Integrated Information Theory (IIT) with Simple Maths

David Rudrauf, **Tonglin Yan** [1], Nils Ruet, Kenneth Williford, **Grégoire Sergeant-Perthuis** [2]

[1] CIAMS, Université Paris-Saclay, [2] LCQB, Sorbonne Université

ASSC 27
Tokyo, 4 July, 2024

For the ‘preprint’ go here: here

- What we do:
  - $\rightarrow$ **Reinforcement learning** + **phenomenological** aspects of consciousness

Reproduce the experience of *space* in 'robots'

Consciousness involves a subjective perspective, characterized by viewpoint-structured organization, a sense of unity (holistic world), embodiment, and an internal representation of the world in perspective from a specific standpoint.

$\hookrightarrow$ the **Projective Consciousness Model**

- Initiated by D. Rudrauf, K. Williford, D. Bennequin, K. Friston [RBG+17, WBFR18, RBW20]

$\longrightarrow$See K. Williford's MoC4 presentation of phenomenological motivation
`https://www.youtube.com/watch?v=eHyVZWZMqzg&t=853s`

- What do I do?
  - $\rightarrow$ Background: mathematical physics (PhD)
  - $\rightarrow$ Machine learning $\cap$ geometry $\rightarrow$ computational biology
  - $\rightsquigarrow$ Embarked on the *consciousness* adventure
    - $\hookrightarrow$ Geometric structure of the world model [RSPB$^+$20, SPRR$^+$23]
- Why this work?
  - $\rightarrow$ Disseminate formal and computational models of consciousness
  - $\rightarrow$ Simplify entering into the mathematical details of the models: IIT, active inference, etc.

Based on a seminar given at:

- PMMC2: Paris Mathematical Models of Cognition and Consciousness (PMMC2, link to the seminar page)
  - → On mathematical and computational models of consciousness and cognition
  - → Longer version on YouTube: link to the talk
  - ↪ Come and give a talk

Today's presentation is a simplification of the excellent presentation:

→ *The mathematical structure of integrated information theory*, Johannes Kleiner and Sean Tull. In Frontiers in Applied Mathematics and Statistics, 2020.

Definition (Probability measures)

Let E be a finite set, denote by $\mathbb{P}(E)$ the set of probability measures on E,

$$p \in \mathbb{P}(E) \iff \forall x \in E, p(x) \geq 0 \text{ and } \sum_{x \in E} p(x) = 1 \tag{0.1}$$

Definition (Markov kernels (stochastic maps))

A Markov kernel or stochastic map T from E to F, denoted as $T : E \to \mathbb{P}(F)$, sends any point $x \in E$ to a probability measure $T_x \in \mathbb{P}(F)$. For $x \in E$ and $y \in F$, we will denote $T_x(y)$ as $T(y|x)$.

- We will be considering multiple variables $(X_i, i \in S)$ denotes X_S
- When S is a finite set:
  - $\rightarrow$ enumerate S, $S \simeq [1, ..., N]$
- We will consider sub-collections of variable:
  - $\rightarrow$ $a \subseteq S$
  - $\rightarrow$ $\overline{a}$ the complement of a in S,

$$a \cup \overline{a} = S \qquad\qquad a \cap \overline{a} = \emptyset$$

- a subset $a \subseteq S$ is associates to a random variable $(X_i, i \in a)$
  - $\rightarrow$ X_i takes values in E_{X_i}
  - $\rightarrow$ Denote $(X_i, i \in a)$ as X_a
  - $\rightarrow$ X_a takes values in $\prod_{i \in a} E_{X_i}$ denoted as E_{X_a}

Definition (Conditional expectation)

Let $(Y_i, i \in S_1)$ be a collection of variables taking values in $F = \prod_{i \in S_1} F_{Y_i}$. Each F_{Y_i} is a finite set. Let $p \in \mathbb{P}(F)$. For any $a \subseteq S_1$, and any function $f : F \to \mathbb{R}$, one defines the conditional expectation with respect to Y_a as;

$$\forall y_a \in F_a, \quad \mathbb{E}[f|Y_a](y_a) = \sum_{y_{\overline{a}} \in Y_{\overline{a}}} \frac{f(y_{\overline{a}}, y_a) p(y_{\overline{a}}, y_a)}{\sum_{y_{\overline{a}} \in Y_{\overline{a}}} p(y_{\overline{a}}, y_a)}$$

Example with two random variables X, Y, assume:

- $\rightarrow$ $X \in E$, $Y \in F$, where E, F are finite sets
- $\rightarrow$ $x \in E$, $y \in F$

then,

- $\rightarrow$ $\mathbb{E}[Y = y|X = x] := P(y|x)$
- $\rightarrow$ and

$$P(y|x) = \frac{P(X = x, Y = y)}{P(X = x)} \qquad \text{(Bayes' Rule)}$$

We want to see the effect of the stochastic dynamic T on sub-collections of variables.

- We start with $T : E_{X_1} \times \ldots \times E_{X_N} \to \mathbb{P}(F_{Y_1} \times \ldots \times F_{Y_M})$
- Choose $a = (1, \ldots, n_1)$ and $b = (1, \ldots, m_1)$ with $n_1 \leq N$ and $m_1 \leq M$
- How to deduce a transition

$$T^{a,b} : E_{X_1} \times \ldots \times E_{X_{n_1}} \to \mathbb{P}(F_{Y_1} \times \ldots \times F_{Y_{m_1}})$$

Definition (Building a Markov kernel $E_{X_a} \to \mathbb{P}(F_{Y_b})$ from a prior Q)

Let $X_S := (X_i, i \in S)$, $Y_{S_1} := (Y_i, i \in S_1)$ and $E = \prod_{i\in S} E_{X_i}$, $F = \prod_{i\in S_1} F_{Y_i}$.
Let $T = E \to \mathbb{P}(F)$ be a Markov kernel. For any $a \subseteq S$ and $b \subseteq S_1$, a choice of $Q \in \mathbb{P}(E)$ allows us to derive from T the kernel denoted $T^{Q,a,b} : E_{X_a} \to \mathbb{P}(F_{Y_b})$, which encodes the effect of the variables X_a on Y_b. It is defined as,

$$\forall y_b \in F_{Y_b}, \forall x_a \in E_{X_a} \quad T^{Q,a,b}(y_b|x_a) := \mathbb{E}[Y_b = y_b | X_a = x_a]$$

– How? $\hookrightarrow$ Joint distribution,

$$P(Y_{S_1} = y, X_S = x) := T(y|x) \times Q(x)$$

$\hookrightarrow$ Sum out (*marginalize*) $X_{\overline{a}}$, $Y_{\overline{b}}$

‘Cutting’ interactions: the central operation

$$
\begin{array}{ccc}
X_1 \dots X_{n_1} & X_{n_1+1} \dots X_N & X_1 \dots X_N \\
a & \overline{a} & S \\
\Big\Downarrow T^{a,b} & \Big\Downarrow T^{\overline{a},\overline{b}} & \Big\Downarrow T \\
b & \overline{b} & S_1 \\
Y_1 \dots Y_{m_1} & Y_{m_1+1} \dots Y_M & Y_1 \dots Y_M
\end{array}
$$

Left: Cutting the interactions. Right: Overall interaction.

- We quantify the effect of ‘cutting’ interactions between variables
- $(T^{a,b}, T^{\overline{a},\overline{b}})$ should be in the same space as T

Definition (Product of local kernels)

For any two probability kernels, $T^{a,b} : E_{X_a} \to \mathbb{P}(F_{Y_b})$ and $T^{\overline{a},\overline{b}} : E_{X_{\overline{a}}} \to \mathbb{P}(F_{Y_{\overline{b}}})$ posit,

$$(T^{a,b} \otimes T^{\overline{a},\overline{b}})(y|x) := T^{a,b}(y_b|x_a).T^{\overline{a},\overline{b}}(y_{\overline{b}}|x_{\overline{a}}) \qquad (0.2)$$

$\rightarrow$ We want to quantify how far $T^{a,b} \otimes T^{\bar{a},\bar{b}}$ is from T

Definition (Informal definition of divergence)

For a finite space Y, we define a divergence D on $\mathbb{P}(Y)$ as a function $D : \mathbb{P}(Y) \times \mathbb{P}(Y) \rightarrow \mathbb{R}_{\geq 0}$ such that, for any two probability distributions P and P_1 in $\mathbb{P}(Y)$, $D(P, P_1)$ decreases as the two distributions P and P_1 get 'closer'; and it reaches its minimum value of 0 when and only when $P = P_1$.

- The dissimilarity is on *probability distributions*.
- Fix $x \in E$ then, $T^{a,b} \otimes T^{\bar{a},\bar{b}}(.|x)$ is a probability distribution.
- Similarly $T(.|x) \in \mathbb{P}(F)$

$\rightarrow$ For a fixed x

- $\rightsquigarrow$ denote $T^{a,b} \otimes T^{\bar{a},\bar{b}}(.|x)$ as $T^{a,b}_{x_a} \otimes T^{\bar{a},\bar{b}}_{x_{\bar{a}}}$
- $\rightsquigarrow$ denote $T(.|x)$ as T_x

- Little φ_e focusing on effects.
  - $\rightarrow$ $S \simeq [1, ..., N]$ and $X_1 ... X_N$
  - $\rightarrow$ $M \subseteq S$, 'old' X_S is now X_M
  - $\rightarrow$ $P \subseteq S$ 'old' Y_{S_1} is now X_P
  - $\rightarrow$ To remember that we start with X_M and we go to X_P denote T_M^P the associated kernel

Definition

For any $M, P \subseteq S$ and $x_M \in X_M$,

$$\varphi^P_{M,x_M} := \inf_{\substack{a \subseteq M \\ b \subseteq P}} D(T^P_{M,x_M} | T^{P,(a,b)}_{M,x_a} \otimes T^{P,(\bar{a},\bar{b})}_{M,x_{\bar{a}}}) \tag{0.3}$$

And

$$\varphi^*_{M,x_M} := \max_{P \subseteq S} \varphi^P_{M,x_M} \tag{0.4}$$

→ One more step to compute *big* Φ, focusing on effects (see here)

$$\Phi_{M,x,b} := \sum_{a \subseteq M} \varphi^{*}_{a,x_a} D(T^{\psi(a,x_a)}_{a,x_a} | T^{b_a}_{a,x_a} \otimes T^{\bar{b}_a}_{a,x_a})$$

with $\psi(M,x) := \mathbf{argmax}\, \varphi^{P}_{M}$

$$\Phi_{M,x} = \text{argmin}_{b \subseteq \psi(M,x_M)} \Phi_{M,x,b}$$

→ IIT 4.0 introduces an important difference in how information is quantified $\rightsquigarrow$ dissimilarity function.
  → Go check their paper [ABF+23].

Thank you very much for your attention

Thank you very much for your attention!

References I

- Larissa Albantakis, Leonardo Barbosa, Graham Findlay, Matteo Grasso, Andrew M. Haun, William Marshall, William G. P. Mayner, Alireza Zaeemzadeh, Melanie Boly, Bjørn E. Juel, Shuntaro Sasai, Keiko Fujii, Isaac David, Jeremiah Hendren, Jonathan P. Lang, and Giulio Tononi, *Integrated information theory (iit) 4.0: Formulating the properties of phenomenal existence in physical terms*, PLOS Computational Biology **19** (2023), no. 10, 1–45.
- D. Rudrauf, D. Bennequin, I. Granic, G. Landini, , K. Friston, and K. Williford, *A mathematical model of embodied consciousness*, Journal of Theoretical Biology (2017).
- David Rudrauf, Daniel Bennequin, and Kenneth Williford, *The Moon illusion explained by the projective consciousness model*, Journal of Theoretical Biology (2020).

References II

- David Rudrauf, Grégoire Sergeant-Perthuis, Olivier Belli, Yvain Tisserand, and Giovanna Di Marzo Serugendo, *Modeling the subjective perspective of consciousness and its role in the control of behaviours*, arXiv:2012.12963, 2020.
- Grégoire Sergeant-Perthuis, Nils Ruet, David Rudrauf, Dimitri Ognibene, and Yvain Tisserand, *Influence of the geometry of the feature space on curiosity based exploration*, NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems, 2023.
- Kenneth Williford, Daniel Bennequin, Karl Friston, and David Rudrauf, *The projective consciousness model and phenomenal selfhood*, Frontiers in Psychology (2018).