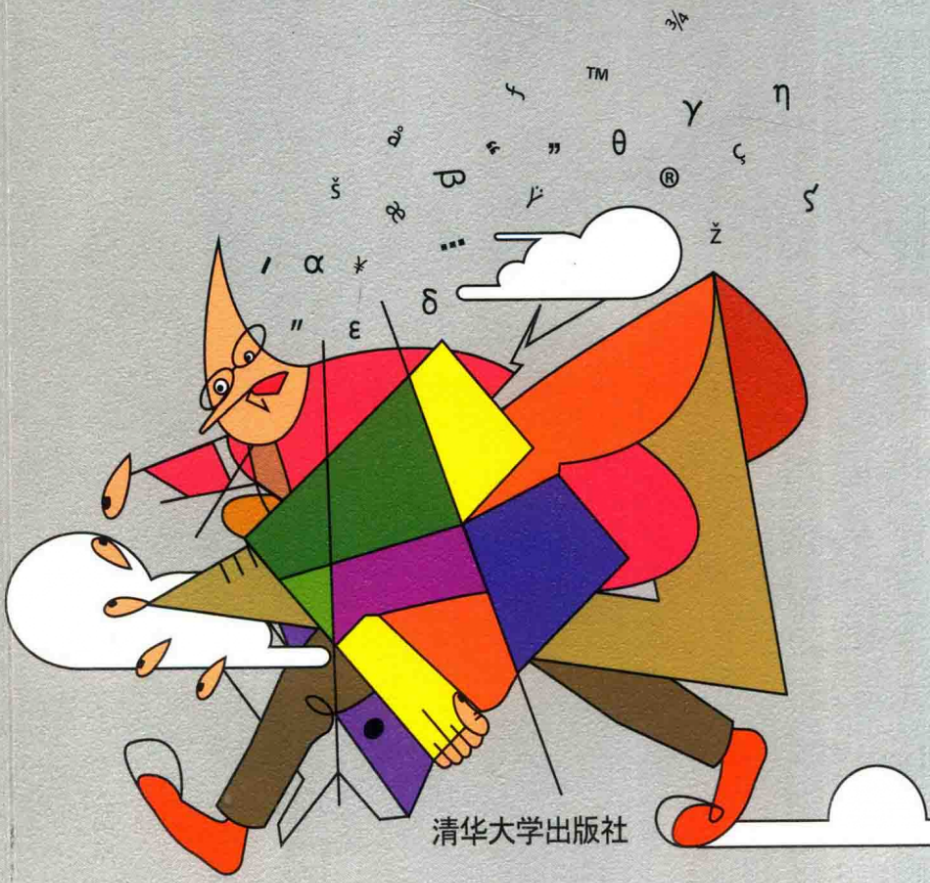
Errors favor intelligent beings

爱犯错的智能体

张军平 著

从开脑到烧脑的科普： 有诗歌，有画，有乡愁，不乏幽默的人工智能科普；

霍金说过： 科普书每多一个数学公式，书的销量将减少一半。**这书没公式！**



清华大学出版社

Errors favor intelligent beings

爱犯错的智能体

张军平 著

清华大学出版社
北京

图书在版编目（CIP）数据

爱犯错的智能体 / 张军平著. — 北京：清华大学出版社，2019
ISBN 978-7-302-53042-8

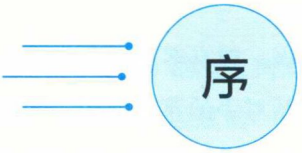
Ⅰ. ①爱…　Ⅱ. ①张…　Ⅲ. ①人工智能－普及读物　Ⅳ. ①TP18-49

中国版本图书馆CIP数据核字（2019）第094449号

责任编辑：胡洪涛　王　华
封面设计：徐腾赫
责任校对：赵丽敏
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市龙大印装有限公司
经　　销：全国新华书店
开　　本：145mm × 210mm　　印　　张：7.75　　字　　数：197千字
版　　次：2019年7月第1版　　印　　次：2019年7月第1次印刷
定　　价：55.00元

产品编号：082524-01



智能体和程序体的对话

我在科学网上第一次看到张军平教授写的系列文章《爱犯错的智能体》时，我还以为这里“智能体”指的是人工智能理论或编程中提到的专业名词 agent。但当我仔细读其内容时，特别是在从头浏览其内容时，才发现这里的智能体主要指的是人，尤其是生物学上的人。作者说的没有错，人确实易犯错误。书中从分析生物人的感知功能谈起，以生动的例子介绍了人的视觉、听觉、触觉和体觉的相关知识及其基本原理。之后又进入人的感情世界，从人的情感、回忆、梦境，一直谈到灵感和错觉。在这个过程中，作者又适时讨论计算机在处理人的感知世界时会遇到的麻烦及处理原则，甚至还不忘介绍一下讨论对象的数学背景。高斯、黎曼、莱布尼茨、庞加莱、爱因斯坦、图灵等大师级人物的名字频频出现。作者不费力地游弋于生命、计算机、数学、物理等几大学科之间，让读者经历一次目不暇接的跨学科科学旅游。再加上一个个有趣的故事，还有诗，画，歌，甚至还有乡愁！以这样的方式来做科普，我还是第一次读到，感觉很新鲜、很解惑，又易于接受。

本书的主角是被称作“智能体”的人，暂时称之为人智能体。人是万物之灵，却也不能避免犯错误，“人非圣贤，孰能无过？”。作者为何用一整本书来讨论这类问题呢？看来除了分析人的感知和认知功能本身外，作者还试图用人智能体犯的错误来考核另外一个智能体，即机器智能体，简称程序体。如果后者遇到了同样的问题，它能避免犯错吗？人智能体犯错

误一般有客观和主观两方面的原因。客观原因可能是面临复杂的环境，包括对手的蓄意欺骗，主观原因则往往可以归结为经验不足或经验有偏差。为什么会把美女看成老太太呢？因为不知道看一幅画可以从多个角度看。为什么会对隐藏在背景中的目标物视而不见呢？因为没有想到画中还会有画。一个正常的人会吃一堑，长一智，变得越来越聪明。这是什么？这就是人智能体积累的经验，以及从经验中提取的理性认识。对此，程序体有一个很好的工具——贝叶斯理论。犯错误好比一个结论不当的贝叶斯推理，说明不是先验有问题，就是驱动先验的似然有问题。一个有丰富先验和可靠选择机制的程序体就不大会犯类似的错误。所以人智能体在感到困惑的时候不妨咨询一下程序体。

把贝叶斯模型比喻为人智能体的经验有一个问题，就是程序体编写的贝叶斯模型都是针对有限前提的，即它只在程序体为它设定的某一类特定环境有效，而人智能体则以它的全部生活和终身经历为其经验支撑。试问程序体能够构造出这样的贝叶斯模型来吗？这可能就是程序体不及人智能体的地方吧。不过贝叶斯理论至今仍是一个活跃的研究领域。随着研究者们向它提出的问题越来越难，要求越来越苛刻，程序体也在一步步赶上来，更深刻的理论和技术不断诞生。2006 年有人提出了结构化先验的概念，力图把程序体中贝叶斯先验涉及的众多概念按人智能体的认知结构组织起来。先验不再局限于某个有限的图结构，而可以是一个时间上无穷的随机过程。更进一步，复旦大学的李斌提出了可学习先验的思想，直接挑战原本属于人智能体的“活到老，学到老”概念。

当然教益还不止这一点。人智能体可以请教程序体的地方还很多。例如我们可以再讨论一下人智能体对隐藏在背景中的目标物视而不见的问题。这次我们考察那条斑点狗。公正地说，斑点狗之所以未能被发现，是因为组成斑点狗的那些斑点是一个离散集合，它们没有连成线条，并且与其他

斑点混杂在一起。结果，本来是“庞然大物”（相对于该图像）的斑点狗消失在斑点之中。这是什么问题？这是知识表示粒度的问题。大粒度的一条狗用稀疏的小粒度斑点表示，当然就看不见了。若问程序体这个问题该怎么办？程序体可能回答：“你怎么不用粒度计算呀？”正如张钹院士指出的：“人类智能的一个公认的特点，就是人们能从极不相同的粒度上观察和分析同一问题。人们不仅能在不同粒度的世界上进行问题的求解，而且能够很快地从一个粒度世界调到另一个粒度世界，往返自如，毫无困难。”粒度计算，这个当年扎德（Zadeh）开辟的新领域，如今已经成为人工智能研究者乐此不疲的探索地。适当地调整计算的粒度，或者灵巧地处理大、小粒度之间的互动，也许可以让那只隐藏在斑点中的狗露出原形。

我们再看看本书中所说的视觉自举原理。动物的眼睛在差异巨大的光强变化之间能够迅速自我调整以适应多变的外来光。我原来一直以为人和猫的眼睛在光强变化下的自适应原理是一样的。感谢本书作者指出这两者之间的区别，使我增加了知识。书中也提到了光强的瞬间变化对交通安全的影响。这个问题可能和粒度计算有关，也可能和贝叶斯先验有关。但无论是粒度计算或贝叶斯先验都无法解决它，因为这不是一个简单的光强调度问题或光强转化问题，而是人智能体同时面对强光和弱光，甚至还有微光时的应对问题。试想，面对漆黑的夜晚里忽然出现的一辆开着远光灯的大卡车，你还能看清楚一只萤火虫吗？幸好，类似的问题计算数学家们早就想到了。有一门学问叫多尺度计算，就是为解决此类问题而诞生的。这对程序体是个好消息。在传统计算中有时会同时出现极大的数（如几百亿）和极小的数（如几百亿分之一）。如按常规方法则在计算进行时不是前者造成溢出，便是后者被按忽略不计处理。如何使量级差异巨大的数能够恰当地同时处理，这是多尺度计算要解决的难题。当然，数值计算和光强调度（物理）、光强感知（生物物理）之间并没有直接联系，这只是一个类比。但也

许可以给我们以某种启发。

通过这些例子或更多的例子，我们可以看到，人智能体和程序体对事物的认知和处理能力实际上是互有短长的。本书作者提到的可解释性问题是一个极好的例子。在求解各种实际问题时，人们往往希望能有一个既能通用建模，又能提供最优解的方法，作者用那位 116 岁的老奶奶做比喻说明鱼和熊掌不可兼得，这个比喻非常贴切。回想在可计算性理论中我们学到过一些“不可计算”定理（不可解问题的另一种说法）。我认为 116 岁老奶奶的例子给出了不可计算定理的一种崭新版本：“通用建模和最优求解不可同时计算定理”，或者直接称为“鱼和熊掌不可兼得定理”，又称“平猫不确定原理”。作者还提到了扎德在 40 多年前提出的复杂系统“预测和可解释性不相容原理”。由此可以解释深度学习的“最优求解和理性解释不可兼得定理”。上述第一个原理说了一个数学事实，可能会长久存在下去。第二个原理则可能是受我们目前的认识能力所限，不知道将来有没有突破的可能性，至少在某种意义上的突破。

本书谈论智能。虽然并没有正面给出智能或人工智能的定义，但是通过很多生动的例子，作者已经透露了对于此类问题的一些观点。读者可能会注意到，人智能体会做的事情很多，会犯错误的场合也很多，而许多常见的错误却没有收入书中。例如棋手错判对方意图，下棋输了；学生没有领会题意，写作文跑题了；投资者错估形势，炒股大亏，等等。为什么呢？我认为作者表明了这样一个观点。人智能体的智能并不局限于理性思维这样的高级形式。学术界常用的公式：数据→信息→知识→智能（或智慧）只是程序体的一种智能公式。从作者罗列的大量视觉、听觉、触觉、体觉的实例来看，该公式并非对生物人智能的一般性概括。如果注意到视觉、听觉、触觉、体觉并非人类独有，则它们还表明了人以外的生物也可以有智能。另一方面，如果我们仔细推敲“体感”这一节，可以发现本书并不

认为大脑是生物智能的唯一产地。文献中报道的著名仿生机器“大狗”能够在复杂地形上负重快跑，它对身体平衡能力的掌控就模拟了人类小脑的功能。在更广的意义上，人类的脑是一个复杂结构，它的各个部分各司其职。例如脑干要负起维持所在人生命的多种重要责任，包括心跳、呼吸、消化、体温、睡眠等重要生理功能。还有许多条件反射和无条件反射。如果要用人工智能技术构造一个人工生命，对脑干功能的模拟就是必不可少的。这令我们想起了布洛克斯主张的“没有表示的智能”。他凭此还获得了 1991 年国际人工智能联合大会的计算机与思维奖。

可能有一种解释是：脑干是一种生命现象，它却与智能无关。但是脑干模拟功能是人工生命的一部分，它与人工智能有关。这种解释使我们意外地得到了一个推论：人工智能模拟的是否不仅仅是智能，而可能也泛指某种生命现象？机器鱼不也是这样吗？但是如果这个观点能够成立的话，就会产生一个问题：它是否管得太宽了？人工智能究竟是我们努力的目标？还是我们应该遵循的方法学？我在《人工智能》一书的前言中曾提到学界对于人工智能的态度有愚公派和智叟派之分。愚公派认为总有一天会把人工智能这座大山完全搬走（到那时机器像人一样聪明），智叟派则认为努力挖山不应懈怠，但挖尽之日永远不会来到。我愿意站在智叟派一边，认为人工智能既是一种（无止境推进的）目标，更是一种（应该持之以恒的）方法学。

在结束序言之前我还想说一句公平话。本书名曰《爱犯错的智能体》。人智能体在这里被一系列的故事批得灰头土脸。但是号称万物之灵的人智能体，其智能真的就那么不堪吗？我在这里只指出一点，人智能体固然爱犯错，但是更能容错。为什么某甲能够一眼认出某乙？尽管某乙外表已与当年初见时很不一样。为什么某丙能解决一个复杂的问题？尽管他从来没有遇到过类似的情况。为什么不同的程序体被设计来处理不同的智能问题？

而人智能体却能够处理各种各样的智能问题，尽管他只有一个大脑，其结构还是固定的。所有这些和他们的容错能力关系极大。为公平起见，我建议作者在本书出版后再写一本《能容错的智能体》，至少和本书一样精彩，或者更精彩。

陆汝钤

中国科学院数学与系统科学研究院

2019 年 3 月于北京

前言

“军平，我觉得你不妨用科普的形式把你的观点写出来！”

看完我写的技术报告，我的博士生导师王珏研究员对我说道。

1 萌芽

那是2006年11月，我博士已毕业3年。小朋友刚2岁，每天抱着她闲逛，看着她日渐成长，痛并快乐着。她虽尚不能流畅交流，但我相信，多跟她说些话，她总能潜移默化吸收一些，也许对她今后的智力发育会有大的帮助。出于天生的好奇心重，我也顺便观察着她的智力发育变化，比如发错音的问题、颜色辨识困难的现象，诸如此类。那段时间，我对人的认知心理也有些兴趣，顺便看了点皮亚杰的《儿童发展心理学》、华生的《行为主义》等心理学方面的书籍。有一阵子，经常为自己在智能发育方面的一些奇思怪想激动不已。为了能方便总结，我向陆汝钤老师申请了去北京的中科院数学所访问2个月。陆老师很快就答应了，并将中科院计算所他的办公室借给我使用。在那里，我完成了图0.1所示的技术报告，还见到了就在隔壁房间办公，我一直很仰慕的人工智能老前辈史忠植老师。偶尔也会去隔壁办公室，跟当时正在用传统机器学习和计算机视觉方法在人脸识别领域奋斗着的山世光，以及在生物信息学领域钻研着的贺思敏闲聊。

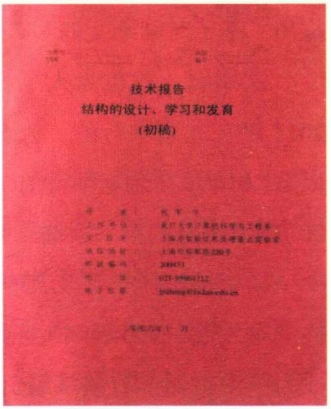


图0.1 最初的技术报告——科普雏形

我把技术报告给陆老师看后，他说还

不够深入。王珏老师也说，缺乏实验在计算机领域是站不住脚的。如果只是想表达自己的观点，不如用科普的形式写出来，就像厦门大学的集禅宗、古琴、机器作曲于一身的周昌乐教授写的《无心的机器》那样。

仔细想了想，感觉工作确实也不是太完整，不如放一放，再多积累点，多看看世界，也许会更丰满。

2 修身

时间过得飞快。2007 年 9 月，我去加州大学圣迭戈分校访问了半年，旁听了不少课程和报告，如提出 Adaboost 算法的约法夫·弗洛德（Yoav Freund）的“机器学习”课，也听到杰弗里·希尔顿（Geoffrey Hinton）介绍他 2006 年发表在 *Science*（《科学》）上，在深度玻尔兹曼机方面的研究进展。不过当时，大家多还处在对“深度学习”将信将疑的阶段，毕竟第二波人工智能的低潮把大家打击坏了。回国后，我继续做我的博士研究方向——基于流形学习的降维研究。我们小组针对高维数据降维后如何做统一的客观评估提出了一套准则，也基于代数拓扑中的持续同调思想，构造了主单纯复形的监督学习算法。在远距离身份识别这一块，基于人可以根据人的走路轮廓而不需要对图像进行细粒度的分析就能识别行人这一特点，提出了基于时间不变的步态模板的行人识别算法。

这期间，我也为人工智能两个主流会议做了些服务性工作，包括给 2013 年在北京举行的“人工智能国际会议”（IJCAI）做学生志愿者主席，以及 2014 年在北京举行的“机器学习国际会议”（ICML）做当地组委会成员。2013 年 11 月的第一个周末，我和西安电子科技大学高新波教授承办了“第十一届中国机器学习及其应用研讨会”。该会议 2002 年始于复旦，由陆汝钤老师发起，2005 年转至南京大学，在王珏老师、周志华教授的推动下，成为国内机器学习领域最负盛名的研讨会。通过参加这些会议，我对人工智能的认识也深入了一些。

2014 年 8 月—2015 年 8 月再次赴美国访问，并被宾夕法尼亚州立大学聘为研究助理（Research Associate），在信息科学技术学院王则 (James Z. Wang) 教授的指导下做些机器学习和气象预测等相关的研究。临行前，特地去拜见了我的博士导师，他又给我讲了一些他对人工智能发展近况的思考，并给我建议了一些值得关注的方向。而我到美国后，也利用这一年的时间，安安静静地思考了人工智能的发展与不足。

回国后，发现深度学习已经如火如荼，不跟进几乎很难在预测性能上占得优势。在发现单块显卡处理能力的问题后，我们便开始陆续购入了更多的 GPU（图形处理器）显卡，来帮助增加计算能力，目前已经有了 22 块像样的显卡。对我来说，与以往最明显的区别，就是发论文的成本高了，这让人多少有些心痛。以前一支粉笔、一块黑板、一个仿真程序可能解决的事，现在靠大数据、GPU、硬盘存储系统，一篇论文的成本可能接近 10 万元。更何况，还经常会碰到参数调了半天，算法不收敛的状况。

2017 年 5 月左右，应邀去西安参加了两次郑南宁老师主持的人工智能的相关研讨会，并参与筹备中国自动化学会混合智能专委会。7 月，国务院发布了《新一代人工智能规划》，其中谈到了人机回路。8 月，混合智能专委会成立，西安交通大学薛建儒老师任主任，我很荣幸当选为副主任之一。同年，我们开通了专委会的微信公众号。

3 科普

2018 年上半年，机缘巧合在《科技日报》上写了一篇关于对抗生成网的访谈报道。想着要给大众读，趣味性就得加强一点。所以，我在报道中讲了两个小故事，一是奥地利小说家斯蒂芬·茨威格写于 1941 年的小说《象棋的故事》里，一个囚犯在监狱里自己跟自己下国际象棋的故事，另一个是金庸的《射雕英雄传》里周伯通被困桃花岛后的双手互搏。虽然内容与对抗生成网的目的有一定的出入，但这两个故事都比较形象地讲述了同一

个模型里存在对抗的事实。后来从阅读量来看，反响还不错。

于是感觉自己可以尝试写点科普文章。刚巧专委会也需要对微信公众号进行宣传，而我又一直对论文成本上升太快、经费有点撑不住耿耿于怀。某天在“追寻长寿之道”时突然发觉个体和统计的差异如此明显，觉得这个道理似乎能解释深度学习的优异和不稳定性，便给专委会微信公众号连续写了两篇文章，《深度学习，你就是那个 116 岁的长寿老奶奶》和《童话（同化）世界的人工智能》，科普深度学习之现状以及对现在产业和学术界带来的同化效应。在文章的最后，我留了个尾巴，我认为现在的研究尚不能完全解开智能的谜团。也许，答案要从犯错中去寻找。

这两篇文章的反响也是出奇的大，我便跟我博士导师的好朋友、对我有很大帮助的中国科学院自动化所的王飞跃老师在微信上说了这件事。开心之余，他建议我不妨写个科普系列。

要写科普系列，我想起了 2000—2003 年在北京读博士期间经常逛北京大学校区二手书摊时偶得的一本科普书《哥德尔，艾舍尔，巴赫——集异璧之大成》。王珏老师告诉我这是本好书，要好好看看。然后跟我讲了译者之一、严勇的导师、北京大学马希文教授（也是周昌乐的导师）的一些轶事，比如精通六国语言，对此书在信达雅的翻译处理方面赞叹不已。2016 年，陆汝钤老师来复旦时又给我补充了一些对马希文教授的回忆。再说说这本书，它在美国一直是本科普畅销书。不过，不足在于，这本书第一版发行的时间是 1979 年，正处在人工智能第一波的寒冬中，对于 1979 年以后的人工智能的进展、观点变更没有涉及。其次，书还是太厚了。真心有兴趣把这本书细细读完的，十有八九是与人工智能相关的科研工作者或从业人员。另一本是 2015 年我在美国买的畅销图画书，克莱夫·吉福德（Clive Gifford）撰写的 *Eye Benders*: *The Science of Seeing and Believing*（《眼睛弯管：看和相信的科学》），书中讲了不少视觉错觉的例子，但并没有从人

工智能的角度去做深入分析。

所以，我想结合两本书的一些背景知识，再加上我 2006 年以来对智能结构发育的一些认识和再认识，以及近年来对人工智能的许多更新理念来撰写。

而在写作手法上，我希望能做到专业和引人入胜。俗话说，科学家上报纸，就会少一圈朋友；科学家上电视，就没有朋友了。所以，写的内容我都反复斟酌过，确保逻辑通畅、无漏洞，防止没朋友。但人不是神，总有可能会出错。如果仍有遗漏和问题，后面我会做个勘误表。而关于如何引人入胜，我采用了与我所读过的科普书不太一样的风格，即小故事加严肃科普的形式，偶尔会穿插几个科学笑话，当然还要有点中国特色，这个风格基本贯穿了全书。

不过，万事开头难。虽然每一节要写的基本路线我都清楚，但怎么开头都挺头痛的。所以，我想了一些办法。比如跑跑步、遛遛狗，期望缓解之余还能释放点多巴胺来启发一下，一有好的点子就赶紧记下。当然，还得有充足的时间投入。所以，在完成这个系列的过程中，我把很多朋友的讲座邀请、登门讨论都无情地拒绝了。没办法，有时候创作和研究都需要一个人有不间断的独立思考时间。这应该就是做科研要有的狂热吧。另外，有的时候短时间高强度的集中思考，确实能促进思维、帮助人更深入细致地思索问题的可能答案。虽然这段时间平均每天睡眠约 5 个半小时，确实很累，每当想放弃时，我就会想起美国作家罗伯特·卡尼格尔写的《知无涯者：拉马努金传》中描述过的印度数学家拉马努金追求数学真理的过程，就会想起因玻尔兹曼方程和朗道阻尼的工作而于 2010 年获得“数学界的诺贝尔奖”——菲尔兹奖、主攻最优输运理论的塞德里克·维拉尼（Cédric Villani）在其书《一个定理的诞生：我与菲尔茨奖的一千个日夜》中提到的坚持和努力。也会想起近代著名学者王国维在《人间词话》中提及的古今

成大事业、大学问者必经之三重境界：

1. 昨夜西风凋碧树。独上高楼，望尽天涯路。——北宋晏殊《蝶恋花·槛菊愁烟兰泣露》

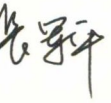
2. 衣带渐宽终不悔，为伊消得人憔悴。——北宋柳永《蝶恋花·伫倚危楼风细细》

3. 众里寻他千百度。蓦然回首，那人却在，灯火阑珊处。——南宋辛弃疾《青玉案·元夕》

总之，写这个系列对我来说，是物超所值的，因为在科普的同时，我也在其中总结了不少我对人工智能诸多问题的观点和探讨，希望能给那些对人工智能感兴趣的人有所启发。

最后感谢中国自动化学会混合智能专委会薛建儒主任、陈德旺副主任、王晓师妹对本科普系列在微信公众号传播的大力支持，感谢科学网连续20余次推荐本科普系列文章至科学网头条，也感谢众多微信公众号如中国工程院院刊、中国自动化学会等的推荐。这些支持，让更多对人工智能感兴趣的人了解了这个科普系列的工作。另外，我也衷心感谢很多朋友在本书撰写中提出的宝贵意见，尤其是与我一同从2002年“第一届中国机器学习及其应用研讨会”出道的、北京交通大学的于剑教授对本书一些概念的讨论。感谢家人和我的学生们的理解和默默支持。没有他们在生活和科研上的顺畅配合，我也不可能有多余的时间来写这个科普系列。也感谢国家自然科学基金（资助号:61673118）、上海市“脑与类脑智能基础转化应用研究”市级科技重大专项资助（项目编号：NO.2018SHZDZX01）和张江实验室对本书的支持。

仅以此书献给我的博士导师：王珏研究员



写于2018年12月24日

目录