Cartesian Frames

Scott Garrabrant

Machine Intelligence Research Institute

October 25, 2020

https://www.lesswrong.com/s/2A7rrZ4ySx6R8mfoT

Outline I

Motivation

Conceptual Parents:

- Hutter's Cybernetic Agent Model
- Pearl's Causality
- Game Theory

Definition

Definition 1 (Cartesian Frame)

A Cartesian frame C over a set W is a triple $(A, E, \cdot)$, where A and E are sets and $\cdot : A \times E \to W$.

Definition

Definition 1 (Cartesian Frame)

A Cartesian frame C over a set W is a triple $(A, E, \cdot)$, where A and E are sets and $\cdot : A \times E \to W$.

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

Definition

Definition 1 (Cartesian Frame)

A Cartesian frame C over a set W is a triple $(A, E, \cdot)$, where A and E are sets and $\cdot : A \times E \to W$.

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

Definition 2 (Agent, Env, World, Eval)

If $C = (A, E, \cdot)$ is a Cartesian frame over W, we say $\text{Agent}(C) = A$, $\text{Env}(C) = E$, $\text{World}(C) = W$, and $\text{Eval}(C) = \cdot$.

Controllables Are Like Outputs

Definition 3 (Ensurables)

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

i.e. $S \in \text{Ensure}(C)$ if there exists a row entirely contained in S.
(We will often identify C with its agent, so we will say things like "C can ensure S," and that will mean that C's agent can ensure S.)

Controllables Are Like Outputs

Definition 3 (Ensurables)

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

i.e. $S \in \text{Ensure}(C)$ if there exists a row entirely contained in S.
(We will often identify C with its agent, so we will say things like "C can ensure S," and that will mean that C's agent can ensure S.)

Theorem 4

Ensure(C) is closed under supersets.

Controllables Are Like Outputs

Definition 3 (Ensurables)

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

i.e. $S \in \text{Ensure}(C)$ if there exists a row entirely contained in S.
(We will often identify C with its agent, so we will say things like "C can ensure S," and that will mean that C's agent can ensure S.)

Theorem 4

Ensure(C) is closed under supersets.

Definition 5 (Preventables, Controllables)

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

Observables Are Like Inputs

Definition 6 (Conditional Policies)

Given $C = (A, E, \cdot)$, $S \subseteq W$, 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 $(a \cdot e \notin S) \to (a \cdot e = a_1 \cdot e)$.

Observables Are Like Inputs

Definition 6 (Conditional Policies)

Given $C = (A, E, \cdot)$, $S \subseteq W$, 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 $(a \cdot e \notin S) \to (a \cdot e = a_1 \cdot e)$.

Definition 7 (Observables)

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

Observables Are Like Inputs

Definition 6 (Conditional Policies)

Given $C = (A, E, \cdot)$, $S \subseteq W$, 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 $(a \cdot e \notin S) \to (a \cdot e = a_1 \cdot e)$.

Definition 7 (Observables)

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

Theorem 8

Obs(C) is closed under Boolean combinations.

Example 1

Environment is either rain or sun. $(E = \{r, s\})$
Agent independently either carries an umbrella or not. $(A = \{u, n\})$
Four worlds tracking umbrella vs none and rain vs sun.
$(W = \{ur, us, nr, ns\})$

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

Example 1

Environment is either rain or sun. $(E = \{r, s\})$
Agent independently either carries an umbrella or not. $(A = \{u, n\})$
Four worlds tracking umbrella vs none and rain vs sun.
$(W = \{ur, us, nr, ns\})$

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

Ensure$(C_1) = \{\{ur, us\}, \{nr, ns\},$
$\{ur, us, nr\}, \{ur, us, ns\}, \{nr, ns, ur\}, \{nr, ns, us\}, W\}$

Example 1

Environment is either rain or sun. $(E = \{r, s\})$
Agent independently either carries an umbrella or not. $(A = \{u, n\})$
Four worlds tracking umbrella vs none and rain vs sun.
$(W = \{ur, us, nr, ns\})$

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

$\text{Ensure}(C_1) = \{\{ur, us\}, \{nr, ns\},$
$\{ur, us, nr\}, \{ur, us, ns\}, \{nr, ns, ur\}, \{nr, ns, us\}, W\}$
$\text{Ctrl}(C_1) = \{\{ur, us\}, \{nr, ns\}\}$

Example 1

Environment is either rain or sun. $(E = \{r, s\})$
Agent independently either carries an umbrella or not. $(A = \{u, n\})$
Four worlds tracking umbrella vs none and rain vs sun.
$(W = \{ur, us, nr, ns\})$

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

$\text{Ensure}(C_1) = \{\{ur, us\}, \{nr, ns\},$
$\{ur, us, nr\}, \{ur, us, ns\}, \{nr, ns, ur\}, \{nr, ns, us\}, W\}$
$\text{Ctrl}(C_1) = \{\{ur, us\}, \{nr, ns\}\}$
$\text{Obs}(C_1) = \{\{\}, W\}$

Example 2

Now, the agent knows the weather. ($A = \{u, n, u \leftrightarrow r, u \leftrightarrow s\}$, E and W are unchanged)

$$C_2 = \begin{matrix} \\ u \\ n \\ u \leftrightarrow r \\ u \leftrightarrow s \end{matrix} \begin{matrix} r & s \end{matrix} \begin{pmatrix} ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{pmatrix}$$

Example 2

Now, the agent knows the weather. ($A = \{u, n, u \leftrightarrow r, u \leftrightarrow s\}$, E and W are unchanged)

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

$\text{Ensure}(C_2) = \{\{ur, us\}, \{nr, ns\}, \{ur, ns\}, \{nr, us\}, \{ur, us, nr\}, \{ur, us, ns\}, \{nr, ns, ur\}, \{nr, ns, us\}, W\}$

Example 2

Now, the agent knows the weather. ($A = \{u, n, u \leftrightarrow r, u \leftrightarrow s\}$, E and W are unchanged)

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

$\text{Ensure}(C_2) = \{\{ur, us\}, \{nr, ns\}, \{ur, ns\}, \{nr, us\}, \{ur, us, nr\}, \{ur, us, ns\}, \{nr, ns, ur\}, \{nr, ns, us\}, W\}$

$\text{Ctrl}(C_2) = \{\{ur, us\}, \{nr, ns\}, \{ur, ns\}, \{nr, us\}\}$

Example 2

Now, the agent knows the weather. ($A = \{u, n, u \leftrightarrow r, u \leftrightarrow s\}$, E and W are unchanged)

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

$\text{Ensure}(C_2) = \{\{ur, us\}, \{nr, ns\}, \{ur, ns\}, \{nr, us\}, \{ur, us, nr\}, \{ur, us, ns\}, \{nr, ns, ur\}, \{nr, ns, us\}, W\}$
$\text{Ctrl}(C_2) = \{\{ur, us\}, \{nr, ns\}, \{ur, ns\}, \{nr, us\}\}$
$\text{Obs}(C_2) = \{\{\}, \{ur, nr\}, \{us, ns\}, W\}$

Example 3

Now, there is a third possible environment, in which a meteor strikes and the agent is never born. ($E = \{r, s, m\}$, $W = \{ur, us, nr, ns, m\}$)

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

Example 3

Now, there is a third possible environment, in which a meteor strikes and the agent is never born. ($E = \{r, s, m\}$, $W = \{ur, us, nr, ns, m\}$)

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

$\text{Ensure}(C_3) = \{\{ur, us, m\}, \{nr, ns, m\}, \{ur, ns, m\}, \{nr, us, m\}, \{ur, us, nr, m\}, \{ur, us, ns, m\}, \{nr, ns, ur, m\}, \{nr, ns, us, m\}, W\}$

Example 3

Now, there is a third possible environment, in which a meteor strikes and the agent is never born. ($E = \{r, s, m\}$, $W = \{ur, us, nr, ns, m\}$)

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

$\text{Ensure}(C_3) = \{\{ur, us, m\}, \{nr, ns, m\}, \{ur, ns, m\}, \{nr, us, m\}, \{ur, us, nr, m\}, \{ur, us, ns, m\}, \{nr, ns, ur, m\}, \{nr, ns, us, m\}, W\}$

$\text{Ctrl}(C_3) = \{\}$

Example 3

Now, there is a third possible environment, in which a meteor strikes and the agent is never born. ($E = \{r, s, m\}$, $W = \{ur, us, nr, ns, m\}$)

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

$\text{Ensure}(C_3) = \{\{ur, us, m\}, \{nr, ns, m\}, \{ur, ns, m\}, \{nr, us, m\},$
$\{ur, us, nr, m\}, \{ur, us, ns, m\}, \{nr, ns, ur, m\}, \{nr, ns, us, m\}, W\}$
$\text{Ctrl}(C_3) = \{\}$
$\text{Obs}(C_3) =$
$\{\{\}, \{ur, nr\}, \{us, ns\}, \{m\}, \{ur, nr, us, ns\}, \{ur, nr, m\}, \{us, ns, m\}, W\}$

Controllables and Observables Are Disjoint!

Theorem 9

If $Env(C)$ *is nonempty,* $Ctrl(C) \cap Obs(C) = \{\}$.

Controllables and Observables Are Disjoint!

Theorem 9

If $Env(C)$ *is nonempty,* $Ctrl(C) \cap Obs(C) = \{\}$.

Definition 10 (Image)

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

Theorem 11

$S \in Ensure(C) \cap Obs(C)$ *iff* $Image(C) \subseteq S$ *and* $Agent(C)$ *is nonempty.*

Controllables and Observables Are Disjoint!

Theorem 9

If $Env(C)$ is nonempty, $Ctrl(C) \cap Obs(C) = \{\}$.

Definition 10 (Image)

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

Theorem 11

$S \in Ensure(C) \cap Obs(C)$ iff $Image(C) \subseteq S$ and $Agent(C)$ is nonempty.

Both follow from the following lemma:

Lemma 12

If $S \in Obs(C)$, then for all $e \in E$ and $a_0, a_1 \in A$,
$a_0 \cdot e \in S$ if and only if $a_1 \cdot e \in S$.

I.e., every column is either entirely in S or entirely outside of S.

A Category of Cartesian Frames

Definition 13 (Chu(W))

Chu(W) is a category.
Objects are Cartesian frames over W.
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 $f \in F$.
Composition is given by $(g_1, h_1) \circ (g_0, h_0) = (g_1 \circ g_0, h_0 \circ h_1)$.

A Category of Cartesian Frames

Definition 13 (Chu(W))

Chu(W) is a category.
Objects are Cartesian frames over W.
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 $f \in F$.
Composition is given by $(g_1, h_1) \circ (g_0, h_0) = (g_1 \circ g_0, h_0 \circ h_1)$.

A morphism from C to D can be thought of an interface that allows the agent of C to interact with the environment of D.
The existence of a morphism from C to D can also be thought of as saying that (the agent of) D is at least as strong as (the agent of) C.

Chu(W) is Self-Dual

Definition 14 ($-^*$)

Let $C = (A, E, \cdot)$ be a Cartesian frame over W. Then C^* denotes the Cartesian frame over W, $(E, A, \star)$, where $e \star a = a \cdot e$.

This takes the transpose of the matrix.

Chu(W) is Self-Dual

Definition 14 ($-^*$)

Let $C = (A, E, \cdot)$ be a Cartesian frame over W. Then C^* denotes the Cartesian frame over W, $(E, A, \star)$, where $e \star a = a \cdot e$.

This takes the transpose of the matrix.

Theorem 15

Chu(W) is isomorphic to its opposite category, Chu(W)^{op}. The isomorphism sends C to C^* and (g, h) to (h, g).

This illustrates the symmetry between agent and environment in this framework.

Sums

Definition 16 ($\oplus$)

Let $C = (A, E, \cdot)$ and $D = (B, F, \star)$ be Cartesian frames over W. $C \oplus D$ is the Cartesian frame $(A \sqcup B, E \times F, \diamond)$, where $a \diamond (e, f) = \begin{cases} a \cdot e & \text{if } a \in A \\ a \star f & \text{if } a \in B \end{cases}$

Sums

Definition 16 ($\oplus$)

Let $C = (A, E, \cdot)$ and $D = (B, F, \star)$ be Cartesian frames over W. $C \oplus D$ is the Cartesian frame $(A \sqcup B, E \times F, \diamond)$, where $a \diamond (e, f) = \begin{cases} a \cdot e & \text{if } a \in A \\ a \star f & \text{if } a \in B \end{cases}$

Definition 17 (0)

Let 0 be the Cartesian frame $0 = (\{\}, \{e\}, \cdot)$ with empty agent, singleton environment, and trivial evaluation function.

Sums

Definition 16 ($\oplus$)

Let $C = (A, E, \cdot)$ and $D = (B, F, \star)$ be Cartesian frames over W. $C \oplus D$ is the Cartesian frame $(A \sqcup B, E \times F, \diamond)$, where $a \diamond (e, f) = \begin{cases} a \cdot e & \text{if } a \in A \\ a \star f & \text{if } a \in B \end{cases}$

Definition 17 (0)

Let 0 be the Cartesian frame $0 = (\{\}, \{e\}, \cdot)$ with empty agent, singleton environment, and trivial evaluation function.

Theorem 18

$\oplus$ is commutative and associative, 0 is the identity of $\oplus$.

Sums

Definition 16 ($\oplus$)

Let $C = (A, E, \cdot)$ and $D = (B, F, \star)$ be Cartesian frames over W. $C \oplus D$ is the Cartesian frame $(A \sqcup B, E \times F, \diamond)$, where $a \diamond (e, f) = \begin{cases} a \cdot e & \text{if } a \in A \\ a \star f & \text{if } a \in B \end{cases}$

Definition 17 (0)

Let 0 be the Cartesian frame $0 = (\{\}, \{e\}, \cdot)$ with empty agent, singleton environment, and trivial evaluation function.

Theorem 18

$\oplus$ is commutative and associative, 0 is the identity of $\oplus$.

Theorem 19

$\oplus$ is the coproduct in Chu(C). 0 is the initial object.

Example 4

$$C_4 = \begin{array}{c} \\ a_0 \\ a_1 \end{array} \begin{array}{c} \begin{array}{cc} e_0 & e_1 \end{array} \\ \begin{pmatrix} w_0 & w_1 \\ w_2 & w_3 \end{pmatrix} \end{array} \quad \text{and} \quad D_4 = \begin{array}{c} \\ b_0 \\ b_1 \end{array} \begin{array}{c} \begin{array}{cc} f_0 & f_1 \end{array} \\ \begin{pmatrix} w_4 & w_5 \\ w_6 & w_7 \end{pmatrix} \end{array}$$

$$C_4 \oplus D_4 = \begin{array}{c} \\ a_0 \\ a_1 \\ b_0 \\ b_1 \end{array} \begin{array}{c} \begin{array}{cccc} e_0 f_0 & e_0 f_1 & e_1 f_0 & e_1 f_1 \end{array} \\ \begin{pmatrix} w_0 & w_0 & w_1 & w_1 \\ w_2 & w_2 & w_3 & w_3 \\ w_4 & w_5 & w_4 & w_5 \\ w_6 & w_7 & w_6 & w_7 \end{pmatrix} \end{array}$$

Products

Definition 20 (&)

Let $C = (A, E, \cdot)$ and $D = (B, F, \star)$ be Cartesian frames over W. $C \& D$ is the Cartesian frame $(A \times B, E \sqcup F, \diamond)$, where $(a, b) \diamond e = \begin{cases} a \cdot e & \text{if } e \in E \\ b \star e & \text{if } e \in F \end{cases}$

Definition 21 ($\top$)

Let $\top$ be the Cartesian frame $\top = (\{a\}, \{\}, \cdot)$ with singleton agent, empty environment, and trivial evaluation function.

Theorem 22

$\&$ is commutative and associative, and $\top$ is the identity of $\&$.

Theorem 23

$\&$ the product of C and D in $Chu(C)$. $\top$ is the initial object.

Example 5

$$C_5 = \begin{matrix} & r \\ u & ur \\ n & nr \end{pmatrix} \quad \text{and} \quad D_5 = \begin{matrix} & s \\ u & us \\ n & ns \end{pmatrix}$$

$$C_5 \& D_5 = C_2 = \begin{matrix} uu = u \\ nn = n \\ un = u \leftrightarrow r \\ nu = u \leftrightarrow s \end{matrix} \quad \begin{matrix} r & s \\ ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{pmatrix}$$

$\oplus$ and & Are De Morgan Dual

Theorem 24

$C\&D = (C^* \oplus D^*)^*$

$C \oplus D = (C^*\&D^*)^*$

$\top = 0^*$

$0 = \top^*$

$C = (C^*)^*$

In Part Two of This Talk

- Homotopy Equivalence
- Small Cartesian Frames
- Redefining Controllables and Observables Categorically
- Functors and Coarse World Models
- Subagents
- Multiplicative Operations