总要告一段落：评价学中文版 1.0 补充序言

詹剑锋博士，2026 年 8 月 9 日

立秋了，我却有点伤感：我从南方移植的桂花树刚刚发出两片新芽；我花费了大量心血种植的鸡蛋花、蒜香藤、炮仗花，倒是很翠绿，但开花却遥遥无期。可夏天已经过了。

两个多月前，我就为评价学中文版写完了序言，可没有想到我们又多花了这么久才完稿。感觉有点疲倦。

我其实在挣扎，要不要给这本书改个名字，另一个备选的名字是《求索科学：智能生命如何认识和改造世界》——这其实是本书的主要目的。我将求索科学总结为四个学科的相互支撑体系：计量学（Metrology）、测试学（Testology）、推理学（Inferology）和评价学（Evaluatology）。我很幸运地命名了三个学科：测试学、推理学和评价学（Evaluatology），创造了三个英文单词：Testology、Inferology 和 Evaluatology。

这四个学科都建立在一个精简和统一的基础概念体系之上。其中最重要的概念是对象，我将它定义为拥有结构、属性、机制和行为的一类实体。我将世界定义为所有对象及其关系的总和。

我和王磊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。我重新定义了溯源等概念和方法，它们可以基于评价学重新构造。

我和王磊提出了统一的测试学，它可以应用于软件和硬件测试、假说检验、理论验证以及其他领域。例如最新的高能物理发现——胶球——我觉得就可以用测试学来检验它，我会单独写一篇文章。我将测试学定义为检验和测试的科学：检验是唯一能验证[1]命题和模型有效性的方法，测试基于检验，定义、实现和赋值对象的行为量。测试学是求索科学体系下的计量学在测试领域的应用。

范帆达和我提出了推理学。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——视推理为基础求索方法之一。我们将推理学定义为生成新命题和新模型的科学。

我提出了评价学，将其定义为揭示原因和影响的科学。传统的评价学相当于本书中阐述的狭义评价学。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的定义、实现和赋值的科学。狭义评价学是计量学在价值领域的应用。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）。

[1]本书中，验证和检验的含义相同。

最重要的是这四个学科相互支撑：计量学负责定义量；测试学负责检验命题和模型的有效性；推理学负责生成新命题和模型；评价学负责揭示原因和影响。评价学可以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；测试学可以检验评价结果的有效性；推理学可以为计量学构建更严谨的公理系统；计量学可以定义测试学的量等。

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评价对象的量赋值；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。

在书中，我更关注世界的基本机制：包括真随机性（微观世界的机制）、因果律（被动对象世界的主导机制）、内在机制（生命世界的主导机制）以及自由意志（智能生命的根本机制）。我认为被动对象世界的随机是一种表面随机性，它实际上确定的，受因果律支配。基于此，我在对象概念的基础上定义了因果、随机和确定等概念，并提出了因果律公理，它是本书构建的求索科学体系的基础。世界的变化是真随机性、因果律、内在机制以及自由意志这些机制共同作用的结果。

我也提出了几个有意思的猜想，首先是随机性、独立性和计算复杂性关系猜想，我认为它们是同一个问题的不同方面，是同一个衍生对象不同的量，求索结果的空间具有相似性和等价性。智能生命因为自由意志而精彩，我提出了自由意志第一猜想和第二猜想，为自由意志的建模建立了基础。

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。

存在算子是评价学的精华，它不仅可以用于实验，而且我甚至觉得它也可以用来识别诈骗、作弊或者魔术，我成功地实验了很多次。

我还提出了对象因果模型，它不仅有独创性，而且我很乐观地相信它能统一所有的评价方法。

我还用故乡的山命名了我提出的求索模型。我故乡有一座山，名叫铁林寨。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。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。但只有通过了检验，命题或模型才能成为对象的真理或者事实。

我要感谢我所在的团队：王磊高级工程师、高婉铃副研究员、王晨曦、范帆达博士、康国新博士，杨郑鑫博士、罗纯杰副研究员、李鸿霄、田郑书媛、周昱彤、苏宇晨、彭睿思、陈思敏、敖翔、王晓睿、杨宇轩、彭路洲、兰川鑫、刘可、尚源峰等。我们经常一起讨论，碰撞出思想的火花。没有他们和她们，这个工作很难这么顺利地开展。在本书中，我尽量忠实地记录这一过程。我每次提出新想法，甚至在不成熟的时候，我的同事和学生就积极地尝试用它们来解决不同的问题。我特别开心，评价学已经在不少领域得到了应用，例如开源芯片 RISC-V、大模型和时序预测系统领域。这个版本也收录了两篇发表在人工智能顶级会议 ICML 上的工作。

很遗憾，这本书没有完成存在算子演算理论。我有些基本想法，不过，我准备把它留给博士生来做。留在下一版本吧，毕竟总要告一段落。

评价学中文版 1.0 序言

詹剑锋博士，2026 年 5 月 27 日

2025 年 11 月，我完成了评价学英文版。原以为过完春节中文版就会顺利面世。我常常在节日期间工作，一是因为热爱研究，另外的原因就是节日期间自由，没有人打扰。没想到一拖就拖到现在，这有几个方面的原因。

在写作期间，我又思考了一些基本的问题，例如概率是不是一个量？什么是事件？事件和对象有什么关系？什么又是关系？我领会到了对基础概念的定义是基础理论突破的关键。这个版本解决了这个问题，我将事件定义为对象或者对象关系变化的结果。事件可以视为衍生的对象，概率是事件这个衍生对象的量。关系则是从更高一层实体的属性或者从主体求索的角度来描述实体之间的相对位置。

我也努力为评价学增加一个哲学基础。我花了两三个月时间阅读了西方基础的哲学体系，尤其是罗素的西方哲学史。我很快有了思考和收获，苦于时间，我没有完成这个任务。相对英文版，中文版重新总结了贡献。从这一点，读者应该能了解到哲学的影响。我在总结本书贡献的时候首先提到了三点：描述世界的基础概念体系，认识世界的基础方法，解释世界变化的机制，这些都是哲学的基础问题。不过，不同于以往的哲学家，我认为哲学应该成为科学与人文的基础，因此我决定单独写一本书：《作为科学与人文基础的哲学》。

我在写作过程中，发现了英文版 [340] 的一个章节有重大的错误，这当然是我的责任。这让我意识到了一个问题，专著或者重大理论的构建最好由一两个人来完成。我以前读西方学者的大部头著作，发现只有一两个作者，深深震惊。我现在明白了，这是最好的方式。我现在也相信自己具有了这个能力。

中文版是英文版 [340] 的重大升级。我还是决定将它命名为评价学 1.0 版本。它实际上包含了四大块内容：计量学（Metrology）、测试学（Testology）、推理学（Inferology）和评价学（Evaluatology），后面三个单词都是我自己创造出来的。我也重新定义了计量学，将其定义为量的定义、实现和赋值的科学。这四个学科互相支撑，我将它统称为求索的科学。不过，评价学还是占据了本书的主要篇幅，因此，本书的标题仍然保留为评价学：解释原因和影响的科学。跟英文版不同，中文版标题，我增加了“原因”。

我未来会重新写一本《求索的科学》的书。不过，除了求索，我还会增加一部分内容：人类是如何创造的。

评价学英文版 1.0 前言：学术江湖任我行–评价学诞生记

詹剑锋博士撰写，2025 年 11 月 28 日

谨以此文献给无数为梦想挣扎的年轻人和曾经的年轻人，大家终将得解放。

早上我从梦中醒来，非常懊悔。我已经在梦里写了两个小时的前言。要是早起两个小时，就完成了这个艰巨而幸福的任务。

2009 年，我接到一个艰巨的任务，来自当时的领导孙凝晖研究员。孙凝晖研究员 2019 年当选中国工程院院士。他让我负责撰写新型计算基础设施的报告。开始在内部把这种新型计算叫高吞吐计算，后来改名为高通量计算。当时我们所在的高性能计算机中心跃跃欲试，准备向新型数据中心转型。孙凝晖研究员把我介绍给了著名的普林斯顿大学教授李凯。请他指导我撰写这份报告。同时期，加入讨论和撰写报告的有中国科学院计算技术研究所（中科院计算所）和中国科学院研究生院（现在的中国科学院大学）近二十位不同方向的研究人员，详细名单可以从图 1的致谢里找到。当中不少人已经成了知名专家和教授。

李凯是个传奇人物。他 1980 年代毕业于中科院计算所，不仅后来成为美国工程院院士，而且也是成功的创业家。他创办的 DataDomain 公司被存储巨头 EMC 以几十亿美金收购。

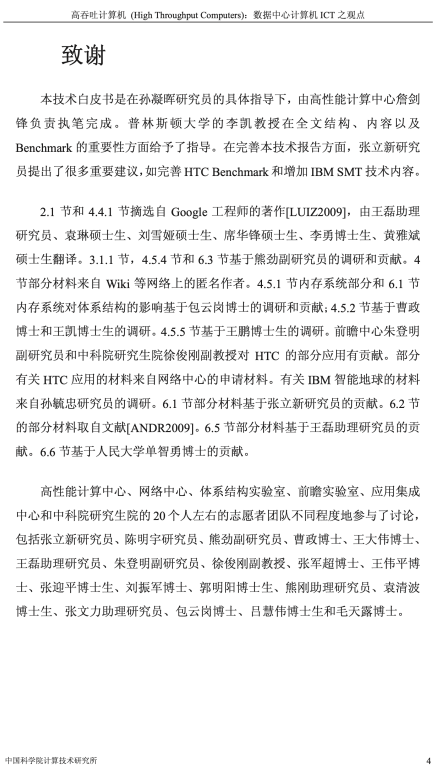
李凯教授给我的两点印象非常深刻。一是谈到创新，他喜欢算一算最新技术的成本。最新的内存多少钱？处理器多少钱？他了如指掌。这让我很惊讶。在国内，算成本似乎显示不了水平，离科学家很远。李凯教授在做研究之前，还喜欢先用最先进的工业界技术搭建一个实验系统，看看“油水”有多大。油水是他的原话。

还有一点，李凯教授至少做成功了两个有影响力的评价基准 (benchmark)。一个是评价处理器的 PARSEC，芯片巨头英特尔资助。还有一个是和传奇华裔女性教授李飞飞合作的 ImageNet。学术界和工业界普遍认为后者是促进人工智能繁荣的力量之一。

我钦佩李凯的影响力。回过头看。和李凯的交流恐怕是我创建评价学理论的开端。我将评价学定义为揭示万物影响的科学。在这里，我也用同样的方法，回顾影响我和这本书的人和事。

在这十几年岁月里，有两件事值得回顾，一件充满着幸运，另外一件则没有那么幸运。放在一起，就像评价学里面提到的对照组。

第一件是我的第一个有影响力的评价基准工作：大数据系统基准 BigDataBench。2013 年，我当时的博士生王磊和我以及别的合作者一起工作。这篇论文的撰写我们只花了两个星期时间，不过应该熬了两个夜。最终这篇文章被计算机体系结构会议 HPCA 2014 工



高吞吐计算机 (High Throughput Computers)：数据中心计算机 ICT 之观点

致谢

本技术白皮书是在孙凝晖研究员的具体指导下，由高性能计算中心詹剑锋负责执笔完成。普林斯顿大学的李凯教授在全文结构、内容以及 Benchmark 的重要性方面给予了指导。在完善本技术报告方面，张立新研究员提出了很多重要建议，如完善 HTC Benchmark 和增加 IBM SMT 技术内容。

2.1 节和 4.4.1 节摘选自 Google 工程师的著作[LUIZ2009]，由王磊助理研究员、袁琳硕士生、刘雪娅硕士生、席华锋硕士生、李勇博士生、黄雅斌硕士生翻译。3.1.1 节，4.5.4 节和 6.3 节基于熊劲副研究员的调研和贡献。4 节部分材料来自 Wiki 等网络上的匿名作者。4.5.1 节内存系统部分和 6.1 节内存系统对体系结构的影响基于包云岗博士的调研和贡献；4.5.2 节基于曹政博士和王凯博士生的调研。4.5.5 节基于王鹏博士生的调研。前瞻中心朱登明副研究员和中科院研究生院徐俊刚副教授对 HTC 的部分应用有贡献。部分有关 HTC 应用的材料来自网络中心的申请材料。有关 IBM 智能地球的材料来自孙毓忠研究员的调研。6.1 节部分材料基于张立新研究员的贡献。6.2 节的部分材料取自文献[ANDR2009]。6.5 节部分材料基于王磊助理研究员的贡献。6.6 节基于人民大学单智勇博士的贡献。

高性能计算中心、网络中心、体系结构实验室、前瞻实验室、应用集成中心和中科院研究生院的 20 个人左右的志愿者团队不同程度地参与了讨论，包括张立新研究员、陈明宇研究员、熊劲副研究员、曹政博士、王大伟博士、王磊助理研究员、朱登明副研究员、徐俊刚副教授、张军超博士、王伟平博士、张迎平博士生、刘振军博士、郭明阳博士生、熊刚助理研究员、袁清波博士生、张文力助理研究员、包云岗博士、吕慧伟博士生和毛天露博士。

中国科学院计算技术研究所 4

图 1: 2009 年我牵头撰写的技术报告。致谢部分提到的不少人已成为著名专家和教授。

业版录用。我当时在美国旧金山机场的一架即将起飞的飞机上，颤悠悠地提交论文。飞机上网络信号很差，点完提交按钮，五分钟后，飞机起飞，回北京。

我们非常幸运。文章得到高分。工业版主席来自芯片巨头英特尔，给了我们很大的鼓励和支持。一开始，我们就觉得这篇文章会有影响力。后来，我们团队的贾禛博士去李凯教授那做博士后。他告诉我，普林斯顿的一个著名计算机教授让李凯推荐大数据基准。李凯教授毫不犹豫地推荐了 BigDataBench。

差不多相同的时间，我另外一名博士生陆钢向 ASPLOS、SOSP 和 OSDI 等顶级会议提交了一篇操作系统论文。我们一共花了六年时间，投了六次稿。前几次，第一轮分数很高，第二轮就被人干掉了。我仍记得有一年投 ASPLOS，一位审稿人甚至在第一轮给出了” 强烈接收” 的评价——这在系统会议中相当罕见。最终，我决定不发表这篇文章。

第二个工作是 AIBench，它是我们有关人工智能系统的评价基准。当时的主要贡献者是高婉铃博士生。她非常聪明和勤奋。论文提交给 HPCA 2020 的工业版。婉铃在提交论文之前，三天三夜几乎没有睡觉。

我对文章的录用充满着信心，因为我们更有经验了，英文写作水平也提高了很多。期待着论文有更大的影响力。评审结果也如我所料，收获了审稿人非常高的评价。五个审稿人：四个接收和一个弱拒绝。

可是我注意到一些异常。给弱拒绝的评审人提到，“人工智能系统评价基准有一个 MLPerf 就够了，没有必要有第二个工作”。MLPerf 聚集了几乎所有的美国巨头公司和顶尖大学，例如斯坦福、谷歌等，是我们的竞争者。MLPerf 当时还没有发表论文，而且也举步维艰，牵头的科学家我也认识，他还跟我抱怨得到的支持很少。

我决定给工业版主席写信，他同样来自英特尔。我感谢他付出的努力，并且提醒他：AIBench 和 MLPerf 是竞争者。我也强调同一个领域有两个独立的竞争性工作，对社区有利。这不是我的一厢情愿。当时美国著名的信息技术媒体记者 Jack Clark在评论 [59] 中将我们的 AIBench 和谷歌主导的MLPerf标准之争上升到中美科技竞争的高度。Jack Clark后来成为OpenAI Strategy and Communications Director和斯坦福AI Index的共同发起人。Jack Clark的评论见图 2.


2025/12/1 15:50 Import AI 197: Facebook trains cyberpunk AI; Chinese companies unite behind 'AIBench' evaluation system; how Cloudflare us...

IMPORT AI

ABOUT WORK

Import AI 197: Facebook trains cyberpunk AI; Chinese companies unite behind 'AIBench' evaluation system; how Cloudflare uses AI

by Jack Clark

Want to analyze real-world AI performance? Use AIBench instead of MLPerf, says AIBench developers:

...Chinese universities and 17 companies come together to develop AI measurement approach...

A consortium of Chinese universities along with seventeen companies – including Alibaba, Tencent, Baidu, and ByteDance – have developed AIBench, an AI benchmarking suite meant to compete with MLPerf, an AI benchmarking suite predominantly developed by American universities and companies. AIBench is interesting because it proposes ways to do fine-grained analysis of a given AI application, which could help developers make their software more efficient. It's also interesting because of the sheer number of major Chinese companies involved, and in its explicit positioning as an alternative to MLPerf.

End-to-end application benchmarks: AIBench is meant to test tasks in an end-to-end way, covering both the AI and non-AI components. Some of the tasks it tests against include: recommendation tasks, 3D face recognition, face embedding (turning faces into features), video prediction, image compression, speech recognition, and more. This means AIBench can measure various real-world metrics, like the latency of a given task that reflects the time it takes to execute the AI part, as well as the surrounding infrastructure software services, and so on.

Fine-grained analysis: AIBench will help researchers figure out what proportion of time their systems spend doing different things while executing a program, helping them figure out, for instance, how much time is spent doing data arrangement for a task versus running convolution operations, or batch normalization, and so on.

The politics of measurement: It's no coincidence that most of AIBench's backers are Chinese and most of MLPerf's backers are American – measurement is intimately tied to the development of standards, and standards are one of the (extraordinarily dull) venues where US and Chinese entities are currently jockeying for influence with one another. Systems like AIBench will generate valuable data about the performance of contemporary AI applications, while also supporting various second-order

https://jack-clark.net/2020/05/11/import-ai-197-facebook-trains-cyberpunk-ai-chinese-companies-unite-behind-aibench-evaluation-system-how-clo... 1/8

图 2: Jack Clark 有关 AIBench 和 MLPerf 竞争的评论。2025 年 12 月 1 日截图。截图的目的是提供证据，以防访问链接消失。

这次，主席没有给我回信。我感觉不妙，一定有事情在悄悄发生。因为工业版的评审委员会里有 MLPerf 的主要贡献者。果然不出我所料，我们的文章最后因非技术原因被拒，有人在后面悄悄运作。

我很愤怒，给大会组织者以及更高层委员会写信。高层委员会的两个美国教授我都认识，有一个可以说很熟悉。有一个女教授还专门给我写信，了解详细的细节。主席给我写信，表示歉意，说让我感到不公平。但最终委员会的决定还是倾向于主席。我毫不奇怪。在差不多同样的时间范围内，计算机体系结构圈内发生了更糟糕的事。因为评审人串通、运作，一名高大的华裔博士生的论文在某著名顶级会议上录用。内容有明显的错误，他想撤稿，导师不同意，最后他选择在导师办公室自杀，震惊全球计算机体系结构圈子，而且影响在外溢。即使是这样震惊的事件，博士生自杀了才换来了姗姗来迟的调查和处罚。

MLPerf 的论文半年后被 ISCA 录用，成了一颗超级明星。我们的论文后来又投给了 Micro 2020，同样是高分。据程序委员会主席给我的回信，在最后讨论时，被全体评委投票所拒。后来又投给 PACT 2021(都快两年过去了)，又注定被拒时，我愤怒了，给主席写信。主席是个法国科学家，为我们不平。他再找了些评委，把论文录用了。但文章的影响已消耗殆尽，我们的论文发现刚开始让评阅人很吃惊，后来慢慢成为常识。

这是一个典型的案例，证明了我们科学技术社区有多残酷；每一个从事科学技术研究的人，他或她的遭遇可能会有多么不公平。因为这件事，高婉铃有几年都感觉到深深地愤怒。有一次聊天，我提到我们有些地方做得不对。高博士跟我说，即使做对了，他们也不会录我们的论文。如果比较两篇文章或两个作者的命运，公开条件几乎一样，最终决定结果的是那些未知或已知却在背后运作的因素。这是我们每个人不得不面对的残酷现实，这也是评价学这本书的主题之一：准实验。

我已经不再愤怒了。我超脱愤怒，靠关注更有意义的事。我觉得不受人制约的最佳方法就是做一个更深刻的工作，不随大流。这也是我开始评价学研究的动力之一。

2021 年，我意识到评价基准尽管在计算机、人工智能、金融领域广泛使用，却没有严谨的科学方法。即使是最著名的人工智能评价基准 ImageNet 也同样没有严谨的方法。我听过李凯教授的公开报告。他说，当年和李飞飞一起申请项目的时候，被人嘲笑过。李凯当时已经是非常资深的著名教授。我觉得应该发展严谨的评价方法。21 年，我在自己创办的期刊 TBench 上写了一篇社论，呼吁建立 Benchmark 科学与工程 [337]。

但什么是 Benchmark 呢？从来没有人严谨地定义过，也不太被人重视。我曾经的博士生汤飞曾经跟我说，别人问他的研究方向，他都不好意思说自己是做 Benchmark 的。他觉得看起来好像很低级，至少一点都不高级。汤飞是一个非常勤奋、有天赋和诚恳的年轻人。他现在在浪潮工作。

22 年，我意识到了 Benchmark 和评价之间的关系。我问了自己一个非常的关键问题，“为什么药物评价用随机控制实验 (RCTs)，但计算机领域仍在用经验方法，例如评价芯片的 SPEC CPU?” 每个 CPU 厂商都喜欢用 SPEC CPU [284] 报告性能数据。我的博士生王晨曦早就证明：同样一个处理器，完全遵循 SPEC CPU 的方法来评价，性能数据之间可以相差超过 10%，甚至达到 400% 的性能差异 [321]。也就是说，两个公司生产性能相同的 CPU，在报告数据时，一个公司可以说自己的性能比另外一个公司好数倍，而遵循的评价方法是大家认可的。

我开玩笑说，报告一个处理器的性能数据，反正不会死人。评价药物则不同，它真的会杀死人。在其他领域，情况更糟糕。比如说大学排名，它甚至会让一些青春期的男孩、女孩甚至他们的父母苦恼或者抑郁。多么讽刺!

2023 年，我正式创造了一个英文单词 Evaluatology 来命名评价学，并且写了一篇文章，评价学：评价科学与工程 [339]。2024 年，我们还在广州组织了一次评价学研讨会，很多学者热情参会，参与讨论和交流。

24 年至 25 年，我几乎无休息地投入了本书的写作。最初计划一个人独立完成，但很快意识到单打独斗的困难，也意识到带同事与学生一起工作更有价值。我第一批邀请了王磊博士、高婉铃博士和博士生李鸿霄加入。一个月后，进度还是滞后，我又邀请博士生王晨曦，特别研究助理范帆达博士加入。上个月，特别研究助理康国新博士对基于评价学的人工智能产生了浓厚兴趣，并付出大量努力，我认为邀请她加入才公平。

在两年研究和写作过程中，有几件事值得提及：

一位熟人受我的评价学公开演讲的启发，快速撰写了一篇文章，通过特殊渠道很快发表在一篇知名杂志上，其中有些观点明显受我的工作的影响。在我公开报告之前，我的论文早发表一个月。他是主动跟我交流的，说他的工作很早就录用了，希望我能参考他的工作。我知道这些杂志的运作方式，还认识他们的责任主编。看到我发了一些关键证据后。他又强调曾在另一篇文章中匿名提及过我的工作；另外想邀请我在他负责的一个杂志专栏写篇文章。我拒绝了他的提议，但也没有向委员会投诉，以免影响他的职业发展。

科技生涯中常遭遇这些令人失望的情况。例如我曾经写过一份技术报告，有人问我有没有发表，如果没有，他想在我的报告的基础上出本书。我对此非常反感，因此在这本专著里，添加了不少脚注，明确标注灵感的来源或者是谁的贡献，以示对他人贡献的尊重。

许多人轻视评价学，认为它是“软领域”。我完全反对此观点，坚信评价与设计同样重要，甚至可能催生基于评价学的新的人工智能范式。在一次交流中，罗纯杰博士提出评价与设计是对偶问题，通俗地讲就是同一硬币的两面。他的论证很有说服力。我在脚注里清楚地提到这个想法受他的影响。

25 年某一天，我突然想到：评价学可被定义为“揭示事物影响的科学”。这一灵感源于朱迪亚 · 珀尔博士和达娜 · 麦肯齐的著作《为什么之书》。

在评价学发展过程中，有两位中外科学家应该提及。一位是美国的迈克尔 · 斯克里文（Michael Scriven）教授，他在 1966 年发表的文章里，就认为对不同事物的评价应该是相似的 [272]。之后他一直致力于建立评价学科，发表了大量有影响力的文章。他是最早提出要建立评价学科的人之一，我追溯到的最早文献是 1966 年。中国学者邱均平教授和他的学生们在 2000 年出版了一本评价学的书 [80]，他可能是最早在中文使用评价学这一词的人。但他们完全忽略了迈克尔 · 斯克里文的大量著作。

我本来一开始不想使用评价学这个词，用的是评价科学和工程 (Evaluatology)。但邱教授的书更多的是各个学科领域方法的汇总，没有统一的方法。另外，他更多的是从社会学、文献计量学的角度出发，完全忽略了实验设计（DoE），随机控制实验（RCTs）等基础性的科学方法。我将评价学定义为揭示原因和影响的科学，评价学包括三个基本问题：评价问题：揭示对象的影响；逆评价问题：解释现象的原因；基于评价学的设计问题：在统一的评价和设计空间实现更优设计 [2]。由于内涵和邱教授定义的评价学完全不同，所以仍然选择使用评价学（Evaluatology）。为了区分，我将中文评价学和英文（Evaluatology）一起使用。我在书中也阐述了过去建立评价学科失败的原因；过去达成共识的研究和实践将评价定义为确认一个事物的优点和价值，容易掉入主观的陷阱；也没有提出的统一的概念、公理、核心问题和方法。当然我的尝试可能也会失败。

最后，我要真诚地感谢家人。用评价学的视角看，我的妻子显然是过去二十年来对我影响最大的人。我们 2001 年 10 月在香山徒步时相识，次年便在没有举办婚礼的情况下

[2]英文版 1.0，原文的中文翻译为“我将评价学定义为解释万物影响的科学，内涵完全不同”。此处于 2026 年 5 月 28 日根据中文版 1.0 做了更新。

旅行结婚，直接前往九寨沟度蜜月。此后一直过着简单而幸福的生活。

女儿是我最珍爱的宝贝。非常想念你，愿你在波士顿的生活充满快乐。每个生命都是独特的，拥有与生俱来的尊严——这种尊严不在外表，而存于内心。

还要感谢过去一年里我照料的小动物、树木和花草。每当精疲力竭时，静静地观察你们的成长总能带给我深沉的平静与力量。