

第三篇

评价学核心

第八章

评价问题定义和基础问题

在本章，我将揭示评价的本质，基于评价的基本概念定义**评价基准**。同时，我将正式定义评价问题及其逆问题 (*inverse problem*) 和对偶问题 (*dual problem*)，并在此基础上定义评价学 (*Evaluatology*)。最后，我将对**评价问题**进行分类，并总结评价学的基础问题。

8.1 评价的实质

我将在本节定义广义的评价概念，并正式阐述评价问题。

我把评价的概念定义在第 3.5 节定义的**原因 (cause)** 和**影响 (effect)** 以及第 三章、第 五 章和第 六 章分别定义的观察、**测量**和**测试**基础之上。为了简化，我先以一个简单的场景来定义评价。

评价学概念定义 1 (评价对象，影响对象，影响，评价，总体影响，直接影响，间接影响)。

假定只存在一个原因**对象** A (*cause*) 和一个**显式对象** B (*manifest object*)， A 称为被评价对象，简称评价对象 (*Evaluated Object*) 或者 EO ， B 称为被影响的对象，简称影响对象 (*Affected Object*) 或者 AO 。

当 A 存在时，相对于 A 不存在， B 可观察、测量或测试到**属性**、**结构**、**机制**或行为的差异， B 的消失，或者新的对象产生，这些可观察、测量或测试的差异称为 A 对 B 产生的影响。本质上，评价问题 可以正式表述为：如何揭示评价对象对影响对象的影响。用第 三章定义的存在算子，评价对象对影响对象的影响可以表示为 $B | ext(A)$ 。

在更复杂的环境下，除了评价对象 A 外，影响对象 B 也会受**混杂对象 (confounding objects, 简称 COs)** 的影响。评价对象和混杂对象共同对影响对象产生的影响称为**总体影响 (overall effect)**。总体影响可以表示为 $B | ext(A, COs)$ 。总体影响可以被直接观察、测量或测试。

A 对 B 的直接影响可以用存在算子表示为 $B | ext(A)$ 。 A 对 B 的间接影响可以用存在算子表示为 $B | (COs | ext(A))$ 。 A 对 B 的直接影响和间接影响均只能通过推理得到。

如果 B 是一个**隐藏对象**，对象的属性、结构、机制或者行为发生**变化**，不可直接观察、测量或者测试，即使是总体影响也只能通过**推理**得到。

一个评价对象会对许多其他对象产生影响。我将其分为两种不同类型：直接影响对象和间接影响对象。

评价学概念定义 2 (直接影响对象, 间接影响对象). 直接影响对象 是指 评价对象直接影响的对象, 间接影响对象 是评价对象通过直接影响对象或者间接影响对象影响的对象。

在上面的例子中, 当 A 直接影响 B 时, B 是直接影响对象。当 A 通过混杂对象影响 B 时, B 又成为了间接影响对象。

根据第 三章的定义, 评价对象对直接影响对象的影响可以视为**衍生对象**, 该衍生对象对间接影响对象的影响本质上是影响的影响, 也就是评价对象对间接影响对象的衍生影响。在此背景下, 评价的本质在于 揭示原因对象对直接影响对象的影响及对间接影响对象的衍生影响。

如果一个对象的实例化过程不会产生不同的影响, 我们则不区分该对象与其实例。

8.2 什么是评价基准？

请注意, 在本书中, 在不同的背景下, *Benchmark* 可以翻译为评价基准、测试基准或者比较基准。本书第三部分和第 四部分讨论评价学, 在这个上下文中, 除非特别声明, 否则 *Benchmark* 指评价基准。

评价基准 (benchmark) 在各学科中广泛使用, 但一直缺少严格的定义。本节定义什么是评价基准。

评价学概念定义 3 (评价基准). 对象 A 和 B , 当研究 $A \Rightarrow B$ 的影响时, 评价基准 就是混杂对象。

在这个定义框架下, 一个评价基准包含两个核心组成部分, 如图 8.1 所示。第一部分是影响对象或者利益相关方角度定义的评价指标, 包括衡量与评价基准相关的潜在风险; 差异阈值, 用于确定可接受的与评价真值的差异; 考虑评价成本以及开展评价所需要的资源。通过考虑这些因素, 评价基准能够更好地满足不同**主体**的评价目标。

第二部分是 COs 的结构、属性与机制, 它们是决定评价基准**有效性**的关键元素。具体包括: 定义简化后的 COs; 配置 COs 关键的结构、属性与机制; 确保 COs 不同的等价程度。通过考虑这些 COs 配置和机制, 评价基准能够为不同的评价问题提供全面和标准化的方法。

8.3 和其他评价定义的不同

我已在第 一章和第 二章总结了评价学与其他研究在概念上的独特差异。

本节我再次强调两点: 第一, 我将“广义的评价”定义为揭示评价对象 (EO) 的影响及其衍生影响, 这与以往研究中将评价等同于判断优点、价值或意义的做法不同——后者可见于 [272, 274, 275, 276, 79, 96, 127, 300, 299] 等文献。

在本书里, 传统的评价问题被认为是狭义的评价问题, 见第 十一章, 它是广义的评价问题的衍生问题, 即在推测的对象影响基础上定义、实现和赋值价值量。在本书的定义中, 价值是评价对象和其他对象之间的影响差异对利益相关方产生的影响。

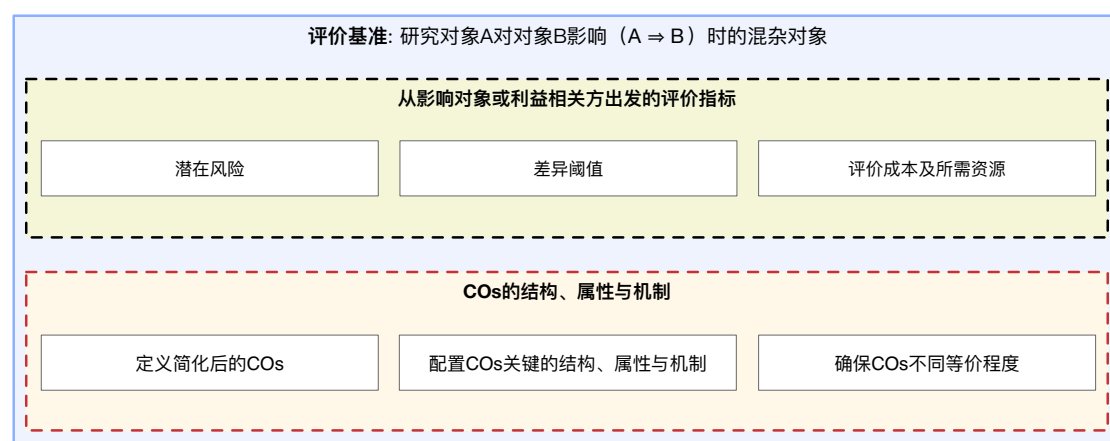


图 8.1: 一个评价基准包含两个核心组成部分。

第二，我定义了简洁明了的评价概念，这些概念定义在第 3 章定义的基础概念体系之上。而其他学者则倾向于采用百科全书式的方法，在单一著作中定义数百个评价概念或术语，如 [274, 190] 所示。

我把评价的概念定义在第 3.5 节的“原因 (cause)”和“影响 (effect)”这两个基本概念之上。

8.4 评价问题的形式化及其对偶问题和逆问题

本节中，我仅在一个简单的情况下正式定义评价问题，并引入评价的对偶问题——设计，以及评价的逆问题——逆评价 (de-evaluation)。

为了表达的清晰起见，我假定对象之间的影响仅限于量的改变，而不考虑其他更复杂的情况，例如对象属性、结构或机制的变化或对象的消失和新对象的出现。

图 8.2¹ 展示了评价、设计与逆评价三者之间的关系。

8.4.1 评价问题的形式化

本章对评价问题进行形式化定义 (Formalization of Evaluation Problem)²。理想情况下，已知一个评价对象 EO，记为 O ，一个显式影响对象 AO，记为 A ，混杂对象集合 COs，记为 S 。

评价的实质可以表示如下：给定 O 、 A 和 C 的有限配置以及相应配置下可测量或可测试的量，推测 O 对 A 的影响。

记 O 的元素为 o 。 O 有 n 个组件 F_i ， n 为常数，记为 $O = F_1 \times \cdots \times F_i \times \cdots \times F_n$ 。每个组件 F_i 有 $\sigma_i \geq 1$ 个配置，其中 σ_i 为常数。记 F_i 的元素为 f_{ij_i} ，其中 $j_i \geq 1$ 且 $j_i \leq \sigma_i$ 。 $o = (f_{11}, \cdots, f_{1\sigma_1}, \cdots, f_{i1}, \cdots, f_{i\sigma_i}, \cdots, f_{n1}, \cdots, f_{n\sigma_n})$ 。 o 量的集合记为 $\mathbb{Q}_o$ 。

¹此图由王晨曦在与我的讨论基础上绘制

²王晨曦和李鸿霄也对本章有贡献。

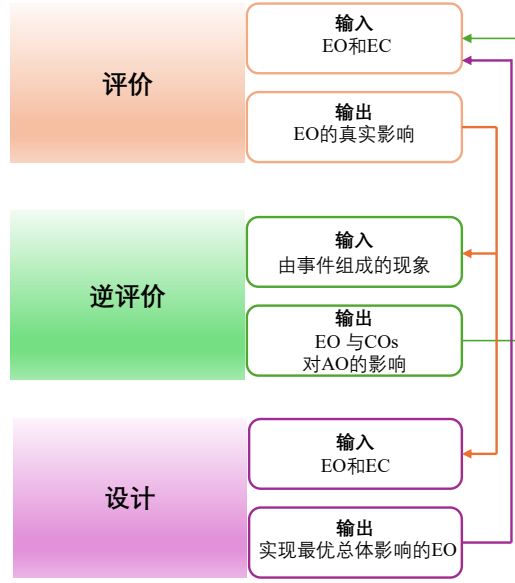


图 8.2: 评价、逆评价以及对偶问题—设计—之间的关系。

记 A 的元素为 a 。 A 有 m 个组件 G_i , m 为常数, 记为 $O = G_1 \times \cdots \times G_i \times \cdots \times G_m$ 。 每个组件 G_i 有 $\phi_i \geq 1$ 个配置, 其中 ϕ_i 为常数。 记 G_i 的元素为 g_{ij_i} , 其中 $j_i \geq 1$ 且 $j_i \leq \phi_i$ 。 $a = (g_{11}, \cdots, g_{1\phi_1}, \cdots, g_{i1}, \cdots, g_{i\phi_i}, \cdots, g_{m1}, \cdots, g_{m\phi_m})$ 。 a 量的集合记为 Q_a 。

记 C 的元素为 c 。 c 有 l 个组件 H_i , l 为常数, 记为 $C = H_1 \times \cdots \times H_i \times \cdots \times H_l$ 。 每个组件 H_i 有 $\omega_i \geq 1$ 个配置, 其中 ω_i 为常数。 记 C_i 的元素为 c_{ij_i} , 其中 $j_i \geq 1$ 且 $j_i \leq \omega_i$ 。 $c = (c_{1j_1}, \cdots, c_{ij_i}, \cdots, c_{lj_l})$ 。 c 的量的集合记为 Q_c 。

评价的实质是如何选择有限的 O 、 A 、 C 对应的元素 o 、 a 和 c , 以及相应配置下的量 Q_o 、 Q_a 以及 Q_c , 在此基础上推测 O 对 A 的影响, 记为函数 $f(O, A)$ 。

8.4.2 设计：评价的对偶问题

在本节中, 我将解释为什么设计 (*design*) 问题是评价的对偶问题 (*Dual Problem of Evaluation*)³。

评价和设计互为对偶问题。在设计上下文里, EO 即设计对象 (*designed object*, 简称 *DO*), 可以设计 DO 的属性、结构或者机制。总体影响是由 DO 和 COs 共同施加在 AO 上引起的。评价旨在从 AO 上可观察、可测量或可测试的变化上揭示 DO 的真实影响, 而设计则是在 DO 的设计空间中寻找特定的属性、结构或者机制, 以实现最优的影响。

8.4.3 逆评价：评价的逆问题

³这个想法是受我的同事罗纯杰博士在一次小组会议上讨论的启发。

基于第 3 章对对象、事件和因果的定义, 本节定义评价的逆问题 (*inverse problem of evaluation*), 即所谓的逆评价 (*de-evaluation*)。

交通拥堵的诊断可以作为一个逆评价问题的示例。城市中普遍存在交通拥堵现象, 而拥堵本身是一种复合现象。复合现象是一组事件的集合, 体现了一组对象改变的结果。诊断交通拥堵的目的是将一组事件—即现象—归因于原因对象 (UO) 和 COs 对 AO 的影响。本质上, 它是评价的逆问题—如何准确地将现象归因于其原因。但在交通拥塞里, 每个对象同时扮演着 UO, AO 和 COs 的角色, 并且相互影响随着时间发生变化。

让我根据 3.5 节因果的定义来解释评价的逆问题。观察到一个现象 P 由许多对象 O_i 组成, 在每个 t 时刻, 它们分别扮演着 UO, AO 或者 COs 的角色。这些对象 O_i 有不同属性 (量)、结构和机制组成, $Q_{ij}(t)$, 及其变化 $\delta Q_{ij}(t)$, 它们均随着时间 (t) 发生改变。

评价的逆问题是追溯影响对象 O_i 的 $Q_{ij}(t)$ 及其变化 $\delta Q_{ij}(t)$ 到 UO (原因) 和 COs。由于每个对象同时可以成为 UO, AO 和 COs, 这将成为一个非常复杂的动态过程。

8.5 什么是评价学 (Evaluatology) ?

我在书的标题里已经将评价学 (Evaluatology) 定义为揭示原因与影响的科学。

正如第 8.4 节阐述的, 评价学包含三个基本的问题: 评价问题、逆评价问题、基于评价学的设计问题。

评价问题的核心是揭示评价对象的影响或者衍生影响。如果评价对象有利益相关方, 评价问题还需要揭示评价对象的价值。

逆评价问题的核心是揭示事件或者现象背后的原因。

基于评价学的设计则是在统一的评价空间和设计空间约束下, 对设计对象 (具体包括对象的结构、属性、机制和行为) 的设计空间进行探索。

我和我的团队正在研究基于评价学的智能设计, 本书未来的版本将会持续更新这个主题的研究进展。

评价学也提供了一个天生 (native) 的解释框架。

评价学概念定义 4 (解释). 解释就是揭示观察到的对象或者对象关系现象背后的原因。

在第 11 章, 我将讨论狭义评价学。

8.6 评价问题的分类

在第 8.4.1 节形式化的基础上, 本节从不同的角度对评价问题进行分类 (categories of evaluation problems)。

8.6.1 依据 EO 性质的评价问题分类

根据第 3.2 节的定义, 对象 (object) 是拥有结构、属性、机制和行为的一类实体, 它有六个维度的性质: 未知和已知, 真随机性和表面随机性, 确定和不确定, 明确定义 (*well-defined*) 和未明确定义, 内生和外生以及显式和隐藏。每个对象都可以被评价, 因此根据 EO 的性质对评价问题进行分类是合理的。

第一类评价问题是当 EO 为自然对象时的情形。可以根据对象的性质划分,进一步分为若干子类:第一子类是该对象是未知的;第二子类是该对象是已知的,但具有真随机性。第三个子类是该对象是已知的,不具有真随机性,可以被明确定义,即能够被准确复制;第四个子类是该对象是已知的,不具有真随机性,可以被明确定义,还能由主体自由改变。

第二类评价问题是对人造物 (artifact) 的评价。这类对象可以是一个想法、对某种不存在事物的想象、虚假信息,也可以是一项新设计及其具体实现。

由于人造物是由主体(可能是自动对象或智能生命)创造的,这类问题的独特性在于它如何对主体产生影响,这种影响机制往往与自然对象的影响机制有显著不同。

8.6.2 依据 AO 及 COs 的的评价问题分类

第一类评价问题是当 AO 或者 COs 未知时。例如,当我们评价平行宇宙或暗物质以揭示其影响时,我们甚至不知道哪些是影响对象 (AOs)。

第二类问题是当 AO 或者 COs 仅部分已知时,例如在宇宙学和天文学中评价某个对象时。此时,影响对象 (AO) 的确切数量尚不明确,也无法确定是否存在混杂对象 (COs)。

评价问题的第三类情形是当 AO 或者 COs 全部已知时。在这种情况下,又可分为四个不同的子类:第一子类:AO 或者 COs 非常复杂,无法被明确定义。第二子类:AO 或者 COs 全部已知且可被明确定义,但由于各种原因(如实现上的限制、成本过高、后果不可承受或伦理问题),无法对其进行任意操控。正如文献所述:“如果一个系统可以用函数建模,那么任意操控就意味着可以将其自变量设置为定义域内的任意数值” [338]。第三子类:AO 或者 COs 不仅全部已知、可明确定义,而且允许任意操控。例如,计算机系统几乎就属于这一类别 [338]。最后一类问题是, AO 是隐藏对象,它的量不可以直接测量或者测试。

8.6.3 依据影响机制的评价问题分类

第一维度是评价对象 (EO) 与影响对象 (AO) 之间的关系。有一类评价问题具有值得重视的共同特征,即:EO 是 AO 的一个组成部分。例如,在第 11.4 节讨论的案例中,EO 是一个 CPU,而它正是直接影响对象——计算机系统——的组成部分,CPU 的性能变化会直接影响整个系统的运行表现。

评价对象 (EO) 与影响对象 (AO) 之间的关系在第 13.6 节中更为复杂,因为产生了新的对象。亲代豌豆 (AO) 的基因 (组成部分),即 EO,会对子代豌豆 (AO) 产生不同的影响。而在另一种情况下,EO 与 AO 相互独立。例如,当 EO 是一种药物时,它会影响特定患者或一群患者。EO 独立于 AO 之外。

第二个维度是不同的影响机制。例如,地球可以对苹果施加引力;CPU 与多个不可或缺的组件协同工作,最终构成一个计算机系统,其影响机制在于协作;一名犯罪嫌疑人则可能通过恐吓、侮辱性或恶意言语对受害者造成伤害。

影响对象 (AO) 之间的相互作用是否会影响评价结果,这是评价问题的第三个维度。这种情况在政策评价中尤为常见,如第 13.1.6 节所述。例如,在评价一项遏制药物成瘾的政策时,影响对象,如个体成瘾者之间的互动,尤其是他们的态度和行为传播,会显著影响政策效果的最终评价结果。

8.7 评价学基础问题

本章，我提出评价学的四大基础问题（Fundamental Issues in Evaluatology）。

8.7.1 评价问题的结构

根据第 四 章提出的因果律公理，除非对象及其影响机制具有真随机性，否则其影响是确定的。

尽管因果律公理指出了在什么样条件下对象的影响是确定的，但其影响并不一定可以计算。如第 8.6 节讨论，评价问题的结构非常多样，例如混杂对象是否全部已知，或者有多少未知。对象的影响是线性的还是非线性的？什么样的评价问题是可以计算的，又有哪些评价问题具有真值？评价问题的结构是评价学的一个基础问题。

8.7.2 建模和计算对象之间的影响

如第8.1节所述，评价的本质是揭示 EO 对 AO 的影响。

在这个上下文里，如何理解对象之间的相互影响并且进行建模几乎是任何学科中的根本问题。通过对不同 EO、AO 及其影响机制的系统性理解和建模，我们可以更深入地认识评价问题。

另外，当一个评价问题是可计算的时候，如何具体进行计算也是一个基础的问题。

8.7.3 评价方法的检验

即使满足因果律公理，揭示一个对象的影响仍可能是在一个复杂的系统条件下完成的。

在不同的条件下，哪些评价方法会产生有效的评价结果？例如不同的方法能不能或在多大程度上能获得接近真值的评价结果？这是一个根本性的问题。遗憾的是，这一问题在此前几乎没有被专门讨论过，直到我在 [339] 中首次正式提出。

8.7.4 元评价

元评价的目的是评价不同的评价方法。对于元评价任务而言：特定的评价方法（a specific evaluation approach）是 评价对象（EO）；影响对象 包括不同条件下的评价对象的评价结果或者评价对象的利益相关方，元评价的核心目标是：揭示这一特定评价方法对影响对象的影响，也就是该评价方法揭示出的 EO 的影响与真值的偏差对影响对象的影响。

8.8 总结

本章正式定义了评价的基本概念，在其基础上定义了什么是 benchmark？形式化定义了评价问题及其逆问题和对偶问题，定义了评价学（Evaluatology）。最后，对评价问题进

行了分类，并总结了评价学的基础问题。

第九章

求索模型

本节，从[求索方法学](#)的角度，总结[主体](#)可以用于揭示原因和影响的基本方法，并将其总结为[求索模型](#)。

我故乡有一座山，名叫铁林寨。我出生直到离开家乡去上大学，一出门，抬头就会看到它。山上有数亿年前地壳运动留下的痕迹——石船。去铁林寨，有好几条道路，从任何一条道路都能到达它，这和[求索](#)很像。所以，我将我总结的求索模型称之为求索铁林寨模型（Interrogation Iron Forest Fortress Model）。

9.1 求索方法学上的两个天然矛盾

人类在认识世界和构建知识体系的求索的过程中，存在方法学上的两个天然矛盾。

一方面，在构建知识体系的过程中，典型的知识体系如牛顿力学，化学的燃烧理论、生物学的遗传规律，就是一个经受过[检验（testing）](#)的确定性的知识体系。然而，当对象数目激增，影响机制变得多样，相互影响变得复杂后，又经常最终诉诸于随机理论。以中学教科书里经常的提及的抛掷硬币或者掷骰子而言，这本身完全属于经典牛顿力学解释的范畴，每一次[结果](#)完全是确定的，但它却经常当做最经典的随机性的例子。

另外一方面，在不少领域又过于推崇[实验](#)，尤其是随机实验，对基于观察（observation）的结果持有根本性的怀疑态度。在第 3.4 节，我定义了观察、实验和随机实验这三个概念。但有趣的是，在人类建立的知识体系里，大多数科学发现都不是借助随机实验获得。相反，不少经典的科学发现，例如开普勒三大定律正是基于观察得到的，不可能对星体开展所谓的实验，更不用提随机实验。

天文学研究的对象多为质量巨大、时空尺度宏大的天体，其运行过程通常具有不可操纵性，[研究者](#)难以像传统物理实验那样开展受控实验与随机实验，而是依赖于长期观察的数据（测量数据）。开普勒正是基于第谷二十多年积累的高精度行星位置观察数据，通过数学计算，最终归纳出行星运动的三大经验规律（模型）。

更有趣的是，古希腊的原子论者（atomist）甚至没有观察的手段，提出的原子论理论框架却与现代科学惊人相似 [261]，核心思想包括：万物皆有原子构成，它们在物理而不是几何上是不可分的；原子之间存在虚的空间（empty space）；原子不可毁灭（indestructible）；原子始终处于永恒运动状态；原子数量及种类无限，形态与体积各异。

而万有引力常量的测量实验则是通过精巧的实验设计和实现，巧妙地排除无关或者[混杂对象](#)的影响。

9.2 揭示原因和影响的基本方法

在主体求索过程中，可用于揭示原因和影响的基本方法总结如下。

9.2.1 测量和测试

只有通过测量和测试，才可以获得显式对象结构的量或者行为的量。测量和测试是揭示原因和影响的起点，可用于不同的求索条件。第 5 章和第 6 章分别阐述了测量学和测试学。

9.2.2 观察

不同于实验，观察无法改变求索对象或求索条件。观察可以借助测量与测试。直接的观察结果无法揭示原因对象的影响，因为它通常反映了原因对象和混杂对象的影响。

可以测量或者测试一个显式对象，获得它的可观察结果。可以借助感官或者技术方法（后者通常是间接地）观察对象的结构或者行为。通过技术方法间接地观察对象的结构或者行为，可以认为是一种模型的方法。在观察的过程中，可以通过测量或者测试来了解一个机制在特定输入情况下的输出。

在观察过程中，也可以借助通过检验的模型来模拟实验。

9.2.3 实验

实验可以改变求索对象或求索条件。我将基础的物理实验方法总结为随机化（randomization）、隔离（isolation）和存在算子（ext operator）。

我将其表述为揭示原因和影响的物理实验方法猜想。

猜想 6（揭示原因和影响的物理实验方法猜想）。有且仅有三种物理实验方法可以用于揭示原因和影响：随机化、隔离与存在算子。

我从对象的角度来阐述这三种基本的物理实验方法。

假如要研究对象 A （原因对象）对对象 B （影响对象）的影响。同时存在混杂对象的干扰。

随机化是从原因对象和影响对象出发，通过真随机化的机制消除混杂对象的影响。隔离则是从混杂对象的出发，消除它对原因对象和影响对象的影响。存在算子则是从原因对象出发，模拟原因对象存在与否，从而测量或者测试原因对象带来的影响差异。我将在第 9.3 节进一步详细阐述。

这些方法在发展过程中进行了一定程度的虚拟化。

实验学概念定义 1（虚拟化实验方法）。虚拟化实验方法是指实验无法物理执行，也就是无法在物理世界操作，只能虚拟执行，也就是只能在数学模型上操作。

我统一将物理实验和虚拟化实验方法称为实验学（Experimentology）方法。我也给出了实验学的定义。

实验学概念定义 2（实验学）。实验学（Experimentology）是设计、实现实验，并获得、分析和检验实验结果的科学。

9.2.4 建模

建模是一种数学方法，它以数学模型的方式表达原因与影响之间的关系。我以第 十九 章将详细阐述的结构因果模型（简称 SCM）来介绍建模方法。

在 SCM 中，首先要对所有内生变量 $V = \{V_i\}$ 使用有向无环图（DAG）进行因果建模，这个 DAG 称为因果图 [221, 224]。对于图上两个节点代表的内生变量，如果研究人员能够确定其中一个变量（原因变量）是另一个变量（结果变量）的直接原因，那么在因果图上就用一条从原因变量节点指向结果变量节点的有向边表示这种因果关系。

DAG 表示的因果模型本质上是一种假设，需要进行检验。

9.2.5 推理与检验

推理是实验中广泛使用的方法，可以在已有前提、事实或者模型基础上推断新的模型。推理的基础可以来自猜想、观察、测量、测试、实验或者建模。推理得到的模型需要通过检验。第 六 章和第 七 章分别阐述了推理和检验。

9.3 三种基础的物理实验方法

9.3.1 随机化

随机化可以直接在物理世界中进行操作。它是一种物理实验方法，能确保无混杂性假设成立。它依赖于两个基本条件。

首先，随机化机制能实现真随机化。真随机化的定义见本文第 三 章。潜在结果理论（见第 十七 章）将它表达为随机分配机制与潜在结果之间独立 [253]。

当研究对象 A_1 （称为干预）和它的比较对象 A_2 （称为控制）对对象 B_i 的影响时，随机化（randomization）机制可以表述如下，针对对象 A_1 和 A_2 ，建立两个比较组，分别称为干预组 and 对照组。对象 B_i 随机分配到对照组或者干预组 [295, 224]。

在这种情况下，当干预分配与潜在结果是独立时。形式上，

$$(Y(1), Y(0)) \perp\!\!\!\perp T, \quad (9.1)$$

其中 $Y(1)$ 和 $Y(0)$ 分别是干预组和对照组下的潜在结果， T 是干预分配指示符 [295]。

第二，对象 B_i 有足够长的队列。在这两个条件满足的情况下，即使 B_i 存在未观察到的混杂因素，干预组和对照组在统计上具有等价性。因此，观察到的结果中的任何差异都是干预的影响。

满足这两个条件，随机化能够推测出平均的影响，它实际上是原因对象对不同的影响对象的平均影响，是一个分布。这种结果分布的本质并不是真随机。

随机化机制将在第 十六 章进一步阐述。

9.3.2 隔离

当研究对象 A 和对象 B 的影响时，隔离（isolation）是一种有效的物理方法，用于屏蔽 COs 的混杂。不同于随机化方法，它不需要建立对照组，对象 B 可以只是一个对象实例。

在物理和化学和其他学科的实验中，隔离是被广泛使用的方法，我用第 十三章将详细介绍的万有引力实验和燃烧实验进一步解释。

亨利·卡文迪许在 1797 年测量引力常数 (G) 的实验 [47, 167]，通过一系列精密的实验技巧实现了对混杂对象的隔离。

实验装置 (Apparatus)。

- 一个扭秤，由一根六英尺长的木杆组成，木杆通过一根金属丝悬挂，两端各固定一个较小的铅球（每个重 1.6 磅）。
- 两个较大的铅球（每个重 350 磅）被放置在较小铅球的附近，两者相距约九英寸。
- 金属丝上附有一面镜子，可将一束光线反射到刻度尺上，从而放大并精确测量因引力作用引起的微小扭转运动。

实验隔离的多种机制

隔离地球引力的影响

- 对称布置：扭秤两侧受到地球引力相互平衡，从而抵消其对扭转运动的净影响。
- 局部质量对称：大铅球在小铅球周围对称放置，以减小引力梯度的不对称性。

隔离临近的对象

- 密闭外壳：实验装置被安置在密闭的房间内，以消除空气流动并减少温度变化的影响。
- 非磁性材料：选用铅是因为其具有非磁性特性，可避免来自外部磁场等磁性干扰对实验造成影响。

拉瓦锡的经典燃烧实验是其对物质转化和空气成分研究的核心实验，该实验设计巧妙且执行严谨，它以封闭空间为基础，隔离混杂对象的影响。

在这个实验中，他将汞 (AO) 密封在蒸馏瓶中加热，观察密闭空间内气体体积的变化，并收集燃烧产生的气体进行分析。

9.3.3 存在算子

我在第 三章定义了存在算子。存在算子从机制和结果来说都不同于隔离机制。存在算子是从原因对象出发，通过实现或者模拟原因对象的有和无，从而观察、测量或者测试原因对象导致的影响差异，而隔离则是从混杂对象的角度出发，屏蔽混杂对象的影响。

另外，从结果来说，这两种机制获得的结果并不相同。

假设只存在原因对象 A ，影响对象 B 和混杂对象 C ，对象 B 是显式对象。隔离 C 后获得的对象 A 对对象 B 的影响为 $B \mid \text{ext}(A)$ 。

使用存在算子得到的影响除了 A 的直接影响， $B \mid \text{ext}(A)$ ，还有通过 C 产生的间接影响， $B \mid (C \mid \text{ext}(A))$ 。

9.4 基础的虚拟化实验方法

在三种物理方法的基础上，我进一步阐述三种虚拟化实验方法，它们是基础的物理实验方法在数学模型上的扩展，包括虚拟隔离、模型化的随机机制和虚拟存在算子。

猜想 7 (虚拟化实验方法猜想). 有且仅有三种虚拟实验方法可以用于揭示原因和影响：虚拟隔离、模型化的随机机制和虚拟存在算子。

9.4.1 虚拟隔离：借助数学模型进行隔离

实验学概念定义 3 (虚拟隔离). 虚拟化隔离是指实验无法物理执行隔离，也就是无法在物理世界进行隔离操作，只能虚拟执行，也就是在数学模型上进行隔离操作。

第 十九章详细介绍的 do 演算 [221, 224] 就是一种虚拟隔离，它难以在物理世界实现，需要借助数学模型实现隔离。

do 算子是在一种理想化情况下隔离其他混杂变量的影响。请注意，我定义的混杂对象和混杂变量是不同的概念。

除此以外，第 十九章基于 do 算子定义的影响也与存在算子定义的影响不同。当研究变量 X 对变量 Y 的影响时，假如存在其他变量 Z_i 对变量 X 存在影响，而对 Y 则没有影响。则通过将 X 设置为其范围内的固定值，可以获得 X 对 Y 的影响。但 X 的取值范围可能有来自 Z_i 的影响¹，这与存在算子定义的影响不同，存在算子会排除掉来自其他对象传递的影响。

9.4.2 模型化的随机机制

实验学概念定义 4 (模型化的随机机制). 模型化的随机机制是通过建立确定干预分配与潜在结果的模型来模拟近似随机化机制。

潜在结果理论（见第 十七章）阐述的协变量调整 [133, 241] 就是一种模型化的随机机制。

根据 PO 的无混杂假设 [241, 246]，在给定观察到的协变量 X 的条件下，干预分配独立于潜在结果 $(Y(0), Y(1))$ ：

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid X. \quad (9.2)$$

¹詹剑锋加了这个脚注，因为王晨曦觉得应该给出并提供了这个脚注。假设存在因果路径 $Z_i \rightarrow X \rightarrow Y$ 。根据 SCM 的定义，当 $\text{do}(X=x)$ 时，路径 $Z_i \rightarrow X$ 会被阻断，按照解释 X 确实是取了其取值范围内的一个确定值，但此时 X 的取值完全不由 Z_i 决定，也即 Z_i 对 X 没有影响（ Z_i 与 X 的取值在统计上是独立的）。此时 X 对 Y 的影响确实完全来自于 X ， do 算子会阻断所有指向 X 有向边的路径。

该假设是从观察数据中建立因果关系的基石。“|”表示“在……条件下”的数学符号，而“ $A|B$ ”表示“在 B 条件下的 A ”。符号“ $\perp$ ”表示“相互独立”，其中“ $A \perp B$ ”表示 A 与 B 相互独立。

它的基本思路是，确定同时决定干预分配与潜在结果的干预前协变量 X ，对它们进行控制。在这样的条件下，决定一个单元 i 接受干预 ($T_i = 1$) 还是对照 ($T_i = 0$) 的机制，可以认为是“近似随机”的。因此，在控制 X 后比较两组结果差异，便可将其归因于干预本身，而非由 X 所代表的系统性差异 [133, 241]。

与此同时，建立潜在结果与干预前协变量 X 的模型。在给定干预与协变量条件下直接估计潜在结果的条件期望。其核心思想并非在样本层面构造可比较的单元，而是在模型层面刻画结果的产生机制，也就是哪些协变量确定了结果 [133, 241]。

具体情况，参见第 十七章。

9.4.3 虚拟存在算子

实验学概念定义 5 (虚拟存在算子). 当无法物理执行存在算子时，虚拟操作算子可以虚拟执行存在算子，也就是在数学模型上模拟物理执行存在算子。虚拟操作算子用 $vert()$ 来表示。

以 11.4 节阐述的 CPU 评价为例。如果直接将 CPU 从一个计算机系统中删除，则系统无法运行。因此，CPU 的影响可以用虚拟存在算子来表达： $| vert(CPU)$ 。

9.5 求索铁林寨模型

本节以揭示原因对象影响这一具体任务，阐述主体在求索过程中，不同求索方法之间的关系，我称之为求索铁林寨模型 (*Interrogation Iron Forest Fortress Model*)。图 9.1² 给出了求索的铁林寨模型。

在求索铁林寨模型里，测量、测试、猜想、观察、实验、建模是平行方法，就像我小时候我攀登铁林寨走过的不同的小路——任何一条道路都可以到达铁林寨的著名景点：石船——它们都可以获得对象的命题或者模型，在此基础上还可以进一步通过推理，获得新的命题或者模型。但这些命题或者模型的有效性取决于检验。只有通过了检验，模型才能成为对象的真理或者事实。

猜想、建模和推理属于心智的方法，从猜想、建模到推理，方法更为严格。观察与实验属于物理的方法，同样，从观察到实验，也意味着方法更为严格。但在求索铁林寨模型里，并不排除通过各种方法获得对象命题或者模型的可能性。

求索的铁林寨模型不是一个严格的递进模型，而是“条条小路通铁林寨”，强调各种方法的互补性，为主体的求索提供了可能。

9.6 总结

本章总结了可用于揭示原因对象影响的方法，包括三个基础的物理实验方法和三个基础的虚拟实验方法。我进一步将它们整体称为求索铁林寨模型，它们不是一个严格的递进模型，相反，不同方法之间是互补的。

²基于詹剑锋提供的草图，进一步讨论后，王晨曦绘制了该图。

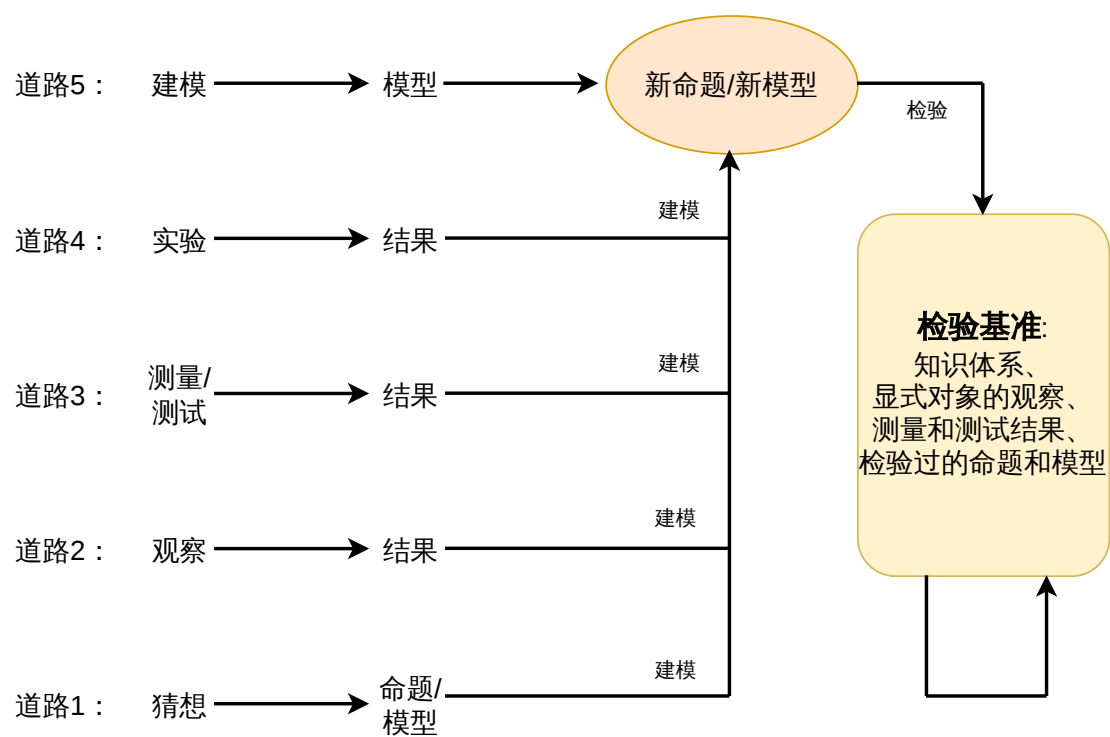


图 9.1: 求索铁林寨模型：条条道路通往铁林寨石船。

第十章

评价学核心概念和评价结果的性质

本章介绍评价学的三个核心概念：现实世界系统、自包含系统和对象因果模型，图 10.1¹给出了它们之间的关系。然后讨论评价结果的基本性质。

10.1 现实世界系统

评价学概念定义 5 (现实世界系统). 现实世界系统是**主体**正在观察的现实世界 (*real-world*) 里的**原因对象**、**影响对象**和**混杂对象**的集合，它们通过相互**影响机制**构成一个系统。

通过现实世界系统揭示对象的影响或者现象的原因有三个显著的障碍。

一、在现实世界识别 EO、AO 与 COs 十分困难；同时，系统中大量无关对象的存在构成重要干扰。

二、现实世界系统难以被操控，存在安全风险，几乎无法为对象或者现象建立可控的**实验**环境，缺少实验手段。各个对象之间互相影响，它们与系统中其他无关对象的**交互**进一步加剧了这一挑战。

三、现实世界系统可获得的观察配置有限，获得的难度也不同。对象实例或其配置空间的分布特征存在现实的倾斜性，这会限制对整个可能性空间的全面探索与理解。

10.2 自包含系统

我在一个简单的情况下阐述什么是自包含系统：原因对象 A ，影响对象 B 以及混杂对象 COs 都是显式的。

即使是这种最简单的情况，也必须采取整体方法，将 EO (A)、AO (B) 以及混杂对象 COs (A_j) 视为一个整体系统，在此基础上推断 EO 对 AO 的真实影响 (true effect)。这有两方面的原因，一方面，混杂对象的影响可能无法隔离；另外一方面，在更为实际的环境里，必须考虑到这些混杂对象的影响。我正式将这个整体系统命名为自包含系统 (*self-contained system*, 简称 *SCS*)。图 10.2 展示了 SCS 的基本组成部分。

¹高婉铃绘制了该图。

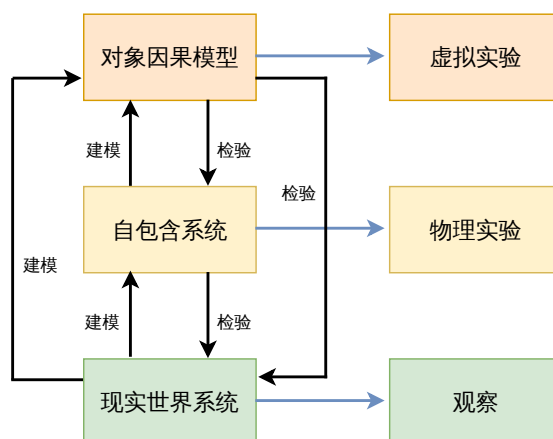


图 10.1: 现实世界系统 (Real-world System, RWS)、自包含系统 (Self-contained System, SCS)、对象因果模型 (Object Causal Model, OCM) 的关系。

评价学概念定义 6 (自包含系统). 自包含系统是研究对象影响或者现象原因的实验系统，它是现实世界系统的物理模型，由原因对象、影响对象以及无法隔离的混杂对象组成。

评价学概念定义 7 (评价条件). 从 SCS 中移除 (EO) 后，所得到的衍生系统称为 评价条件 (evaluation condition, 简称 EC) 。

在构建自包含系统时，有四个需要遵循的基本原则。

第一个原则是同一性原则，自包含系统包含的对象以及对象之间的相互影响机制应与现实世界系统的对应物完全相同。如果违反了这个原则，建立的自包含系统就不能认为是现实世界系统的物理模型。

第二个原则是简单性原则。自包含系统是现实世界系统的物理模型，也就是系统的简化表示。构建的自包含系统应尽量减少对象的数目，对象的具体配置应简单，从而降低系统复杂性。如果违反了这个原则，自包含系统就没有独立于现实世界系统存在的必要。现实世界系统非常复杂，依据它建立自包含系统时，如果需要增加混杂对象，应在复杂性和建模的准确性之间权衡。同样，违反了这个原则，自包含系统也没有独立于现实世界系统存在的必要。

第三个原则是递进性原则。自包含系统相对于现实世界系统，应提供后者缺少的实验手段。违反这个原则，自包含系统同样没有独立于现实世界系统存在的必要。

第四个原则是可观察、可测量或可测试原则。自包含系统里应包含显式对象，主体可以对它进行测量或者测试。如果违法了这个原则，自包含系统无法被主体求索，缺少检验的手段。

如果一个 EO 具有间接影响对象 (AO)，需要解决的一个衍生问题是如何推断评价对象的衍生影响。上述提及的相同方法可用于揭示该衍生影响。

评价学概念定义 8 (衍生自包含系统，衍生评价条件). 衍生自包含系统是研究衍生 EO (EO 的影响) 对间接影响对象影响的实验系统，它由衍生 EO、间接影响对象以及无法隔离的混杂对象组成。从衍生 SCS 中移除衍生 EO 后，所得到的衍生系统称为 衍生评价条件 。

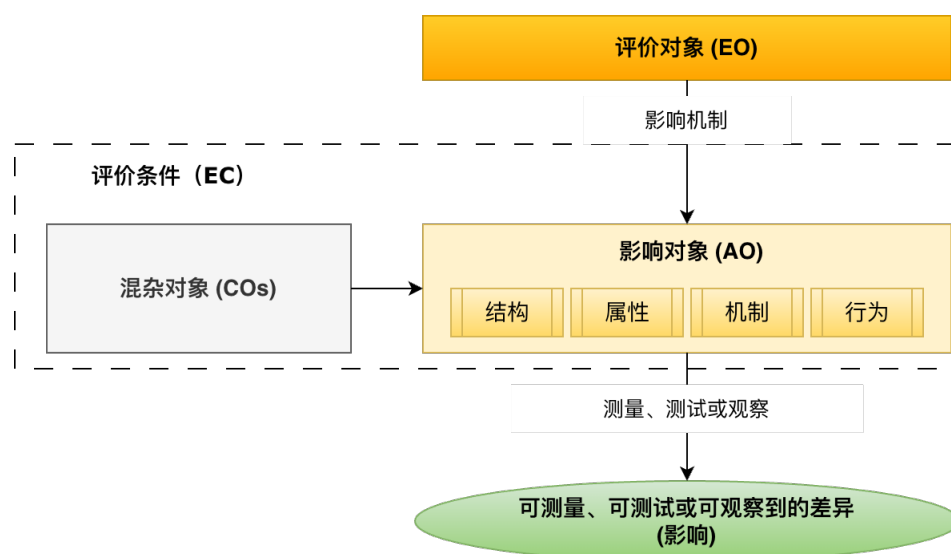


图 10.2: 自包含系统的基础组件。

在该衍生自包含系统的上下文里，我们可以推断出衍生影响。然而，它可能包含许多直接或间接的 AO；EO 的影响可能会通过多个间接 AO 进行传递，且链条可能很长。

10.3 对象因果模型

评价学概念定义 9 (对象因果模型 (object causal model, OCM))。对象因果模型 是基于对象之间的影响机制建立的因果模型，它是现实世界系统和自包含系统的**数学模型**。

自包含系统定义了 EO, AO 和 COs，在此基础上，对象因果模型进一步刻画对象之间多样化的影响机制。它可以用于预测对象的**变化**。对象因果模型的检验依赖于对自包含系统或者现实世界的测量或测试。

10.3.1 符号与建模语言

我采用以下符号表示对象之间的影响，请注意每个领域的影响机制都可能不同，第 [十三章](#) 将详细介绍不同领域的影响机制。 A 和 B 分别为独立的对象， $A \Rightarrow B$ 表示 A 对 B 有影响。我使用了符号 $\Rightarrow$ 表示对象之间的影响， $\Rightarrow$ 不同于 $\rightarrow$ ，本书在其他场景下使用 $\rightarrow$ 。 $A \nRightarrow B$ 表示 A 对 B 没有影响。 $A \Leftrightarrow B$ 表示 A 和 B 之间互相有影响。

如果 $A \Rightarrow B$, $A \nRightarrow C$, $A \Rightarrow B \rightsquigarrow C$ 表示 A 通过 B 给 C 施加的间接影响，可以用存在算子表示为 $C \mid \text{ext}((B \mid \text{ext}(A)))$ 。间接影响可以进一步传递，例如 $A \Rightarrow B \rightsquigarrow C \rightsquigarrow D \rightsquigarrow E$ 。

刻画对象之间互相影响需要新的建模语言。在 SCM 方法里广泛使用的**有向无环图**(DAG) 无法胜任刻画对象之间相互影响的建模语言。例如力学中两个对象互相施加作用力和反作用力，作用力就是一种典型的影响机制，用 DAG 无法描述。另外，DAG 结构也无法描述对象的结构、机制或者行为，对象的消失与新对象的出现。

10.3.2 基本构造块

OCM 以对象作为基本构造块 (building block) 对现实世界系统或者自包含系统进行建模, 不同于第 十九 章阐述的结构因果模型以及其他章节定义的其他评价方法: 它们都基于第 三章定义的**事件**——也就是对象的变化结果——进行建模, 建立**变量**之间的关系。

这种新的建模方式带来了变化, 例如需要新的原语构造块来刻画对象之间的相互影响。基于对象建模, 存在以下几种基本构造块。

最简单的基本构造块是 $A \Rightarrow B$, B 是一个显式对象。通过观察、测量和测试可以获得 A 对 B 的影响。

第二种基本构造块是 $A \Rightarrow B$, 但 B 不是一个显式对象, 则需要识别一个最小的显式对象 C , 使得 $B \subset C$ 。在这种情况下需要推断 A 对 B 的影响, 因为 C 中除 B 以外的其余组成 (记为 $C - B$) 也可能存在对 B 的影响。

第三种基本构造块是 $A \Rightarrow B$, 其中 A 和 B 均是显式对象, 但存在混杂对象 A_i , 对对象 B 有影响, 并且无法隔离。可直接观察、测量或者测试的影响是 A 和 A_i 对 B 的总体影响, 需要推断 A 对 B 的影响。

第四种基本构造块是对象 $A_1, A_2, A_3, \dots$ 均不是显式对象, 它们一起成为一个第三方可测量和可测试的显式对象 A 。需要推断 A_i 对 A 的影响。

第五种基本构造块是 $A \Rightarrow B$, A 和 B 均是显式对象。对象 $A_1, A_2, A_3, \dots$ 均不是显式对象, 它们是 A 的组成部分。需要推断 A_i 对 B 的影响。

第六种基本构造块是 $A \Rightarrow B$, A 是显式对象, B 不是显式对象, C 是最小的显式对象, 使得 $B \subset C$ 。对象 $A_1, A_2, A_3, \dots$ 均不是显式对象, 它们是 A 的组成部分。需要推断 A_i 对 B 的影响。

第七种基本构造块是 $A \Rightarrow B$, A 和 B 均是显式对象。对象 $A_1, A_2, A_3, \dots$ 均不是显式对象, 它们是 A 的组成部分。相应地, 对象 $B_1, B_2, B_3, \dots$ 均不是显式对象, 它们是 B 的组成部分。需要推断 A_i 对 B_i 的影响。

结构因果模型 (SCM) 方法 [221, 224] 的基本构造块, 例如**链** (chain) 或者**中介** (mediation), **叉** (fork) 和**碰撞** (collider) 在 OCM 里不再成立。在 SCM 里, 这三个原语构造块的属性是前门准则、后门准则或者更通用的 do 演算能工作的前提条件。我仍然以 [224] 里讨论的经典例子来阐述这个问题。

一个典型的链或者中介的例子 [224] 是 火 $\rightarrow$ 烟 $\rightarrow$ 警报。根据 SCM 里的经典**结论**, 如果控制烟, 即有烟或者没有烟, 则 火 和 警报 条件独立。

按照 OCM 来分析, 火 和 警报器 都是对象, 烟 是对象 火 的一种**属性**。也可以进一步分析火, 认为它是多种对象相互影响的一种现象 (事件集合), 例如可燃烧物质、氧气、诱发燃烧的对象, 如废弃的燃烧的烟头, 或者诱发燃烧的某个事件, 例如电路发生短路。这里按照最简单的情况进行建模, 只考虑火和警报器。如果 火 有 烟 (一种属性), 则对象 警报器 发出 警报 (对象的一个机制)。如果 火 没有 烟 (一种属性), 则对象 警报器 不发出 警报。

另一个典型的例子 [224] 是关于叉构造块的。鞋码尺寸 $\leftarrow$ 小孩年龄 $\rightarrow$ 阅读能力。根据 SCM 的分析, **混杂变量**小孩年龄 使得鞋码尺寸 和 阅读能力 具有统计相关性。如果控制小孩年龄, 则鞋码尺寸和阅读能力条件独立。

按照 OCM 来分析, 小孩是一个对象, 鞋码尺寸、年龄、阅读能力是他或她的三个属性或机制, 均随着时间发生变化。鞋码尺寸受**内部机制** (生长发育) 和外部对象 (环境和影响) 的影响, 阅读能力受内部机制 (智力水平)、**自由意志** (例如内部驱动力) 和外部对象的影响, 例如家庭和学校等。

最后一个典型例子 [224] 是有关碰撞 (collider) 构造块的。演艺天赋 $\rightarrow$ 明星 $\leftarrow$ 美貌。根据 SCM 的分析, 对于普通人来说, 演艺天赋和美貌本来是独立的, 如果对明星 =1 进行控制, 则演艺天赋和美貌变得相关, 要么这个明星具有演艺天赋, 要么具有美貌。

按照 OCM 来分析, 一个普通人或者明星都是一个对象, 演艺天赋 和美貌是一个具体对象的两个属性。一个普通人, 需要在一个时间范围内和一系列对象发生相互影响。成为明星实际上是这个人的影响。他或者她 (或者其作品) 对别人产生的影响足够大后, 就可以称之为明星。

10.3.3 基本建模方法

OCM 基于对象及其对象关系对世界进行建模。OCM 区分被动对象, 普通生命和智能生命。OCM 在对象层次区分六种不同维度的对象性质, 未知和已知, 真随机性和表面随机性, 确定和不确定, 明确定义 (*well-defined*) 和未明确定义, 内生和外生以及显式和隐藏。OCM 在机制方面区分影响机制、内部机制和自由意志, 采用不同的方法建模。OCM 基于这样的概念框架, 对对象的变化进行建模。

评价学概念定义 10 (内生, 外生, 内生对象, 外生对象)。内生和外生是相对于研究的系统而言的, 内生指属于系统内部, 外生指系统以外。当自包含系统未能包含所有的对象时, 称自包含系统内部的对象为内生 (*endogenous*) 对象, 自包含系统外的混杂对象为外生 (*exogenous*) 对象。

OCM 可以隔离每个对象来分析系统内其他对象对它的影响, 这种分析天然具有局部性。也就是说, 可以针对每一个对象来设立对象影响方程。对象影响方程的意义在每个领域都不同, 需要用专业知识来解读。OCM 针对每个对象, 建立一系列的方程。通过求解方程, OCM 推断一个对象对其他对象的影响。方程的求解依赖于领域知识。

在满足因果律公理的前提条件下, OCM 是一个确定性的方法, 它刻画每个被动对象如何受其他内生和外生对象的影响而发生变化, 这种影响通常会随时间发生变化。

如果能对内部机制或者自由意志建模, OCM 也能刻画每个生命如何受其内部机制以及其他内生和外生对象的影响而发生变化, 或者进一步刻画每个智能生命如何受其内部机制、自由意志以及其他内生和外生对象的影响而发生变化。

10.4 评价结果的四个基本性质

评价的实质是揭示对象的影响或者衍生影响。影响最终会体现在属性、结构、机制或者行为的变化, 或者对象的消失或者新对象的产生。

本节, 我将讨论评价结果的四个基本性质, 它们源自评价的本质, 见第 8.1 节的讨论。这四个基本性质关注评价结果的真值、评价结果差异的可溯源性、评价结果的可比较性和一致性。

在后续版本中, 我会给出严格的证明。

10.4.1 评价结果的真值

这个性质宣称: 当对象及其影响机制不具有真随机性时, 对象的影响具有真值。

这个属性来自因果律公理。根据因果律公理，当对象及其影响机制不具有真随机性时，EO 对 AO 的影响是确定的。确定意味着存在真值，但这个真值是广义的真值，包含属性、结构、机制和行为，以及对象的消失或者出现。

10.4.2 评价结果差异的可溯源性

该性质宣称：在满足因果律公理的前提下，对于同一个（衍生）评价对象，总体（衍生）影响的差异可归因于（衍生）评价条件的差异，从而确立了评价的可溯源性（Evaluation Traceability）。

该性质关注影响差异的可追溯性。当对象及其影响机制不具有真随机性时，根据评价结果的第一性质，其影响存在真值。总体影响的差异可归因于评价条件的变化。若无此公理，总体影响中观察到的差异将无法解释，这违背我们的科学和工程直觉。

对于衍生评价对象，衍生影响的差异可归因于衍生评价条件的变化。该性质同样适用，原因相同。

10.4.3 评价结果可比较性

该性质宣称：当满足因果律公理的条件下，不同组的（衍生）EO 和（间接）AO 在相同的混杂对象（COs）条件下，它们的影响（或衍生影响）是可以比较的。

不言而喻，这个性质与评价结果的可比较性相关。相同的 COs 提供了统一的参照系。评价结果能在相同条件下反映不同 EO 和 AO 之间的影响，因而影响之间具有可比较性。

同样，当两组衍生评价对象与间接评价对象在相同的混杂对象（COs）下进行比较时，它们的衍生影响能够准确反映在相同条件（相同的混杂对象（COs））下衍生 EO 对间接 AO 的影响，从而使这些结果具有可比较性。

10.4.4 评价结果一致性

该性质宣称：当满足因果律公理的条件下，当使用来自混杂对象总体的不同样本，来揭示（衍生）评价对象对直接（间接）对象的影响时，其（衍生）影响始终收敛于真值，称为评价结果一致性（Consistent Evaluation Outcomes）。该性质为（衍生）影响在混杂对象总体下的真值提供了估计的可能。

根据评价结果的第一性质，一对评价对象和直接影响对象在混杂对象总体下，其影响具有真值。从混杂对象总体中抽取一个样本，它作为整个总体的近似。因此，不同样本产生一致的影响结果，这些结果收敛于整个混杂总体下的真值。这种收敛受到样本能力的影响，而样本能力由所选择的采样策略决定，该策略决定了样本能否准确代表总体的基础特征。

当一对衍生评价对象和间接影响对象配备混杂对象总体时，其衍生结果具有真值。该公理同样适用，原因相同。

10.5 总结

本章介绍了评价学三个核心概念：现实世界系统、自包含系统与对象因果模型。进一步在自包含系统的概念基础之上，讨论了评价结果的四个基本性质。

第十一章

狭义评价学：价值量的定义、实现与赋值

本章给出了价值的解释，定义了狭义评价学，提出了狭义评价学的基本方法。最后给出了两个狭义评价学的案例。

11.1 价值的解释

不同于其他对象，利益相关方具有[自由意志](#)，可以在不同的评价对象之间做出自由选择。不同评价对象对利益相关方产生不同的影响或者衍生影响，它们之间的影响差异对利益相关方产生的影响是价值产生的来源。

评价学概念定义 11 (评价对象的价值 (the value of the EO))。评价对象的价值是评价对象和其他评价对象之间的影响差异对利益相关者产生的影响。

一个 EO 或者其他 EO 可以对同一个利益相关方产生不同维度的影响或者衍生影响，因而也会产生不同维度的影响差异。

评价学概念定义 12 (价值函数 (value function))。价值函数是将不同维度的影响差异映射为一个复合评价指标的机制，这个复合评价指标就是价值。价值本质上是一种[量](#)，价值函数映射的量就是价值量。其定义、实现和赋值应遵循[计量学](#)的基本原理和方法。

11.2 什么是狭义评价学？

评价学概念定义 13 (狭义评价学)。狭义评价学是价值量定义、实现与赋值的科学，是计量学在价值领域的应用。

狭义评价学则揭示对象之间的影响差异对利益相关方的影响。尽管利益相关方具有自由意志，但在神经科学基础方面，自由意志的产生仍具有客观性。因此，价值本身也可以被认为具有客观性。

猜想 8 (自由意志的第二猜想)。

自由意志在神经科学基础上具有客观性，因而价值也具有客观性。

实验学提供了揭示原因和影响的基础方法，尽管狭义评价学也采用相同的方法，但仍存在一些差异。

第一，狭义评价学关注一个对象对利益相关方的价值。一个对象的影响会扩散，但扩散的影响最终会聚集在利益相关方。在确定一个评价对象对利益相关方的价值时，狭义评价学的主要任务之一是确定需要研究的直接影响对象以及间接影响对象的范围。

第二，狭义评价学揭示的是影响差异的影响。自包含系统和对象因果模型的方法仍然适用，但需要引入不同的评价对象以及它们的直接影响对象和间接影响对象。

第三，尽管自由意志具有客观性，但它对行为的影响在不同的个体之间差异较大。因此，需要对利益相关方的不同行为模式进行建模。

第四，价值的赋值是一个总体的影响，依赖于利益相关方本身的结构、属性、行为和机制，但也受其他混杂对象的影响。

11.3 基本方法

狭义评价学的基本方法包括价值量的定义、价值量的实现以及价值量的赋值。

价值量可以通过参考评价对象以及参考利益相关方来定义。

问题定义 4 (价值量定义的问题定义). 存在评价对象 O 的实例 o_i 以及利益相关方 S 的实例 s_j ，如何选择 $n \geq 2$ 个参考对象 o_i 以及 $m \geq 1$ 个参考利益相关方 s_j ，在此基础上定义价值量及其价值函数。

如何选取代表性的参考对象以及参考利益相关方是价值量定义的关键。只有选择至少两个参考对象才能定义影响差异。

不同类型的对象对不同类型的利益相关方产生影响的机制不同，因此不能脱离具体类型的价值对象和利益相关方来定义价值量。也就是说，不同类型的评价对象和利益相关方需要重新定义价值量。

作为测量标准，价值量的实现需要提供一个可工作的系统，它是对价值量赋值的参考。

最后一步是价值量的赋值。该工作在自包含系统的概念基础下开展。狭义评价学背景下的自包含系统需要引入不同的评价对象以及不同评价对象的直接影响对象和间接影响对象，此外还包括混杂对象以及作为测量标准的价值量的实现。

11.4 狭义评价学案例

本节阐述价值量定义、实现和赋值的整体过程，不详细区分每一个阶段。第一个案例由王晨曦撰写，第二个案例由康国新撰写。

11.4.1 处理器评价

一个组织正在采购计算机的关键组件—CPU。为了做出正确的决定，该组织在相应 CPU 基础上搭建的计算机上执行应用程序，评价各种备选 CPU。在此评价过程中，该组织将收集有关性能和能效的大量数据。

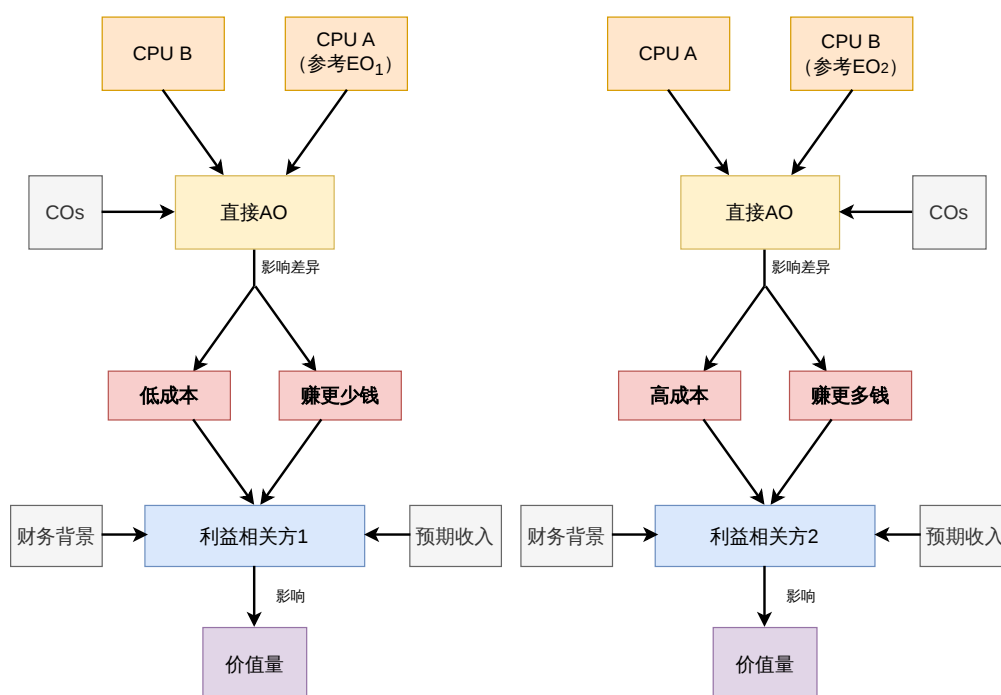


图 11.1: 典型的利益相关方如何为 CPU 赋予价值，也就是 CPU 价值量的赋值。评价学将价值定义为评价对象相对于其他评价对象的影响差异对利益相关方的影响。

利益相关方可能是采购计算机的组织、具有不同需求的特定用户、负责创建 CPU 设计规范的设计人员以及制造 CPU 的生产商。在这个例子里，我假定利益相关方是使用 CPU 的应用服务商。

图 11.1¹ 展示了该利益相关方，它经营一项业务，应用程序的运行速度直接影响盈利能力——更快的性能带来更高的收入。然而，性能更快的 CPU 功耗更高，购买价格更高，带来更高的购买与维护成本。

假定已经获得了不同维度的 EO 影响。仍存在一个问题：不同的 CPU 在各类应用中性能表现各异，购买与维护成本也各不相同，利益相关方该如何权衡选择？

根据 11.1 节的讨论，EO 的价值反映了 EO 以及其他 EO 的影响差异对利益相关者的影响。请注意，EO 的影响可以传播到多个直接或者间接 AO，最终聚集到利益相关方。

根据收集到的数据，该利益相关方将形成一个显式或隐式的价值函数，该函数可以将 EO 相对于参考 EO 的影响差异映射到一个综合评价指标上。

价值量的定义可以基于两款参考 CPU A（参考 EO₁）和 B（参考 EO₂）时，它们的性能选项呈现出明显的权衡：A 以更高的功耗代价提供更快计算速度，而 B 则以较低的计算速度为代价优先考虑能耗效率。由于 A 在速度上优于 B，因此它可为利益相关方带来了更高的收入。影响的差异可以被量化，即图 11.1 中的“高成本”或者“赚更少钱”。

利益相关方对这些备选 CPU 价值的赋值取决于利益相关方的状态及其他混杂对象，

¹在和我讨论的基础上，王晨曦绘制了本图。

例如财务背景和预期收入，这些因素会影响利益相关者方何为备选 CPU 的价值赋值。价值函数将系统地考虑这些因素。前文阐述的自包含系统仍然可以应用于这个场景。

11.4.2 药物评价

在评价一种针对特定疾病（例如某种恶性肿瘤）的新药物时，首先要确定影响对象（例如患有该特定疾病的患者或患者群体）。可以在这些影响对象上测量或测试出该新药物（原因对象，UO）或其竞争药物（不同 UO）的影响差异。在此过程中，医疗或研究机构将收集有关临床疗效和安全性的海量数据。

在这种情况下，针对该疾病的新药物是原因（Cause，即 UO）对象，现有的标准治疗药物是竞争 UO，特定患者或一组患者是**直接影响对象**（AO），而特定患者身上可测量或可测试的差异（如肿瘤缩小率、不良反应发生率）则是该药物引起的影响（Effect）。在生物学中，确定药物影响的底层**机制**通常极具挑战性。同时，对于治疗效果而言，混杂对象（COs）可能包括工作条件（这可能会给患者带来不同的心理或生理压力）以及生活条件（如不同的空气或水质）。在现实中，我们无法将工作条件和生活条件的影响完全隔离开来。因此，我们需要构建一个自包含系统（SCS），该系统同时包含了不同的 UO 及其它们的 AO 和混杂对象（COs），在此基础上我们得以揭示该药物的影响链。

假定我们已经获得了该药物（UO）及其竞争治疗方案在不同维度的直接影响。仍存在一个核心问题：针对该特定疾病，不同的备选药物在临床表现上各异，采购成本与医疗资源消耗也各不相同，利益相关者该如何权衡选择？

在药物评价案例中，利益相关方可以是患者家属、医院、医保机构或政府。在这个例子里，我假定利益相关方是公共医保机构。公共医保机构致力于提升整体人群的健康水平，药物的疗效直接影响其目标的达成——更显著的治疗效果带来更高的公共健康收益。然而，疗效更好的创新药物（UO）往往研发成本极高，导致更高的采购价格，或者可能伴随需要特殊重症监护的副作用，带来更高的后续医疗与维护成本。

根据 11.1 节的讨论，药物的价值反映了该针对特定疾病的药物（UO）以及其他现有药物的影响差异对利益相关者产生的影响。请注意，药物的影响首先作用于特定患者（直接 AO），进而可能改变医疗资源的占用和家庭的经济负担（间接 AO），最终触及作为支付方和决策者的利益相关者。

根据收集到的临床与卫生经济学数据，该利益相关者将形成一个显式或隐式的**函数**，该函数可以将新药物（UO）与候选药物（竞争 UO）对直接 AO 的影响差异，映射到一个综合评价指标。这就是狭义评价学中的**价值函数**（value function），它将临床层面的药物影响转化为了决策层面的价值。

当决策者比较两款治疗该疾病的药物 A（参考 UO，例如作为 UO 的某新型靶向药）和 B（参考 UO₁，例如作为候选 UO 的传统化疗药）时，呈现出明显的权衡：A 以其高昂的经济代价提供了更好的疗效，而 B 虽然疗效不如 A，但药品价格更易被患者接受。不同 UO 的影响差异对利益相关方的影响是可以量化的，即经济支出和寿命延长。利益相关方对这些备选药物价值的赋值，高度取决于利益相关方自身的状态（例如预算或患者数量）及其他混杂对象（例如当地该疾病的流行病学特征），这些因素会直接影响利益相关方如何为备选药物的价值赋值。价值函数将系统地考虑这些因素。

11.5 本章小结

本章给出了价值的定义，在计量学定义的基础上，将狭义评价学定义为价值量定义、实现与赋值的科学。最后给出了狭义评价学的基本方法和两个典型案例。