第一篇

本书贡献和相关工作

第一章

本书的贡献

在本章，我将开门见山地回答一个核心问题：本书的关键贡献是什么？

本书自始至终，当使用“我”时，指詹剑锋。当提到其他人时，本书会直接使用他或者她的名字。“我们”则包含多个人，具体的信息会在上下文里提供。另外，为了更准确的表达，第一次出现专门术语时，我会在括号里添加对应的英文。另外，每一个概念在每一章第一次出现时，会加上该概念定义的链接。

1.1 提出理论构建困境猜想

任何理论的构建都依赖于概念的定义，明确定义的概念是构建和描述理论的语言的基本组成单元。本书会从头开始定义所有的概念。不幸的是，没有概念是不言而喻的，都需要定义，这就会面临一个逻辑的悖论：如何定义第一个概念？因此，我首先阐述构建理论的困境，这也是本书的贡献之一。

构建理论有两个困境：首先，定义一个概念时需要借助一个或一组明确定义的概念，一组概念的定义之间还可能相互依赖。这样会陷入概念定义的嵌套或者相互依赖的陷阱。我发现在定义的过程中无法打破这一陷阱；其次，任何理论的构建都必须借助理论以外的假说（*assumption*）。用后文正式的定义概念可以将构建理论的困境描述如下：

猜想 1 (理论构建困境猜想)**.** *任何理论的构建都存在一个困境：必须依赖于外生的（exogenous）概念和假说，理论自身不能自洽。*

这些外生的概念和假说可以认为是一个公理系统（*axiom system*）。

1.2 构建的知识体系

图 1.1 [1]总结了本书构建的知识体系。我将主体认识世界的四个基础求索方法：测量、测试、推理和评价升华为四个互相支撑的学科：计量学（Metrology）、测试学（Testology）、推理学（Inferology）和评价学（Evaluatology）。在本书的第二部分，我统称它们为求索科学。

[1]在我绘制的草图基础上，王磊绘制了该图。

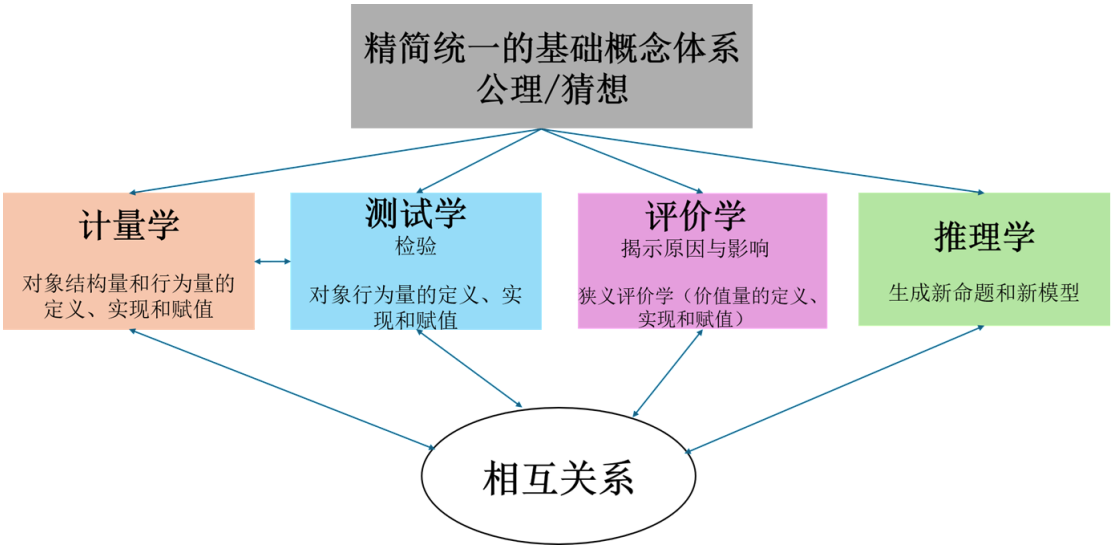

图 1.1: 求索科学体系下四个学科之间的关系。

1.2.1　定义新的计量学

我在求索科学的体系下重新定义了计量学。求索科学体系下的计量学是定义、实现和赋值对象的结构量和行为量的科学。对象典型的结构量有长度、重量，行为量则有正确率、安全性等。传统的计量学定义是测量的科学及应用。

王磊和我初步实现了这一想法。

1.2.2　提出测试学

我将测试学定义为检验与测试的科学：检验是唯一能验证命题和模型有效性的方法；测试基于检验，对测试对象的行为量进行定义、实现和赋值。实质上，它是求索科学体系下的计量学在测试领域的应用。我用 Testology（测试学）一词来命名这一新领域。

我和王磊提出了统一的测试原理和方法学（testing principles and methodologies），可以跨越不同学科。例如验证一个物理学理论（theory）和模型（model）、命题或者假设的检验（hypothesis testing）以及人造物 [2]的测试（如软件和硬件的测试）。”

王磊和我实现了这一想法。

1.2.3　提出推理学

范帆达和我提出了推理学，我创造了英文单词 Inferology 来描述这一学科。我们的目标是在求索科学的体系下，把不同学科中关于推理的概念、问题和方法统一到一个更基础的框架下——把推理视为智能生命的基础求索方法之一。我们将推理学定义为生成新命题和新模型的科学。

[2]中文里，通常将 artifact 翻译为人工制品，我翻译为人创造的物体，简称人造物，可以是人造卫星、艺术等。

1.2.4　提出评价学

我正式引入评价学（Evaluatology）学科，我创造了英文 Evaluatology [3]这一术语来涵盖评价学这门令人振奋的科学。我将评价学正式定义为“揭示原因与影响的科学”。评价学揭示原因对象的影响和解释现象的原因，并且能够用于设计，它的核心议题之一是设计和实现实验。我将实验学（Experimentology）定义为设计、实现实验，并获得、分析和检验实验结果的科学。评价学是综合地应用实验学、计量学、推理学和测试学来揭示原因和影响的科学。

我将一个对象的价值解释为它和其他对象之间的影响差异对利益相关方的影响，在此基础上，将狭义评价学定义为价值量（value quantity）的定义、实现和赋值的科学，这也是计量学在价值领域的应用。传统的评价研究与实践相当于本书中阐述的狭义评价学。评价学的范畴更广，包括揭示对象的影响（评价）、解释现象的原因（逆评价）以及设计（评价的对偶问题）。

为了突出我创建的学科与已有相关学科的差异，在本书中，我始终将中文“评价学”和英文“Evaluatology”这两个词语作为一个整体使用。

我提出，检验统一的评价学（Evaluatology）学科是否建成，至少有三个检验基准。首先，需要明确定义不同评价问题的结构，指出哪一类评价问题是可以计算的，哪些评价问题具有真值。其次，需要提出统一的概念、公理、数学或图形表示和方法。最后，统一的理论可以在不同领域应用。

1.2.5　定义四门学科的相互关系

如图 1.1所示，计量学、测试学、推理学和评价学四个学科在智能生命求索的过程中相互支撑：计量学负责定义、实现和赋值量、推理学基于已有命题和模型或者新求索结果生成对象或者对象关系的新命题和新模型。评价学基于测量、测试和推理揭示现象的原因和原因对象的影响，并生成模型。狭义评价学是价值量的定义、实现与赋值。测试学不仅定义、实现和赋值对象的行为量，并且负责检验命题和模型的有效性。

更重要的是这四个学科相互支撑：可以基于评价学来揭示测量系统不同组件、对象的影响，如量的定义，对测量结果的影响；可以基于测试学来检验评价结果的有效性；可以基于推理学为计量学构建更严谨的公理系统；也可以基于计量学来定义测试学的量等。

评价学是集大成者，它紧密依赖计量学、测试学和推理学。评价学依赖计量学和测试学为对象的量赋值；依赖推理学推测不能直接测量或测试的量；依赖测试学检验评价结果的有效性。这是我为什么仍然保留本书标题的主要原因。

1.2.6　从头开始定义精简和统一的基础概念体系

我从头开始定义了一套精简和统一的基础概念体系作为本书的基础，具体包括对象（object）、关系（relationship）、事件（event）、主体（subject）、求索（interrogation）、因果（cause and effect）、命题（proposition）、模型（model）和公理系统（axiom system）。其中，对象是最基础的概念，我将它定义为拥有结构（structure）、属性（property）、机制（mechanism）和行为（behaviour）的一类实体。

[3]我最初创造的英文单词是“Evaluationology”。我的博士生李鸿霄咨询了中国科学院大学的外籍英文教师，后者建议将它改为“Evaluatology”。我接受了他的建议。

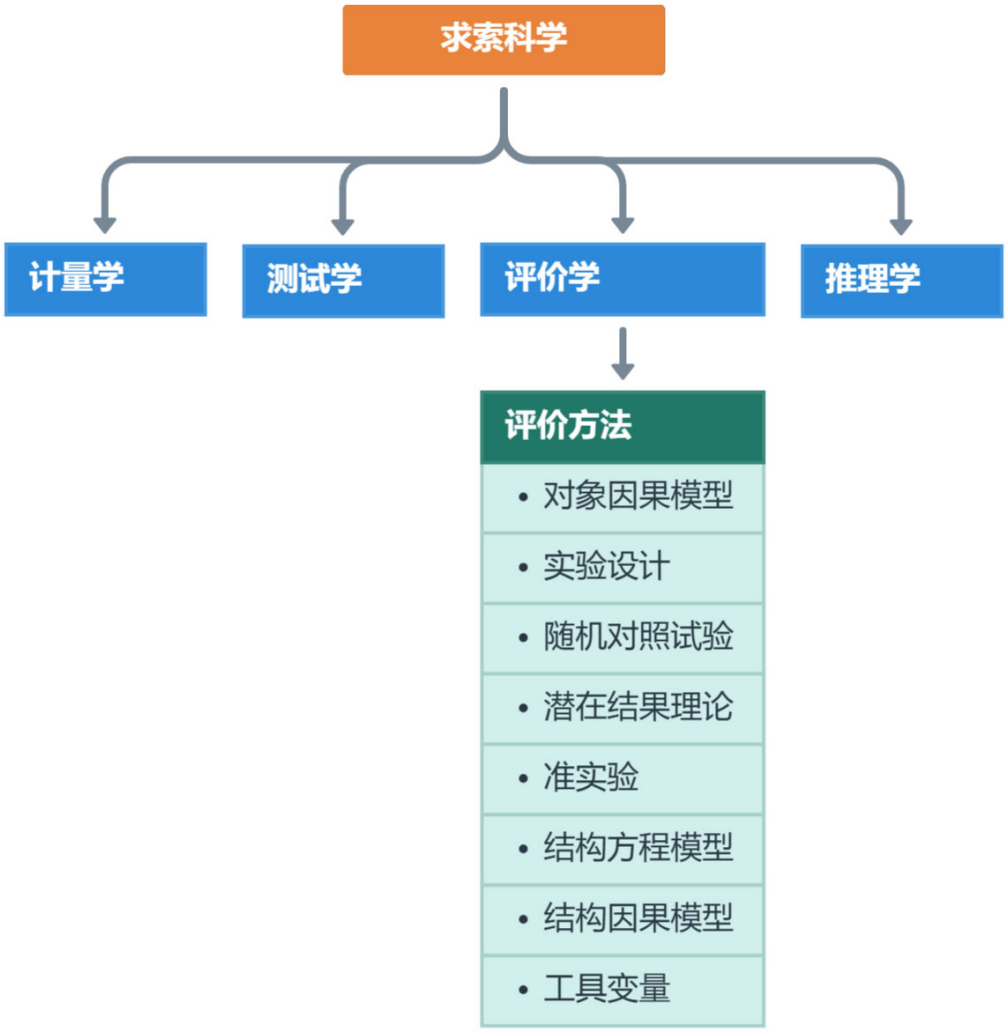

图 1.2: 本书构建的层次化概念体系。

不同学科经常使用名字相同的概念，它们的含义是否一致？例如经典的概率论中概率定义在事件这个基础概念之上，事件在其他学科里也有不同的含义，例如计算机学科中常讲事件驱动，相对论理论中也有事件的概念。这些概念尽管使用了相同的名字，但内涵和外延并不一致。我在本书中尝试建立一套精简和统一的概念体系来描述不同学科需要解决的问题。

不少基础概念及其它们之间的关系缺少明确的定义，例如对象、关系、事件等概念及它们的关系。在本书中，我尝试从头开始定义所有将使用的概念，并阐述它们之间的关系。我坚持用浅显易懂的语言，尽管我认为严谨的数学语言是必要的。

我尝试减少基础概念的数目。在本书里，我主要使用这九个基础的概念和其他辅助或者衍生的概念来描述我尝试解决的问题。从这个意义上来说，基础概念体系相当于公理系统的地基。

我对一些基础概念进行了新的定义，例如因果，我从对象相互关系的角度进行了新的定义，这是本书的基础之一。

我定义了一些新的概念，例如因果独立性，因果条件独立性，这需要在将来发展新的数学模型和方法。

根据这四门学科的关系，相应地，我也提出了层次化的概念体系，如图 1.2 [4]。第一层是求索科学的概念，第二层是计量学、测试学、推理学和评价学的概念，第三层是不同评价方法的概念。如果是同一个名字的概念定义，它们的作用效果取决于具体的上下文。例如在两个场景（求索的方法和某一个具体的评价方法）下出现相同概念不同的定义时，当上下文是某一个具体的评价方法时，后者会覆盖（override）前者。

1.3　提出因果律公理

我定义了因果、随机和确定等概念，并在此基础上提出了因果律公理，作为求索科学体系的公理基础。

1.3.1　因果影响的定义

我正式定义了因（cause）和果（effect），并且定义了简洁的数学和图形符号。

考虑由对象 A 和 B 组成的理想环境（setting）里，不存在其他对象，或者其他对象的影响可以被忽略；相对于不存在 A 时，存在 A 会导致 B 中可测量、可测试或可观察到差异。我定义 A 为*原因对象 (cause object, UO)*, 简称原因或者因 (cause)，B 为*被影响对象*（*affected object*），简称 AO 或影响对象，B 中可观察、可测量、可测试的差异称为 A 对 B 的*影响*（*effect*）。

A 对 B 的影响可以是 B 的结构、属性、机制或者行为的变化，或者 B 的消失，或者新对象 C 的产生。我将*影响机制*定义为原因对象对影响对象施加影响的方式。影响机制是一种外部机制。

我用存在算子（existence operator）ext() 来表达某一个对象的影响。符号 $|\ \mathrm{ext}(A)$ 表示相对于不存在 A 时，存在 A 产生的差异。符号 $B\ |\ \mathrm{ext}(A)$ 表示相对于不存在 A 时，存在 A 导致对象 B 可观察、可测量、可测试的差异。

[4] 在我绘制的草图基础上，何谦绘制了该图。

1.3.2 因果律公理

在定义确定与随机等概念的基础上，我显式地提出了因果律公理：如果一个对象及其影响机制不具有真随机性，则对象的影响是确定的。因果律公理是本书提出的理论体系的基础。

1.4 提出新的世界模型

我提出了新的世界模型。

1.4.1 定义内部机制和自由意志

我定义了内部机制和自由意志。

考虑仅由对象 B 组成的理想环境（setting）里，不存在其他对象，或者其他对象的影响可以被忽略；受内部机制的影响，B 发生变化，可测量、可测试或可观察到差异，B 中可测量、可测试或可观察的差异称为内部机制对 B 的影响（effect）。可以表达为 $B \mid \mathrm{ext}(B)$。

自由意志（*free will*）是做出自由和有意图的选择的潜力（capacity）和实际能力（capability）。自由意志有不同的自由度（degree）。我提出了自由意志第一猜想和第二猜想，为自由意志的建模奠定了基础。

1.4.2 提出三个世界假说

我提出了三个世界的假说。

微观对象世界（microscopic object world）由微观对象构成，通常是原子和亚原子尺度的粒子，遵循量子力学原理，具有真随机性（true randomness）。

被动对象世界（passive object world）由被动对象构成，受因果律（the axiom of cause and effect）支配，也就是说任何被动对象的量的改变都取决于其他对象的影响，但影响机制可以多样。在实际观察中，被动对象的随机性是一种表面随机性，产生的根源是因为未知对象或者已知对象未知的影响，以及主体和求索系统与对象交互产生的随机性。

自由意志世界（free will world）由被动对象、普通生命和智能生命构成，不仅由因果律支配，也受自由意志以及其他内部机制的支配。也就是说，智能生命不仅被动地接受原因对象的影响，还具有做出自由和有意图选择的潜力与实际能力，并且受内部机制的影响。

1.4.3 解释世界变化的理论

我提出了一个简单的理论来解释世界的变化。

微观对象具有真随机性。被动对象的变化受因果律支配。生命世界中普通生命的变化受因果律和内部机制的影响；智能生命的变化受因果律、内部机制和自由意志的支配。

我进一步提出了进化猜想。

图 1.3: 我故乡的山–铁林寨的石船，有很多条乡间小路可以到达它。这和求索很像，启发我提出了求索铁林寨模型。我还写过一首诗《铁林寨，大山妈妈》，讨论不同时代人、动物和植物命运的际遇；我对铁林寨的观察结论是很不幸的：在历史洪流里，人的命运有时甚至不及动物和植物。这首诗见我的个人诗集《秋日已尽》。

1.4.4　提出智能生命区别于被动对象和普通生命的独特机制

我提出智能生命有两个区别于被动对象和普通生命的独特机制：自由意志（free will）和求索能力（interrogation）。在此基础上，我提出了实现人工智能路径的构想。康国新博士和高婉铃博士与我共同对这些想法进行了深入探讨、阐述和实现。

1.5　实验方法猜想以及随机、独立和计算复杂性关系猜想

我提出了揭示原因与影响的物理实验方法猜想和虚拟化实验方法猜想。

有且仅有三种物理的实验方法可用于揭示原因与影响：随机化、隔离与存在算子。

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。

我进一步提出了随机、独立和计算复杂性关系猜想。

1.6　提出新的求索模型：求索铁林寨模型

我故乡有一座山，名叫铁林寨。我在家的时候，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，如图 1.3。去铁林寨，有好几条山间小路，我小时候，走不同的道路，都可以登上石船，这和求索很像。这启发我提出了新的求索模型：求索铁林寨模型（Interrogation Iron Forest Fortress Model）。

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，基于它们，主体都可以获得对象的命题或者模型，就像我小时候爬铁林寨，从任何一条小道都可以通往石船。在此基础上还可以进一步通过推理，产生新的命题或者模型。但这些命题或者模型最终依赖于检验。只有通过了检验，模型才能成为对象的事实或真理。

求索铁林寨模型并不是一个严格的递进模型，而是阐述了各种方法的互补性，它们都为主体的求索提供了可能。

1.7 揭示评价的本质

我提出评价的本质是揭示一个原因对象的影响或者衍生影响。

对象间的引力表现为一种普遍影响；燃烧体现了氧气的影响；遗传源于遗传物质及其他物质因素的影响；犯罪行为必然产生有害影响；法律实施会产生社会影响；计算机的处理器（CPU）和大语言模型（LLM）也会产生可测量或可测试的影响。

为系统揭示这些多样的影响，我建立了一个统一的概念框架（conceptual framework）、模型（model）和方法学（methodology）。这就是我提出将“评价学”（Evaluatology）作为一门新学科的主要动机。

1.7.1 定义评价问题、对偶问题及逆问题

评价和设计（design）是一对对偶问题（dual problem）。一个原因对象和混杂对象（confounding objects，简称 COs）共同对影响对象（affected object）产生总体影响（Overall effect）。评价问题可以定义为如何从总体影响中推测原因对象的真实影响，而设计问题可以定义为如何探索设计对象（designed object）的结构、属性和机制空间，以实现最优的总体影响。

逆评价的目的是将现象追溯到原因对象和混杂对象的影响。我将在后文正式定义这些概念。

1.8 解释为什么过去创建评价学科的努力失败了？

我系统地解释了为什么过去建立评价学科（discipline of evaluation）的努力失败了。

评价领域的现有共识是将评价定义为“确定事物（thing）优点（merit）和价值，包括内在价值（worth）和可量化的价值（value）的过程”[272, 274, 275, 276, 79, 96, 127, 300, 299]。然而，这一定义未能抓住评价的本质（essence），因为“优点”和“价值”这些术语本质上是主观的——它们的解释在各个主体（subject）之间存在显著差异，这构成了该定义的根本性缺陷。

在本书里，我将已有达成共识的评价领域视为狭义评价学，我将其定义为价值量的定义、实现与赋值的科学。求索科学，具体包括计量学和评价学，为狭义评价学提供了严谨的方法。

过去的工作没有明确地识别和定义评价在认识论（Epistemology）体系里的位置。大多数学者只是在自己熟悉的领域讨论评价，大部分方法也是经验性和领域特定的（ad-hoc）。迈克尔 · 斯克里文（Michael Scriven）是少数几个例外之一，他认为评价是一个工

具学科，但他很遗憾没有准确区分评价和（理论）验证的本质差别，实际上他把后者认为是前者的一部分。

过去的工作没有提出统一的评价概念 (concept)、问题定义 (problem statement)、公理 (axiom)、数学表达 (mathematical formulation) 和方法学 (methodology)。相反，评价领域的权威学者们依赖百科全书式（encyclopedic）的方法，定义成百上千的概念或术语 [274, 190]。

1.9 解释价值的本质

我将一个对象的（可度量）价值 (value) 解释为它和其他对象之间的影响差异对利益相关方（Stakeholder）的影响 (effect)。基于求索科学体系下的计量学和评价学，我提出了价值量的定义、实现和赋值的方法。

1.10 提出了评价学的通用方法，其他方法可视为特例

我基于现实世界系统、自包含系统和对象因果模型三个核心概念，提出了评价学的通用方法。评价学的通用方法是一个综合性的方法，它包括基于现实世界系统的观察方法、基于自包含系统的实验方法和基于对象因果模型的建模方法。其他合作者系统性地总结了其他基础的评价方法，包括每种方法的基本概念、问题定义、假设和方法原理。

我将其他方法处理为了评价学通用方法的特例或者补充方法，包括实验设计 (design of experiments, DoE)、随机对照实验（randomized control trials, RCTs)、准实验（quasi-experiments）、结构因果模型（structural causal models）、结构方程模型（structural equation models)、工具变量(instrumental variables)和潜在结果理论(potential outcome theory）。

1.11 定义基准和实验床

我将英文 benchmark 翻译为基准，根据不同的上下文，benchmark 可能意指评价基准、测试基准或者比较基准。将 testbed [5] 翻译为实验床，以取更广泛的含义。我在后文会正式定义什么是实验。

“基准”和“实验床”在工程中被广泛使用，却缺乏正式定义和严谨的方法论。我正式地给出了“基准”和“实验床”的操作性定义（operational definition)。高婉铃博士与我共同提出了实验床的设计与实现原理和方法。

1.12 在不同领域应用评价学（Evaluatology）

许多合作者积极将评价学（Evaluatology）应用于不同领域，展现了 (Evaluatology) 的强大潜力。这个版本收录了两篇发表在人工智能顶级会议 ICML 2026 的两个工作：高婉铃副研究员、博士生田郑书媛和洪学海研究员将评价学应用于大语言模型；范帆达博士和硕士生王晓睿将评价学应用于时序预测系统。

[5] testbed 的中文翻译包括，试验台; 测试平台; 试验床; 实验床; 测试床

第二章

评价：古老的实践，欠发展的学科

在本章中，我将总结评价领域最先进的理论（state-of-the-art）和实践（state-of-the-practice）。请注意，评价本身有不同的定义，因此不同学者眼里的评价领域也存在差异。在本书，我将计量、测试、评价和推理视为智能生命认识世界并构建知识体系的四大基础求索方法，并将评价学定义为揭示原因和影响的科学。第 2.1节将解释为何评价是一项古老实践，并整合了迈克尔 · 斯克里文（Michael Scriven）博士的观点。第 2.2节总结了不同的评价概念和理论。第 2.3节回答了为何过去的努力未能成功建立评价这一学科，并节总结了评价学和之前研究有什么不同？第 2.4节总结了各领域特定 (ad-hoc) 的评价实践。第 2.5节总结了不同学科中对因果的定义。第 2.6节总结了本章内容。

2.1 评价是古老的实践

斯克里文 [274] 将评价定义为 “确定事物的优点（merit）、内在价值（worth）和可量化价值（value）的过程，或该过程的结果” [272, 274, 275, 276]。基于这一定义，他认为评价是一项古老的人类实践，比人类其他实践更为悠久。我在本节中吸收了他的观点。

斯克里文（Scriven）认为，逻辑 (logic)（本书第七章阐述的推理的基础之一）和评价是 “两个基础工具学科，在不同学术领域普遍应用”。他得出结论：非形式逻辑 (informal logic) 和评价的实际应用比任何正式学科（formal academic disciplines）或其早期形式（early forms）的建立都早。

根据斯克里文 [274]，非形式逻辑和语法（grammar）与语言共同进化，其应用早于任何正式学科建立之前，这两个学科从根本上都依赖于语言结构（linguistic structures）作为其进一步发展的基石。然而，评价实践甚至在语言能力出现之前就已存在，这是由于它们在早期人造物（artifact）的创造中有着不可或缺的作用。例如，最早的记录显示，工匠（石器工人）的材料质量和设计复杂性上有一致的趋势：设计越复杂，材料质量越高。这一现象不是在单个考古遗址中发现的，而且贯穿了数千年的历史。这说明石器工人在制作工具时，会不断评价材料质量和设计效果。

在第 二部分，我将求索定义为理解对象及其对象关系的潜力（capacity）、实际能力（capability）和过程（process）。请注意，不同于计算机软件工程里面向对象的对象概念。我定义的对象是一组具有不同结构、属性、机制和行为的一类实体，它是世界的基本组

成单元。我系统分析了四种基础的求索能力：测量、测试、推理和评价。

我将评价定义为揭示一个原因对象的影响（effect）和衍生影响（derived effect）。如果一个对象有利益相关方（stakeholder），我将一个对象的可量化价值 [272, 274, 275, 276] 解释为它相对于参考对象的影响差异对利益相关方的影响。

对于任何智能生命，不仅仅是人类，甚至包括动物，评价是一种基本的求索能力，因为它们大多数都能知道和预测某些原因的影响。例如，一只鹿知道狮子的出现会危及它的生命。因此，我得出结论，评价实践比迈克尔 · 斯克里文博士提到的时期要早得多，甚至在人类不存在的时期就已经出现。

2.2 已有评价理论

在本节，我将总结不同领域的理论。

斯塔佛比姆（Stufflebeam）提出要对评价进行概念化（conceptualize evaluation）[300]：提出假设，对评价进行系统性定义，建立评价的理论框架。对评价进行概念化实际上就是建立评价的理论体系。

斯塔佛比姆（Stufflebeam）[300] 认为需要解决的八个问题。内沃（Nevo）[202, 203] 在斯塔佛比姆的基础 [300] 上扩展，提出要解决十个主要维度的问题。我利用这个结构 [202, 203]）总结已有的评价理论。在这个过程中，我也吸收了斯克里文 [275] 的讨论。

2.2.1 定义

除了人事评价（personnel evaluation）以外，教育评价 (educational evaluation) 可能是评价学科最早开始的研究领域之一。教育评价先驱拉尔夫 · 泰勒（Ralph Tyler）[314] 将评价视为“确定教育目标在何种程度上得以实现的过程”。根据我的定义，这一定义可重新表述为 “揭示教育过程中任何对象的影响，以确定其是否达到预期影响的过程”。

另一种被广泛接受的评价定义由克龙巴赫（Cronbach）[67]、斯塔佛比姆（Stufflebeam）[298] 和阿尔金（Alkin）[5] 等评价学者提出，即评价“为决策（decision-making）提供信息”。根据我的定义，这一定义可重新表述为 “为了支持决策，揭示并报告对象的影响”。

斯克里文将评价定义为“确定事物的优点（merit）、内在价值（worth）和可量化价值（value）的过程，或该过程的结果” [272, 274, 275, 276]。在这里，我区分了两个意思相近的英文词汇：worth 和 value，前者翻译为内在价值，后者翻译为可量化价值。但如果在有些上下文里，只单独出现一个词汇时，例如 worth 或者 value，我则简化翻译为价值。

社会科学领域的评价学者对评价的定义达成了相当大的共识，即对优点或内在价值的评估（assessment）[79, 96, 127, 300] [1]，或是由描述 (description) 和判断 (judgement) 组成的活动 [105, 288]。一个由 17 名成员组成的评价标准（standard）联合委员会（他

[1] 在本书中，大多数时候，我不区分评价（evaluation）和评估（assessment）。如果有区分，我将评估视为评价这一任务或者学科的一个技术性的过程（procudure）。

们代表了 12 个与教育评价相关的组织）将评价定义为“对某些对象的内在价值或优点进行系统调查” [299]。

社会科学领域的评价学者对评价的定义也体现了这一学科面临也必须解决的困境。他/她们把评价定义为优点、内在价值的评估 [79, 96, 127, 300]，或者是由描述或者判断组成的活动 [105, 288]。那又如何定义“优点”、“价值”、“评估”或者“判断”呢？如何区分内在价值和可量化价值？

根据我的定义，任何事物或者任何评价过程或者其结果都可以成为评价的对象或者衍生对象。对象价值的产生是因为其利益相关方具有自由意志，能选择不同的对象候选者。对象价值的本质是对象相对于参考对象的影响差异对利益相关方的影响。此外，比较同一类别中不同评价对象的影响或衍生影响是可行的。正如第 二十三 章所分析的，我认为比较（comparison）是最原始的求索（interrogation）形式。

一些团体或学者反对评价的判断性定义（judgemental definition）。例如，斯坦福评价联盟小组将评价定义为“对某个项目过程中发生的事件或者及其结果进行系统检查——这种检查旨在帮助改进该项目和具有相同总体目标（purpose）的其他项目” [68]。根据我的定义，某个项目过程中发生的事件或者其结果可以是一个衍生评价对象。揭示其影响自然有助于改进项目。

从教育评价的角度出发，克龙巴赫（Cronbach）及其合作者 [68] 用比喻的方式阐述了评价的内涵：“一位教育者的成功要由受教育者所学到的来判断”，而不是一位“类似篮球比赛的裁判”，后者的职责是决定谁对谁错。根据我的定义，受教育者所学到的是教育者的一种影响；对同一类不同评价对象的影响或衍生影响进行比较是可行的。正如第二十三章所分析的，我认为比较是最原始的求索能力。不同受教育者的影响是可以比较的，这实际上是一种判断。

罗斯（Rossi）等人在其著作中提出了评价概念框架 [242]。贯穿全书 [242]，“评价”、“项目评价（program evaluation）”和“评价研究”这几个术语可以互换使用。尽管他们专注于社会项目（social program）的评价，但他们声称评价研究并不局限于那个领域 [242]。

罗斯等 [242] 将项目评价定义为“应用社会科学研究方法系统调查社会干预项目（social intervention programs）的有效程度（effectiveness），这种方式适应其政治和组织环境，旨在为改善社会状况（social conditions）的社会行动（social actions）提供信息”。在这个定义中，社会干预涵盖了“健康、教育、就业、住房、社区发展、贫困、刑事司法和国际发展”等领域的人类服务项目。

根据我的定义，社会干预项目是评价对象。在揭示不同项目的影响后，我们可以调查它们的有效程度（effectiveness）。

2.2.2 评价的本质或观点

在斯克里文 [275] 的著作中，他总结了评价的几种不同观点，包括 A：“强决策支持（strong decision support）”；B：“弱决策支持（weak decision support）”；C：“相对主义（relativistic）”；D：“丰富描述（rich description）”；E：“社会过程（social process）”；F：“建构主义（constructivist）”或“第四代（fourth generation）”。在原文中，斯克里文使用的是“评价模型”一词。我更愿意使用“本质或观点”而非“模型”有两个原因：本质或观点更为准确；而“评价模型”这个单词在评价学（Evaluatology）中有其他用途。

A: “强决策支持”观点

这种观点在拉尔夫 · 泰勒（Ralph Tyler）的工作中得到了体现，尽管没有被明确阐述，但在 CIPP（背景（context）、输入（input）、过程（process）、产品（product））评价方法 [2] [298] 中得到了广泛的阐述。遗憾的是，这种观点并未揭示评价的本质。相反，它聚焦于评价的目的（purpose）：“将项目评价作为合理项目管理过程的一部分，评价人员进行调查以得出评价结论，旨在协助决策者。”

B. “弱决策支持”观点

这种观点以马文 · 阿尔金（Marv Alkin）[5] 等评价理论学者为代表，他们将评价定义为“为决策者服务，收集事实数据（factural data），由决策者得出所有评价结论”。同样，这种观点并未揭示评价的本质。相反，它聚焦于评价的一个子过程：提供与决策相关的数据，甚至止步于得出评价结论之前。

C. “相对主义”观点

这种观点源自两位社会科学家的研究，他们代表了这一类方法 [243]：认为评价应基于客户（client）的价值（value）[3]，评价人员不对这些价值做出任何判断，也不参考其他价值。遗憾的是，它并未解释“价值”是什么。在评价学（Evaluatology）中，如果一个评价对象有利益相关方，那么价值就被解释为该对象的影响相对于参考对象的影响的差异对利益相关方的影响。请注意，我在第 8.1节对利益相关方进行了正式定义。此外，客户只是众多利益相关方之一，这并不能成为排除其他利益相关方的理由。

D. “丰富描述”方法。

这一观点得到了包括鲍勃 · 斯塔克（Bob Stake）[290]、北达科他州学派（North Dakota School）和许多英国理论家等学者的广泛支持。该观点认为，评价可以作为一种“人类学或者新闻事业 [4]：评价者报告他们所看到的内容，而不试图做出评价陈述（evaluative statement）或推断（infer）出评价结论——甚至不涉及客户的价值（如相对主义者那样）”。

我将“作为一种人类学或新闻事业”理解为我定义的“求索”的一种。然而，以上观点非常模糊，没有严格地定义，并且说明什么是有效的求索方法，例如在不同的求索条件（Interrogation condition）下如何进行测量、测试、推理或评价。

E: “社会过程”学派，

由斯坦福大学学者李 · 克隆巴赫（Lee Cronbach）领导的一个学术小组（此处称为 C&C，即克隆巴赫及其同事的观点 [68]）否认了评价功能的重要性，即“(i) 为项目的外部决策提供支持，或 (ii) 确保问责制（accountability）”。

[2] 原文用的是评价模型，我认为更准确的是评价方法。

[3] 在这个定义里，并没有区分内在价值（worth）和可量化价值（value），所以我在中文里直接直接使用价值。

[4] 英文原文为 ethnographic or journalistic enterprise，ethnographic 中文通常翻译为民族志，是人类学的一种核心方法，跨领域通常难以理解，我翻译为人类学方法。

该观点强调否认评价功能的重要性，但未能揭示评价的本质。

F：“建构主义”或“第四代”方法

根据埃贡 · 古巴（Egon Guba）和伊冯娜 · 林肯（Yvonna Lincoln）[106] 的观点，它拒绝将评价视为对质量（quality）、优点（merit）、价值（value）等的追求（search）。相反，他们声称评价结果是“由个体 (individuals)构建并通过群体 (group) 协商的结果”。

这种观点忽略了任何对象的影响都是客观的。似乎他们试图在解释有关质量、优点、价值等的价值函数是如何被赋予一个评价对象的，我在第8.1节有详细阐述。在评价学（Evaluatology）中，影响和衍生影响都是客观的。价值被解释为对象相对参考对象的影响差异对利益相关方的影响。

G: 跨学科（transdiciplinary）观点

斯克里文 [274] 持跨学科观点，将评价视为一个工具学科。

斯克里文声称，该观点具有四个区别于先前研究的特征 [275]。

首先，它是一种“客观主义的评价观（an objectivist view of evaluation），类似于观点 A，认为评价旨在确定项目、人员或产品等的优点或价值”。不幸的是，当将评价定义为确定价值、优点，而缺少客观解释（objective interpretation）时，很容易陷入主观性（subjectivity）的泥潭。

其次，该方法是一种“以消费者为导向的观点，而非以管理为导向（或中介为导向、治疗师为导向）的项目评价方法——相应地适用于人员评价（personnel evaluation）、产品评价（product evaluation）等”。从评价学（Evaluatology）的角度来看，消费者、管理方、中介或治疗师是不同的利益相关方；不同利益相关方的角度之间并不矛盾。相反，可以使用相同的方法来推断评价对象对不同利益相关方的影响。

第三，该方法是一种“泛化（generalized）的观点”。它“不仅是一种泛化观点，而且是在整个人类知识和实践领域对评价的概念进行泛化”。不幸的是，斯克里文并未提出泛化的评价概念（concept）、公理（axiom）、问题定义（problem statement）、问题结构分类（problem structure and category）、数学和图形符号（mathematical and graphical notation）和通用方法（universal methodology）。相反，他仍然依赖百科全书式的方法来定义成百上千个概念或术语 [274]，正如其他评价社区在 [190] 中所做的那样。

第四，跨学科的观点是“一种技术性的观点（a technical one）”。不幸的是，斯克里文未能找到一种合适的技术语言（如数学或者图形符号）来定义评价问题，也没有讨论不同的评价问题结构，更没有提出解决不同评价问题的通用方法。

2.2.3　功能

斯克里文 [272] 是第一个指出“形成性评价（formative evaluation）”与“总结性评价（summative evaluation）”之间区别的人，尽管他并非第一个意识到这种重要性的人。形成性评价和总结性评价是评价的两种主要角色（role）或功能（function）。

斯塔佛比姆 [297] 也提到了相同的两种功能，他建议“区分旨在服务于决策的前瞻性评价（proactive evaluation）和旨在服务于问责（accountability）的回顾性评价（retroactive evaluation）”。因此，评价可以服务于两种功能：“形成性”和“总结性”。

罗伯特 · E · 斯塔克 [288] 从不同利益相关者的角度，以类比的方式区分了形成性评价与总结性评价："当厨师品尝汤时，这是形成性的；当客人品尝汤时，这是总结性的。"在形成性功能中，评价用于改进和发展正在进行的活动（或项目、人员、产品等）。在总结性功能中，评价用于问责（accountability）、认证（certification）或选择（selection）。

还有其他关于评价功能的讨论是从不同角度进行的。

过程评价（*Process evaluation*）侧重于"项目或干预措施期间的活动和事件，通过记录和收集数据来调查项目或干预措施如何以及为何产生结果（results）"[166, 190]。

"效果[5]评价或评估 (Impact evaluation or assessment)"聚焦于"评价对象（如项目、干预措施、政策、组织或技术）的结果或效果，旨在建立评价对象（evaluand，evaluated object）与结果（outcome）之间的因果推断（causal inference）"[166, 190]。

从评价学（Evaluatology）的角度来看，评价只有一个独特的功能：揭示一个（评价）对象的影响或衍生影响。从这个角度来看，对于形成性评价或过程评价，评价对象是设计或者生产一个人造物的不同阶段中的中间对象，或创造人造物的整个过程。而对于总结性评价或效果评价，评价对象是被交付的对象。尽管评价对象不同，方法论是相同的。

2.2.4　角色

在斯克里文 [274] 看来，评价是"跨学科"中最强大且多才多艺的工具学科之一。斯克里文认为，这些工具学科还包括逻辑、设计和统计。我认为他列的这些工具学科不在一个层次。他声称，"科学本身只能通过评价来与伪科学相区分，即通过评价证据的质量、研究设计、工具、解释等"[274]。

我赞同斯克里文关于评价的基础性角色的观点。我认为评价是基础的求索方法之一，就像测量、测试和推理一样。然而，斯克里文错误地将验证（verification）（在我的分类体系里属于测试的一种）归类为评价。正如我们在第六章中所讨论的，区分科学与伪科学是测试的根本作用。

2.2.5　数据收集

先前的研究广泛讨论了如何收集数据，而没有深入思考评价对象的不同性质，因此这些想法是领域特定的。

例如，斯塔佛比姆的 CIPP 模型[6] [297] 建议，评价应关注"每个评价对象的四个变量：(a) 目标 (goal)，(b) 设计 (design)，(c) 实施过程 (process of implementation)，(d) 结果（outcome）"。根据这种方法，对一个教育项目的评价将是"对（a）目标的优点（merit），（b）计划的质量（quality），（c）计划正在实施的程度（extent），（d）结果的价值（worth）的评估"。

斯塔克 [288] 在他的全貌模式（Countenance Model）中建议，关于评价对象应收集两套信息：描述性信息和判断性信息（descriptive and judgmental）。描述性信息应关注"可能影响结果、交易、实施过程和结果的意图和关于先置条件（prior conditions）的观察"。判断性信息包括"关于相同先置条件的标准和判断，这些条件可能影响结果和交易"。

[5]在本书里，区分效果（impact）与本书定义的影响（effect）。

[6]这里的评价模型相当于评价学（Evaluatology）的（广义的）方法。在评价学 (Evaluatology) 里，评价模型是评价系统的简化表示。

埃贡 · 古巴（Egon Guba）和伊冯娜 · 林肯（Yvonna Lincoln）[105] 扩展了斯塔克的响应式教育模型（Responsive Education Model）[289]，并应用了自然主义（naturalistic）范式。古巴和林肯 [105] 建议，评价人员应关注“五种信息：（a）关于评价对象及其设置（setting）和周围条件（surrounding conditions）的描述性信息，（b）回应相关受众（audience）关切（concern）的信息，（c）关于相关问题（issue）的信息，（d）关于可量化的价值（value））的信息，（e）与内在价值（worth）和优点（merit）评估相关的标准的信息”。

2.2.6　标准与准则

必须选择用于判断（judge）评价对象优点（merit）和价值（worth）的标准（standard）与准则（criteria），是评价被视为具有主观性的根本原因之一。

在本书中，我认为基准（benchmark，oracle），标准（standard）和准则（criteria）的性质相同的，它们本质上都是一种参考（reference），见本书第 5.2节和第 23.1节的讨论。

一些学者不认为评价有判断性质（nature）。一些学者将评价定义为服务于决策或其他目的的信息收集活动 [5, 67, 296]，因此他们无需选择评价准则。

其他学者将“目标实现（goal achievement）”作为评价准则，却没有证明其为何是一个适当的准则 [314, 233]。他们直接忽略评价准则的问题。

评价领域已多次尝试为教育和社会项目的评价制定标准和准则 [62, 299, 298, 303, 215]。尽管一些学者 [68, 291] 批评整个标准制定工作的理论依据在目前的发展水平下还不太成熟（premature），但人们对其范围和内容似乎达成了相当大的共识。

博鲁奇（Boruch）和科德雷（Cordray）[33] 分析了六套此类标准，并得出结论：它们之间存在大量的重叠和相似之处。由丹尼尔. 斯塔佛比姆（Daniel Stufflebeam）博士主持的“教育评价标准联合委员会”[299] 制定并发布了最详尽、最全面的一套标准。该标准委员会由 17 名成员组成，代表了与教育评价相关的 12 个专业组织。所提出的 30 个标准分为四大类：“效用标准（utility standard）确保评价满足实际信息需求；可行性标准（feasibility standard）确保评价现实且谨慎；正当性标准（propriety standard）确保评价合法且合乎伦理；准确度标准（accuracy standards）确保评价揭示并传达了技术上充分的信息。”

大多数评价专家似乎都同意，用于评价特定对象的标准（或准则）必须根据该对象的具体背景（context）和对它进行的评价功能（function of its evaluation）来确定。已有文献中提出的评价准则有着不同的接受程度，包括：实际和潜在客户已识别的需求 [299, 219, 272]、理想的或社会价值观 [7] [105, 127]、由专家或其他相关群体制定的已知标准 [105, 288]，或替代对象的质量 [127, 272]。

罗斯等人 [242] 讨论了项目绩效（program performance）的标准或者准则，它们可能以不同形式体现在项目绩效的各个方面。这些标准或准则可能是“目标人群的需求（needs）或愿望（wants）、陈述的项目目标（goals）和目的（objectives）、专业标准（professional standards）、惯例做法（customary practice）、其他项目的规范（norms）、法律要求（legal requirements）、伦理或道德价值观（ethical or moral values）、社会正义（social justice）、公平（equity）、过去的绩效（past performance）、历史数据（historical data）、项目管理者设定的目标（targets set by program managers）、专家意见（expert opinions）、目标人群干预前的基线水平（baseline levels）、在没有项目的情况下预期的情况（反事实，

[7] 价值观对应的英文是 values，复数，不同于单数的 value，可量化的价值

counterfactual)、成本或相对成本（cost or relative cost)”。社会项目的有效程度是通过其对结果（outcome）的改变来衡量的，这些结果代表了其旨在改善的社会条件的预期改善。

在第 二十三章中，我认为比较是最基本（primitive）的求索方法，测量和测试的原理和方法学都依赖于比较。从这个角度来看，比较与主观性无关，因为测量和测试被认为是客观的。根据评价学（Evaluatology），评价基准实质上是混杂对象，它们是客观存在的，在相同的评价基准下，我们可以对评价结果进行比较。

2.2.7 过程

已有产业界和研究中的评价过程都是领域特定的，因评价观点的不同而有所差异。

一种评价理论方法将评价视为确定目标是否实现的活动 [314]。在这种方法中，推荐了以下评价过程：“(a) 以行为术语（behavioral terms）陈述目标，(b) 开发测量工具(instruments)，(c) 收集数据，(d) 解释发现，以及 (e) 提出建议。”

根据斯塔克（Stake）的全貌模型（Countenance Model）[288]，评价过程应包括：“(a) 描述一个项目（program)，(b) 向相关受众报告描述的信息，(c) 获取和分析他们的判断，以及 (d) 将分析后的判断报告给受众。”

后来，在他的反应式评价模型（Responsive Evaluation Model，斯塔克（Stake）[289] 提出在评价人员与评价对象相关的各方之间持续的“对话”。他明确规定了评价人员与受众在评价过程中进行动态交互（interaction）的 12 个步骤。

普罗佛斯（Provus）[233] 提出了“一个五步评价过程，包括 (a) 澄清项目设计，(b) 评估项目的实施，(c) 评估其中期结果（result)，(d) 评估其长期结果，以及 (e) 评估其成本和效益。”

斐陶斐（Phi Delta Kappa）[8]评价研究委员会 (Phi Delta Kappa Study Committee on evaluation) [298] 提出了一个三步评价过程。它包括“(a) 通过与负责决策的受众交互来界定信息需求，(b) 通过正式的数据收集和分析过程获取所需信息，以及 (c) 以方便交流的方式向决策者提供信息。”

斯克里文 [270] 在他的路径比较模型（Pathway Comparison Model）中提出了九个步骤。古巴和林肯 [105] 建议，一个自然主义的反应式评价应通过一个包括以下四个阶段的流程来实施：“(a) 启动和组织评价，(b) 识别关键问题和关切，(c) 收集有用信息，以及 (d) 报告结果并提出建议。”

我提出的评价学（Evaluatology）有一个统一的评价过程，无论评价对象是什么。

2.2.8 方法

本书的第 四部分系统地总结了其他广泛使用的评价方法，包括实验设计（design of experiments）、随机对照实验（randomized control trials）、工具变量（instrumental variables）、潜在结果理论（potential outcomes thoery）、准实验（quasi-experiments）、结构因果模型（structural causal models）和结构方程模型（structural equation models）。遗憾的是，目前缺乏对这些评价方法的验证。

我提出了评价学通用的方法，它基于现实世界系统、自包含系统和对象因果模型三个核心概念，实际上是三个平行方法的综合体：基于现实世界系统的观察方法、基于自包

[8] “Phi Delta Kappa”是专有名词，采用音译“斐陶斐”（遵循希腊字母 Φ Δ κ 的发音）

含系统的实验方法和基于对象因果模型的模型方法。其他方法可以视为评价学通用方法的特例，或者作为观察、实验和建模方法的补充。

2.3 为什么过去建立统一评价学科的努力没有成功？

本章由我和范帆达撰写。

斯克里文（Scriven）是国际上最早倡导建立评价学科（discipline of evaluation）的学者。在 1966 年发表的文章 [272] 里，斯克里文提出评价是一种“逻辑活动，其本质是相似的，无论是评价咖啡机、教学工具、房屋建设计划还是课程计划。该活动简单来说就是收集和结合绩效数据与一组加权的目标指标（原文是 scales)），以产生比较或数值的评分或者评级（rating)。”这是建立统一评价学科的最早想法。

斯克里文长期强调评价可以成为类似逻辑学的跨学科（transdiscipline）的工具学科，并将评价定义为确定事物的优点（merit)、内在价值（worth）或可量化价值（value）的过程或结果 [272, 274, 275, 276, 277]；此后，从 1960 年代到 2010 年代，斯克里文发表了大量文章，目标是建立评价学科 [272, 270, 271, 273, 274, 275, 276, 277]。在后文里，我将斯克里文的评价学科研究简称为斯氏评价学。

在 2010 年出版的中文著作中，中国学者邱均平等讨论了“科学评价”的含义：也就是从文献计量学出发，研究如何对国家竞争力、大学、学术期刊和科研成果等进行测量与排序 [80]。他们最早在中文语境中使用了“评价学（evaluation science)”这个术语，后文简称邱氏评价学。邱氏评价学本质上是“科学评价”的学科化 [80]，邱均平从学科建设的角度讨论了“科学评价”的概念范围、研究对象和理论基础；但其学科体系以“科学评价”为中心，总结了社会科学、教育和文献计量学中的许多概念和领域特定的方法。遗憾的是，他们完全忽略了斯克里文发表的大量早期文献，也未能注意到我们在第 四部分将讨论的许多严谨的评价方法。邱均平等将邱氏评价学的学科位置在情报学与管理科学的框架之下。

在 2016 年的一篇总结文章 [277] 里，斯克里文声称评价学科已经建立；然而，我认为他夸大了这一情况。我认为检验统一的评价学科是否建成至少有三个检验基准。首先，需要明确定义不同评价问题的结构，指出哪一类评价问题是可以计算的，哪些评价问题具有真值。其次，需要提出统一的概念，公理、数学或图形表示和方法等。最后，统一的理论可以在不同领域应用。以任意一条检验标准来判断，过去所有的工作都没有实现。

在王和詹等人 [322, 339] 的研究中，对特定的人造物，如计算机组件（处理器 CPU）的评价仍然在采用经验性的方法。同一处理器，采用工业界的标准方法，不同的人获得的评价结果差异极大。假如评价学科已经建成，如斯克里文所说，像处理器这样相对药物来说更为简单的评价对象 [9]，不同的人获得的评价结果应该具有一致性，接近真值。不幸地的是，采用斯氏或邱氏评价学，甚至无法评价处理器、药物。实际上，不同领域仍然在采用不同的经验性方法（见第 2.4节），缺少理论支持。即使是第 四部分讨论的严谨的评价方法也有不同的假设和适用范围。

我认为，过去努力建立评价学科失败的根本原因有三个。

首先是缺少对测量、测试、评价和推理等基本活动的辨析和严谨的定义，经常将这些基础的求索活动混为一谈。

[9] 我在后文中讨论了不同评价问题的结构，相对药物来说，计算机的影响对象完全是可以复制的，因而这一类评价问题更加简单。

其次，现有评价领域的共识是将评价定义为“确定事物的优点（merit）、可量化价值（value）或内在价值（worth）的过程” [272, 274, 275, 276, 79, 96, 127, 300, 299]。斯克里文声称 [275]，正是科技共同体“价值中立”的教条（value-free doctrine）阻碍了评价学科的发展，即科学和工程界拒绝将“价值”一词引入其领域。“价值中立” 指的是一种主张在科学研究、社会分析或决策过程中，应排除主观价值，力求保持客观中立的原则。

然而，更深层次的原因是评价领域当前有关评价的定义未能揭示评价的本质。“优点（merit）、可量化价值（value）或内在价值（worth）”这些词在本质上是主观的——因人而异——这是导致失败的根本原因之一。

最后，尽管斯克里文设想了评价学科有一个核心主题 [274, 275, 276]。然而，在斯氏评价学里，评价的核心主题从未用一种技术语言明确定义。过去的工作，包括斯氏评价学，从未提出统一而适用于不同领域的评价概念、问题定义、问题结构分析、公理、数学符号、公式和方法。例如，由于没有这些统一的概念，斯氏评价学仍依赖于百科全书的方法定义了成百上千个概念或术语 [274]，正如其他评价学者所做的那样 [190]。

在本书里，我将评价对象的价值定义为它和其他对象之间的影响差异对利益相关方的影响，在此基础上，我将狭义评价学定义为价值量的定义、实现与赋值的科学。求索科学体系下的计量学和评价学为狭义评价学提供了严谨而统一的方法，这是斯氏评价学和邱氏评价学未见的。

邱氏评价学属于斯氏评价学的范畴，二者均属于狭义评价学，后者是本书阐述的评价学的一个分支。遗憾的是，斯氏评价学和邱氏评价学均采用的是经验方法，缺少严谨的理论支撑。在实验学领域，已经发展了不少严谨方法（见第 四部分），斯氏评价学和邱氏评价学受研究的视野和方法的限制，无法或者未能有效地采纳这些方法。

2.3.1 评价学 (Evaluatology) 和过去研究努力的不同

第一，学科位置不同。过去的评价研究属于本书定义的狭义评价学，常被视为教育、政策、管理、科研评价、项目评估或社会项目研究中的应用方法；即使在强调评价是跨学科工具的斯氏评价学里，评价也常常与判断价值、判断证据质量、制定标准和支持决策混合在一起。本书提出的求索科学则将评价与计量、测试和推理并列为四类基本求索方法。

在求索科学体系下，计量学回答如何定义、实现和赋值对象的量，包括结构量和行为量；测试学基于检验对测试对象的行为量进行定义、实现和赋值，并验证（证实或者证伪）对象的命题或模型是否有效；推理学回答如何从已有事实、前提和求索结果生成新命题或模型；评价学回答一个对象对另一个对象产生了什么影响。四者相互支撑但不能相互替代：评价常常需要测量和测试结果、检验结论和推理过程，但评价本身不是测量、不是测试，也不是一般意义上的推理。

就学科定位而言，斯克里文评价学把评价视为与逻辑、统计并列的“跨学科工具学科”（transdiscipline） [274, 275]，但他把验证、证据质量判断也并入评价（参见本章第 2.2.4节对评价“角色”的讨论），在求索科学体系下后者属于测试。斯克里文评价学没有明确区分和定义计量、测试、推理与评价，而在求索科学体系下，它们是四类相互独立又相互支撑的求索方法。

邱氏评价学从文献计量学出发，本质上是“科学评价”的学科化 [80]，邱氏评价学止步于传统文献计量与管理测量，甚至都没有将这些概念建立在传统的计量学基础上。邱氏评价学的学科位置在情报学与管理科学的框架下，而在求索科学体系下，评价学具有与计量学、测试学、推理学并列的基础学科地位。

已有评价研究的学科定位也有助于解释为什么过去的评价实践容易产生混乱。在一些领域，评价被简化为指标的测量或打分，甚至都缺少计量学的理论支持；在另一些领域，评价被视为与测试、排名、认证或证据判断等同。这样的做法在工程实践中大行其道，但缺少理论的严谨性和方法的科学性。

第二，研究问题不同。现有的评价研究和实践属于本书定义的狭义评价学，主要围绕五类问题展开：如何判断一个对象的价值，如何支持决策，如何改进项目或课程，如何规范评价实践，以及如何在特定领域构造评价标准。这些问题当然重要，但它们并没有直接回答评价学最基础的问题：如何揭示对象的影响？哪一类评价问题可以计算？哪些评价问题具有真值。在这一方法学框架下，价值判断不是评价的起点，而是进一步回答“评价对象与其他对象之间的影响差异对利益相关方的影响”这一衍生问题。

具体到斯克里文等人的代表性工作。斯克里文将评价定义为“确定事物的优点(merit)、内在价值（worth）或可量化价值（value）的过程”，并把这一相对主观的量化“打分”当作评价学的核心问题 [276, 274]，但他始终没有把“价值”还原为一个对象对另一个对象（利益相关方）的客观影响，因此回答的是“如何对对象作出价值判断”，而非“影响在什么条件下以什么样的机制产生”。

邱均平等则把核心问题定位为“科学评价”，即对国家竞争力、大学、学术期刊和科研成果等对象进行测量与排序 [80]，这一取向关心“如何测量并比较对象的优劣”。如何测量是计量学要解决的问题。本书系统地分析了不同对象的性质，在大部分情况下对象的量及其影响无法直接被测量，只能被推理。同样邱均平等学者没有追问评价结果在什么条件下具有真值。

第三，基础概念体系不同。过去评价研究积累了大量领域术语、领域特定的方法和实践规范。例如，斯克里文的评价辞典定义了大量评价术语 [274]，项目评价理论也形成了丰富的观点（见第 2.2.2节） [279, 6]。这些概念对具体实践很有帮助，但不同领域之间往往难以统一：教育评价、项目评价、科研评价、产品评价、政策评价、计算机系统评价和人工智能基准测试和评价使用不同的术语，有时术语相同却有不同的内涵。

本书从头建立精简和统一的基础概念体系，从对象、关系、事件、原因、影响、主体、命题和模型出发，进一步定义评价对象（evaluated object, EO）、影响对象（affected object, AO）、混杂对象（confounding objects, COs）、评价条件（evaluation condition, EC）、自包含系统（self-contained system, SCS）和对象因果模型（object-oriented causal model, OCM）等概念。

在概念体系上，斯克里文的《评价辞典》（Evaluation Thesaurus） [274] 按字母顺序收录了数百条评价术语，是一部评价辞典而非公理化体系 [274]，这种词典式罗列缺少统一的基础概念体系（如本书的对象、事件、EO、AO、EC），因而无法跨领域统一表达评价问题。邱均平等在“学科构建”部分讨论了评价学的概念范围、研究对象与理论基础，尝试给出学科定义 [80]，但其概念体系仍以“科学评价”为中心、由定性、定量与综合评价方法的术语构成，缺少能够跨越并且迁移到自然科学、工程与社会科学的精简而统一的基础概念体系。

从头定义统一而精简的概念体系的目标不是增加术语数量，而是建立可跨领域并可迁移的统一语言。无论评价对象是一个教育项目、一种药物、一项政策、一篇论文、一个处理器、一套人工智能模型，还是测试或者评价过程本身，这一统一语言都要求明确定义 EO、AO 和 COs 以及整体的评价模型。只要这些问题能够被统一表示和定义，不同领域的评价实践就不再只是经验性方法的并列，相反可以在同一个理论框架下进行比较和分析。

第四，研究范畴不同。过去方法多从具体评价场景出发：教育评价关心课程和学习结果，项目评价关心项目目标、实施过程和影响，科学评价关心科研成果、学术影响和指标体系，工程评价关心性能、可靠性和可比较性。本书则将评价学的基本问题概括为三类。第一类是评价问题，即在给定评价条件下揭示 EO 对 AO 的影响；第二类是逆评价问题，即从现象、结果或影响推测可能的原因；第三类是基于评价学的设计问题，即在理解影响机制的基础上设计更优的对象。这一框架使评价学不仅能解释已有评价实践，也能连接因果归因、机制分析、工程设计和基准构造。

对照来看，斯克里文提出的“评价的逻辑”（the logic of evaluation）包含确立准则、构造标准、测量比对、综合的评价判断这四步 [276]，这是相对“揭示影响”而言较为狭窄的价值判断领域，更不包含从现象反推原因的逆评价问题，也不含评价的对偶问题：设计。

邱均平等的研究框架按“理论—方法—实践”展开，方法层落在定性、定量与综合评价方法，实践层落在各类竞争力评价与学术排名 [80]，其框架服务于评估与排序这一单一目的，忽略了揭示原因和影响这一深刻的基础问题。

第五，过去的评价研究缺少严谨的方法，普遍忽略或者不重视严谨的实验方法。

斯氏评价学发展过因果探究方法，如“作案手法”（modus operandi）法与无目标评价（goal-free evaluation） [273, 271]，但对随机对照实验等方法持保留态度，也从未把实验设计（DoE）、随机对照实验（RCTs）、潜在结果理论（PO）、结构因果模型（SCM）与工具变量（IV）纳入视野。

邱氏评价学依赖的则是经验性的方法，称为定性、定量与综合评价方法（如指标体系、赋权与综合评分） [80]，完全没有讨论实验设计（DoE）、随机对照实验（RCTs）、潜在结果理论（PO）、结构因果模型（SCM）或工具变量（IV）等严格的评价方法，其方法学仍停留在测量与排序，而非揭示影响。

第六，过去的评价研究停滞于领域特定的（ad-hoc）方法，未能提出统一的通用方法。本书提出的评价学通用方法并不是随机对照实验（RCTs）、实验设计（DoE）、结构因果模型（SCM）、潜在结果理论（PO）或工具变量（IV）等方法之外的又一种补充方法，而是一个重新组织这些方法的理论框架。它以现实世界系统、自包含系统和对象因果模型这三个核心概念为支点，分别形成了基于现实世界系统的观察方法、基于自包含系统的实验方法和基于对象因果模型的建模方法，详见第 十二章。随机化、隔离和存在算子是基础的物理实验方法，而模型化的随机机制、虚拟隔离和虚拟存在算子则对应的基础虚拟实验方法。

本书并不把随机对照实验（RCTs）、实验设计（DoE）、潜在结果理论（PO）、结构因果模型（SCM）、结构方程模型（SEM）、工具变量（IV）等严谨的方法和准实验（quasi-experiments）、基准（benchmark）和标杆工程（benchmarking）等经验性方法视为彼此孤立的方法。相反，它们都可以被放入评价学中重新理解：这些方法都试图在不同条件下揭示某种对象、干预、因素、机制或系统配置的影响，只是它们依赖的假设、可控制条件、混杂处理方式、因果模型、采样方式和评价结果解释不同。因此，本书不仅讨论如何使用这些方法，也讨论这些方法自身在什么条件下能够产生有效评价结果。

因此，评价学通用方法的作用是把测量、测试、观察、实验、建模、推理和检验（属于测试）放入同一个求索模型（求索铁林寨模型）里，说明在不同的求索条件下如何揭示 EO 对 AO 的影响。其他方法可以被理解为评价学通用方法的特例或者补充。

已有的评价研究和实践已经充分说明评价活动具有跨领域的重要性，也为评价学的建立提供了重要思想资源。但本书提出的评价学并不是对已有评价概念、评价模型和领

域方法的汇编，也不是将“评价学”作为一个宽泛标签重新命名既有实践。它的核心目标在于：以“对象之间的影响”为核心研究问题，以 EO、AO、COs、EC、SCS 和 OCM 等概念为统一基础，以评价问题、逆评价问题和基于评价学的设计问题为研究框架，以随机化、隔离和存在算子为揭示影响的基础物理实验方法，并以现实世界系统、自包含系统和对象因果模型三个核心概念为支点提出评价学通用方法。正是在这个意义上，本书试图建立一门能够跨自然科学、工程科学、社会科学和智能科学使用的评价学。

2.4 不同领域特定的评价实践的总结

本小节简要介绍评价在不同领域特定的（ad-hoc）[10]实践，部分基于我们先前的工作 [339]。

2.4.1 商业领域

尽管英文 benchmarking 在不同领域广泛使用，例如计算机，人工智能和商业，其目标基本相同。但在商业领域，中文通常将 benchmarking 翻译为标杆工程。坎普（Camp）[41]”将标杆管理（Benchmarking）定义为“寻找那些能引领公司卓越绩效（superior performance）的最佳实践（best practices ）”。标杆管理包含两个主要步骤：（1）基于行业最佳实践建立运营目标（operation targets）；（2）“一个积极、主动、结构化的过程，旨在改变运营，最终实现卓越绩效和竞争优势”。安德森等人 [9] 的研究将标杆管理的本质总结为“知识的追求以及向他人学习”。

从评价学（Evaluatology）的角度来看，过程、个体、政策或业务中产生的中间对象均可被评价。所谓的最佳实践也是一个评价对象。运营的目标可以看作是评价对象期望的影响。“一个积极、主动、结构化的过程”本质上是一个揭示不同评价对象（例如政策）的影响并进行试验的过程，从而“改变运营，最终实现卓越绩效和竞争优势”。

2.4.2 金融领域

在金融和教育领域，指数（index）被广泛用作基准（benchmark），用于评价所研究的个人或系统的整体表现（performance）。这里指数本质上一种比较基准。选定一组个人或系统（参考对象），这些指数通过计算它们的加权平均值得出，最后作为比较的基准。

例如，在金融领域，股票市场指数被用作评估股票市场表现的基准。这些指数通过计算一组代表性股票的加权平均值得出 [82]。一些广为人知的股票市场指数包括道琼斯工业平均指数（Dow Jones Industrial Average）、标准普尔 500 指数（S&P 500）、纳斯达克综合指数（NASDAQ Composite）和上证综合指数（Shanghai Stock Exchange Composite Index）。

不同的指数采用不同的计算方法。最常见的方法是加权平均法，该方法根据成分股价格的加权平均值确定指数值。另一种方法是几何平均法，该方法计算股票价格的几何平均值，并使用基期价格（base period price）进行调整。通常，股票市场指数在每个交易日收盘时发布。一些指数提供商提供实时指数数据，使投资者能够了解最新的市场情况。

[10]ad-hoc 在中文中常见的翻译有“临时的”“特设的”“专门的”“特定目的的”以及“即席的”等，本书将它翻译为“领域特定的”。因为不少方法已经在特定领域用了很久，翻译为临时的，不妥。

布伦特基准（Brent benchmark）用于确定布伦特原油的价格 [268]。布伦特原油是一种来自北海地区油田生产的轻质低硫原油，在这里，它本质是一种参考对象。由于其供应相对稳定且质量高，布伦特原油已成为国际石油市场的重要基准。世界各地的交易员、投资者和行业参与者参考布伦特基准来跟踪和评估布伦特原油的价格。

在金融领域，指数或基准本质上是我们进行比较的参考对象，正如我们在第 二十三章中所讨论的。指数本质上是一种比较基准。

2.4.3 社会科学领域

根据罗斯等人的研究 [242]，社会科学领域的评价研究工作最早可追溯至托马斯 · 霍布斯（Thomas Hobbes）及其同时代学者的工作，他们试图“在社会科学学科中运用数值测量（numerical measures）来评估社会状况（social conditions），并识别死亡率、发病率和社会解体的原因”。

根据罗斯等人 [242] 的定义，项目评价是使用社会科学方法系统评估旨在“改善社会条件以及我们个人和集体福祉”的项目的过程，旨在为利益相关者提供答案。他们 [242] 总结出评价问题和方法的五个领域，这些领域表现出强烈的相互作用：（1）项目的需求，（2）项目理论和设计，（3）项目过程，（4）项目效果（impact），（5）项目效率（efficiency）。

从评价学（Evaluatology）的角度来看，项目评价的本质是揭示社会项目对社会状况、个人和集体福祉的影响。然而，从评价学的角度来看，罗斯等人提出的评价问题的五个领域应通过不同的求索方式来解决，而非单一的评价。项目的必要性可通过测量来量化；项目理论的正确性需在实施前测试或者验证；项目过程或其中间过程产生的的对象可通过评价来揭示其影响。

2.4.4 计算机科学领域

在计算机科学领域，存在不同的观点和视角。例如，轩尼诗（Hennessy）等人强调了 [117] 基准（benchmark）的重要性，并将其定义为“专门为测量计算机性能 [11] 而选择的程序”。另一方面，约翰 (John) 等人 [143] 编写了一本关于性能评价和基准测试或者评价（Benchmarking）的书籍，但并未为这些概念提供正式定义。在计算机或者人工智能领域，我将 Benchmarking 翻译为基准测试或者基准评价，其具体的含义取决于不同的上下文。总而言之，当有清晰定义的检验基准（Test Oracle）时，它指的是基准测试；而用于揭示影响时，它指的是基准评价。

科内夫（Kounev）等人 [160] 提出了基准（Benchmark）的正式定义，即“基于性能、可靠性或安全等特定特征，评价（evaluating）和比较（comparing）系统或组件的方法及工具”。美国计算机协会测量评价专业组（ACM SIGMETRICS）[39, 158] [12] 将性能评价定义为“基于一系列有序且明确定义的分析和定义步骤，生成数据以显示计算机系统组件的频率和执行时间”。

SPEC CPU 基准（统常简称为 SPEC CPU） [284] 被广泛认为是用于 CPU 性能评价的最著名的基准套件。在其发展历史中，已经发布了七个版本的 SPEC CPU 基准套

[11] 英文 performance 一词，在不同学科里中文翻译不同，计算机领域通常翻译为性能，而在社会科学领域，通常翻译为绩效。

[12] 英文全称为 Special Interest Group on Measurement Evaluation

件，最新版本为 SPEC CPU2026 [286]。SPEC CPU 工作负载（workload）涵盖了范围广泛的计算密集型任务。

为确保结果的可信度（credibility of the results），总体指标计算为各个负载下参考对象（参考机）/评价对象（测试机）运行时间的比率的几何平均值。每个比率基于三次运行的中值（median）执行时间或两次运行中较慢的一次 [285]。

杰克 · 唐加拉（Jack Dongarra）等人 [75] 提出了用于评价高性能计算（HPC）系统的 LINPACK 基准。LINPACK 基准旨在求解 n 阶稠密线性方程组，表示为方程 $Ax = b$。它起源于 1970 年代发展的 LINPACK 软件包。

从评价学（Evaluatology）的角度来看，计算机学科中的基准本质是混杂对象，我将在本书第 8.2节给出正式的定义。

2.4.5 人工智能领域

如图 2.1 [13], ImageNet 是计算机视觉领域的重要基准，包含 21,841 个不同类别的 14,197,122 张高分辨率图像，每张图像都被人工标注，通常称为 ImageNet-21K [71]。这些类别（category）涵盖了广泛的物体、动物和场景。

ILSVRC 是 ImageNet 大规模视觉识别竞赛的缩写，它是年度计算机视觉竞赛，专注于 ImageNet-21K 的一个子集 ImageNet-1K [260]。其旨在测试深度学习模型在图像分类（image classification）和物体检测（object detection）等任务上的性能。ILSVRC 提供这些特定任务的配置和测试准则。

ImageNet-1K 主要用于图像分类任务，包含 1,281,167 张训练图像、50,000 张验证（validation）图像和 100,000 张测试（test）图像。ILSVRC 中常用的测试指标包括 Top-1 准确度（accuracy），它度量预测类别（predicted category）与图像真实类别（true category）之间的匹配程度，以及 Top-5 准确度，它表示图像真实类别是否包含在模型预测的前五个类别中。

评价学精准地区分了评价和测试这两个概念，运行 ImageNet 可以测试不同算法的准确度，但如果要揭示不同算法的准确度差异的原因，则是一个评价的过程。

从评价学的角度来看，人工智能领域的基准也是混杂对象。

2.4.6 医学领域

医学领域的评价可追溯至早期医学时代，尽管没有书面记录。严格的现代医学评价方法学（Methodology）和体系（system）早在 1938 年就已建立 [48]。临床实验（Clinical trials）是医学评价的主要和广泛认可的方法，其历史可追溯至 250 多年前。临床实验被定义为评价医疗干预（medical interventions）对人类受试者（human subjects）潜在影响的实验设计 [141]。

1948 年，奥斯汀 · 布拉德福德 · 希尔（Austin Bradford Hill）作为首席统计师，设计了一项极为成功的随机对照实验，证明了链霉素（streptomycin）——最早问世的抗生素之一——对结核病有效。他将写有患者名字的相同大小和材质的纸片装入信封中来随机抽取干预组患者和对照组患者。这一分配方案代替了之前使用的存在偏差（bias）的历史对照（*historical controls*）或顺序分配（*sequential allocation*）方案。这项研究是医学史上的一个里程碑，不仅让医生们认识到了“神奇药物”，还巩固了随机对照实验的声誉，这种方法很快便成为流行病学临床研究的标准 [224]。

[13] 罗纯杰贡献了本图，他是我们之前工作 [339] 的作者之一。

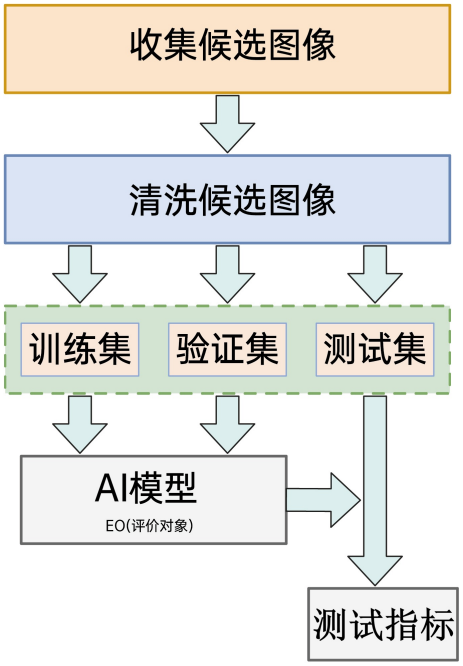

图 2.1: 基于 ImageNet 的评价工作过程。

目前，基于实验设计的临床实验可分为多种类型，包括随机实验（randomized trials）、双盲实验（double-blind trials）、前瞻性实验（prospective trials）和回顾性实验（retrospective trials） [278]。

如图 2.2所示，随机对照实验（RCTs）被视为医学评价的黄金标准，拥有严格可靠的理论框架 [198]。由于方法学的限制，它只能获得干预在影响对象群体（群体成员之间通常存在显著的差异）中的平均影响，无法获得干预对影响对象个体的影响。即使不是不可能，RCT 很难应用于其他学科，例如物理学、化学或生物学领域以解释许多现象的原因 (cause) [14]。我在第 三章定义了什么是现象。RCTs 本质上是一种蛮力法，其高昂的时间和财务成本限制了其应用。

为了弥补 RCTs 的不足，新兴的临床评价方法已被提出 [134, 235]，如真实世界数据评估（Real-World Data assessment）和数字临床实验（digital clinical trials）。这些新型医学评价方法仍处于早期阶段，理论基础存在明显缺陷，例如缺乏严谨性（rigor）和可靠性（reliability）。

2.4.7　心理学领域

在心理学领域，社会人格心理学家（social and personality psychologists）常依赖量表（Scale），例如心理量表、测验或问卷来评估心理测量变量，这些变量包括态度（attitudes）、特质（traits）、自我概念（self-concept）、自我评价（self-evaluation）、信念（beliefs）、能

[14]这个评论受到王磊博士的启发，2025 年的某一天的中饭后，我和他在中科院物理所园区遛弯，我们聊天时，他清晰地表达了这个观点。也就是说 RCTs 不可能用于物理、化学和生物中众多现象产生原因的解释，而评价学 (Evaluatology) 能用于揭示现象的原因。

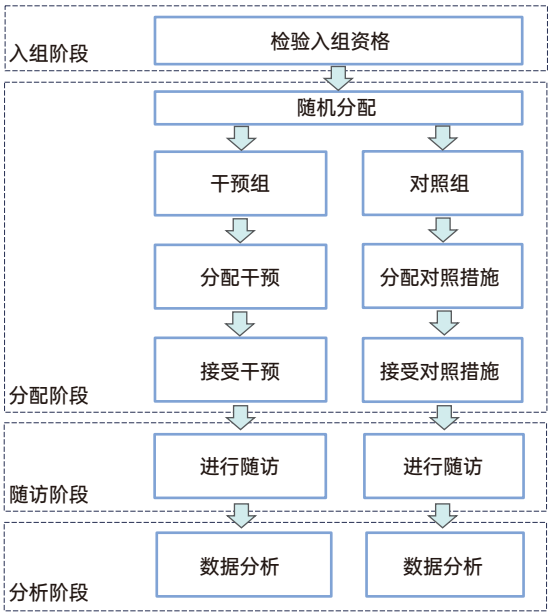

图 2.2: 随机对照实验（RCTs）评价过程。 [196]

力（abilities）、动机（motivations）、目标（goals）、社会感知（social perceptions）等 [93]。这些变量本质上是隐藏变量，无法被直接测量，只能被推测。而且与个体的特性密切相关。

我认为应该基于本书阐述的求索的科学，构建全新的理论和方法。

2.5　不同学科中的因果定义

由于我将评价定义在原因和影响等基础概念之上，本小节总结了不同学科中对因（cause）和果（effect）的定义，贡献者是范帆达。

哲学中因果概念的起源与演进

因果（causation）这一概念最早由亚里士多德（Aristotle）系统地提出。亚里斯多德理解因果并非为一种单一的关系（relation），而是一组用于回答“为什么”的解释性原理（explanatory principle）。在其关于“四因说（Four Causes）” [118] 的理论中——质料因（material）、形式因（formal）、动力因（efficient）与目的因（final）——因果关系被广泛理解为对“某个实体（entity）为何存在”或“某个事件（event）为何发生”这一问题的不同层面的解答。重要的是，亚里士多德将因果关系主要视为一种解释性结构（explanatory structure），而非严格意义上的基于事件的关系（event-based relation），这一观点在之后的数个世纪中深刻影响了科学与哲学思想的发展。

大卫 · 休谟（David Hume）在近代早期哲学中引发了一场决定性的转变，他质疑因果关系是否包含任何可直接观察到的（observable）或逻辑上必然的联系（logically necessary connection） [130]。在休谟看来，经验所揭示的仅仅是事件之间的不变的联结（constant

conjunction）与时间上的先后顺序（temporal succession），而所谓将原因与结果联系起来的必然性，并非源自经验观察本身，而是来自人类习惯性的期待。这一怀疑论分析从根本上削弱了形而上学中关于因果必然性的观念，并由此奠定了一个长期悬而未决的问题，在放弃了形而上学必然性（metaphysical necessity）的观念后，因果知识如何超越简单的规律而具有存在的合理性（justified beyond mere regularity）。

针对休谟提出的挑战，后继哲学家试图以不诉诸形而上学必然性（metaphysical necessity）的方式，对因果关系加以重新概念化，从而保留其解释功能。伊曼努尔 · 康德（Immanuel Kant ）认为，因果性并非源自经验，而是一种先验的必然范畴（a necessary a priori category），正是通过这一范畴，经验本身才得以可能 [148]。尽管康德重新确立了因果关系在经验知识中的不可或缺性，但他的理论也将因果性从世界的结构（the structure of the world）转移到了人类认知结构（the structure of human cognition）之中。

二十世纪的哲学进一步推动了因果理论的多样化发展。规律论（regularity-based）[188] 的观点在休谟思想的基础上加以扩展，将因果关系等同于符合法则的关联模式。与之相对，反事实理论（counterfactual theories） [175]——最具代表性的是大卫 · 刘易斯的工作——从“差异制造（difference-making）”的角度界定因果关系，即：如果在相关的反事实条件下缺少某一因素，结果便不会发生，那么该因素即构成原因。与此同时，机制论（mechanistic theories） [186, 23] 强调因果过程（process）、结构（structure）以及有组织的相互作用（organized interactions），将关注焦点从事件之间的抽象关系（abstract relations），转向影响是通过哪些具体路径（pathway）产生的。

哲学中的因果理论并未收敛为单一的定义，而是形成了一组从不同侧面刻画因果关系的观点谱系。经典理论侧重于因果的解释功能，休谟传统强调经验规律性及其所面临的认识论限制，反事实理论关注因果所体现的差异制造（difference-making），而机制论则突出有组织的过程与结构。这种多元性反映出一个事实：即便在哲学内部，因果性也并非一个单一、同质的概念，而是一个用于理解解释（explanation）、依赖关系（dependence）与变化（change）的多维框架。

其他学科中的因果定义与关注重点

在哲学之外，因果概念在多个学科中被系统性地重新界定。这些学科并未试图回答“因果本身是什么”这一形而上学问题，而是围绕各自的研究目标与实践需求，对因果关系的不同侧面加以取舍和强调。由此，因果逐渐从一种以解释（explanation）为中心的概念，演化为一组面向具体任务的操作性构造（operational constructs）。

在自然科学中，尤其是在物理学、化学与生物学领域，因果通常被理解为由稳定机制和自然规律所支配的生成关系（generative relation）。在这一语境下，因果并不主要以反事实术语（counterfactual terms）来表达，诸如“如果没有某一因素结果是否仍会发生”，而是体现为在给定初始条件和约束下，系统如何必然或概率性地演化出特定状态。自然科学中的因果关注重点包括现象背后的机制、因果关系在不同情境下的稳定性与可复现性，以及解释是否能够嵌入统一的理论体系之中 [264]. 因此，该领域中的因果主要服务于解释（explanation）与理解（understanding），而非直接服务于决策（decision-making）或干预（intervention）。

在医学与流行病学中，因果概念则呈现出显著的行动导向特征。此类研究并不要求因果关系完整揭示底层生物机制，而是关注在现实条件下实施某种干预是否会系统性地改变结果的风险或分布。因果在这里被操作化为干预状态与非干预状态之间的对比，其关注重点包括反事实对照、影响大小（如风险差与风险比）以及不确定性的控制，例如随

机化、偏差校正（bias adjustment）与证据等级（evidence hierarchies） [245, 119]. 因此，医学中的因果本质上是一种以决策为导向的因果。

在工程学与控制论中，因果被明确理解为可操纵并产生可预测系统响应的关系。如果某一输入、参数或结构调整能够稳定地改变系统输出或性能指标，则该关系在工程意义上被视为因果关系。该领域关注因果的重点包括变量的可控性、系统对变化的响应一致性，以及在动态与闭环系统中的稳定性与反馈特性 [15]. 工程因果的核心目标在于控制、设计与优化，而非对因果关系进行形而上学解释。

统计学习与人工智能语境下的因果研究通常以结构化形式语言来刻画“因果问题本身”，而不仅仅是以相关性为基础的统计推断。以结构因果模型（structural causal model）[229] 与因果图为代表的框架，将因果关系表述为结构方程与有向图，并在此基础上区分并统一三类推理任务：基于观察数据的关联（association）、面向干预的影响推断（intervention, $do(\cdot)$） [224] 以及反事实推理（counterfactual reasoning） [220]. 在这一框架中，关键问题首先是识别（identification）：在给定因果结构（causal structure）假设与观察数据分布（observed data distribution）的条件下，目标因果量（target causal quantity，即研究者希望从数据中推断的因果影响，例如干预分布 $P(Y|\mathrm{do}(X))$ 或平均处理影响 ATE 等因果估计量）是否可以被唯一确定；随后才是估计（estimation）：在有限样本下如何近似地获得该因果量的数值。与之相辅相成，面向机器学习的因果推断研究进一步强调从数据中学习因果结构与稳定关系，并关注在分布变化或环境迁移条件下的可泛化性（例如不变性原则与因果发现） [229].

在时间序列分析（time series analysis）与计量经济学（econometrics）中，因果关系通常被操作化为一种基于预测能力的统计关系，其中最具代表性的定义是格兰杰因果（Granger Causality） [103]。其核心思想是：利用变量 Y 自身历史信息预测 Y 未来的值，如果引入变量 X 的滞后信息（lagged values，相对于 Y 的历史信息）能够显著提升对 Y 未来的预测准确度（例如降低预测误差），则认为 X 是 Y 的“格兰杰原因”。这一框架以时间先后性为基本前提，通过滞后项保证“因先于果”，并体现出因果关系的方向性与非对称性。与简单相关性不同，格兰杰因果显式利用时间结构，并通过回归建模与假设检验（如 F 检验 [110]）判断变量是否具有额外的预测贡献。然而，该定义并不诉诸哲学意义上的必然联系，而是建立在可检验的预测性能改进之上，本质上刻画的是一种统计上的可预测依赖关系。因此，其结果可能受到遗漏变量（omitted variables）、非平稳性（non-stationarity）或结构突变（structural breaks）的影响，通常需要结合领域知识或结构化因果方法进行进一步分析。

2.6 本章总结

本章概述了评价学科的现状，深入探讨了其历史根源和理论基础。它将评价视为一种古老的实践但欠发展的学科进行了审视，并总结了不同领域发展的各种评价概念、理论、领域特定的实践。本章也分析了过去未能成功建立评价学科的原因。由于本书将评价的概念定义在原因和影响这两个基础概念之上，本章也总结了不同学科对因果的定义。