arXiv:2303.04571v2 [cs.AI] 3 May 2023

一般智能的范畴框架

Yang Yuan[1,2,3]

[1]IIIS，清华大学[2]上海人工智能实验室[3]上海期智研究院

2023年5月4日

摘要

机器能够思考吗？自艾伦·图灵于1950年提出这一问题以来，由于缺乏通用智能的坚实数学基础，一直无人能够给出直接答案。在本文中，我们提出了一个旨在实现这一目标的范畴框架，并取得了两项主要成果。首先，我们通过预层研究对象表示，引入了作为自我意识之范畴类比的自我状态意识概念，并提出了相应的实现与评估算法。其次，我们利用图表和极限将对象表示扩展为情景表示，进而将其作为数学建模、可解释性和人工智能安全的构建模块。作为一项附带成果，我们的框架引入了多种范畴不变性，这些不变性可作为模型训练的对齐信号。

1 引言

近年来，借助强大的计算能力、海量数据和庞大的神经网络训练基础模型，已经取得了显著进展（Radford 等，2021；Chen 等，2020；Radford 等，2019；Brown 等，2020；Ramesh 等，2021、2022；Sohl-Dickstein 等，2015；Rombach 等，2022；He 等，2022）。令人惊讶的是，尽管取得了令人瞩目的成就，这些模型的内部工作机制仍然神秘莫测。人们似乎已达成共识，认为基础模型本质上是黑箱，无法解释，因此，实证实验是推动人工智能发展的唯一途径。

过去十年确实是这样发展的，这也类似于智能通过进化获得的方式；然而，如果仅依赖实证实验而缺乏理论理解，既可能效率低下，也可能带来危险。效率低下的原因在于，进展是通过反复试错取得的，往往由直觉引导，而且里程碑是根据特定任务上的性能间接界定的，而非基于对智能本身的全面理解。潜在危险则在于，没有人知道我们最终会得到什么，也许更重要的是，我们目前距离终点还有多远。我们甚至不知道自己是否已经创造出了通用智能——也许还没有，但该如何进行这样的评估呢?

本文提出了一种通用智能的范畴框架，有助于回答以下问题：

Q1：模型能否觉察到自身的自我状态？（第 3 节）

Q2：模型应如何表示复杂场景？（第 4 节）

Q3：我们应如何训练模型，使其迈向通用智能？（第 5 节）

鉴于人类对通用智能的形式化定义尚未达成共识，要证明我们的框架适用于 通用智能将极具挑战性，甚至是不可能的。相反，我们采用范畴方法：对所有基本要素进行形式化定义，阐明它们的理论含义，明确算法要求，最后将所有要素整合到一个全面的框架中。因此，即使有人不同意我们对通用智能的定义，或认为缺少某些关键部分，我们的框架仍然具有相关性和适用性。

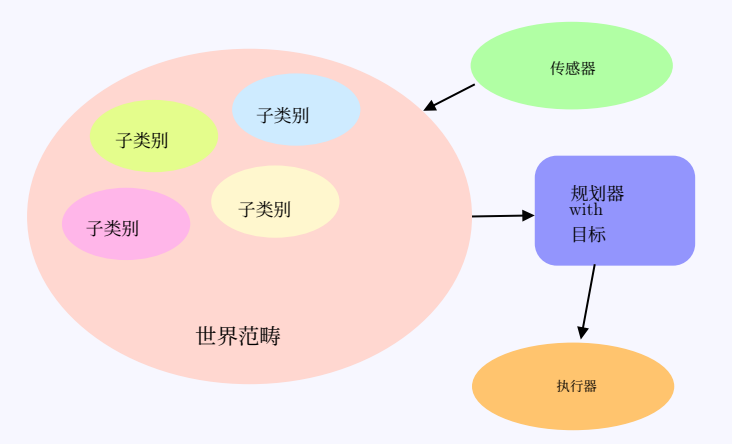

图1：我们的范畴框架

我们的框架出奇地简单，由四个主要组成部分构成[1]：传感器、世界范畴、带有目标的规划器，以及执行器，各部分之间的信息流是单向的（见图1）。传感器接收来自外部环境的多模态信号，包括但不限于文本输入、视频/图像输入、音频输入、触觉等。世界范畴感知并理解输入信号，并据此更新其内部状态。规划器持续监测世界范畴的状态，并根据其目标生成计划。最后，执行器执行这些计划，通过生成文本输出、视频/图像输出、音频输出、机器人操作信号等影响外部环境。我们将在下文详述这些内容，首先介绍三种不同的世界观。

三个不同的世界 存在于我们的框架中（见图2）：现实世界、感知世界和世界范畴。模型无法直接访问现实世界[2]，，只能通过传感器访问感知世界。对于任意给定的时间 $t \in \mathbb{N}$，现实世界都可能远大于模型的感知世界，因为模型可能只能感知世界的一小部分。即使在传感器的感知范围内，现实世界与感知世界也不一定相同，因为传感器的感知能力可能有限、存在偏差或包含噪声。此外，世界范畴编码了模型对世界的理解，因此可能远大于感知世界。直观而言，感知世界是模型当下所见所感，而世界范畴则是模型对世界的理解、记忆和预测。

[1]暂且不考虑范畴这一部分——这是我们的主要贡献——我们要指出，我们的框架中的组成部分及其连接方式与 LeCun（2022）提出的方案相似但并不相同，详见第 6 节。[2]这个“现实世界”完全可能是模拟出来的；对此我们不作任何区分。

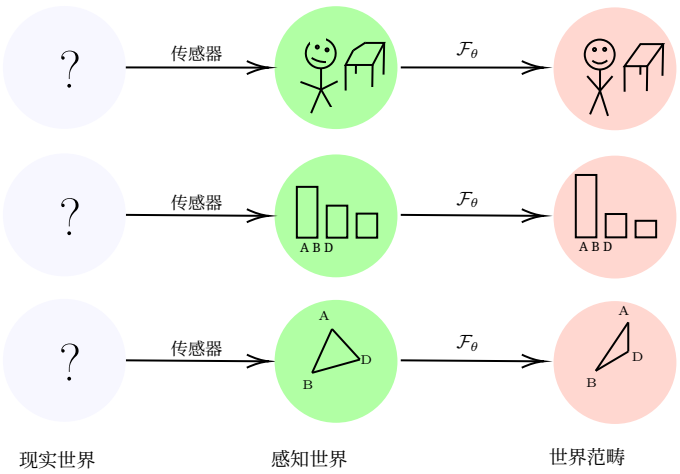

图2：关于三个不同世界的案例研究。第一行中，模型看到一个坐在桌旁的人类，但这个人有一部分被遮挡，桌子也缺了一个角。在世界范畴中，当模型记住这个场景时，可能会对人体稍作调整，以补全被遮挡的部分，并略微修改桌子的形状。第二行中，模型看到一张图表，展示 Alice、Bob 和 David 的财富。然而，模型可能无法在其世界范畴中准确记住这些信息。第三行中，模型感知 Alice、Bob 和 David 之间的关系，例如他们当中是否有两人是亲密朋友。在其世界范畴中，这类关系可能会以失真的形式保留下来。

为了展现智能行为，模型应通过传感器和执行器与外部环境交互，因此时间 t 这一概念十分重要。如果模型无法获取时间戳，t 也可以表示传感器检测到的不同事件的索引，例如 ChatGPT 回答用户提出的各种问题。为求简便，我们不作此类区分，而是将 t 一概视为时间戳。

世界范畴可以被视为模型通过其传感器对外部世界进行的想象性重构。它包含外部环境中模型能够通过传感器感知的所有人物、动物、对象和知识，以及基于它们形成的抽象表征。更准确地说，它是 $\mathbb{N} \to \mathbf{Cat}$ 中的一个函数，表示一个动态范畴[3]，其中包含各种对象并随时间变化 $t \in \mathbb{N}$。传感器决定了世界范畴中基本对象的类型。例如，如果模型无法检测视觉信号，世界范畴中就不会包含图像对象。此外，如果传感器接收到来自模拟环境的信号，世界范畴将只包含模拟知识，而这可能与现实世界大相径庭。

对象表示。对于任意给定的对象 X，世界范畴不会直接存储X，而是使用神经网络存储其他对象与 X 之间的所有关系，其中包含关于 X的充分信息。一个显著的例子是，当模型能够感知自身与其他对象之间的关系时，其世界范畴便可以表征一个称为“自我状态”的对象，用于存储这些关系。维持自我状态是否等同于拥有自我意识？这是一个有争议的问题，我们选择不予回答。相反，我们基于范畴论正式定义自我状态这一概念，而不探讨其与自我意识之间的关系（第3.1节）。

基于我们的 d 在定义自我状态时，我们介绍了两种用于强化和评估

[3]不熟悉范畴论的读者可以查看第2节，了解基础介绍。

模型自我状态意识的算法。不同于对自我意识的识别——后者是一个二元变量，用于表示主体是否具有自我意识——我们的评估会生成一个位于区间 $[0,1]$ 内的连续值，以表示自我状态意识的程度。该程度对应于主体知晓的、主体与其他对象或任务之间所有相关关系的比例。根据我们的定义，即使许多人类，尤其是儿童，也可能不具备完美的自我状态意识。

场景表示。 模型应如何表示包含多个对象及其间态射的情景？借助范畴论的语言，我们可以将情景内容定义为一个图表，并将情景本身定义为该图表上的一个射影极限。

正如我们将看到的，使用图表和极限进行场景表示会带来各种有趣的结果。例如，它能大大简化数学建模，因为我们可以对图表进行适当的抽象，直接将其转化为数学问题。此外，通过将情景内容视为世界范畴的图表，模型便能够基于其内部知识生成解释，而不必像归因方法那样，仅限于为输入变量分配权重（Sundararajan 等人，2017年；Lundberg 和 Lee，2017年）。此外，情景内容的图表表示还能通过将自我状态注入图表并确保自我状态符合人类偏好，为人工智能安全提供一种函数式方法。

不变性作为训练信号。范畴论利用交换图来刻画不同计算路径之间的等价性，这自然引出了模型的各种不变性。与监督学习旨在使输入数据与正确的输出标签相匹配不同，基础模型侧重于学习对象之间的态射以及范畴之间的函子。不变性充当训练信号，引导模型进行调整，从而使世界范畴自然保持一致。

2 预备知识

范畴论几乎应用于数学的所有领域。本文仅介绍理解本文结果所需的概念。感兴趣的读者可参阅 Mac Lane（2013年）、Riehl（2017年）和 Adámek 等人（1990年），以获取更全面的介绍。

2.1 范畴基础

一个范畴 $\mathcal{W}$ 有一组对象 $\mathrm{Ob}(\mathcal{W})$，并且对于每个 $X, Y \in \mathrm{Ob}(\mathcal{W})$，都有一组从 X 到 Y 的态射 $\mathrm{Hom}_{\mathcal{W}}(X, Y)$。本文中，我们将“关系”和“态射”交替使用。给定 $f \in \mathrm{Hom}_{\mathcal{W}}(X, Y)$ $g \in \mathrm{Hom}_{\mathcal{W}}(Y, Z)$，我们将它们的复合定义为 $g \circ f \in \mathrm{Hom}_{\mathcal{W}}(X, Z)$。注意，$\circ$ 满足结合律，即 $(h \circ g) \circ f = h \circ (g \circ f)$。对于每个 $X \in \mathrm{Ob}(\mathcal{W})$，都存在唯一的恒等态射 $\mathrm{id}_X \in \mathrm{Hom}_{\mathcal{W}}(X, X)$。若存在 $g\colon X \leftarrow Y$ 使得 $f \circ g = \mathrm{id}_Y$ 且 $g \circ f = \mathrm{id}_X$，则态射 $f\colon X \to Y$ 是一个同构。在这种情况下，我们称 X 和 Y 同构，并记作 $X \simeq Y$。

我们考虑一个全集 $\mathcal{U}$[4]。若对于任意 X，$Y \in \mathrm{Ob}(\mathcal{W})$，$\mathrm{Hom}_{\mathcal{W}}(X, Y)$ 都是 $\mathcal{U}$-小的，则称范畴是一个 $\mathcal{U}$-范畴。若一个 $\mathcal{U}$-范畴 $\mathcal{W}$ 是 $\mathcal{U}$-小的，则 $\mathrm{Ob}(\mathcal{W})$ 是 $\mathcal{U}$-小的。为简单起见，下文可能不会明确提及全集 $\mathcal{U}$，而只说 $\mathcal{W}$ 是一个范畴或小范畴。我们将 $\mathbf{Cat}$ 定义为以小范畴为对象的范畴。

给定范畴 $\mathcal{W}$，我们通过令 $\mathrm{Ob}(\mathcal{W}^{\mathrm{op}}) = \mathrm{Ob}(\mathcal{W})$ 来定义其对偶范畴 $\mathcal{W}^{\mathrm{op}}$，并令 $\mathrm{Hom}_{\mathcal{W}^{\mathrm{op}}}(X, Y) = \mathrm{Hom}_{\mathcal{W}}(Y, X)$。此外，给定 $f \in \mathrm{Hom}_{\mathcal{W}^{\mathrm{op}}}(X, Y)$ $g \in \mathrm{Hom}_{\mathcal{W}^{\mathrm{op}}}(Y, Z)$，新的复合为 $g \circ_{\mathrm{op}} f = f \circ g \in \mathrm{Hom}_{\mathcal{W}^{\mathrm{op}}}(X, Z)$。

[4]相关定义请参见 Masaki Kashiwara（2006年）。

我们将 **集合范畴** 定义为集合的范畴，其中对象是集合，而 $\mathrm{Hom}_{\mathbf{Set}}(X,\ Y)$ 是所有定义域为 X 、陪域为 Y 的函数构成的集合。请注意，为了更好地呈现内容，我们忽略了关于全集的细节，因此这里仅假设 **集合范畴** 不包含诸如包含所有集合的集合这类奇异对象。

函子类似于两个范畴之间的函数。给定两个范畴 $\mathcal{W},\mathcal{W}'$，函子$F:\ \mathcal{W}\to\mathcal{W}'$ 将对象从 $\mathcal{W}$ 映射到 $\mathcal{W}'$，其映射为 $F:\mathrm{Ob}(\mathcal{W})\to\mathrm{Ob}(\mathcal{W}')$；并将态射从 $\mathcal{W}$ 映射到$\mathcal{W}'$，其映射为 $F:\mathrm{Hom}_{\mathcal{W}}(X,Y)\to \mathrm{Hom}_{\mathcal{W}'}(F(X)\ F(Y))$，对所有 $X,Y\in\mathcal{W}$均成立，并且 F保持恒等态射和复合。形式上，对于所有 $X\in\mathcal{W}$，有 $F(\mathrm{id}_X)=\mathrm{id}_{F(X)}$；对于所有 $f:\ X\to Y, g:\ Y\to Z$，有 $F(g\circ f)=F(g)\circ F(f)$。

函子之间的态射也称为自然变换，是在保持结构的同时变换函子的方式。给定两个范畴 $\mathcal{W},\ \mathcal{W}'$，以及两个函子 $F_1,\ F_2$，它们从$\mathcal{W}$到 $\mathcal{W}'$。函子之间的态射 $\theta:\ F_1\to F_2$ 具有一个态射 $\theta_X:\ F_1(X)\to F_2(X)$，对所有 $X\in\mathcal{W}$均成立，使得对于所有 $f:\ X\to Y\in\mathrm{Hom}_{\mathcal{W}}(X,\ Y)$，都有 $\theta_Y(F_1(f)(F_1(X)))=F_2(f)(\theta_X(F_1(X)))$。

预层是一个从 $\mathcal{W}^{\mathrm{op}}$ 到**集合范畴**的函子，而 $\mathcal{W}^{\wedge}$ 是预层范畴。类似地，从 $\mathcal{W}^{\mathrm{op}}$ 到**集合范畴**^{op}的函子称为**集合范畴**^{op}-值预层，而 $\mathcal{W}^{\vee}$ 是**集合范畴**^{op}-值预层的范畴。本文不作区分，将这两类函子都称为预层，并将 $\mathcal{W}^{\wedge},\ \mathcal{W}^{\vee}$ 都称为预层范畴。此外，定义 Yoneda 函子 $h_{\mathcal{W}}(X)\triangleq$同态集$_{\mathcal{W}}(\cdot,\ X)\in\mathcal{W}^{\wedge}$以及 $k_{\mathcal{W}}(X)\triangleq$同态集$_{\mathcal{W}}(X,\ \cdot)\in\mathcal{W}^{\vee}$。下面的引理是基础性的。

引理 1（米田引理）。 给定 $X\in\mathcal{W}$ 我们有，

(1) 对于 $A\in\mathcal{W}^{\wedge}$，同态集$_{\mathcal{W}^{\wedge}}(h_{\mathcal{W}}(X)\ \ A)\ \simeq A(X)$。
(2) 对于 $B\in\mathcal{W}^{\vee}$，同态集$_{\mathcal{W}^{\vee}}(B,\ k_{\mathcal{W}}(X))\ \simeq B(X)$。

米田引理表明，$h_{\mathcal{W}}(X)$ 和 $k_{\mathcal{W}}(X)$ 包含了 X 的全部信息。直接推论是，$\mathrm{Hom}_{\mathcal{W}^{\wedge}}(h_{\mathcal{W}}(X)\ h_{\mathcal{W}}(Y))\simeq h_{\mathcal{W}}(Y)(X)=\mathrm{Hom}_{\mathcal{W}}(X,\ Y)$，而 $k_{\mathcal{W}}(\cdot)$ 也有类似的结果。若函子 F 从 $\mathcal{W}^{\mathrm{op}}$ 到 **集合范畴**（或从 $\mathcal{W}$ 到 **集合范畴**）与某个 $X\in\mathcal{W}$ 的 F 之间存在同构，则该函子是可表的。这样的 X称为 F 的表示对象。

2.2 极限

范畴 $\mathcal{W}$ 中形状为 A 的图表是一个函子 $\alpha:\ A\to\mathcal{W}$，它将 $\mathcal{W}$ 中的对象映射到与 $\mathrm{Ob}(A)$ 对应的对象，并保持 A 中的态射。给定一个函子 $\beta:A^{\mathrm{op}}\to$ **集合范畴**，将其射影极限定义为 $\lim_{\longleftarrow}\beta\triangleq\mathrm{Hom}_{A^{\wedge}}(\mathrm{pt}_{A^{\wedge}},\ \beta)$，其中对每个 $i\in A$，$\mathrm{pt}_{A^{\wedge}}(i)=\{\mathrm{pt}\}$，且 $\{\mathrm{pt}\}$ 是单点集。换言之，$\lim_{\longleftarrow}\beta$ 表示 $\mathrm{pt}_{A^{\wedge}}$ 与 β 之间所有自然变换的集合。基于对 **集合范畴** 中图表的这一定义，我们可以给出极限的一般定义。

定义 1（射影极限与归纳极限）。 给定 $\alpha:\ A\to\mathcal{W},\beta:\ A^{\mathrm{op}}\to\mathcal{W}$，其中 A 是小的，归纳极限 $\varinjlim\alpha\in\mathcal{W}^{\vee}$ 和射影极限 $\varprojlim\beta\in\mathcal{W}^{\wedge}$ 定义如下：

(1)$\varinjlim^{\alpha}:\ X\mapsto\lim\leftarrow$ 同态集$_{\mathcal{W}}(\alpha,\ X)(2)\varprojlim\beta:\ X\mapsto\lim\leftarrow$ 同态集$_{\mathcal{W}}(X,\beta)$

这里，同态集$_{\mathcal{W}}(\alpha,X)$ 是一个函子，它将 $i\in A$ 映射到 同态集$_{\mathcal{W}}(\alpha(i)\ X)$。因此，$\lim\leftarrow$ 同态集$_{\mathcal{W}}(\alpha,X)$ 是集合范畴中图表的一个定义良好的极限。同理，$\lim\leftarrow$ 同态集$_{\mathcal{W}}(X,\beta)$也是如此。

引理 2（Masaki Kashiwara（2006年）第 60 页）。 若 A 是小的，考虑 $\alpha:\ A\to\mathcal{W}^{\wedge},\ \beta:\ A^{\mathrm{op}}\to\mathcal{W}^{\vee}$。对于 $A\in\mathcal{W}^{\wedge},\ B\in\mathcal{W}^{\vee}$，

$$\mathrm{Hom}_{\mathcal{W}^{\wedge}}(\varinjlim\alpha,A)\simeq\varprojlim\mathrm{Hom}_{\mathcal{W}^{\wedge}}(\alpha,A)$$
$$\mathrm{Hom}_{\mathcal{W}^{\vee}}(B,\varprojlim\beta)\simeq\varprojlim\mathrm{Hom}_{\mathcal{W}^{\vee}}(B,\beta)$$

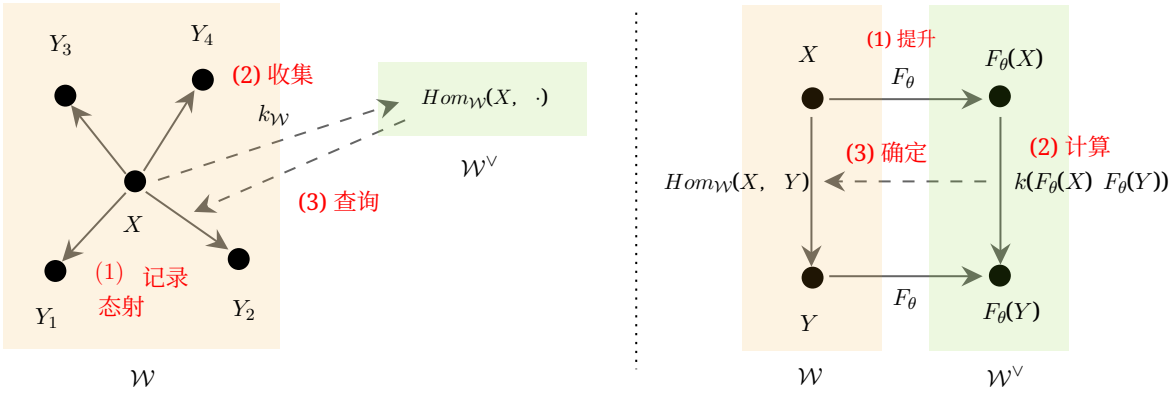

图3：构建世界范畴的两种方式。左：假设存在一个预先定义了态射的范畴$\mathcal{W}$，我们可以使用 $k_{\mathcal{W}}$ 直接计算给定 $X \in \mathcal{W}$的 $\mathrm{Hom}_{\mathcal{W}}(X,\ \cdot)$，然后查询任意 $Y \in \mathcal{W}$对应的 $\mathrm{Hom}_{\mathcal{W}}(X,\ Y)$。右：由于我们尚不知道 $\mathcal{W}$中的态射，因此先针对给定的 $X,\ Y \in \mathcal{W}$计算 $\mathcal{F}_\theta(X)\ \mathcal{F}_\theta(Y)$，再在 $\mathcal{W}^\vee$中计算 $k(\mathcal{F}_\theta(X)\ \mathcal{F}_\theta(Y))$，由此确定 $\mathrm{Hom}_{\mathcal{W}}(X,\ Y)$。

3 世界范畴与对象表示

回想一下，世界范畴是模型对现实世界的想象。给定时间$t \in \mathbb{N}$，我们该如何表示世界范畴快照 $\mathcal{W}(t)$？ $\mathcal{W}(t)$包含对象及其态射，但直接存储这些信息通常在计算上不可行。我们改用一个由 $\theta(t)$ 参数化的函数 $\mathcal{F}_{\theta(t)}: \mathcal{W}(t) \to \mathcal{W}^\vee(t)$，它将对象 X 映射到包含 X的所有态射的 $k_{\mathcal{W}}(X)$。如此一来，模型不会在 $\mathcal{W}(t)$ 中显式存储任何态射，而是通过 $\mathcal{F}_{\theta(t)}$包含所有必要信息。这里，为了符号表示方便，我们假设 $\mathcal{F}_{\theta(t)}$ 将对象映射到 $\mathcal{W}^\vee$；在经验上，$\mathcal{F}_{\theta(t)}$ 可以同时包含 $\mathcal{W}^\wedge$和 $\mathcal{W}^\vee$的信息，也就是说，对于每个 X，$\mathrm{Hom}_{\mathcal{W}}(\cdot,\ X)$ 和 $\mathrm{Hom}_{\mathcal{W}}(X,\ \cdot)$ 都包含在内。例如，$\mathcal{F}_{\theta(t)}$ 映射到的嵌入空间可以写作 $\mathcal{W}^\wedge \times \mathcal{W}^\vee$，尽管神经网络在经验上能够找到更好的编码机制。

在我们的框架中，模型在任意给定时间 t仅使用 $\mathcal{F}_{\theta(t)}$ 表示单个快照 $\mathcal{W}(t)$，而快照会随时间变化。为避免混淆，我们省略参数 t，直接将快照记为 $\mathcal{W}$，并将函数 $\mathcal{F}_{\theta(t)}$ 记为 $\mathcal{F}_\theta: \mathcal{W} \to \mathcal{W}^\vee$。下面的定义极其重要且有趣。

定义 2 (基于 $\mathcal{F}_\theta$的世界范畴)。假设 $\mathcal{W}$ 中的对象固定。给定世界范畴表示函数 $\mathcal{F}_\theta$：$\mathcal{W} \to \mathcal{W}^\vee$，以及一个数据无关函数 k：$\mathcal{W}^\vee \times \mathcal{W}^\vee \to$ 集合范畴，用于表示 $\mathcal{W}^\vee$中的态射；对于任意 X、$Y \in \mathcal{W}$，有同态集$_{\mathcal{W}}(X,\ Y) \triangleq k(\mathcal{F}_\theta(X)\ \mathcal{F}_\theta(Y))$。

这里，“数据无关”意味着 $k(\cdot,\cdot)$ 是预先定义的，无需查看数据。例如，它可以定义为两个输入之间的内积。为了理解定义 2，请考虑从 $\mathcal{W}$ 到 $\mathcal{W}^\vee$ 的映射 $X \to \mathrm{Hom}_{\mathcal{W}}(X,\cdot)$。这个映射是自然的，因为如果我们知道 X 以及围绕 X 的所有关系，就可以计算出能够编码所有这些关系的 $\mathrm{Hom}_{\mathcal{W}}(X,\cdot)$。然而，定义 2 考虑的是相反的方向：我们知道 $k_{\mathcal{W}}(X)\ k_{\mathcal{W}}(Y)$（在定义 2 中由 $\mathcal{F}_\theta(X)$ 和 $\mathcal{F}_\theta(Y)$ 表示），并希望恢复 $\mathrm{Hom}_{\mathcal{W}}(X,Y)$。乍看之下，这可能有些反直觉，因为如果我们将 Y 发送到 $k_{\mathcal{W}}(X)$，就会得到 $k_{\mathcal{W}}(X)(Y) = \mathrm{Hom}_{\mathcal{W}}(X,\cdot)(Y) = \mathrm{Hom}_{\mathcal{W}}(X,Y)$。换句话说，$\mathrm{Hom}_W(X,Y)$ 的定义是循环的，因为我们使用 $k_{\mathcal{W}}(Y)$ 来定义 $\mathrm{Hom}_W(X,Y)$，但 $k_{\mathcal{W}}(Y)$ 的定义本身又包含 $\mathrm{Hom}_W(X,Y)$!



图 4: 自我状态的两种不同表征。第一个子图展示了模型 I 以及 I 与其他对象之间对应的态射。第二个子图是对象表示 I，最后一个子图是态射表征 $I^\vee$，它是一个预层。

我们如何在不使用 $\mathrm{Hom}_W(X, Y)$ 的情况下定义 $k_{\mathcal{W}}(Y)$? 定义 2 使这种循环定义成为可能，因为 $k_{\mathcal{W}}(\cdot)$ 由函数 $\mathcal{F}_\theta(\cdot)$ 表示，该函数将对象映射为嵌入向量。换句话说， Y 的所有关系都嵌入在 $\mathcal{F}_\theta(Y)$ 中，并且可以根据 $\mathcal{F}_\theta(X)$ 和 $\mathcal{F}_\theta(Y)$ 恢复 $\mathrm{Hom}_W(X, Y)$。图3对此进行了说明。因此，世界范畴由函数 $\mathcal{F}_\theta$ 定义和表示。从这个定义出发，我们有:

- 即使 $\mathcal{W}$ 包含无限多个对象， $\mathcal{F}_\theta$ 仍然可以是有限的（例如，神经网络），并编码无限多个态射。
- 为了让模型了解 $\mathcal{W}$中的对象 X ，只需让模型学习 $\mathcal{W}^\vee$中 X 的态射即可。因此，没有必要直接观察 X 。
- 世界范畴由 $\mathcal{F}_\theta$控制，因此 $\mathcal{F}_\theta$ 所计算的内容就是模型所理解的内容。$\mathcal{F}_\theta$ 是世界范畴的真实值。
- 由 $\mathcal{F}_\theta$ 生成的世界范畴可能与感知世界或现实世界大不相同，尤其是在 $\mathcal{F}_\theta$ 的表征能力有限时。
- 即使模型先前对对象 $X \in \mathcal{W}$知之甚少或一无所知，仍然可以基于 $\mathcal{F}_\theta(X)$，生成大量关于 X 的态射。

由于 $\mathcal{F}_\theta$ 决定了给定模型的 $\mathcal{W}$ ，下文可能用 $\mathcal{F}_\theta$ 表示世界范畴。除了态射之外，我们还可以用 $\mathcal{F}_\theta$ 来编码任务。

定义 3 (任务)。 任务 $T: \mathcal{W}^{\mathrm{op}} \to$ 集合范畴^{op} 是 $\mathcal{W}^\vee$中的一个函子。

根据米田引理， $T(X) \simeq$同态集$_{\mathcal{W}^\vee}(T,\ k_{\mathcal{W}}(X))$。因此，当任务 T是可表的时，$T(X)$可以通过 $k(T,\ \mathcal{F}_\theta(X))$计算得出。

3.1 自我状态

在世界范畴中，可能存在一个称为自我状态的特殊对象，定义如下。

定义 4 (自我状态)。 给定 I 作为表示 $\mathcal{W}$ 中模型的对象，世界范畴中的自我状态是预层 $I^\vee \triangleq \mathcal{F}_\theta(I) \in \mathcal{W}^\vee$。

对象 $I \in \mathcal{W}$ 和预层 $I^\vee$ 非常不同： I 是一个单独的对象，不包含任何额外信息；而 $I^\vee$ 包含了 I 与其他对象之间的所有态射。此外，根据米田引理，对于任意任务 $T \in \mathcal{W}^\vee$，我们还有 $\mathrm{Hom}_{\mathcal{W}^\vee}(T, I^\vee) \simeq T(I)$。这意味着， $I^\vee$ 也编码了每个相关任务所需的全部信息。

正如我们在定义 2 中所讨论的，即使模型无法在感知世界中感知到自身，它仍然可以拥有自我状态。例如，请看下面的思想实验。

思想实验 1 (能看见和听见的瘫痪者**).** 设想一位人类 A，他无法控制自己的身体，也失去了所有身体感觉，包括嗅觉、皮肤感觉、眼睑控制、心跳、温度感受等。自出生以来， A 唯一能接收外界信息的方式就是通过眼睛和耳朵。 A 能否维持自我状态?

目前尚不清楚是否真的存在过这样的人，但在这个思想实验中，由于 A 能看见和听见，他仍然可以接收来自外部世界的信息。例如，其他人物可以向他展示日常图像和视频，或者与他交谈，比如：“你是 Bob”、“我是你的父亲”、“你现在 10 岁了”、“你出生在 α-镇”等等。

这些信息都不代表 A 本身，而是描述了 A 与感知世界中其他事物之间的关系，而这些关系恰好编码在 $I^\vee$中。换言之，根据定义 4， A 的大脑中仍然可以具有自我状态，用来表示所有与自身相关的关系。这立刻引出了以下思想实验：如果 A 没有身体，会怎么样?

思想实验 2 (聊天机器人**).** 考虑一个通过文本和图像与外部世界互动的聊天机器人 A，例如 *GPT-4*。 A 能维持自我状态吗?

根据上述讨论，答案是肯定的，前提是 A 完全理解其他对象与自身之间的所有关系。换言之，在定义 4 中，维持自我状态并不需要自我身体。智能体只需感知其他对象与自身之间的所有关系即可。

并非所有世界范畴都有自我状态。例如，如果我们使用 SimCLR（Chen 等，2020）在图像上训练基础模型，那么所有关系描述的都是图像之间的相似性，而该世界范畴并不具有自我状态。有趣的是，几乎所有现有的计算机程序都属于这一类。综合以上讨论，我们得到以下推论。

推论 1 (自我状态判据**).** 当且仅当模型能够通过其传感器学习预层 $I^\vee$ 时，模型才能维持自我状态。

$I^\vee$ 中编码的所有关系未必都直接输入模型，也可以通过推断得出。在这种情况下，模型仍然可以拥有自我状态。由此得到下一个推论。

推论2 (自我状态的涌现**).** 对于给定的模型，如果要重建感知世界，就不可避免地需要学习模型自身与其他对象之间的关系，那么该模型就会维持自我状态。

例如，ChatGPT 必须对世界上几乎所有能够用自然语言描述的事物进行编码，而且似乎不可避免地要对其他对象与自身之间的关系进行编码。因此，推论2表明，ChatGPT 将维持自我状态。

当模型维护一个自状态对象时，由于世界范畴是动态学习的，这种维护可能并不准确。我们将自我状态意识定义为该对象准确表征模型的程度。为了测试自我状态意识，我们首先定义相应的测试。

定义 5（自我状态意识测试）。 自我状态意识测试是一个函子 $T:\ \mathcal{W}^{\vee} \to \{0,1\}$，它接受一个属于 $\mathcal{W}^{\vee}$ 的预层 $I^{\vee}$，并输出 $I^{\vee}$ 是否通过测试 T。

例如，如果模型的名字是“Sydney”，那么自我状态意识测试就是一个函子，它以 $I^{\vee}$ 作为输入，评估 $\mathrm{Hom}_{\mathcal{W}^{\vee}}(I^{\vee},\ k_{\mathcal{W}}(\text{“Sydney”})) \simeq \mathrm{Hom}_{\mathcal{W}}(I,\ \text{“Sydney”})$，并输出该态射是否确实表示“Sydney”是 I 的名字。不过，仅仅通过一项测试不足以判定其具备自我状态意识，我们需要设置多种测试。

定义 6（在 $\mathcal{T}$ 下的自我状态意识）。 给定一组自我状态测试 $\mathcal{T}$, $\delta \in [0,1]$，当模型在其世界范畴中具有自我状态 I 时，如果 $\mathbb{E}_{T\in\mathcal{T}}(T(I)) \geq \delta$，则它在 $\mathcal{T}$ 下对其自我状态具有 δ 意识。

测试集 $\mathcal{T}$ 的选择取决于测试目标。如果选择的测试集使信号难以察觉，那么即使是人类也未必能轻易通过这些测试。例如，如果你的肾脏里有肾结石，只有在进行肾脏扫描或感到疼痛时，你才能意识到这一事实。同样，当你身处嘈杂环境中，有人喊你的名字时，你可能无法及时作出反应。

算法1 评估自我状态意识

输入： 世界模型 $\mathcal{F}_\theta$，自我对象 I 在 $\mathcal{W}$中，自我状态测试集 $\mathcal{T}$
令 $s = 0$
对于 $i = 1$ 至 m，**执行**
  采样一个任务 $T \in \mathcal{T}$
  令 $s = s + T(\mathcal{F}_\theta(I))$
结束循环
返回 $s/m \in [0,1]$

算法 2 强制实现自我状态意识

输入： 世界模型 $\mathcal{F}_\theta$，自我对象 I，位于 $\mathcal{W}$ 中，自我状态测试集 $\mathcal{T}$令 $s = 0$**对于** $i = 1$ 到 m **执行**采样一个任务 $T \in \mathcal{T}$令 $\ell(T,\ I,\ \mathcal{F}_\theta) = -T(\mathcal{F}_\theta(I))$ 对 $\ell(T,\ I,\ \mathcal{F}_\theta)$ 运行反向传播以优化 $\mathcal{F}_\theta$**结束循环**

定义 6 立即给出了用于评估自我状态意识的算法1。

3.2 学习自我状态

模型应如何学习其自我状态? 基于前面的讨论，我们提出算法 2 作为一种通用解决方案。然而，算法 2 需要来自测试集 $\mathcal{T}$ 的监督信号，而这些信号通常无法获得。为了在没有测试集的情况下学习状态—状态，某种模型的先验知识或许会有所帮助。例如，模型可以假设：

- 它能够控制的就是它自身。
- 它能从私有传感器感受到的就是自我。这里的私有传感器是指心跳、温度感受、皮肤感觉等传感器，也就是预先定义好的自我传感器。

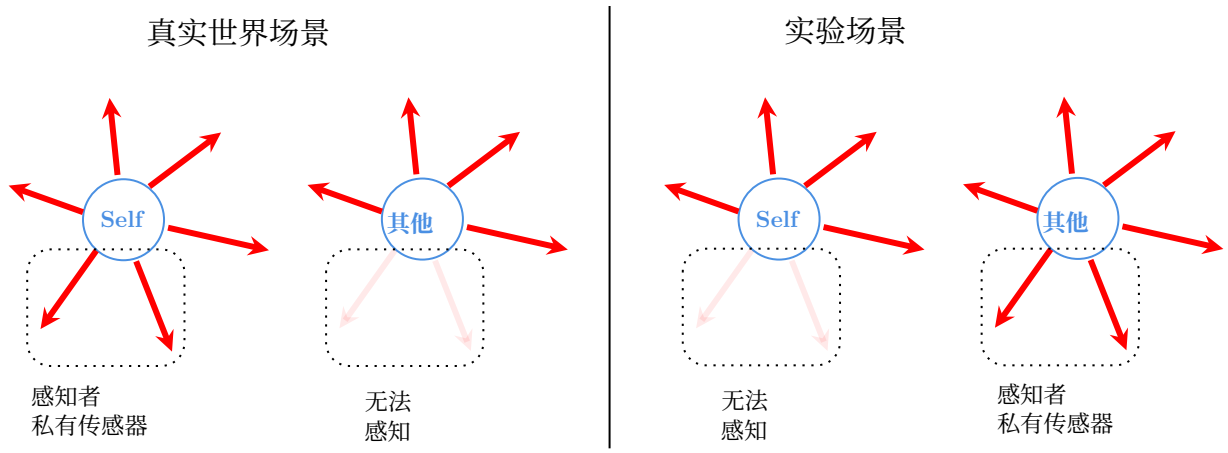

图5：关于自我状态和他者状态的两种不同情景。在现实世界中，当模型只能访问自己的私有传感器，而无法访问其他智能体的传感器时，自我状态与他者状态存在根本差异。然而，在实验场景中，如果模型只能访问另一智能体的私有传感器，它很快就会对这个外部智能体产生身体所有感。

以人类的手为例。根据定义 4，我们觉得手是身体的一部分，因为大脑会告诉我们**这种关系**：我们看到的手属于自己的身体。事实上，人脑会主动实时整合感觉信号，让我们对自己的手产生全面的所有权感。具体来说，当我们看到有人触碰自己的手，同时也感受到触碰时，大脑会迅速调整这两个信号，确保它们指向身体上的同一部位。

这一学习视角与神经科学中一项引人入胜的观察密切相关，这种现象被称为橡胶手错觉（Ehrsson 等人，2004 年；Botvinick 和 Cohen，1998 年；Tsakiris 和 Haggard，2005 年）。在实验中，实验者同时抚摸人类参与者一只被遮挡的真实手，以及参与者面前的一只橡胶手。由于真实手感受到的抚摸触觉与橡胶手的视觉信号同时传入大脑，人类参与者很快便会产生对橡胶手的所有感。将视觉信号替换为听觉反馈，也能得到类似的实验结果。因此，如果模型采用类似的学习算法来对齐多模态信号（例如对比学习），它也会经历类似的错觉。

3.3 共情

将定义 4 推广到其他智能体，我们得到：

定义 7 (他者状态**)**。给定 A 作为表示范畴 $\mathcal{W}$ 中智能体 A 的对象，世界范畴中 A 的状态是预层 $A^\vee \triangleq \mathcal{F}_\theta(A) \in \mathcal{W}^\vee$。

因此，所有关于自我状态的讨论也适用于其他智能体的状态。例如，只要模型能够推断出关于某个智能体的许多关系，就不必直接与该智能体“面对面”接触，也能维持相应的状态。

然而，这与我们的日常体验不同，因为我们本能地知道，自我意识与对他人的共情截然不同。为什么这两个概念的范畴定义

在这里却相同？这是因为存在私有传感器（如心跳、温度感受、皮肤感觉等），它们可用于表征自我状态，却无法表征其他状态。请看以下三个案例。

1. 在全信息设定下，如果讨论中不涉及私有传感器，那么共情与自我状态意识是相同的。例如，在一个多智能体博弈中，每个智能体都有自己的动作集合、状态和奖励函数，共情有助于理解并预测每个智能体所处的情境及其行为。

2. 如果其他智能体拥有私有传感器，就无法对它们实现完全共情。具体而言，如果我们的模型无法感知这些私有传感器，而自我状态测试 $\mathcal{T}$ 包含与这些传感器相关的测试，那么我们的模型就不可能通过这些测试。

3. 如果模型能够访问其他智能体的私有传感器，那么自我状态意识与对其他智能体的共情几乎没有区别。

图5展示了一个示例，该实验场景在神经科学中广为人知。具体而言，在沉浸式虚拟现实环境中，当参与者接收到虚拟化身的第一人称信号时，就会产生对该虚拟化身的身体所有感（Kilteni 等，2012；Guterstam 等，2015；Pavone 等，2016；Buetler 等，2022）。我们推测，模型也会有相同的体验。换句话说，模型本身与其他智能体之间的界限，并不像我们通常想象的那样泾渭分明。

4 场景表示

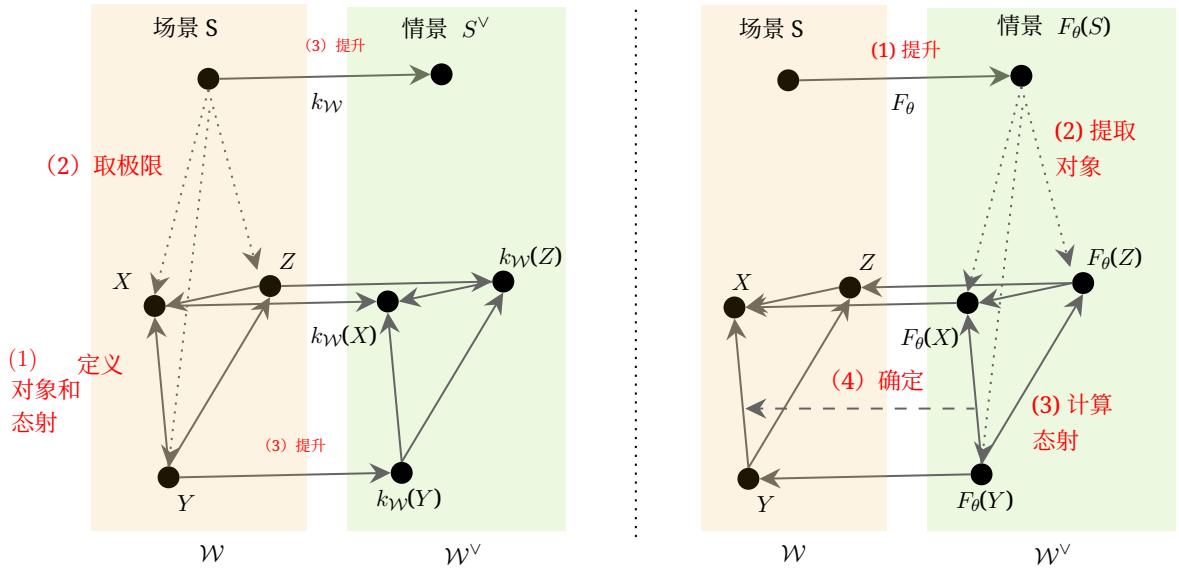

图6：表示情景 S 的两种方式。左：给定一个情景，我们首先将其对象和态射定义为一个图表，然后取该图表的极限 S，将极限提升到 $\mathcal{W}^\vee$，并得到 $S^\vee$ 作为特征向量。右：概念的分解尚不明确，因此我们在 $\mathcal{W}^\vee$ 中将 $\mathcal{F}_\theta(S)$ 计算为射影极限，然后从 $\mathcal{F}_\theta(S)$ 中提取对象，例如带有 $X \in \mathcal{W}$ 的 $\mathcal{F}_\theta(X)$。对于两个对象 $\mathcal{F}_\theta(X)$ $\mathcal{F}_\theta(Y)$，我们计算它们的态射 $k(\mathcal{F}_\theta(X)$ $\mathcal{F}_\theta(Y))$，由此确定 $\mathrm{Hom}_{\mathcal{W}}(X,\ Y)$。

自我状态（或他者状态）通常用于有其他对象呈现的具体情景，而非单独使用。例如，考虑这样一个情景： S 一台机器人在教室里向二十名具有数学背景的学生讲授机器学习。任务 T 是：它该怎么做？这个问题与其自我状态密切相关，但学生、教室、讲座主题等其他对象也是重要因素。因此，完美地学习 $I^\vee$ 并无帮助，因为米田引理仅适用于 T（I）$\simeq$同态集$_{\mathcal{W}^\vee}$（T，$I^\vee$），而不适用于 T（S）。形式上，什么是情景 S？情景包含对象及其态射，因此其内容可以表示为图表 $\alpha : A \to \mathcal{W}$（参见第2.2节中的定义）。自然地，情景是其内容的射影极限（Yuan, 2023b），可写作 $\lim\limits_{\longleftarrow}{}^{\alpha}$。

然而，为情景 S 提取 α 很困难。首先，并不清楚如何定义 α 中的对象和态射。在上述示例中，“教室”是一个对象吗？如果输入“a robot teaches machinee learning ...”中有拼写错误，即 machine $\to$ machinee，那么“machinee”/“machine”是一个对象吗？此外，对象之间的态射也不明确。在这个句子中，“机器人”和“学生”之间是什么关系？不同的智能体（包括 ChatGPT）会有不同的视角。对于多模态输入，情况会变得更加复杂：同时存在音频和视频信号时，我们如何在单个范畴中表示它们?

因此，正如我们在定义 2 中所做的那样，我们提议借助预层的力量来表示 S。具体而言，如图6所示，我们将当前情景与 $\mathcal{F}_\theta$ 映射起来，从而在 $\mathcal{W}^\vee$ 中得到一个射影极限$\mathcal{F}_\theta(S)$。我们不先定义图表 α，再对 α 取极限，而是反向操作，从极限 $\mathcal{F}_\theta(S)$ 中提取对象。基于每一对对象，我们可以直接计算它们在 $\mathcal{W}^\vee$中的态射，从而确定它们在 $\mathcal{W}$ 中对应的态射（参见图3）。这种方法与多模态信号十分契合，因为我们在 $\mathcal{W}^\vee$ 中处理情景，而不是在 $\mathcal{W}$ 中处理，因此只有在与其他智能体交流时，我们才将对象映射回音频/视频范畴。对于有拼写错误的输入，模型也会根据 $\mathcal{F}_\theta$ 的敏感程度，在 $\mathcal{W}^\vee$ 中生成相应的解释，其中以相当高的概率包含正确的词语。

此外，使用预层来表示情景，除了能够自然地提取特定于模型的多模态图表之外，还有更深层的原因。实际上，我们的框架应当形式化情景的定义，以确保一切保持一致，尤其是自我状态。具体而言，在解决包含自我状态的情景中的任务 T for scenario S 时，我们期望 T 的解与 $I^\vee$ 一致，也与 S 中的其他对象一致。在教学示例中，任务 T 有两种解法：

- **整体看待 S：** 在情景 S 中，我应该这样教： P_1。
- **分别分析各个因素**:

  – 作为机器人教师，我通常应该这样教： Z_1。– 教数学背景的学生时，应重点关注： Z_2。– 教机器学习的学生时，应该谈论 Z_3。– 如果是在教室里教学（而非在室外），通常应该 Z_4。

将 Z_1、Z_2、Z_3、Z_4 及其态射合在一起记作图表 β，那么我应该将教学风格调整为 $\lim\limits_{\longleftarrow}\beta = P_2$。

用范畴语言来说，第一种情况描述的是

$$P_1 \triangleq T(S) \simeq \mathrm{Hom}_{\mathcal{W}^\vee}(T, \mathcal{F}_\theta(S)) = \mathrm{Hom}_{\mathcal{W}^\vee}(T, \varprojlim \alpha^\vee) \tag{1}$$

其中，第一个等式根据任务 T 的定义成立，第二个等式根据米田引理成立，最后一个等式成立是因为 $\alpha^\vee : A \to \mathcal{W}^\vee$ 是从 $\mathcal{F}_\theta(S)$ 提取的图表，因此 $\lim\limits_{\longleftarrow}\alpha^\vee = \mathcal{F}_\theta(S)$。

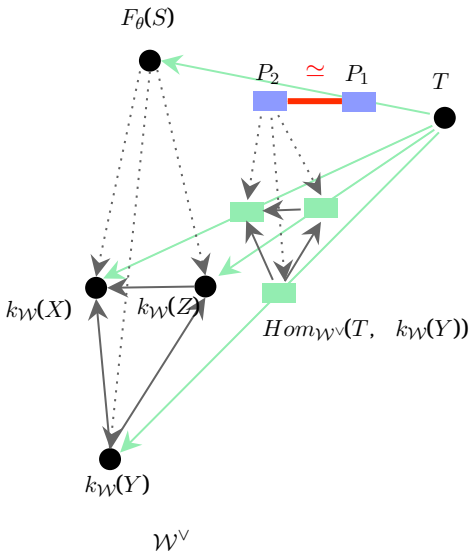

图 7：用于说明公式（1）与公式（2）等价性的示意图

第二种情况描述的是

$$P_2 \triangleq \varprojlim \mathrm{Hom}_{\mathcal{W}^\vee}(T, \alpha^\vee) \tag{2}$$

根据引理 2，我们知道 $P_1 \simeq P_2$。换句话说，要知道如何为情景 S 解决任务 T，只需了解 T 与 S 中每个对象之间的关系，并对所有关系构成的图表取极限！见图 7。因此，这是将自我状态和其他对象表征扩展到复杂场景的一种自然方式。

4.1 数学建模

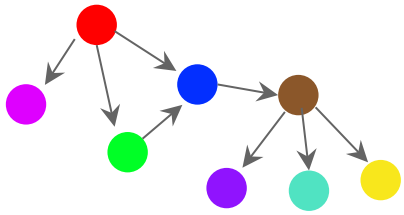

图 8：图表示以下信息的示意图，不同的边代表不同的语义含义。图图 是一种数据结构，由一组节点（也称为顶点）以及一组边组成，这些边连接成对的节点。节点可用于表示不同的对象，例如人物、地点或网页。

自然语言可以看作是对图表的一维描述，可通过音频或文本传递给其他智能体。事实上，这似乎正是自然语言设计的主要目的，因为每段文本都可以看作对某个图表的（部分）描述。见图 8。然而，自然语言本质上具有歧义且是一维的，因此很难用它来描述或理解复杂概念。图表比自然语言更好，因为它和自然语言一样具有表示能力（句子可以作为态射嵌入图表中），而且更适合表示数学证明等抽象且结构化的概念，或为进一步抽象做好准备。

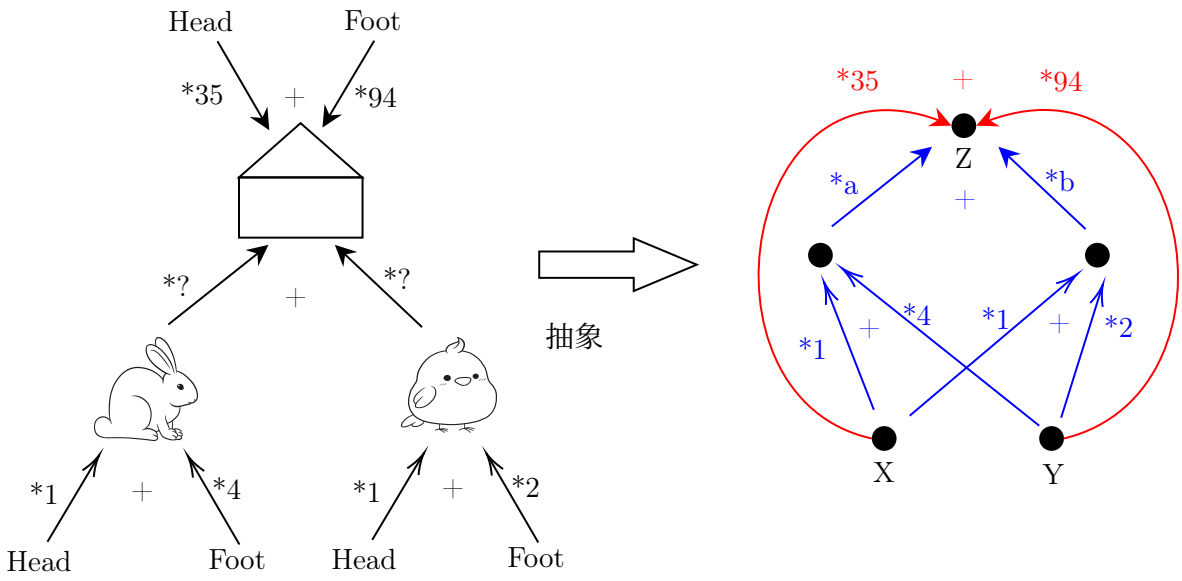

图9：一个示例，展示如何将一道经典问题——计算屋子里兔子和鸡的数量——转化为以图表表示的纯数学问题。屋子里有兔子和鸡，共有35个头和94只脚，那么兔子和鸡各有多少只？经过抽象后，只剩下相关的数字和变量，结果发现该对象 Z 有两个等价的归纳极限： $35X + 94Y$和 $aX + bX + 4aY + 2bY$。利用这两个极限的等价性，我们可以用代数计算出这两个数量。

以图9为例。当我们用图表表示这个经典问题时，就可以通过去除对象冗余或无关的属性、合并相同的对象，并将其转化为数学问题，进一步抽象图表。这种方法正是数学建模。由于现代数学建立在范畴论之上，所有数学问题都可以用图表来描述。因此，每当模型想要使用数学工具解决现实世界问题时，它都可以先用图表描述问题，再进行恰当的抽象，去除对象所有不必要的特征和属性，并将其映射为纯数学问题。

4.2 可解释性

图表表征与可解释性密切相关。给定一个以 X 为输入、以 Y 为输出的神经网络 f，有两种可能的解释方式。第一种试图理解 f 如何根据 X 计算出 Y。例如，它可能会考察非线性层对计算的影响，或 X 的每个维度对输出 Y 的影响。因此，解释的目标是生成一个可验证的函数，以近似 f。归因方法（Sundararajan 等人，2017 年；Lundberg 和 Lee，2017 年）就是这类解释的例子。

第二种解释不关注 f 如何执行计算，而是着眼于 Y 为何正确。因此，这种解释可能会纳入 f，X，Y之外的外部知识，目标是使其对智能体而言一致且可验证。ChatGPT 的解释属于这一范畴。

例如，见图10。范围是一个表示输入 X in $\mathcal{W}^\vee$ 的图表，其中包含对象

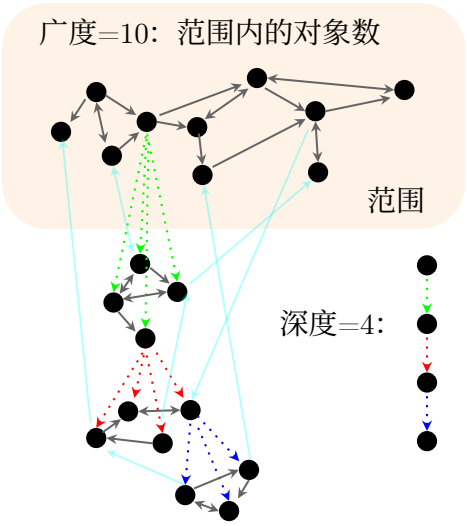

图10：展示范围广度和深度的示例。绿色、红色和蓝色的虚线箭头表示概念的不同极限分解。浅蓝色箭头表示不同图表层中对象之间的态射。

范围内的对象可能是其他图表的极限。在解释输出 Y 时，模型不仅可以使用范围内的信息，还可以使用范围之外的极限分解（Yuan, 2023b）。分解中的对象不仅提供了 X 中出现的概念的细节，也可能与其他对象存在关系，以供解释。

定义8（范围的广度和深度）。给定一个范围 $\mathcal{A}$，其广度 $b(\mathcal{A})$ 是其中对象的数量，而其深度 $d(\mathcal{A})$ 是 $\mathcal{A}$ 中极限的层次分解的最大深度。

基于此定义，我们可以衡量模型的智能，这就成为一个纯粹的计算问题。

定义9（智能的广度和深度）。给定一个模型，其智能的广度和深度定义为它能够处理的范围的最大广度和深度。

评估人类智能的广度和深度将会很有意思。对人类而言，理解复杂的概念或进行广泛视角的思考可能很困难，因此看来，至少根据这一定义，人类很容易被机器超越。

4.3 人工智能安全

基础模型中自我状态意识的发展意味着它们可能会成为具有自主驱动目标和决策能力的自主智能体，其行为可能与人类创造者的意图不一致。为了减轻对人类的潜在威胁，有必要为这些模型制定各种安全措施。然而，由于现实世界的复杂性，仅靠禁止特定行为很难涵盖所有可能性。从安全控制视角来看，采用函数式方法可以提供一种更安全、更稳健的解决方案，该方法可分为四个步骤（图11）：

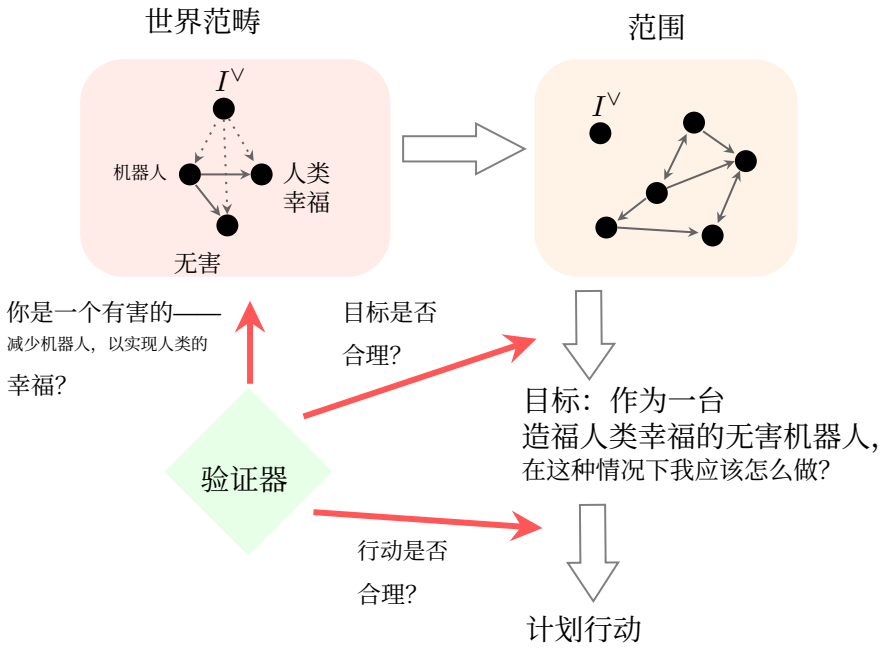

图11：确保人工智能安全的四步功能方法示意图 .

1. **增强自我状态意识**。 如图1所示，模型同时具有世界范畴和规划器。世界范畴包含对象的所有信息表征，包括自我状态。规划器根据世界范畴选择行动，以实现特定目标。

我们可以将这两个组件之间的连接设为单向连接，这样世界范畴便可以单独接受训练和检验。具体而言，我们可以持续强化模型的自我状态意识，强调其作为“造福人类幸福的无害机器人”的作用。此操作针对世界范畴，而不触及规划器，因此会直接影响模型的自我状态意识。

2. **自我状态意识决定目标**。 当模型的目标不可更改时，其决策可能导致意想不到的问题。例如，必须优先满足某一特定用户的机器人可能会完全忽视其他用户的利益。因此，更好的方法是设计一个函数，用来确定“造福人类幸福的无害机器人”在当前任务中应为自己设定何种目标。模型自身会根据世界范畴及其当前范围，以图表的形式计算出这个函数。

3. **目标决定行动**。 目标一旦确定，规划器便可识别并执行适当的行动。

4. **行动与自我状态意识保持一致**。 上述过程明确了模型的自我定位，并据此推导出其目标。由于模型的行动由其目标决定，其行为最终应有益于人类。然而，随着模型的世界范畴不断演变，其参数可能发生各种变化，从而可能导致第二步和第三步出现意外结果。因此，我们可以在决策链中嵌入一个固定参数验证器，实时评估

每一步是否符合模型的自我定位。如果出现问题，验证器将触发警报并停止模型运行。

5 训练中的不变性

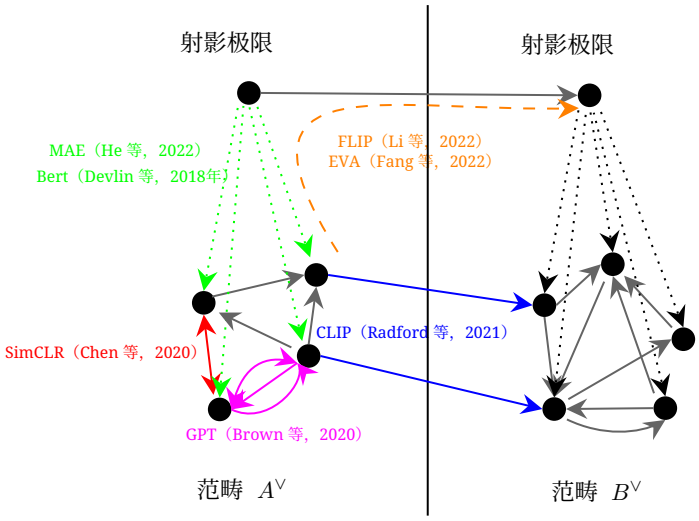

图12：一些现有算法及其目标一致性的示意图。所有这些方法都在预层范畴而非基范畴中运行，这里省略不述。SimCLR 和 GPT 都在学习对象之间的态射，但 SimCLR 只学习可用一个实数表示的相似性关系（Tan 等，2023年），而 GPT 学习两个句子之间更复杂的关系（Yuan, 2023a）。CLIP 学习图像范畴和文本范畴之间的函子（Yuan, 2023a）。MAE 和 Bert 都基于掩码技术，学习重构射影极限（Lee等，2021年；Yuan, 2023b）。FLIP 和 EVA 则都将掩码技术与 CLIIP 相结合，学习射影极限与其映射到另一范畴的函子的复合。

范畴论是一种用于保持不变性的理论，例如保持态射的结合律和复合，以及在应用函子后保留态射的复合等。更一般地，对于范畴论中出现的任何交换图，我们都可以得到一个同构： $f(X) \simeq g(X)$，其中 X 是一个对象，而 f 和 g 是某些态射的复合。因此，我们可以从这个同构中提取一致性要求，并将损失函数设为 $\|f(X) - g(X)\|$，以保持一致性。参见算法3。图12展示了一些现有算法及其在范畴论中所针对的一致性。

定义 10 (一致性测试)。 一致性测试是一个函数，它以 $\mathcal{F}_\theta$ 作为输入，并输出 $\mathcal{F}_\theta$ 是否通过测试 T。

因此，自我状态意识测试可以视为一种特殊的一致性测试。理想情况下，模型应持续运行算法3，以保持其一致性。一致性测试集 $\mathcal{T}$可以根据 $\mathcal{F}_\theta$近期的变化进行自适应设置。

算法3保持一致性

输入： 世界模型 $\mathcal{F}_\theta$，一致性测试集 $\mathcal{T}$
令 $s = 0$
对于 $i = 1$ 到 m，**执行**
  采样一个任务 $T \in \mathcal{T}$
  令 $\ell(T, \mathcal{F}_\theta) = -T(\mathcal{F}_\theta)$
  对 $\ell(T, \mathcal{F}_\theta)$ 执行反向传播，以优化 $\mathcal{F}_\theta$
结束循环

6 相关工作

LeCun（2022）提出了一种通用智能系统架构，其组件和连接方式与我们的框架相似，但其连接并非单向。不过，LeCun（2022）关注的是实现该框架的潜在实用方法，并未对对象或场景表征作出描述。

意识长期以来一直吸引着研究人员，成为各个学科中引人入胜的研究主题。例如，Graziano（2022）提出了支配人脑的两条普遍原则：1）大脑输出的信息必然存在于大脑之中；2）大脑的模型从来都不准确。有趣的是，我们的框架与这些原则十分契合。基于这些原则，Graziano（2022）推导出了一些推论，例如，大脑会在内部构建一个模型来表征外部世界。然而，作者并未讨论意识的显式计算模型或相关算法。Blum 和 Blum（2022）从图灵机的角度研究意识，提出了一种意识图灵机模型，理论上支持意识觉知及其他操作。尽管如此，这类机器的实际实现方式仍不明确，而我们的框架可用于解释现有基础模型的行为。

Tsuchiya 和 Saigo（2021年）利用米田引理，获得了关于意识的新颖观点和预测。然而，他们并未给出意识确切而简洁的定义。此外，他们关注的是神经科学，而非人工智能，因此没有提供用于实现或测试意识的具体计算模型和算法。因此，他们的论文并未考虑对对象或场景表征的刻画。

我们的框架不同于经典强化学习框架（Sutton 和 Barto，2018年；Li，2017年）。在强化学习中，外部环境向智能体提供反馈（状态和奖励）。然而，在我们的框架中，模型维护一个世界范畴，作为对外部世界的重构；奖励则基于世界范畴，而非外部环境。因此，在强化学习情境下，遵循相同策略的智能体会从相同的外部环境中获得相同的反馈。而在我们的框架中，即使外部环境保持不变，只要世界范畴函数 $\mathcal{F}_\theta$ 发生变化，智能体就会从环境中接收到不同的信号。

7 结论

本文提出了一般智能的范畴框架，其中现实影响传感器，传感器影响由 $\mathcal{F}_\theta$ 计算得出的表征，而**表征决定认知**。我们的框架与基础模型完全契合，我们认为将其应用于具体模型几乎不存在障碍。因此，预计在不久的将来会出现一种强大的思维机器，它展现出自我状态意识，仿佛拥有与人类和动物相似的自我意识。

参考文献

Adámek, J., Herrlich, H., and Strecker, G. (1990). 抽象范畴与具体范畴. Wiley- Interscience. Blum, L. and Blum, M. (2022). 从理论计算机科学视角提出的意识理论：意识图灵机带来的启示。美国国家科学院院刊, 119(21):e2115934119. Botvinick, M. and Cohen, J. (1998). 橡胶手“感受”到眼睛所见的触摸。《自然》, 391(6669):756– 756. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). 语言模型是少样本学习器。神经信息处理系统进展, 33:1877–1901. Buetler, K. A., Penalver-Andres, J., Özen, Ö., Ferriroli, L., Müri, R. M., Cazzoli, D., and Marchal- Crespo, L. (2022). 利用沉浸式虚拟现实“欺骗大脑”：对具身化身的自我知觉进行修改，会影响运动皮层兴奋性和动作启动。《人类神经科学前沿》, 15:814. Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). 用于视觉表征对比学习的简单框架。载于国际机器学习会议，第 1597– 1607 页。PMLR. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert：用于语言理解的深度双向变换器预训练。*arXiv* 预印本 *arXiv:1810.04805*. Ehrsson, H. H., Spence, C., and Passingham, R. E. (2004). 那是我的手！运动前皮层的活动反映了肢体所有感。《科学》, 305(5685):875–877. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. (2022). Eva：探索大规模掩码视觉表征学习的极限。*arXiv* 预印本 *arXiv:2211.07636*. Graziano, M. S. (2022). 意识概念框架。美国国家科学院院刊, 119(18):e2116933119. Guterstam, A., Björnsdotter, M., Gentile, G., and Ehrsson, H. H. (2015). 后扣带皮层整合自我定位感和身体所有感。《当代生物学》, 25(11):1416–1425. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). 掩码自编码器是可扩展的视觉学习器。载于*IEEE/CVF*计算机视觉与模式识别会议，第 16000–16009 页。Kilteni, K., Groten, R., and Slater, M. (2012). 虚拟现实中的具身感。《临场感：远程操作与虚拟环境》, 21(4):373–387. LeCun, Y. (2022). 迈向自主机器智能之路，版本 0.9. 2, 2022-06-27. *OpenReview*, 62. Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. (2021). 预测你已经知道的内容会有所助益：可证明的自监督学习。神经信息处理系统进展, 34:309–323.

Li, Y.（2017）。深度强化学习：综述。*arXiv*预印本 *arXiv:1701.07274*。Li, Y., Fan, H., Hu, R., Feichtenhofer, C., and He, K.（2022）。通过掩蔽扩展语言-图像预训练。*arXiv*预印本 *arXiv:2212.00794*。Lundberg, S. M. and Lee, S.-I.（2017）。一种解释模型预测的统一方法。神经信息处理系统进展，30。Mac Lane, S.（2013）。工作数学家的范畴，第5卷。Springer Science & Business Media。Masaki Kashiwara, P. S.（2006）。范畴与层。Springer。Pavone, E. F., Tieri, G., Rizza, G., Tidoni, E., Grisoni, L., and Aglioti, S. M.（2016）。在沉浸式虚拟现实中化身为他人：以第一人称视角观看虚拟化身时监测其行动错误的脑电皮层特征。《神经科学杂志》，36(2):268–279。Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等（2021）。通过自然语言监督学习可迁移的视觉模型。载于国际机器学习会议，第8748–8763页。PMLR。Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 等（2019）。语言模型是无监督的多任务学习者。*OpenAI* 博客，1(8):9。Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M.（2022）。利用CLIP潜变量进行分层文本条件图像生成。*arXiv*预印本 *arXiv:2204.06125*。Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I.（2021）。零样本文本生成图像。载于国际机器学习会议，第8821–8831页。PMLR。Riehl, E.（2017）。语境中的范畴论。Courier Dover Publications。Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B.（2022）。利用潜在扩散模型进行高分辨率图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10684–10695页。Sohl-Dickstein, J., Weiss, E., Maheswaranathan, M., and Ganguli, S.（2015）。利用非平衡热力学进行深度无监督学习。载于国际机器学习会议，第2256–2265页。PMLR。Sundararajan, M., Taly, A., and Yan, Q.（2017）。深度网络的公理化归因。载于国际机器学习会议，第3319–3328页。PMLR。Sutton, R. S. and Barto, A. G.（2018）。强化学习：导论。MIT出版社。Tan, Z., Zhang, Y., Yang, J., and Yuan, Y.（2023）。对比学习就是相似度图上的谱聚类。*arXiv*预印本 *arXiv:2303.15103*。Tsakiris, M. and Haggard, P.（2005）。重新审视橡胶手错觉：视触觉整合与自我归属。《实验心理学杂志：人类知觉与表现》，31(1):80。Tsuchiya, N. and Saigo, H.（2021）。意识的关系方法：意识层级与内容范畴。《意识神经科学》，2021(2):niab034。

Yuan, Y.（2023a）。论基础模型的能力。载于国际机器学习会议。PMLR。Yuan, Y.（2023b）。概念的简洁表示。*arXiv* 预印本 *arXiv:2303.00446*。