B_1 \times B_2$.

Thus $(D_1 \& 1_{R_2 \cup R_3}) \otimes (D_2 \& 1_{R_1 \cup R_3}) \simeq D_1 \& D_2 \& 1_{R_3}$, so C's agent can observe V according to the constructive multiplicative definition, completing the proof. □

You may have noticed that the last part of the proof would have been much simpler if ⊗ distributed over &, but ⊗ does not in general distribute over &. (⊗ distributes over ⊕ and ⅋ distributes over &.)

In this case, however, ⊗ does distribute over &. I do not plan on going over it now, but there is actually an interesting relationship between observables and cases where ⊗ distributes over &.

4.3. Example

Let C_0 be defined as in the previous examples, with $R = \{ur, nr\}$ and $S = \{us, ns\}$. Let $T_X = (W \setminus X) \cap \text{Image}(C_0)$, so that $1_{T_R} = 1_S$ and $1_{T_S} = 1_R$. By the multiplicative definitions of observables, there then exist two frames

$$C_1 = \text{Assume}_R(C) \;\&\; 1_S = \begin{array}{c} \\ r \to u \\ r \to n \end{array} \begin{array}{c} \begin{array}{ccc} r & u\,s & n\,s \end{array} \\ \begin{pmatrix} u\,r & u\,s & n\,s \\ n\,r & u\,s & n\,s \end{pmatrix} \end{array}$$

and

$$C_2 = \text{Assume}_S(C) \;\&\; 1_R = \begin{array}{c} \\ s \to u \\ s \to n \end{array} \begin{array}{c} \begin{array}{ccc} s & u\,r & n\,r \end{array} \\ \begin{pmatrix} u\,s & u\,r & n\,r \\ n\,s & u\,r & n\,r \end{pmatrix} \end{array}$$

such that $C_0 \simeq C_1 \otimes C_2$.

Here, C_1 is an agent that treats the "makes decisions when it's sunny" part of itself as though it were an external process. Similarly, C_2 externalizes its ability to make decisions when it's rainy.

This example illustrates both multiplicative definitions, and also shows the construction used in the constructive multiplicative definition.

Appealing again to the fact that $\text{Assume}_S(C) \cong 1_S \otimes C$, we also have the option of restating $C_0 \simeq C_1 \otimes C_2$ here as $C_0 \simeq ((1_R \otimes C_0) \,\&\, 1_S) \otimes ((1_S \otimes C_0) \,\&\, 1_R)$. In words, this says that $\text{Agent}(C_0)$ is (biextensionally equivalent to) a team consisting of:

1. that very agent, picking an action after the environment either (a) gives it a promise it will rain or (b) makes it powerless and doesn't rain; and
2. that very agent, picking an action after the environment either (a) gives it a promise it won't rain or (b) makes it powerless and rains.

4.4. Updatelessness

The relationship between observables' additive and multiplicative definitions is interesting. You can think of the additive definition as updateful, while the multiplicative definition is updateless.

The C_i in the additive definition are basically given a promise that the world will end up in S_i. The C_i in the multiplicative definition, however, are instead given a promise that their choices have no effect on worlds outside of S_i.

I think the updateless factorization is better, and thus prefer the multiplicative definition in spite of the fact that it is more complicated.

When an updateless agent observes something, it becomes the version of itself that only affects the worlds in which it makes that observation. When an updateful agent observes something, we assume that all the worlds in which it does not make that observation do not exist. The fact that the additive and multiplicative definitions above are equivalent illustrates the equivalence of the updateful and updateless views in the simple cases where there is true observation. However, they diverge as soon as you

want to try to approximate observation. The updateless view approximates better, as it makes sense to think of a subagent that has only a very small effect on worlds in which it does not make the observation that it makes.

Also, note that the C_i in the additive definition are not subagents of C, but they are additive sub-environments. The C_i in the multiplicative definition are multiplicative subagents of C.

5. Internalizing-Externalizing Definitions

Next, we have the nonconstructive internalizing-externalizing definition of observables.

Definition: We say that C's agent can observe a finite partition V of W if either $A = \{\}$ or C is biextensionally equivalent to something in the image of $\text{External}_V \circ \text{Internal}_V$.

Again, we have a constructive version of this definition.

Definition: We say that C's agent can observe a finite partition V of W if either $A = \{\}$ or $C \simeq \text{External}_V(\text{Internal}_V(C))$.

Claim: These definitions are equivalent to each other and to the definitions above.

Proof: The case where $A = \{\}$ is trivial, so we assume that A is nonempty. Clearly if C's agent can observe V under the constructive internalizing-externalizing definition, then V is also observable in C under the non-constructive version.

Next, assume that C is in the image of $\text{External}_V \circ \text{Internal}_V$ (up to biextensional equivalence). Recall that the image of Internal_V up to biextensional equivalence is exactly those Cartesian frames $(B, F, \star)$ such then F is nonempty and for all $f_0, f_1 \in F$ and $b \in B$, we have $v(b \star f_0) = v(b \star f_1)$. Thus, $C \simeq \text{External}_V(B, F, \star)$, where $(B, F, \star)$ is of this form. Let $v_B : B \to V$ send each element $b \in B$ to the unique $v_b \in V$ such that $v(b \star f) = v_b$ for all $f \in F$, and let V_B be the image of v_B. Then, $\text{External}_V(B, F, \star) = (B/X, X \times F, \diamond)$, where $X = \{\{b \in B \mid v_B(b) = v'\} \mid v' \in V_B\}$, and $q \diamond (x, f) = q(x) \star f$.

Let $V_B = \{v_1, \ldots, v_m\}$, and let $B_i = \{b \in B \mid v_B(b) = v_i\}$. Then, we clearly have that $\text{External}_V(B, F, \star) \cong (B_1 \times \cdots \times B_m, V_B \times F, \bullet)$, where $(b_1, \ldots, b_m) \bullet (v_i, f) = b_i \star f$. But this is clearly isomorphic to ${}^{D_1}\&\ldots\&D_m$, where $D_i = (B_i, F, \star_i)$, where $b \star_i f = b \star f$. Thus, C 's agent can observe V according to the nonconstructive additive definition of observables.

Finally, we assume that C's agent can observe V according to the nonconstructive additive definition of observables, and we show that C's agent can observe V according to the constructive internalizing-externalizing definition. Let $C \simeq C_1 \& \ldots \& C_n$, where $C_i \triangleleft \perp_{S_i}$. Let $C_i = (A_i, E_i, \cdot_i)$, and without loss of generality, let $C = C_1 \& \ldots \& C_n = (A, E, \cdot)$, where $A = A_1 \times \cdots \times A_n$ and $E = E_1 \sqcup \cdots \sqcup E_n$.

First, we show that $\text{Internal}_V(C) \simeq C_1 \oplus \cdots \oplus C_n$. Let $C_1 \oplus \cdots \oplus C_n = (A_1 \sqcup \cdots \sqcup A_n, E_1 \times \cdots \times E_n, \star)$. Observe that (since A is nonempty), $\text{Internal}_V(C) \cong (A \times F, B/F, \star')$, where $F = \{E_1, \ldots, E_n\}$, where $(a, f) \star q = a \cdot q(f)$.

We construct

$$(g_0, h_0) : (A_1 \sqcup \cdots \sqcup A_n, E_1 \times \cdots \times E_n, \star) \to (A \times F, B/F, \star')$$

and

$$(g_1, h_1) : (A \times F, B/F, \star') \to (A_1 \sqcup \cdots \sqcup A_n, E_1 \times \cdots \times E_n, \star)$$

as follows. Let $g_1((a_1, \ldots, a_n), E_i) = a_i$. Let $g_0(a_i) = ((a_1, \ldots, a_i, \ldots, a_n), E_i)$, where $a_i \in A_i$, and $a_j \in A_j$ is chosen arbitrarily for $j \neq i$. Let $h_0(q) = (q(E_1), \ldots, q(E_n))$, and $h_1(e_1, \ldots, e_n) = q$, where $q(E_i) = e_i$. Clearly, h_0 and h_1 are inverses.

To see that (g_0, h_0) is a morphism, observe that for all $a_i \in A_1 \sqcup \cdots \sqcup A_n$ and $q \in B/F$, we have

$$
\begin{aligned}
g_0(a_i) \star' q &= ((a_1, \ldots, a_i, \ldots, a_n), E_i) \star' q \\
&= (a_1, \ldots, a_i, \ldots, a_n) \cdot q(E_i) \\
&= a_i \cdot_i q(E_i) \\
&= a_i \star (q(E_1), \ldots, q(E_n)) \\
&= a_i \star h_0(q),
\end{aligned}
$$

where $a_i \in A_i$.

To see that (g_1, h_1) is a morphism, observe that for all $((a_1, \ldots, a_n), E_i) \in A \times F$, and for all $(e_1, \ldots e_n) \in E_1 \times \cdots \times E_n$, we have

$$
\begin{aligned}
g_1((a_1, \ldots, a_n), E_i) \star (e_1, \ldots, e_n) &= a_i \star (e_1 \ldots, e_n) \\
&= a_i \cdot_i e_i \\
&= (a_1, \ldots, a_n) \cdot e_i \\
&= (a_1, \ldots, a_n) \cdot h_1(e_1, \ldots, e_n)(E_i) \\
&= ((a_1, \ldots, a_n), E_i) \star h_1(e_1, \ldots, e_n).
\end{aligned}
$$

It is clear that $(g_0, h_0) \circ (g_1, h_1)$ and $(g_1, h_1) \circ (g_0, h_0)$ are both homotopic to the identity, since $h_0 \circ h_1$ and $h_1 \circ h_0$ are both the identity.

Now, we have that $\text{Internal}_V(C_1 \& \ldots \& C_n) \simeq C_1 \oplus \cdots \oplus C_n$, and so we also have dually that $\text{External}_V(C_1 \oplus \cdots \oplus C_n) \simeq C_1 \& \ldots \& C_n$. Thus, $C \simeq \text{External}_V(\text{Internal}_V(C))$. □

The thing that is going on here is that when C internalizes V, the agent of C then has the full ability to choose how V goes (among ways of V going that were possible in C). $\text{Internal}_V(C)$ might have other choices than just choosing how V goes. If it does, then it can freely entangle those other choices with the choice of V however it wants.

When C then externalizes V, it loses all control over V. However, it preserves the ability to entangle all of its other choices with the way that V goes. This ability for the agent to entangle its choices with V is exactly what it means to say "V is observable."

5.1. Example

Let C_0 be defined as in the previous examples, with $V = \{\{ur, nr\}, \{us, ns\}\}$. By the internalizing-externalizing definitions, there exists a frame

$$
\text{Internal}_V(C_0) \cong \begin{array}{cc} & \left(\begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \end{array}\right. \\ (u, r) & \left|\begin{array}{c} \\ \\ ur \\ us \\ nr \\ ns \\ ur \\ ns \\ nr \end{array}\right| \\ (u, s) & \\ (n, r) & \\ (n, s) & \\ (u \leftrightarrow r, r) & \\ (u \leftrightarrow r, s) & \\ (u \leftrightarrow s, r) & \\ (u \leftrightarrow s, s) & \left.\begin{array}{c} \\ \\ us \end{array}\right) \end{array},
$$

which is biextensionally equivalent to

$$C_1 = \begin{array}{c} \\ ur \\ nr \\ us \\ ns \end{array} \begin{pmatrix} \\ ur \\ nr \\ us \\ ns \end{pmatrix} \begin{array}{c} \\ \\ \\ \\ \\ \end{array} \begin{pmatrix} \\ ur \\ nr \\ us \\ ns \end{pmatrix}.$$

We then have that

$$\text{External}_V(C_1) \cong \begin{array}{l} \\ (r \to ur, s \to us) \\ (r \to nr, s \to ns) \\ (r \to ur, s \to ns) \\ (r \to nr, s \to ur) \end{array} \begin{array}{c} r \quad s \\ \begin{pmatrix} ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{pmatrix} \end{array},$$

which is isomorphic to C_0.

This example illustrates both internalizing-externalizing definitions, and also shows the construction used in the constructive definition.

In our next post, we'll conclude the sequence by showing how to formalize agents that learn and act over time using Cartesian frames.

Time in Cartesian Frames

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the twelfth and final post in the Cartesian Frames sequence. Read the first post here.

Up until now, we have (in the examples) mostly considered agents making a single choice, rather than acting repeatedly over time.

The actions, environments, and worlds we've considered might be extended over time. For example, imagine a prisoner's dilemma where "cooperating" requires pushing a button every day for five years.

However, our way of discussing Cartesian frames so far would treat "push the button every day for five years" as an atomic action, a single element $a \in A$.

Now, will begin discussing how to use Cartesian frames to explicitly represent agents passing through time. Let us start with a basic example.

1. Partial Observability

Consider a process where two players, Yosef and Zoe, collaboratively choose a three-digit binary number. Yosef first chooses the first digit, then Zoe chooses the second digit, then Yosef chooses the third digit. The world will be represented by the three-digit number. The Cartesian frame from the perspective of Yosef looks like this:

$$C_0 = \begin{pmatrix} & & & \\ 000 & 010 & 000 & 010 \\ 001 & 011 & 001 & 011 \\ 000 & 011 & 000 & 011 \\ 001 & 010 & 001 & 010 \\ 100 & 110 & 110 & 100 \\ 101 & 111 & 111 & 101 \\ 100 & 111 & 111 & 100 \\ 101 & 110 & 110 & 101 \end{pmatrix}.$$

Here, $C_0 = (A_0, E_0, \cdot_0)$ is a Cartesian frame over

$W_0 = \{000, 001, 010, 011, 100, 101, 110, 111\}$.

The four possible environments from left to right represent Zoe choosing 0, Zoe choosing 1, Zoe copying the first digit, and Zoe negating the first digit.

The eight possible agents can be broken up into two groups of four. In the top four possible agents, Yosef chooses 0 for the first digit, while in the bottom four, he chooses 1. Within each group, the four possible agents represent Yosef choosing 0 for the third digit, choosing 1 for the third digit, copying the second digit, and negating the second digit.

Consider the three partitions W_1, W_2, and W_3 of W_0 representing the first, second and third digits respectively. $W_i = \{w_i^0, w_i^1\}$, where $w_1^0 = \{000, 001, 010, 011\}$, $w_1^1 = \{100, 101, 110, 111\}$, $w_2^0 = \{000, 001, 100, 101\}$, $w_2^1 = \{010, 011, 110, 111\}$, $w_3^0 = \{000, 010, 100, 110\}$, and $w_3^1 = \{001, 011, 101, 111\}$.

Clearly, by the definition of observables, W_2 is not observable in C_0. But there is still a sense in which this does not tell the whole story. Yosef *can* observe W_2 for the purpose of deciding the third digit, but can't observe W_2 for the purpose of deciding the first digit.

There are actually many ways to express this fact, but I want to draw attention to one specific way to express this partial observability: $\text{External}_{W_1}(C_0)$ can observe W_2.

Indeed, we have

$$\text{External}_{W_1}(C_0) \simeq C_1 = \begin{pmatrix} & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ 000 & 010 & 100 & 110 \\ 000 & 010 & 100 & 111 \\ 000 & 010 & 101 & 110 \\ 000 & 010 & 101 & 111 \\ 000 & 011 & 100 & 110 \\ 000 & 011 & 100 & 111 \\ 000 & 011 & 101 & 110 \\ 000 & 011 & 101 & 111 \\ 001 & 010 & 100 & 110 \\ 001 & 010 & 100 & 111 \\ 001 & 010 & 101 & 110 \\ 001 & 010 & 101 & 111 \\ 001 & 011 & 100 & 110 \\ 001 & 011 & 100 & 111 \\ 001 & 011 & 101 & 110 \\ 001 & 011 & 101 & 111 \end{pmatrix}.$$

It may seem counter-intuitive that when you externalize W_1, and thus take some control out of the hands of the agent, you actually end up with more possible agents. This is because the agent now has to specify what the third digit is, not only as a function of the second digit, but also as a function of the first digit. The agent could have specified the third digit as a function of the first digit before, but some of the policies would have been identical to each other.

The four possible environments of C_1 specify the first two digits, while the 16 possible agents represent all of the ways to have the third digit be a function of those first two digits. It is clear that W_2 is observable in C_1.

This gives us a generic way to define a type of partial observability:

Definition: Given a Cartesian frame C over W, and partitions V and T of W, we say V is observable in C after time T if V is observable in $\text{External}_T(C)$.

2. Partitions as Time

Built into the above definition is the fact that we are thinking of (at least some) partitions of W as representing time. This makes a lot of sense when we think of W as a set of possible complete world histories. For any given time, this gives a partition where world histories are in the same subset if they agree on the world history up to that point in time.

For example, the above partition W_1 was the partition that we got by considering a time after Yosef chooses the first digit, but before Zoe chooses the second digit.

Further, this gives us a sequence of nested partitions, since the partition associated with one time is always a refinement of the partition associated with an earlier time.

Note that this is a multiplicative/updateless view of time. There is also an additive/updateful view of time, in which time is a nested sequence of subsets. In the additive view, possible worlds are eliminated as you pass through time. In the multiplicative view, possible worlds are distinguished from each other as you pass through time. We will focus on the multiplicative view, which I consider better-motivated.

3. Nested Subagents

Let $C = (A, E, \cdot)$ be a fixed Cartesian frame over a world W. Let $T_0, \cdots, T_n$ be a sequence of nested partitions of W, with $T_0 = \{W\}$, $T_n = \{\{w\} \mid w \in W\}$, and T_{i+1} a refinement of T_i.

This gives a nested sequence of multiplicative superagents $C_{T_n} \triangleleft_\times \cdots \triangleleft_\times C_{T_0}$, where $C_{T_i} = \text{External}_{T_i}(C)$, which follows from the lemma below.

Lemma: Given a Cartesian frame C over W, if U and V are partitions of W and U is a refinement of V, then $\text{External}_U(C) \triangleleft_\times \text{External}_V(C)$.

Proof: Let $C = (A, E, \cdot)$, and let $u : W \to U$ and $v : W \to V$ send each element of W to their part in U and V respectively. Let $\text{External}_U(C) = (A/B_U, B_U \times E, \cdot_U)$, where $B_U = \{\{a' \in A \mid \forall e \in E, u(a' \cdot e) = u(a \cdot e)\} \mid a \in A\}$. Similarly, let $\text{External}_V(C) = (A/B_V, B_V \times E, \cdot_V)$, where $B_V = \{\{a' \in A \mid \forall e \in E, v(a' \cdot e) = v(a \cdot e)\} \mid a \in A\}$. Let $b_U : A \to B_U$ and $b_V : A \to B_V$ send each element of A to its part in B_U and B_V respectively.

Since U is a refinement of V, there exists a $v' : U \to V$, such that $v' \circ u = v$. Further, we have that B_U is a refinement of B_V, so there exists a $b'_V : B_U \to B_V$ such that $b'_V \circ b_U = b_V$.

It suffices to show there exist three sets X, Y, and Z, and a function $f : X \times Y \times Z \to W$ such that $\text{External}_U(C) \simeq (X, Y \times Z, \diamond)$ and $\text{External}_V(C) \simeq (X \times Y, Z, \bullet)$, where $\diamond$ and $\bullet$ are given by $x \diamond (y, z) = f(x, y, z)$ and $(x, y) \bullet z = f(x, y, z)$.

We will take X to be A/B_U and Z to be $B_V \times E$. We define Y to be the set of all right inverses to b'_V, $Y = \{y : B_V \to B_U \mid \forall b \in B_U,\ b'_V(y(b)) = b\}$. We will let $f(x, y, (b, e)) = x(y(b)) \cdot e$.

First, we show

$$\begin{aligned} \text{External}_U(C) &= (A/B_U, B_U \times E, \cdot_U) \\ &\simeq (X, Y \times Z, \diamond). \end{aligned}$$

We define

$$(g_0, h_0) : (A/B_U, B_U \times E, \cdot_U) \to (X, Y \times Z, \diamond)$$

and

$$(g_1, h_1) : (X, Y \times Z, \diamond) \to (A/B_U, B_U \times E, \cdot_U)$$

as follows. Let g_0 and g_1 be the identity on $X = A/B_U$, and let $h_0 : Y \times Z \to B_U \times E$ be given by $h_0(y,(b,e)) = (y(b),e)$. Finally, let $h_1 : B_U \times E \to Y \times Z$ be chosen to satisfy $h_1(b,e) = (y,(b_V'(b),e))$, where y is such that $y(b_V'(b)) = b$, and for $b' \neq b_V'(b)$, $y(b')$ is chosen arbitrarily to be any preimage of b' under b_V'.

We have that (g_0,h_0) is a morphism, because for all $x \in A/B_U$ and $(y,(b,e)) \in Y \times Z$,

$$\begin{aligned} g_0(x) \diamond (y,(b,e)) &= f(x,y,(b,e)) \\ &= x(y(b)) \cdot e \\ &= x \cdot_U (y(b),e) \\ &= x \cdot_U h_0(y,(b,e)). \end{aligned}$$

Similarly, (g_1,h_1) is a morphism, because for all $x \in X$ and $(b,e) \in B_U \times E$, we have

$$\begin{aligned} g_1(x) \cdot_U (b,e) &= x \cdot_U (b,e) \\ &= x(b) \cdot e \\ &= x(y(b_V'(b))) \cdot e \\ &= f(x,y,(b_V'(b),e)) \\ &= x \diamond (y,(b_V'(b),e)) \\ &= x \diamond h_1(b,e), \end{aligned}$$

where y is as given in the definition of h_1. Since $g_0 \circ g_1$ and $g_1 \circ g_0$ are both the identity, we have that $(g_0,h_0) \circ (g_1,h_1)$ and $(g_1,h_1) \circ (g_0,h_0)$ are both homotopic to the identity, so $\text{External}_U(C) \simeq (X, Y \times Z, \diamond)$.

Next, we show

$$\begin{aligned} \text{External}_V(C) &= (A/B_V, B_V \times E, \cdot_V) \\ &\simeq (X \times Y, Z, \bullet). \end{aligned}$$

We define

$$(g_2, h_2) : (A/B_V, B_V \times E, \cdot_V) \to (X \times Y, Z, \bullet)$$

and

$$(g_3, h_3) : (X \times Y, Z, \bullet) \to (A/B_V, B_V \times E, \cdot_V)$$

as follows. Let h_2 and h_3 be the identity on $Z = B_V \times E$, and let $g_3 : X \times Y \to A/B_V$ be given by $g_3(x, y) = x \circ y$. To see that $x \circ y$ is in A/B_V, we need to verify that $b_V \circ x \circ y$ is the identity on B_V. Indeed,

$$\begin{aligned} b_V \circ x \circ y &= b'_V \circ b_U \circ x \circ y \\ &= b'_V \circ y, \end{aligned}$$

which is the identity on B_V. Let $g_2 : A/B_V \to X \times Y$ be given by $g_2(q) = (q', b_U \circ q)$, where $q' \in A/B_U$ is chosen such that for all $b \in B_V$, $q'(b_U(q(b))) = q(b)$, and for b' not in the image of $b_U \circ q$, $q'(b') \in b'$. We can do this simultaneously for all inputs of the form $b_U(q(b))$, since $b_U \circ q$ is injective, since it has a left inverse, b'_V.

We have that (g_2, h_2) is a morphism, because for all $q \in A/B_V$ and $(b, e) \in Z$, we have

$$\begin{aligned} g_2(q) \bullet (b, e) &= (q', b_U \circ q) \bullet (b, e) \\ &= f(q', b_U \circ q, (b, e)) \\ &= q'(b_U(q(b))) \cdot e \\ &= q(b) \cdot e \\ &= q \cdot_V (b, e) \\ &= h_2(q) \cdot_V (b, e), \end{aligned}$$

where q' is as in the definition of g_2. Similarly, (g_3, h_3) is a morphism, because for all $(x, y) \in X \times Y$ and $(b, e) \in B_V \times E$, we have

$$
\begin{aligned}
g_3(x,y) \cdot_V (b,e) &= x \circ y \cdot_V (b,e) \\
&= x(y(b)) \cdot e \\
&= f(x,y,(b,e)) \\
&= (x,y) \bullet (b,e) \\
&= (x,y) \bullet h_3(b,e).
\end{aligned}
$$

Since $h_3 \circ h_2$ and $h_2 \circ h_3$ are both the identity, we have that $(g_2, h_2) \circ (g_3, h_3)$ and $(g_3, h_3) \circ (g_2, h_2)$ are both homotopic to the identity, so $\text{External}_V(C) \simeq (X \times Y, Z, \bullet)$, completing the proof. □

The sequence $C_{T_0}, \dots, C_{T_n}$ represents the agent persisting across time, but each subagent C_{T_i} does not really represent a single time-slice of the agent. Instead, C_{T_i} represents an agent persisting across time starting at the time T_i.

I think that this is actually the more natural notion. However, if we want to think about an agent persisting across times as a sequence of single times-slices of the agent, we could also do that. Since $C_{T_{i+1}}$ is a multiplicative subagent of C_{T_i}, $C_{T_{i+1}}$ must have a sister D_{T_i+1} in C_{T_i}, so we could consider the sequence $D_{T_1}, \dots, D_{T_n}$.

4. Controllables Decrease and Observables Increase Over Time

An interesting fact about these sequences $C_{T_0}, \dots, C_{T_n}$ is that controllables decrease and observables increase over time, so for $i \preceq j$ we have $\text{Obs}(C_{T_i}) \subseteq \text{Obs}(C_{T_j})$ and $\text{Ctrl}(C_{T_i}) \supseteq \text{Ctrl}(C_{T_j})$ (and $\text{Ensure}(C_{T_i}) \supseteq \text{Ensure}(C_{T_j})$ and $\text{Prevent}(C_{T_i}) \supseteq \text{Prevent}(C_{T_j})$), which follows directly from the following two lemmas.

Lemma: Given a Cartesian frame C over W, if U and V are partitions of W and U is a refinement of V, then $\text{Ctrl}(\text{External}_V(C)) \supseteq \text{Ctrl}(\text{External}_U(C))$.

Proof: Let $C_V = \text{External}_V(C)$, and let $C_U = \text{External}_V(C)$. We will actually only need to use the fact that $C_U \triangleleft_\times C_V$, and that both C_U and C_V have nonempty agents. C_U and C_V do in fact have nonempty agent, because, as we have shown, externalizing a partition of W always produces nonempty agents.

It suffices to establish that $\text{Ensure}(C_{T_i}) \supseteq \text{Ensure}(C_{T_j})$, and the result for Ctrl follows trivially.

Since $C_U \triangleleft_\times C_V$, there exist X, Y, Z, and $f : X \times Y \times Z \to W$ such that $C_U \simeq (X, Y \times Z, \diamond)$ and $C_V \simeq (X \times Y, Z, \bullet)$, where $\diamond$ and $\bullet$ are given by $x \diamond (y, z) = f(x, y, z)$ and $(x, y) \bullet z = f(x, y, z)$. Let $C_U' = (X, Y \times Z, \diamond)$, and let $C_V' \simeq (X \times Y, Z, \bullet)$. Observe that X and Y are nonempty.

Since Ensure is preserved by biextensional equivalence, it suffices to show that $\text{Ensure}(C_V') \supseteq \text{Ensure}(C_U')$. Let $S \in \text{Ensure}(C_U')$. Thus, there exists some $x_0 \in X$, such that for all $(y, z) \in Y \times Z$, $x_0 \diamond (y, z) = f(x_0, y, z) \in S$. Since Y is nonempty, we can take an arbitrary $y_0 \in Y$, and observe that for all $z \in S$, $(x_0, y_0) \bullet z = f(x_0, y_0, z) \in S$. Thus, $S \in \text{Ensure}(C_V')$. □

Lemma: Given a Cartesian frame C over W, if U and V are partitions of W and U is a refinement of V, then $\text{Obs}(\text{External}_V(C)) \subseteq \text{Obs}(\text{External}_U(C))$.

Proof: Let $C = (A, E, \cdot)$, and let $u : W \to U$ and $v : W \to V$ send each element of W to their part in U and V respectively. Let $\text{External}_U(C) = (A/B_U, B_U \times E, \cdot_U)$, where $B_U = \{\{a' \in A \mid \forall e \in E, u(a' \cdot e) = u(a \cdot e)\} \mid a \in A\}$. Similarly, let $\text{External}_U(C) = (A/B_V, B_V \times E, \cdot_V)$, where $B_V = \{\{a' \in A \mid \forall e \in E, v(a' \cdot e) = v(a \cdot e)\} \mid a \in A\}$. Let $b_U : A \to B_U$ and $b_V : A \to B_V$ send each element of A to its part in B_U and B_V respectively.

Since U is a refinement of V, there exists a $v' : U \to V$, such that $v' \circ u = v$. Further, we have that B_U is a refinement of B_V, so there exists a $b'_V : B_U \to B_V$ such that $b'_V \circ b_U = b_V$.

Let $S \in \text{Obs}(\text{External}_V(C))$. Thus, for every pair $q_0, q_1 \in A/B_V$, there exists a $q_2 \in A/B_V$ such that $q_2 \in \text{if}(S, q_0, q_1)$. Thus, we can define an $f : A/B_V \times A/B_V \to A/B_V$ such that for all $q_0, q_1 \in A/B_V$, $f(q_0, q_1) \in \text{if}(S, q_0, q_1)$.

Our goal is to show that $S \in \text{Obs}(\text{External}_U(C))$. For this, it suffices to show that for any $q_0, q_1 \in A/B_U$, there exists a $q_2 \in A/B_U$ such that $q_2 \in \text{if}(S, q_0, q_1)$.

Let $q_0, q_1 \in A/B_U$ be arbitrary. Given an arbitrary $b \in B_U$, let $q_i^b \in A/B_V$ be any element that satisfies $q_i^b(b'_V(b)) = q_i(b)$. This is possible because $q_i(b) \in b \subseteq b'_V(b)$. It does not matter what q_i^b does on other inputs. Let $q_2 : B_U \to A$ be such that for all $b \in B_U$, $q_2(b) = f(q_0^b, q_1^b)(b'_V(b))$.

To complete the proof, we need to show that $q_2 \in A/B_U$ and $q_2 \in \text{if}(S, q_0, q_1)$.

To show that $q_2 \in A/B_U$, we need that for all $b \in B_U$, $q_2(b) \in b$. Let $b \in B_U$ be arbitrary. Since $q_0(b) \in b$, by the definition of B_U, it suffices to show that for all $e \in E$, $u(q_2(b) \cdot e) = u(q_0(b) \cdot e)$. Further, since $q_1(b) \in b$, we already have that for all $e \in E$, $u(q_1(b) \cdot e) = u(q_0(b) \cdot e)$. Thus, it suffices to show that for all $e \in E$, either $q_2(b) \cdot e = q_0(b) \cdot e$ or $q_2(b) \cdot e = q_1(b) \cdot e$. Indeed, if $q_2(b) \cdot e \in S$, then

$$\begin{aligned} q_2(b) \cdot e &= f(q_0^b, q_1^b)(b'_V(b)) \cdot e \\ &= q_0^b(b'_V(b)) \cdot e \\ &= q_0(b) \cdot e, \end{aligned}$$

and similarly, if $q_2(b) \cdot e \notin S$, then $q_2(b) \cdot e = q_1(b) \cdot e$. Thus, we have that for all $e \in E$, $u(q_2(b) \cdot e) = u(q_0(b) \cdot e)$, so for our arbitrary $b \in B_U$, $q_0(b) \in b$, so $q_2 \in A/B_U$.

Let $(b, e) \in B_U \times E$ be such that $q_2 \cdot_U (b, e) \in S$. We want to show that $q_2 \cdot_U (b, e) = q_0 \cdot_U (b, e)$. Indeed,

$$\begin{aligned}
q_2 \cdot_U (b, e) &= q_2(b) \cdot e \\
&= f(q_0^b, q_1^b)(b_V'(b)) \cdot e \\
&= f(q_0^b, q_1^b) \cdot_V (b_V'(b), e) \\
&= q_0^b \cdot_V (b_V'(b), e) \\
&= q_0^b(b_V'(b)) \cdot e \\
&= q_0(b) \cdot e \\
&= q_0 \cdot_U (b, e).
\end{aligned}$$

Symmetrically, if $(b, e) \in B_U \times E$ is such that $q_2 \cdot_U (b, e) \notin S$, we have $q_2 \cdot_U (b, e) = q_1 \cdot_U (b, e)$. Thus $q_2 \in \text{if}(S, q_0, q_1)$.

Thus, since q_0 and q_1 were arbitrary, we have that $S \in \text{Obs}(\text{External}_U(C))$, completing the proof. □

This result allows us to think of time as a sort of ritual in which control of the world is sacrificed in exchange for ability to condition on the world.

5. Directions for Future Work

As I noted at the start of this sequence, Cartesian frames take their *motivation* from Hutter, attempting to improve on the cybernetic agent model; they take their *angle of attack* from Pearl, using combinatorics to infer functional structure from relational structure; and they take their *structure* from game theory, working with base objects that look similar to normal-form games.

Building up from very simple foundations, we have found that Cartesian frames yield elegant notions of agents making choices and observations, of agents acting over

time, and of subagent relations. At the same time, Cartesian frames allow us to switch between different levels of description of the world and consider many different ways of factorizing the world into variables.

I suspect that this is the last post I will write on Cartesian frames for a while, but I am excited about the framework, and would really like to get more people working on it.

To help with that, I've commented below with various directions for future work: ways that I think the framework could be extended, made better, or applied.

- frames that are partitions into rectangles
- generalizing observability
- preferences and goals
- subagents
- logical time
- logical uncertainty
- formalizing time
- computational complexity
- time and coarse world models
- category-theory-first approaches

I've erred on the side of inclusion in these comments: some may point to dead ends, or may be based on false assumptions.

If you have questions or want to discuss Cartesian frames, I'll be hosting a fourth and final office hours / discussion section this Sunday at 2pm PT on GatherTown.

Cartesian Frames and Factored Sets on ArXiv

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Papers on Cartesian frames and factored sets are now on arXiv.

Cartesian Frames: https://arxiv.org/abs/2109.10996

Factored Sets: https://arxiv.org/abs/2109.11513

The factored set paper is approximately identical to the sequence here, while the Cartesian frame paper is rewritten by Daniel Hermann and Josiah Lopez-Wild, optimized for an audience of philosophers.