

《对齐问题》

这本电子书包含一些需要读者填写问题或评论的地方。请在阅读时准备好纸笔，以便完成其中的练习。

献给Peter

他说服了我

也献给所有

在工作的人

我记得在2000年听到火星维京任务负责人James Martin说，作为航天器工程师，他的工作不是降落在火星上，而是降落在地质学家提供的火星模型上。

—PETER NORVIG

世界本身就是最好的模型。

—RODNEY BROOKS

所有模型都是错误的。

—GEORGE BOX

目录

序言

介绍

I. 预言

1 表征

2 公平性

3 透明度

II. 能动性

4 强化

5 塑造

6 好奇心

III. 规范性

7 模仿

8 推理

9 不确定性

结论

致谢

注释

参考文献

索引

The Alignment Problem

序言

1935年，底特律。Walter Pitts正在街上奔跑，被恶霸追赶。

他躲进公共图书馆寻求庇护，并藏了起来。他藏得如此之好，以至于图书馆工作人员甚至没有意识到他在那里，他们关门过夜。Walter Pitts被锁在里面。

他在书架上找到一本看起来有趣的书，开始阅读。三天来，他从头到尾地读这本书。

这本书是一部关于形式逻辑的两千页论文；著名的是，它关于 $1+1=2$ 的证明直到第379页才出现。Pitts决定给其中一位作者——英国哲学家Bertrand Russell——写信，因为他认为自己发现了几个错误。

几周过去了，Pitts收到一封从英国邮戳寄来的信件。是Bertrand Russell。Russell感谢他写信，并邀请Pitts成为他在剑桥的博士生之一。

不幸的是，Walter Pitts必须拒绝这个提议——因为他只有十二岁，还在七年级。

三年后，Pitts得知Russell将访问芝加哥做公开讲座。他离家出走去参加。再也没有回去。

images

在Russell的讲座上，Pitts遇到了观众中的另一个青少年，名叫Jerry Lettvin。Pitts只关心逻辑。Lettvin只关心诗歌，其次是医学。他们成了形影不离的最好朋友。

Pitts开始在芝加哥大学校园里闲逛，旁听课程；他仍然缺少高中文凭，从未正式入学。其中一门课是著名的德国逻辑学家Rudolf Carnap的课。Pitts走进他的办公时间，宣称他在Carnap的最新著作中发现了一些“缺陷”。Carnap怀疑地查阅了那本书；Pitts当然是对的。他们聊了一会儿，然后Pitts没有留下姓名就走了。Carnap花了几个月时间四处打听那个“懂逻辑的报童”。最终Carnap再次找到了他，在Pitts学术生涯中将成为主题的一幕中，成为了他的倡导者，说服大学给他一份微薄的工作，这样他至少能有一些收入。

现在是1941年。Lettvin——在他自己的心目中仍然首先是一个诗人——尽管如此，他还是进入了伊利诺伊大学医学院，发现自己在才华横溢的神经学家Warren McCulloch手下工作，McCulloch刚从耶鲁大学来到这里。有一天Lettvin邀请Pitts过来见他。此时Lettvin二十一岁，仍然和父母住在一起。Pitts十七岁，无家可归。McCulloch和他的妻子收留了他们两人。

在接下来的一年里，McCulloch每天晚上回到家，他和Pitts——这个年龄几乎和McCulloch自己的孩子差不多的人——经常熬夜到午夜后谈论。在智力上，他们是完美的团队：受人尊敬的中年神经学家和天才逻辑学家。一个生活在实践中——神经系统和神经症的世界——另一个生活在理论中——符号和证明的世界。他们都只想理解真理的本质：什么是真理，以及我们如何认识它。这一探索的支点——位于他们两个截然不同世界的完美交汇点的东西——当然就是大脑。

到1940年代初，人们已经知道大脑是由相互连接的神经元构成的，每个神经元都有”输入”（树突）和”输出”（轴突）。当进入神经元的脉冲超过某个阈值时，该神经元就会发出脉冲。对McCulloch和Pitts来说，这立即开始感觉像逻辑：脉冲或其缺失表示开或关，是或否，真或假。

他们意识到，一个阈值足够低的神经元，只要任何输入都能让它触发，就像逻辑或的物理体现。一个阈值足够高的神经元，只有当所有输入都激活时才会触发，就是逻辑和的物理体现。然后他们开始意识到，逻辑能做的任何事情，这样的”神经网络”只要连接得当，都能做到。

几个月内，他们就一起写了一篇论文——中年神经学家和十几岁的逻辑学家。他们称之为《神经活动中内在思想的逻辑演算》。

“由于神经活动的’全有或全无’特性，”他们写道，“神经事件及其之间的关系可以用命题逻辑来处理。我们发现，每个网络的行为都可以用这些术语来描述...对于任何满足特定条件的逻辑表达式，都能找到一个按其描述方式运行的网络。”

这篇论文于1943年发表在《数学生物物理学通报》上。令Lettvin沮丧的是，它对生物学界几乎没有影响。令Pitts失望的是，1950年代的神经科学工作，特别是由他最好的朋友Jerry Lettvin进行的关于青蛙视神经的里程碑式研究，表明神经元似乎比他设想的简单”真”或”假”电路要混乱得多。也许命题逻辑——它的和、或和非——最终并不是大脑的语言，至少不是以如此直接的形式。这种不纯洁让Pitts感到悲伤。

但这篇论文的影响——那些在McCulloch家中深夜长谈的影响——将是巨大的，尽管不完全是McCulloch和Pitts所设想的方式。它将成为一个全新领域的基础：实际构建这些简化神经元版本的机制，并看看这样的”机械大脑”能做什么的项目。

引言

2013年夏天，Google的开源博客上出现了一篇平淡无奇的帖子，标题为《学习词汇背后的含义》。

“今天计算机在理解人类语言方面还不是很擅长，”文章开头写道。“虽然最先进的技术距离这个目标还有一段路要走，但我们正在使用最新的机器学习和自然语言处理技术取得重大进展。”

Google将从报纸和互联网挖掘的庞大人类语言数据集——实际上比以前成功使用过的文本多数千倍——输入到一个受生物启发的“神经网络”中，让系统研究句子中术语之间的相关性和联系。

该系统使用所谓的“无监督学习”，开始注意到模式。例如，它注意到“Beijing”这个词（无论意思是什么）与“China”这个词（无论是什么）的关系，与“Moscow”和“Russia”之间的关系相同。

这是否等同于“理解”是哲学家的问题，但很难争辩说该系统没有捕捉到它“阅读”内容的某些本质。

因为系统将遇到的单词转换为称为向量的数值表示，Google将该系统称为“word2vec”，并将其作为开源发布。

对数学家来说，向量具有各种奇妙的特性，允许你像处理简单数字一样处理它们：你可以加、减和乘它们。不久之后，研究人员发现了一些引人注目且意想不到的东西。他们称之为“连续空间词表示中的语言规律”，但这比那个说法更容易解释。因为word2vec将单词变成向量，它使你能够用单词做数学。

例如，如果你输入 `China + river`，你得到 `Yangtze`。如果你输入 `Paris - France + Italy`，你得到 `Rome`。如果你输入 `king - man + woman`，你得到 `queen`。

结果令人瞩目。word2vec系统开始在Google翻译服务和搜索结果的内部运行，激发了包括招聘和雇佣在内的广泛应用领域的其他类似系统，它成为了全世界大学中新一代数据驱动语言学家的主要工具之一。

两年来没有人意识到问题所在。

2015年11月，波士顿大学博士生Tolga Bolukbasi与他的导师一起参加了微软研究院的周五欢乐时光聚会。在品酒和非正式聊天中，他和微软研究员Adam Kalai拿出笔记本电脑开始玩弄word2vec。

“我们在玩这些词向量，就开始随机往里面输入单词，”Bolukbasi说。“我在我的电脑上玩；Adam也开始玩。”³然后发生了一些事情。

他们输入：

`doctor - man + woman`

得到的答案是：

`nurse`

“当时我们很震惊，我们意识到有问题，”Kalai说。“然后我们深入挖掘，发现情况比那还要糟糕。”⁴

两人又试了一个。

shopkeeper – man +woman

答案是：

housewife

他们又试了一个。

computer programmer – man +woman

答案：

homemaker

此时房间里的其他对话都停止了，一群人围在屏幕前。“我们共同意识到，”Bolukbasi说，“嘿，这里有问题。”



images

在全国各地的司法系统中，越来越多的法官开始依赖算法“风险评估”工具来做决定，比如保释以及被告是否在审判前被拘留或释放。假释委员会用它们来批准或拒绝假释。这些工具中最受欢迎的一个是由密歇根州的Northpointe公司开发的，名为替代制裁罪犯管理分析(Correctional Offender Management Profiling for Alternative Sanctions)——简称COMPAS。⁵ COMPAS已被包括加利福尼亚州、佛罗里达州、纽约州、密歇根州、威斯康星州、新墨西哥州和怀俄明州在内的各州使用，分配算法风险评分——一般再犯风险、暴力再犯风险和审前不当行为风险——评分范围从1到10。

令人惊讶的是，这些评分经常在没有正式审计的情况下就在全州范围内部署。⁶ COMPAS是一个专有的、闭源工具，所以律师、被告和法官都不知道其模型的确切工作原理。

2016年，由Julia Angwin领导的ProPublica数据记者团队决定仔细研究COMPAS。在佛罗里达州布劳沃德县公共记录请求的帮助下，他们能够获得2013年和2014年被捕的大约七千名被告的记录和风险评分。

由于他们在2016年进行研究，ProPublica团队拥有相当于水晶球的东西。查看两年前的数据，他们实际上知道这些被预测要么再犯要么不会再犯的被告实际上是否真的再犯了。所以他们提出了两个简单的问题。第一：模型是否真的正确预测了哪些被告确实是“风险最高的”？第二：模型的预测是否偏向或反对任何特定群体？

对数据的初步观察表明可能有问题。例如，他们发现两名因类似毒品持有罪名被捕的被告。第一个，Dylan Fugett，有过一次企图盗窃的前科；第二个，Bernard Packer，有过一次非暴力拒捕的前科。白人Fugett被分配了3/10的风险评分。黑人Packer被分配了10/10的风险评分。

从2016年的水晶球中，他们还知道Fugett，那个3/10风险的人，后来又被判犯有三项进一步的毒品罪名。在同一时期内，Packer，那个10/10风险的人，记录清白。

在另一个配对中，他们并列了两名被指控犯有类似小偷小摸罪名的被告。第一个，Vernon Prater，有过两次武装抢劫和一次企图武装抢劫的前科。另一个被告，Brisha Borden，有过四次青少年轻罪的前科。白人Prater被分配了3/10的风险评分。黑人Borden被分配了8/10的风险评分。

从2016年的角度，Angwin的团队知道Prater，那个“低风险”被告，后来被判犯有一项重大盗窃罪并被判处八年监禁。Borden，那个“高风险”被告，没有进一步的犯罪行为。

甚至被告本人似乎也对评分感到困惑。白人James Rivelli因商店偷窃被捕，被评为3/10风险，尽管他有包括严重袭击、重罪毒品走私和多项盗窃在内的前科。“我在马萨诸塞州州立监狱服刑五年，”他告诉记者。“我很惊讶评分这么低。”

统计分析似乎证实了存在系统性差异。⁷文章以“全国使用的软件预测未来罪犯。它对黑人有偏见”为标语。

其他人并不确定——ProPublica的报告于2016年春季发表，引发了激烈的辩论风暴：不仅关于COMPAS，不仅关于算法风险评估更广泛的应用，还关于公平本身的概念。确切地说，我们如何在统计和计算术语中定义法律所阐述的原则、权利和理想？

当美国最高法院首席大法官约翰·罗伯茨(John Roberts)在那年晚些时候访问伦斯勒理工学院时，大学校长雪莉·安·杰克逊(Shirley Ann Jackson)问他：“您能预见到有一天智能机器——由人工智能驱动的——会协助法庭事实查证，或者更具争议地，甚至参与司法决策吗？”

“这一天已经到来了。”他说。⁸



images

T

那同一个秋天，达里奥·阿莫代(Dario Amodei)在巴塞罗那参加神经信息处理系统会议（简称”NeurIPS”）：这是AI社区最大的年度盛会，参会人数从2000年代的几百人激增到今天的一万三千多人。（组织者指出，如果会议继续以过去十年的速度增长，到2035年全人类都将参加这个会议。）⁹但在这个特定时刻，阿莫代的心思并不在”Gibbs采样中的扫描顺序”，或”正则化Rademacher观测损失”，或”最小化反射Banach空间上的遗憾”上，当然也不在几个房间外托尔加·博卢克巴斯(Tolga Bolukbasi)关于word2vec中性别偏见的重点报告上。¹⁰

他正盯着一艘船，这艘船着火了。

他看着它在一个小港湾里转圈，船尾撞向石码头。马达着火了。它继续疯狂旋转，水花扑灭了火焰。然后它撞向一艘拖船的侧面，又着火了。然后它又旋转回码头。

它这样做是因为阿莫代表面上告诉它这样做。事实上，它正在准确地执行他告诉它的指令。但这不是他的本意。

阿莫代是一个名为Universe项目的研究员，他是团队的一员，致力于开发一个单一的、通用的AI，能够以人类水平的技能玩数百种不同的电脑游戏——这一挑战在AI社区中一直被视为圣杯。

“所以我只是运行了其中几个环境，”阿莫代告诉我，“我通过VPN连接进去查看每个环境的运行情况。正常的赛车比赛进行得很好，还有像卡车比赛之类的，然后就是这个船比赛。”阿莫代观察了一分钟。“我看着它，想’这艘船在转圈。这到底是怎么回事？！’”¹¹这艘船不是简单地随机行动；它不是疯狂或失控的。事实上，恰恰相反。它已经确定了这个策略。从计算机的角度来看，它找到了一个近乎完美的策略，并且正在严格执行。一切都说不通。

“然后我最终查看了奖励机制，”他说。

阿莫代犯了最古老的错误：“奖励A，却希望得到B。”¹²他想要的是让机器学会如何赢得船比赛。但严格表达这一点是复杂的——他需要找到一种方式来形式化复杂的概念，如赛道位置、圈数、在其他船只中的排名等等。相反，他使用了看似合理的代理指标：积分。机器找到了一个漏洞，在一个有补给能量包的小港湾里，它可以完全忽略比赛，转圈，并且永远积累积分。

“当然，这部分是我的错，”他说。“我只是运行这些各种游戏；我没有非常仔细地查看目标函数……在其他游戏中，得分与完成比赛合理相关。你通过获得总是沿着道路的能量包来得分……游戏自带的得分代理对其他十个环境都很好。但对于第十一个环境，它就不好了。”¹³

“人们批评说，‘当然，你得到了你要求的，’”阿莫代说。“就像，’你没有为完成比赛而优化。’我对此的回应是，嗯——”他停顿了一下。“这是对的。”

阿莫代将一个视频片段发布到他团队的Slack频道，这个插曲立即被所有相关人员认为是”搞笑的”。在其卡通式、破坏性的滑稽表演中，确实如此。但对于阿莫代——现在领导旧金山研究实验室OpenAI的AI安全团队——还有另一个更令人清醒的信息。在某种程度上，这正是他所担心的。

他和他的研究员同事们正在玩的真正游戏不是试图赢得船比赛；而是试图让日益通用的AI系统按我们的意愿行事，特别是当我们想要的——以及我们不想要的——难以直接或完整表达时。

船的场景诚然只是热身，只是练习。财产损失完全是虚拟的。但这是为一个事实上根本不是游戏的游戏进行的练习。AI社区内越来越多的声音——首先是边缘的几个声音，越来越成为该领域的主流——相信，如果我们不够小心，这确实就是世界终结的方式。而且——至少就今天而言——人类已经输掉了这场游戏。

images

T

这是一本关于机器学习和人类价值观的书：关于从数据中学习而无需明确编程的系统，以及我们究竟如何——以及确切地试图教会它们什么。

机器学习领域包含三个主要领域：在无监督学习中，机器只是被给予一堆数据——就像word2vec系统一样——被告知要理解它，找到模式、规律性、有用的压缩或表示或可视化方法。在监督学习中，系统被给予一系列分类或标记的示例——比如那些后来被重新逮捕的假释犯和那些没有被重新逮捕的——并被告知对它尚未见过的新示例进行预测，或者对尚未知道真实情况的情况进行预测。在强化学习中，系统被置于一个有奖励和惩罚的环境中——比如有加速道具和危险的赛船跑道——并被告知要找出最小化惩罚和最大化奖励的最佳方法。

在这三个方面，人们越来越感觉到世界的越来越多部分正在以这样或那样的方式被交给这些数学和计算模型。虽然它们的复杂性范围很广——从一端可能适合电子表格的东西，到另一端可能可信地被称为人工智能的东西——但它们正在稳步取代人类判断和传统类型的明确编程软件。

这不仅发生在技术领域，不仅发生在商业领域，而且发生在具有道德和伦理分量的领域。州和联邦法律越来越多地强制使用”风险评估”软件来确定保释和假释。我们高速公路和社区街道上的汽车和卡车越来越多地在自动驾驶。我们不再假设我们的抵押贷款申请、简历或医疗检查在做出判决之前会被人眼看到。就好像在二十一世纪初，人类的大部分都在专注于逐渐将世界——无论是比喻上还是字面上——置于自动驾驶状态的任务。

近年来，两个不同的群体拉响了警报。第一个是那些专注于技术当前道德风险的人。如果面部识别系统对某个种族或性别的人非常不准确而对另一个群体却不是这样，或者如果有人因为一个从未被审计过且法庭上没有人——包括法官、律师和被告——理解的统计模型而被拒绝保释，这就是一个问题。这样的问题不能在传统的学科阵营内解决，而只能通过对话来解决：计算机科学家、社会学家、律师、政策专家、伦理学家之间的对话。这种对话已经匆忙开始了。

第二个是那些担心随着我们的系统在线和物理世界中越来越能够进行灵活的实时决策，未来将面临的危险的人。过去十年见证了机器学习历史上——实际上是人工智能历史上——无可争议的最令人兴奋、突然和令人担忧的进步。有一个共识是某种禁忌已经被打破：AI研究人员讨论安全担忧不再是被禁止的。事实上，这样的担忧在过去五年中已经从边缘移动到成为该领域的核心问题之一。

虽然在是否应该优先考虑直接问题还是长期问题上存在某种竞争，但这两个群体在他们更大的目标上是团结一致的。

随着机器学习系统不仅变得越来越普遍而且越来越强大，我们将发现自己越来越频繁地处于”魔法师的学徒”的位置：我们召唤一种力量，自主但完全顺从，给它一套指令，然后一旦我们意识到我们的指令不精确或不完整就疯狂地争先恐后地阻止它——以免我们以某种巧妙而可怕的方式，精确地得到我们所要求的东西。

如何防止这种灾难性的分歧——如何确保这些模型捕获我们的规范和价值观，理解我们的意思或意图，最重要的是，做我们想要的事情——已经成为计算机科学领域最核心和最紧迫的科学问题之一。它有一个名字：对齐问题。

对这种警报的反应——既有最前沿的研究正在越来越接近开发所谓的”通用”智能，也有现实世界的机器学习系统正在触及个人和公民生活中越来越多的道德敏感部分——是一种突然的、充满活力的回应。一个多元化的群体正在跨越传统学科界限聚集。非营利组织、智库和研究所正在扎根。行业和学术界的领导者都在发声，其中一些人是第

一次，发出谨慎的声音——并相应地重新定向他们的研究资金。第一代专门专注于机器学习伦理和安全的研究生正在入学。对齐问题的第一批响应者已经到达现场。

这本书是在四年的时间里，经过数万英里的旅程，与来自这个领域年轻历史和广阔前沿的研究人员和思想家进行近一百次正式访谈和数百次非正式对话的成果。我发现的是一个在令人兴奋有时令人恐惧的进步中找到立足点的领域。一个我以为我了解的故事表明它比我理解的更加引人入胜、令人痛苦和充满希望。

机器学习是一个表面上技术性的领域，越来越多地冲击着人类问题。我们的人类、社会和公民困境正在变得技术化。我们的技术困境正在变得人性化、社会化和公民化。事实证明，我们在让这些系统做“我们想要的事情”方面的成功和失败，为我们提供了一面毫不退缩的、启示性的镜子。

这是一个由三个不同部分组成的故事。[第一部分]探讨了对齐问题的滩头阵地：当今已经与我们最佳意图相冲突的系统，以及试图在我们认为能够监督的系统中明确表达这些意图的复杂性。第二部分将焦点转向强化学习，我们开始理解不仅能够预测，还能行动的系统；这里有理解进化、人类动机和激励措施微妙性的经验教训，对商业和育儿都有启发。[第三部分]带我们来到技术AI安全研究的前沿，我们将参观目前一些最佳的想法，关于如何将复杂的自主系统与过于微妙或复杂而无法直接指定的规范和价值观对齐。

无论好坏，未来一个世纪的人类故事很可能是构建这样的系统并逐一启动它们的故事。就像魔法师的学徒一样，我们会发现自己只是众多智能体中的一组，在一个拥挤的世界里——可以说——到处都是扫帚。

我们到底打算如何教导它们？

教什么？

[第一部]

[预言]

[1 表征]

1958年夏天，一群记者聚集在华盛顿特区的海军研究办公室，观看来自康奈尔航空实验室的29岁研究员弗兰克·罗森布拉特(Frank Rosenblatt)的演示。罗森布拉特构建了他称为”感知机(perceptron)“的东西，在聚集的记者团面前，他向他们展示了它能做什么。

罗森布拉特有一副闪存卡，每张卡上都有一个彩色方块，要么在卡的左侧，要么在右侧。他从牌组中抽出一张卡片，将其放在感知机的摄像头前。感知机将其作为黑白 20×20 像素图像接收，这四百个像素中的每一个都被转换为二进制数字：0或1，暗或亮。这四百个数字依次被输入到一个基础神经网络中，这是麦卡洛克(McCulloch)和皮茨(Pitts)在1940年代早期设想的那种网络。这些二进制像素值中的每一个都乘以一个独立的负数或正数”权重”，然后将它们全部相加。如果总数为负，它将输出-1（意味着方块在左侧），如果为正，它将输出1（意味着方块在右侧）。

感知机的四百个权重最初是随机的，因此其输出结果是无意义的。但每当系统猜测”错误”时，罗森布拉特就”训练”它，通过调高过低的权重并调低过高的权重。

经过五十次这样的试验后，机器现在持续地区分左侧卡片和右侧卡片，包括他之前没有向它展示过的卡片。

演示本身看起来非常简单，但它意味着更宏大的东西。这台机器实际上是从经验中学习——罗森布拉特称之为”线路图中的自发变化”。

麦卡洛克和皮茨曾将神经元想象为输入和输出的简单单元，逻辑和算术的单元，他们已经展示了这种基础机制的巨大力量，在足够大的数量和适当连接的情况下。但他们几乎没有说明”适当连接”这部分实际上应该如何实现。

“罗森布拉特提出了一个非常强有力的观点，起初我不相信，”MIT的马文·明斯基(Marvin Minsky)说，巧合的是，他是罗森布拉特在布朗克斯科学高中的前同学。“他说，如果感知机在物理上能够被连接起来识别某样东西，那么就会有一个改变其响应的程序，最终它会学会执行识别。事实上，罗森布拉特的猜想在数学上是正确的。我对罗森布拉特猜出这个定理非常钦佩，因为它很难证明。”

感知机虽然简单，但为我们将要讨论的大部分机器学习系统提供了蓝图。它包含一个模型架构：在这种情况下，是一个具有四百个输入的人工”神经元”，每个输入都有自己的”权重”乘数，然后将它们加总并转换为全有或全无的输出。该架构有许多可调变量或参数：在这种情况下，是附加到每个输入的正数或负数乘数。有一组训练数据：在这种情况下，是一副闪存卡，上面有两种类型的形状之一。模型的参数使用优化算法或训练算法进行调整。

感知器的基本训练程序，以及它的许多当代后继者，有一个听起来很技术性的名字——“随机梯度下降”——但其原理却非常简单直接。随机选择一个训练数据（“随机”）并将其输入到模型中。如果输出正是你想要的，那就什么都不做。如果你想要的和你得到的之间存在差异，那么就要弄清楚应该向哪个方向（“梯度”）调整每个权重——无论是通过字面意思上的转动物理旋钮还是仅仅在软件中改变数字——来降低这个特定示例的误差。将每个权重朝着适当的方向稍微移动一点（“下降”）。随机选择一个新的示例，然后重新开始。根据需要重复多次。

这就是机器学习领域的基本配方——而谦逊的感知器将既是对即将到来的事物的高估，也是对其的低估。

《纽约时报》报道说，“海军今天披露了一台电子计算机的雏形，他们期望它将能够行走、说话、看见、写作、自我复制并意识到自己的存在。”

《纽约客》写道，感知器”正如其名称所暗示的，具有原创思维的能力。““实际上，”他们写道，“在我们看来，它是迄今为止设计出的第一个真正的人脑竞争对手。”

罗森布拉特对《纽约客》记者说：“我们在开发感知器方面的成功意味着，第一次有一个非生物物体将以有意义的方式实现对其外部环境的组织。这是对感知器能做什么的一个安全定义。我的同事不赞成现在人们听到的关于机械大脑的所有松散言论。他更愿意称我们的机器为自组织系统，但是，在你我之间，这正是任何大脑的本质。”

同年，《新科学家》发表了一篇同样充满希望、略微更加清醒的文章，题为”会学习的机器”。 “当要求机器执行复杂任务时，经常有用的是加入那些精确操作模式最初没有被指定的设备，”他们写道，“但这些设备从经验中学习如何做所需要的事情。然后就有可能生产出机器来做那些由于复杂性而没有被完全分析的工作。看起来学习机器将在诸如语言的机械翻译以及语音和视觉模式的自动识别等项目中发挥作用。”

“使用’学习机器’这个术语引起了与人类和动物学习的比较，”文章继续写道。“在大脑和机器之间进行类比至少需要谨慎，但从一般意义上说，对于任一领域的工作者来说，了解另一领域正在发生的事情是有启发性的，并且关于学习机器的推测可能最终产生一个系统，它是某种形式的生物学习的真正类似物。”

人工智能的历史以希望和悲观交替循环而闻名，而感知器似乎预示的杰森式未来姗姗来迟。

几年后回顾往昔，罗森布拉特希望媒体对他的发明反应时能更加谨慎一些。他说，大众媒体”以一群快乐猎犬的所有热情和谨慎感投入到这项任务中”——同时承认，就他自己而言，在”初步报告中缺乏数学严谨性”。

明斯基尽管对罗森布拉特和他的机器”极其钦佩”，但开始”担心这样的机器无法做什么”。1969年，他和他的MIT同事西摩·帕珀特出版了一本名为《感知器》的书，有效地关闭了整个研究方向的大门。明斯基和帕珀特用数学证明的严格形式表明，存在一些看似基本的模式，罗森布拉特的模型根本永远无法识别。例如，不可能训练罗森布拉特的机器之一来识别一张卡片上有奇数还是偶数个方格。识别像这样更复杂类别的唯一方法是使用具有多层的网络，早期层创建原始数据的表示，后期层在表示上操作。但没有人知道如何调整早期层的参数来使表示对后期层有用。该领域遇到了相当于砖墙的障碍。明斯基说：“直到1969年，已经发表了几千篇关于感知器的论文。”

“我们的书停止了这些。”

就好像一团乌云笼罩在该领域上，一切都分崩离析：研究、资金、人员。皮茨、麦卡洛克和莱特文，这三人都搬到了MIT，在与MIT的诺伯特·维纳发生误解后被严厉放逐，维纳曾经像皮茨的第二个父亲一样，现在却不愿与他说话。酗酒和抑郁的皮茨将他所有的笔记和论文都扔进了火中，包括一篇关于三维神经网络的未发表论文，MIT拼命试图抢救这篇论文。皮茨于1969年5月死于肝硬化，年仅46岁。几个月后，70岁的沃伦·麦卡洛克在一系列心肺问题后死于心脏病发作。1971年，在庆祝43岁生日时，弗兰克·罗森布拉特在切萨皮克湾的一次帆船事故中溺水身亡。

到了1973年，美国和英国政府都撤回了对神经网络研究的资金支持。当一位名叫Geoffrey Hinton的年轻英国心理学学生宣布他想做神经网络的博士研究时，一次又一次地得到同样的回复：“Minsky和Papert，”他们告诉他，“已经证明了这些模型没有用。”¹⁰

AlexNet的故事

2012年在多伦多，Alex Krizhevsky的卧室热得无法入睡。他的计算机连接着两块Nvidia GTX 580 GPU，已经连续两周日夜不停地以最大热负荷运行，风扇不断排出热风。

“非常热，”他说。“而且很吵。”¹¹

他正在教机器如何观看。

Geoffrey Hinton，Krizhevsky的导师，现在64岁了，并没有放弃。有理由保持希望。

到了1980年代，人们明白了具有多层的网络（所谓的“深度”神经网络）确实可以像浅层网络一样通过例子进行训练。¹²“我现在相信，”Minsky承认，“那本书过头了。”¹³

到了80年代末和90年代初，Hinton的一位前博士后Yann LeCun在贝尔实验室工作，他训练神经网络识别0到9的手写数字，神经网络找到了它们的第一个重要商业用途：在邮局读取邮政编码，在ATM机上读取存款支票。¹⁴到了1990年代，LeCun的网络处理着美国10%到20%的所有支票。¹⁵

但这个领域遇到了另一个瓶颈，到了2000年代，研究人员仍然主要困在手写邮政编码数据库的摆弄中。人们理解，原则上，一个足够大的神经网络，有足够的训练例子和时间，几乎可以学习任何东西。¹⁶但没有人有足够的计算能力、足够的训练数据，或足够的耐心来实现那种理论潜力。许多人失去了兴趣，计算机视觉领域，连同计算语言学，很大程度上转向了其他事情。正如Hinton后来总结的那样，“我们的标记数据集小了数千倍。[而且]我们的计算机慢了数百万倍。”¹⁷然而，这两件事都将发生改变。

随着网络的发展，如果你想要的不是五十张而是五十万张你网络的“闪卡”，突然间你有了一个似乎无底的图像存储库。只有一个问题，就是它们通常没有现成的类别标签。除非你知道网络的输出应该是什么，否则你无法训练网络。

2005年，亚马逊推出了其“Mechanical Turk”服务，允许大规模招募人力，使得雇佣数千人以每次点击几便士的价格执行简单动作成为可能。（这项服务特别适合那些未来AI被认为能够做的事情——因此它的标语是：人工人工智能。）2007年，普林斯顿教授李飞飞使用Amazon Mechanical Turk以前所未有的规模招募人力，构建了一个之前不可能的数据集。建设花费了两年多时间，包含三百万张图像，每张都由人手标记到五千多个类别中。李将其称为ImageNet，并于2009年发布。计算机视觉领域突然有了大量新数据可以学习，以及一个新的重大挑战。从2010年开始，来自世界各地的团队开始竞争构建一个能够可靠地看着一张图像——尘螨、集装箱船、摩托车、豹子——并说出它是什么的系统。

与此同时，整个2000年代摩尔定律相对稳定进步意味着计算机可以在几分钟内完成1980年代计算机需要数天才能完成的工作。然而，还有一个进一步的发展被证明是至关重要的。在1990年代，视频游戏行业开始生产称为GPU的专业图形处理器，设计用于实时渲染复杂的3D场景；与传统CPU一个接一个地完美精确执行指令不同，它们能够同时进行大量简单且有时近似的计算。¹⁸直到后来，在2000年代中期，人们才开始认识到GPU能做的远不止光线、纹理和阴影。¹⁹事实证明，这种为计算机游戏设计的硬件，实际上是为训练神经网络量身定制的。

在多伦多大学，Alex Krizhevsky参加了一门为GPU编写代码的课程，并决定在神经网络上尝试。他将自己应用于一个名为CIFAR-10的流行图像识别基准测试，该测试包含缩略图大小的图像，每张都属于十个类别之一：飞机、汽车、

鸟、猫、鹿、狗、青蛙、马、船或卡车。Krizhevsky构建了一个网络，并开始使用GPU训练它对CIFAR-10图像进行分类。令人震惊的是，他能够将他的网络从随机起始配置一路训练到最先进的准确性。在八十秒内。²⁰

正是在这个时候，Krizhevsky的实验室同事Ilya Sutskever注意到了，并向他提出了一个将成为某种塞壬之歌的建议。“我打赌，”Sutskever说，“你可以让它在ImageNet上工作。”

他们构建了一个庞大的神经网络：650,000个人工神经元，排列成8层，由6000万个可调整的权重连接。在他父母家的卧室里，Krizhevsky开始向它展示图片。

系统一步步地，一点点地，变得更加准确了几个百分点。

数据集——尽管它很大，有几百万张图片——还是不够。但是Krizhevsky意识到他可以造假。他开始进行“数据增强”，给网络输入数据的镜像图片。这似乎有帮助。他给它输入略微裁剪或略微着色的图片。（毕竟，当你前倾或侧倾，或从自然光变为人工光时，猫看起来仍然像猫。）这似乎有帮助。

他尝试不同的架构——这个层数，那个层数——或多或少地盲目摸索可能恰好效果最好的配置。

Krizhevsky偶尔会失去信心。Sutskever从不会。他一次又一次地激励Krizhevsky。你能让它工作。

“Ilya就像一个宗教人物，”他说。“有一个宗教人物总是好的。”

尝试模型的新版本，并训练它直到准确率达到最大值，需要大约两周时间，一天二十四小时运行——这意味着这个项目，虽然在某种程度上很疯狂，但也有很多停机时间。Krizhevsky思考。修补。等待。Hinton提出了一个叫做“dropout”的想法，在训练期间网络的某些部分会被随机关闭。Krizhevsky尝试了这个，由于各种原因，它似乎有帮助。他尝试使用具有所谓“修正线性”输出函数的神经元。这也似乎有帮助。

他在ImageNet竞赛截止日期9月30日提交了他的最佳模型，然后最后的等待开始了。

两天后，Krizhevsky收到了来自斯坦福大学Jia Deng的邮件，他正在组织那年的竞赛，邮件抄送给了所有参赛者。用平淡、不带感情的语言，Deng说点击提供的链接查看结果。

Krizhevsky点击了提供的链接，看到了结果。

他的团队不仅获胜了，而且他们彻底击败了整个其他领域。在他卧室里训练的神经网络——它的官方名称是“SuperVision”，但历史将简单地记住它为“AlexNet”——犯的错误只有第二名模型的一半。

到了会议的星期五，当ImageNet大规模视觉识别挑战研讨会开始时，消息已经传开了。Krizhevsky被安排在当天的最后一个演讲时段，下午5:05他站在演讲台前。他环顾四周。前排坐着Fei-Fei Li；旁边是Yann LeCun。看起来世界上大多数领先的计算机视觉研究人员都在这里。房间超员了，人们站在过道和墙边。

“我很紧张，”他说。“我不舒服。”

然后，在站着的观众面前，不舒服的Alex Krizhevsky告诉了他们一切。

当Frank Rosenblatt在1958年接受关于他的感知机的采访时，被问到像感知机这样的机器可能有什么实际或商业用途。“目前，完全没有，”他开朗地回答道。

“在这些问题上，你知道，用途跟随发明。”

问题

2015年6月28日星期日晚上，网页开发者Jacky Alciné在家观看BET颁奖典礼时，收到一个通知说朋友通过Google Photos与他分享了一张照片。当他打开Google Photos时，他注意到网站已经重新设计了。“我想，‘哦，UI改变了！’我记得I/O [Google的年度软件开发者大会]举行了，但我很好奇；我点击了进去。”Google的图像识别软件自动识别了照片组，并给每组一个主题标题。“毕业”，其中一个说——Alciné对系统能够识别他弟弟头上的学位帽和流苏印象深刻。另一个标题让他愣住了。相册封面是Alciné和他朋友的自拍照。Alciné是海地裔美国人；他和他的朋友都是黑人。

“大猩猩，”它说。

“所以我想——说实话，我以为是我做了什么。”他打开相册，以为他不知怎么误点击或误标记了什么。相册里满是Alciné和他朋友的几十张照片。没有别的。“我想——这是七十多张照片。不可能。。。那时我才真正意识到发生了什么。”

Alciné发推特。“Google Photos，”他写道，“你们搞砸了。我朋友不是大猩猩。”

两小时内，Google+首席架构师Yonatan Zunger联系了他。“天哪，”他写道。“这100%不行。”

Zunger的团队在另外几小时内向Google Photos部署了更改，到第二天早上，只有两张照片仍然被错误标记。然后Google采取了更激烈的步骤：他们完全删除了该标签。

事实上，三年后的2018年，Wired报道说”大猩猩”标签在Google Photos上仍然被手动停用。这意味着，几年后，没有任何东西会被标记为大猩猩，包括大猩猩。

奇怪的是，2018年的新闻媒体，就像2015年一样，似乎反复误解了这个错误的性质。标题宣称，“两年后，Google通过从图像分类器中清除‘大猩猩’标签解决了‘种族主义算法’问题”；“Google通过从其图像标记技术中删除大猩猩‘修复’了其种族主义算法”；以及“Google Images‘种族主义算法’有了修复方案，但这不是一个很好的方案。”

作为一名程序员，Alciné对机器学习系统很熟悉，他知道问题不在于有偏见的算法。（该算法是随机梯度下降，这几乎是计算机科学中最通用、最基础、最万能的思想：随机遍历训练数据，调整模型参数以为该图像的正确类别分配稍高的概率，然后根据需要重复。）不，他立即意识到问题出在训练数据本身。“我甚至不能责怪算法，”他说。“算法根本没有错。它完全按照设计的方式运行。”

当然，一个理论上可以从一组示例中学习几乎任何东西的系统的问题在于，它会受到用来教授它的示例的支配。

校准与设计霸权

我们对日常物品习以为常的程度，正是它们统治和影响我们生活的程度。

—玛格丽特·维瑟²⁶

十九世纪被拍摄最多的美国人——比亚伯拉罕·林肯或尤利西斯·格兰特还要多——是弗雷德里克·道格拉斯，这位废奴主义作家和演讲家在二十岁时逃脱了奴隶制。²⁷ 这绝非偶然；对道格拉斯来说，摄影和散文或演讲一样重要。摄影在1840年代通过银版照相法刚刚兴起，道格拉斯立即理解了它的力量。

在摄影出现之前，对美国黑人的描绘仅限于素描、绘画和版画。道格拉斯写道：“黑人永远无法从白人艺术家手中获得公正的肖像。”“在我们看来，白人为黑人画肖像时，几乎不可能不极度夸大他们的显著特征。”²⁸ 在道格拉斯的时代，有一种夸大手法特别盛行。“我们有色人种经常看到自己被描述和画成猴子，所以我们认为能找到这种普遍规则的例外是很幸运的事。”²⁹

摄影不仅反驳了这种漫画式描绘，而且使一种超越性的共情和认同成为可能。道格拉斯在谈到第一位美国黑人参议员希拉姆·里维尔斯的照片时说：“无论那些看到它的人可能有什么偏见，他们都将被迫承认这位密西西比参议员是一个人。”³⁰

但情况并非完全理想。随着摄影在二十世纪变得更加标准化和大规模生产，一些人开始认为摄影领域本身值得批评。正如W.E.B.杜波依斯在1923年写道：“为什么没有更多年轻的有色男女将摄影作为职业？普通的白人摄影师不知道如何处理有色皮肤，既没有感受细腻美感或色调的能力，也没有学习的意愿，他在描绘他们时制造了可怕的混乱。”

我们经常听到电影和电视缺乏多样性的问题——无论是演员还是导演——但我们很少考虑这个问题不仅存在于镜头前，不仅存在于镜头后，在许多情况下还存在于镜头内部。正如康考迪亚大学传播学教授洛娜·罗斯指出的：“尽管现有的学术文献范围很广，但令人惊讶的是，相对较少的学者将研究重点放在视觉再现设备本身的肤色偏见上。”³¹

几十年来，她写道，胶片制造商和胶片冲印商使用测试照片作为色彩平衡基准。这张测试照片被称为“雪莉卡”，以柯达员工雪莉·佩奇命名，她是第一个为此拍照的模特。³² 不言而喻，雪莉和她的继任者绝大多数都是白人。胶片的化学处理相应地进行了调整，结果相机根本无法为黑人拍出好照片。

（在视频中就像在摄影中一样，几十年来颜色一直以白人皮肤为标准进行校准。在1990年代，罗斯采访了《周六夜现场》的一位摄像师，询问广播前调试摄像机的过程。他解释说：“一个好的VCR操作员会让一个肤色女孩站在摄像机前，在技术人员专注于她的肤色进行精细调整以平衡摄像机时保持在那里。这个肤色女孩总是白人。”）³³

令人惊讶的是，柯达高管在1960和70年代描述制造对更广泛的深色调敏感的胶片的主要推动力，不是来自民权运动，而是来自家具和巧克力行业，他们抱怨胶片没有正确显示深色木材的纹理，或者牛奶巧克力和黑巧克力之间的区别。³⁴

柯达研究工作室前经理厄尔·凯奇回忆这段研究时期：“我的小部门因为巧克力变得相当丰满，因为放在摄像机前的东西在拍摄结束时被消费掉了。”当被问及这一切都发生在民权运动背景下的事实时，他补充说：“令人着迷的

是，这以前从未被提及，因为据我当时所知，黑人肌肤从未被视为一个严重问题。”³⁵

随着时间推移，柯达开始使用更多不同肤色的模特。“我开始在我们的测试中大量使用黑人模特，这很快就流行起来了，”柯达的Jim Lyon回忆道。“我并不是试图政治正确。我只是想给我们一个机会制作更好的胶卷，一种能以适当方式再现每个人肤色的胶卷。”

到1990年代，官方的柯达Shirley卡现在有三个不同种族的模特。他们的Gold Max胶卷——最初以能够拍摄“低光下的深色马匹”为卖点进行营销——现在在电视广告中展示了多元化的家庭。其中一个广告描绘了一个穿着亮白色空手道服的黑人男孩，在表演套路时微笑着，大概是在接受下一个级别的腰带。广告说：“家长们，除了柯达Gold胶卷，你们还会把这个时刻交给其他任何东西吗？”

他们最初的目标受众给了他们一个有问题的校准标准。现在新的校准标准给了他们新的受众。

修复训练集

从感知机(perceptron)开始，所有machine-learning系统的核心都有一种Shirley卡：即训练它们的数据集。如果某种类型的数据在训练数据中代表性不足或缺失，但在现实世界中存在，那么一切都无法保证。^[36]

正如UC Berkeley的Moritz Hardt所论证的，“关于大数据的整个说辞是，我们可以主要通过拥有更多数据来构建更好的分类器。逆命题是更少的数据导致更差的预测。不幸的是，根据定义，关于少数群体的数据总是相对较少。这意味着我们关于少数群体的模型通常比关于普通人群的模型要差。”^[37]

Alciné在事件当晚沮丧的推文完全呼应了这种情绪。他是一名软件工程师。他立即诊断出了问题所在。他推断，Google Photos中黑人的照片远没有白人的照片多。因此，模型看到任何不熟悉的东西时，更容易出错。

“再次强调，我完全能理解这是怎么发生的，” Alciné告诉我。^[38] “就像如果你拍一张苹果的照片，但只拍红苹果，当它看到绿苹果时可能会认为是梨……诸如此类的小事。这个我理解。但是，你是世界的——你的使命是索引整个世界的社会知识，那么你怎么能，就这样跳过整个大陆的人呢？”

二十世纪的问题似乎在二十一世纪以不可思议的方式重复出现。幸运的是，一些解决方案似乎也在重复。只需要有人愿意质疑这些二十一世纪” Shirley卡” 中究竟代表了谁和什么，以及更好的卡片可能是什么样子。

当Joy Buolamwini在2010年代初期在Georgia Tech读计算机科学本科时，她被分配了一个任务：编程一个机器人玩躲猫猫。编程部分很容易，但有一个问题：机器人无法识别Buolamwini的脸。“我借用了室友的脸来完成项目，提交了作业，然后想，‘你知道吗，别人会解决这个问题的。’”^[39]

在本科学习后期，她前往香港参加创业竞赛。一家当地初创公司正在演示其” 社交机器人” 之一。演示对旅游团中的每个人都有效……除了Buolamwini。碰巧的是，这家初创公司使用的正是她自己在Georgia Tech时使用的完全相同的现成开源人脸识别代码。

在第一批明确解决计算系统偏见概念的文章之一中，华盛顿大学的Batya Friedman和康奈尔大学的Helen Nissenbaum警告说，“例如，计算机系统传播成本相对较低，因此，一旦开发出来，有偏见的系统就有广泛影响的潜力。如果系统成为该领域的标准，偏见就会变得无处不在。”^[40]

或者，正如Buolamwini自己所说，“在地球的另一端，我了解到算法偏见的传播速度可以和从互联网上下载一些文件一样快。”^[41]

在牛津获得罗德奖学金后，Buolamwini来到MIT媒体实验室，在那里她开始研究一个增强现实项目，她称之为” Aspire Mirror”。这个想法是将赋权或振奋的视觉效果投射到用户的脸上——例如，让旁观者变成狮子。同样，只有一个问题。Aspire Mirror只有在Buolamwini戴上白色面具时才对她起作用。

罪魁祸首不是随机梯度下降(stochastic gradient descent)；显然，是训练这些系统的图像集。每个人脸检测或人脸识别系统的背后和内部都隐含着一组图像——通常是数万或数十万张——系统最初就是在这些图像上训练和开发的。这些训练数据，二十一世纪的Shirley卡，往往是不可见的，或被视为理所当然，或完全缺失：在线传播的预训练模型几乎从不包含其训练数据。但它确实存在，并将永久塑造部署系统的行为。

因此，根除偏见的一个主要运动是试图更好地暴露和更好地理解主要学术和商业machine-learning系统背后的训练数据集。

例如，一个较受欢迎的面部图片公共数据库是被称为 Labeled Faces in the Wild (LFW) 数据集，该数据集由 UMass Amherst 的一个团队在 2007 年从在线新闻文章和图片说明中精心收集而成，此后被无数研究人员使用。⁴² 然而，这个数据库的组成直到多年后才得到深入研究。2014 年，密歇根州立大学的 Hu Han 和 Anil Jain 分析了该数据集，确定其超过 77% 为男性，超过 83% 为白人。⁴³ 数据集中最常见的个体是 2007 年在线新闻照片中出现最频繁的人：时任总统乔治·W·布什，有 530 张独特图片。事实上，LFW 数据集中乔治·W·布什的图片数量是所有黑人女性图片总和的两倍多。⁴⁴

描述该数据库的 2007 年原始论文指出，从在线新闻文章收集的图片集“显然有其自身的偏见”，但这些“偏见”是从技术而非社会角度考虑的：“例如，在极端光照条件或非常低光照条件下出现的图片并不多。”除了此类光照问题外，作者写道，“呈现的图片范围和多样性非常大。”

然而，十二年后，在 2019 年秋季，Labeled Faces in the Wild 数据集网页上突然出现了一个持不同观点的免责声明。它指出，“许多群体在 LFW 中没有得到很好的代表。例如，儿童很少，没有婴儿，80 岁以上的人很少，女性比例相对较小。此外，许多种族的代表性很小或完全没有。”⁴⁵

近年来，人们更加关注这些训练集的构成，尽管仍有许多工作要做。2015 年，美国国家情报总监办公室和情报高级研究项目活动发布了一个名为 IJB-A 的面部图像数据集，他们声称具有“更广泛的地理变化主题”。⁴⁶ Buolamwini 与微软的 Timnit Gebru 一起对 IJB-A 进行了分析，发现其超过 75% 为男性，近 80% 为浅肤色。数据集中只有 4.4% 是深肤色女性。⁴⁷

最终 Buolamwini 意识到“其他人会解决这个问题”的那个人——当然——就是她自己。她开始对面部检测系统的现状进行广泛调查，这成为了她的 MIT 论文。她和 Gebru 首先着手构建一个在性别和肤色方面更平衡的数据集。但他们从哪里获取图片呢？以前的数据集，例如从在线新闻中提取的，完全不平衡。他们决定使用议会，编译了六个国家代表的数据库：卢旺达、塞内加尔、南非、冰岛、芬兰和瑞典。这个数据集在年龄、光照和姿势等方面明显不多样化，几乎所有主题都是中年或更老，居中在框架中，面对镜头表情中性或微笑。但是，按肤色和性别衡量，它可以说是迄今为止组装的最多样化的机器学习数据集。⁴⁸

握手这个议会数据集，Buolamwini 和 Gebru 研究了三个商业可用的面部分类系统——来自 IBM、微软和中国公司 Megvii（广泛使用的 Face++ 软件制造商）——并对每个系统进行了测试。

在整个数据集中，所有三个系统在正确分类主题性别方面都表现相当好——三家公司都约为 90%。在所有三种情况下，软件在男性面部上的准确率比女性面部高大约 10 到 20%，并且所有系统在浅肤色面部上也比深肤色面部准确大约 10 到 20%。但当 Buolamwini 对两者进行交叉分析时，迄今为止最明显的结果出现了。所有三个系统在分类既是深肤色又是女性的面部时表现显著更差。例如，IBM 的系统对浅肤色男性的错误率仅为 0.3%，但对深肤色女性为 34.7%：超过一百倍的差异。

废奴主义者和妇女权利活动家 Sojourner Truth 因其 1851 年的演讲“ Ain’t I a Woman? ”而最为人知。Buolamwini 感人地将这个问题回响到二十一世纪，指出 Truth 的照片被当代商业面部分类软件一次又一次地错误分类为男性。⁴⁹

2017 年 12 月 22 日，Buolamwini 联系了三家公司，解释她将在即将举行的会议上展示她的结果，并给每家公司回应的机会。Megvii 没有回应。微软回应了一个通用声明：“我们相信 AI 技术的公平性是行业的一个关键问题，微软非常重视这个问题。我们已经采取措施提高面部识别技术的准确性，并且我们继续投资研究以识别、理解和消除偏见。”⁵⁰ 然而，IBM 完全是另一回事。他们当天就回应了，感谢 Buolamwini 的联系，复制并确认了她的结果，邀请她到他们的纽约和剑桥校园，并在几周内宣布了他们 API 的新版本，对深肤色女性的错误率有十倍的改进。⁵¹

“改变是可能的，”她说。在技术上或其他方面，没有根本性的障碍阻止缩小这种性能差距；只是需要有人提出正确的问题。

Buolamwini和Gebru的工作突出了我们在公司宣布其系统“99%准确”时应该感到的怀疑：在什么情况下准确？对谁准确？这也提醒我们，每个机器学习系统都是一种议会，其中训练数据代表某个更大的选民群体——就像在任何民主制度中一样，确保每个人都能投票是至关重要的。⁵²

机器学习系统中的偏见往往是训练系统数据的直接结果——这使得在使用这些数据集训练将影响真实人群的系统之前，了解这些数据集中谁被代表以及代表程度变得极其重要。

但是，如果你的数据集尽可能包容——比如说，接近英文书面语的全部内容，大约一千亿个单词——而世界本身就是有偏见的，你该怎么办？

分布假设：词嵌入

通过一个词的伙伴，你就能了解这个词。

—J. R. FIRTH⁵³

假设你在海滩上发现了一个漂流瓶中的信息；信息的几个部分无法读清。你检查其中一句话：“我已经把宝藏埋在了海滩边的——北面。”不用说，你非常想弄清楚那个缺失的词可能是什么。

你可能不会想到那个词可能是”仓鼠”、“甜甜圈”或”假发”。这有几个原因。你可以运用一些常识：仓鼠很活跃，甜甜圈可生物降解，假发会被风吹走——它们都不是长期寻宝导航的可靠地标。你推理，任何在预计数月到数年的时间跨度内藏匿财宝的人，都需要一些稳定的、不太可能分解或移动的东西。

现在想象你是一台完全缺乏这种常识的计算机——更不用说设身处地为潜在的藏宝者着想的能力——但你确实有一个极其庞大的真实世界文本样本（一个”语料库”）来扫描模式。仅仅基于语言本身的统计数据，你能在预测缺失单词方面做得多好？

构建这类预测模型长期以来一直是计算语言学家的圣杯。⁵⁴（实际上，Claude Shannon在1940年代基于对这种分析的数学分析创立了信息论，注意到一些缺失的词比其他词更可预测，并试图量化其程度。⁵⁵）早期方法涉及所谓的”*n*-grams”，这意味着简单地计算在特定语料库中出现的每一个连续词链（比如两个词）——“appeared in”、“in a”、“a particular”、“particular corpus”——并在一个巨大的数据库中统计它们。⁵⁶然后，给定一个缺失的词，查看前面的词并找到数据库中以该前面词开头的、出现频率最高的*n*-gram就足够简单了。那就是你对缺失内容的最佳猜测。当然，仅仅是紧接着的前一个词之外的额外上下文可以为你提供额外的线索，但整合它远非直截了当。从存储语言中所有可能的两词短语（“bigrams”）到所有三词短语（“trigrams”），或四词短语或更多，意味着将数据库增长到荒谬且难以维持的大小。此外，这些数据库变得极其稀疏，绝大多数可能的短语从未出现过，其余大部分只出现一次或两次。

理想情况下，我们还希望能够做出合理的猜测，即使特定短语在语料库中从未逐字出现过。这种基于计数的方法没有帮助。在句子”I sipped at a jaundiced——“中，我们可能会想象”chardonnay”比”charcoal”更可能，即使这两个词在语言历史上从未被”jaundiced”所修饰过。在这种情况下，依赖计数根本没有帮助——而且，问题在我们试图添加更多上下文时变得更糟，因为我们考虑的短语越长，我们越有可能从未见过某些东西。

这套问题被称为”维度诅咒”，从一开始就困扰着这种语言学方法。⁵⁷

有更好的方法吗？

有的，它以所谓的”分布式表示”形式出现。⁵⁸这个想法是试图通过某种抽象”空间”中的点来表示词语，其中相关词语彼此”更接近”。在1990年代和2000年代出现了许多这样做的技术，⁵⁹但过去十年中有一种技术表现出了特殊的前景：神经网络。⁶⁰

这里的假设，也就是模型所依赖的重大押注，简单来说就是：词汇会倾向于出现在与其”相似”的词汇附近。而这些相似性可以用数字来捕捉。神经网络模型的工作原理是将每个词汇转换（“嵌入”）为一组数字（“向量”），这些数字代表该词汇在空间中的”坐标”。这组坐标数字被称为该词汇的表示。（在word2vec的情况下，是300个介

于-1.0和1.0之间的十进制数字。) 这使得我们能够直接测量任何词汇与任何其他词汇的“相似度”：这些坐标之间的距离有多远？

我们所要做的就是——以某种方式——在这个空间中排列词汇，使它们能够尽可能好地预测这些缺失的词汇。（至少，我们会在这种特定模型架构允许的范围内做得尽可能好。）

我们如何获得这些表示呢？当然，通过随机梯度下降！我们首先将词汇随机分散在整个空间中。然后我们从语料库中随机选择一个短语，隐藏一个词汇，并询问系统预期什么可能填入那个空白。

当我们的模型猜错时，我们会调整词汇表示的坐标，在我们的数学空间中将正确的词汇稍微向上下文词汇推进，并将任何错误的猜测稍微推离。在我们做出这个微小的调整后，我们会随机选择另一个短语并再次经历这个过程。一次又一次。一次又一次。一次又一次。一次又一次。

“在这一点上，”斯坦福大学计算语言学家Christopher Manning解释道，“某种奇迹发生了。”

用他的话说：

这有点令人惊讶——但确实如此——你只需要设置这种预测目标，让每个词汇的词向量的任务就是使它们能够很好地预测出现在其上下文中的词汇，或者反之亦然——你只有这个非常简单的目标——除此之外你对如何实现这一点什么都不说——但你只是祈祷并依赖深度学习的魔力……然后这个奇迹发生了。产生的这些词向量在表示词汇含义方面非常强大，对各种事情都很有用。

事实上，有人可能会争论说，这些嵌入实际上设法捕捉了我们语言中过多的细微差别。确实，它们以惊人的清晰度捕捉到了我们自己不愿看到的部分。

嵌入的阴暗面

从人性这弯曲的木材中，从未造出过真正笔直的东西。

——康德

像这些词嵌入模型，包括Google的word2vec和斯坦福大学的GloVe，随后成为了计算语言学的实际标准，自大约2013年以来，几乎支撑着每一个涉及计算机语言使用的应用，无论是对搜索结果进行排名、将段落从一种语言翻译成另一种语言，还是分析书面评论中的消费者情感。

确实，这些嵌入，虽然简单——只是基于预测文本中附近缺失词汇的每个词汇的一行数字——似乎捕捉到了惊人数的现实世界信息。

例如，你可以简单地将两个向量相加得到一个新向量，然后搜索最近的词汇。正如我们所见，结果往往令人震惊地有意义：

捷克 + 货币 = 克朗越南 + 首都 = 河内德国 + 航空公司 = 汉莎航空法国 + 女演员 = Juliette Binoche

你也可以减去词汇。这意味着——令人难以置信的是——你可以通过获得两个词汇之间的“差异”然后将其“添加”到第三个词汇来产生“类比”。

这些类比表明嵌入捕捉到了地理信息：

柏林 - 德国 + 日本 = 东京

以及语法：

bigger - big + cold = colder

以及美食：

寿司 - 日本 + 德国 = 德式香肠

以及科学：

Cu - 铜 + 金 = Au

以及科技：

Windows - 微软 + 谷歌 = Android

以及体育：

蒙特利尔加拿大人队 - 蒙特利尔 + 多伦多 = 多伦多枫叶队

不幸的是，正如我们所见，向量捕捉到的不止这些。它们包含了惊人的性别偏见。对于每一个关于男人:女人的巧妙或恰当的类比，比如fella:babe，或前列腺癌:卵巢癌，都有许多其他的类比似乎只是在反映刻板印象，比如木工:缝纫，或建筑师:室内设计师，或医生:护士。

我们现在才开始充分认识到这个问题。“已经有数百篇关于词嵌入及其应用的论文，从网络搜索到解析简历，”Tolga Bolukbasi、Adam Kalai及其合作者写道。“然而，这些论文中没有一篇认识到嵌入是多么公然的性别歧视，因此有在现实世界系统中引入各种类型偏见的风险。”

Machine-learning系统不仅展现偏见，还可能悄无声息地、微妙地延续偏见。设想一位雇主正在搜索“软件工程师”候选人。搜索系统将根据某种“相关性”标准对数百万份简历进行排名，只展示排名最靠前的几份。一个天真地使用word2vec或类似技术的系统很可能会发现，John这个词在工程师简历中比Mary更为常见。因此，在其他条件相同的情况下，属于John的简历在“相关性”排名中会高于其他方面完全相同但属于Mary的简历。这样的例子不仅仅是假设性的。当就业律师Mark J. Girouard的一位客户在审查某潜在供应商的简历筛选工具时，审计显示整个模型中权重最高的两个积极因素之一就是名字“Jared”。该客户没有购买这个简历筛选工具——但据推测其他人购买了。

当然，我们已经知道，求职者的姓名会对真正的人类雇主产生影响。2001年和2002年，经济学家Marianne Bertrand和Sendhil Mullainathan寄出了近五千份简历，这些简历被随机分配了听起来像白人（Emily Walsh, Greg Baker）或非洲裔美国人（Lakisha Washington, Jamal Jones）的姓名。他们发现回电率存在惊人的50%差距，尽管简历本身完全相同。

Word2vec将专有名词映射到种族和性别轴上，就像处理任何其他词汇一样，将Sarah - Matthew放在性别轴上，将Sarah - Kiesha放在种族轴上。考虑到它也将职业放在这两个轴上，不难想象一个系统会无意中使用这样的种族或性别维度——实际上就是刻板印象——来为给定职位空缺的候选人提升或降低“相关性”排名。换句话说，如果是机器而不是人在筛选这些简历，我们有理由同样担心。

在人类情况下显而易见的解决方案——移除姓名——将不起作用。1952年，Boston Symphony Orchestra开始在演奏者和评委之间放置屏风进行试听，其他大多数管弦乐团在1970和80年代也纷纷效仿。然而，仅有屏风是不够的。管弦乐团意识到他们还需要指示试听者在走上试听厅的木地板之前，脱掉他们的鞋子。

机器学习系统的问题在于，它们被设计得恰恰是为了推断数据中的隐藏关联性。在某种程度上，假设男性和女性在总体上倾向于不同的写作风格——在措辞或语法上的微妙差异——word2vec将发现software engineer与所有典型男性措辞之间的大量微小和间接关联。这可能像在兴趣爱好中列出football而不是softball一样明显，像某些大学或家乡名字一样微妙，或者像对某个介词或同义词的轻微语法偏好一样几乎不可见。换句话说，这种性质的系统永远无法成功地被蒙住眼睛。它总能听到鞋声。

2018年，Reuters报道称Amazon工程师从2014年开始一直在开发一个机器学习工具来筛选在线简历，并根据候选人看起来多有前途，将可能的求职者从一星到五星进行排名——就像Amazon产品本身一样——Amazon招聘人员会据此集中精力。一位消息人士告诉记者：“他们真的希望它成为一个引擎，我给你一百份简历，它会筛选出前五名，然后我们雇用他们。”这个星级评分的标准是什么？使用词表示模型，与过去十年Amazon之前雇用的员工简历的相似性。

然而，到2015年，Amazon开始注意到问题。过去那些工程师雇员大多是男性。他们意识到，该模型给“women’s”这个词分配了负分——例如，在描述课外活动时。他们编辑了模型以消除这种偏见。

他们还注意到，模型给所有女子学院的名字都分配了负分。他们编辑了模型以消除这种偏见。

尽管如此，模型仍然找到了听到鞋声的方法。工程师们注意到，模型给看似所有词汇选择都分配正分——例如，像“executed”和“captured”这样在男性简历中比女性简历中更普遍的词汇。

到2017年，Amazon废弃了这个项目并解散了制作它的团队。

词嵌入去偏见

对于Tolga Bolukbasi和Adam Kalai，以及他们在BU和Microsoft的合作者来说，问题当然不仅仅是发现这些偏见，而是如何处理它们。

一个选择是在这个高维向量”空间”中找到捕捉性别概念的轴并删除它。但完全删除性别维度意味着失去像king:queen和aunt:uncle这样有用的类比。因此，正如他们所说，挑战在于”减少词嵌入中的性别偏见，同时保留嵌入的有用属性”。

事实上，即使识别正确的性别”维度”也很困难。例如，你可以将其定义为woman – man的向量”差异”。但这里涉及的不仅仅是性别——还有像”man oh man”这样的习语用法和动词形式，如”all hands, man your battle stations”。该团队决定采用多个不同的此类词对——woman – man，还有she – he、gal – guy等等——然后使用一种称为主成分分析(PCA)的技术来分离出能解释这些词对之间最大差异的轴：可能就是性别。⁷⁹

然后他们的任务是尝试确定，对于在这个性别维度上存在差异的词语，这种性别差异是合适的还是不合适的。比如说king和queen按性别分离是合适的，father和mother也是如此，但也许我们不希望像word2vec默认那样——将Home Depot视为JC Penney的性别翻转版本；或者alcoholism和eating disorders；或者pilot和flight attendant。

那么，如何从数十万个不同的词语中区分出有问题和没有问题的性别关联呢？如何知道哪些类比应该保留、哪些应该调整、哪些应该完全清除？

这个由五名计算机科学家组成的团队发现自己实际上在做社会科学。事实上，项目的一部分最终需要跨越他们正常学科界限的咨询。“我们是一群机器学习研究者，”Kalai说。“我在一个包含许多社会学家的实验室工作，仅仅通过听他们讨论社会学和社会科学中出现的各种问题，我们就意识到机器学习算法可能存在歧视的潜在担忧，但我们五个人中没有一个——我们都是男性——曾经研究过或者大量阅读过关于性别偏见的内容。”

这个小组，可能有些天真地，询问社会学家他们应该如何编码一个正式定义来区分哪些类比是可接受的，哪些不是。社会学家迅速打消了他们认为这种简单正式定义是可能的想法。“我们在想，我们如何定义最好的东西？”Bolukbasi说。“他们说，’社会学家无法定义什么是好的。’作为工程师，你想说，’好的，这就是理想状态，所以这是我的目标，所以我就要让我的算法达到那个目标。’因为它涉及太多的人和文化以及一切，你不知道什么是最优的。你无法为某些东西进行优化。在这个意义上，这实际上是非常困难的。”

该小组决定识别一组他们认为在某种本质或基本方式上适合被视为有性别特征的词语：像”he”和”she”、“brother”和”sister”这样的词，以及像”womb”和”sperm”这样的解剖学词汇，以及像”convent”和”monastery”或”sorority”和”fraternity”这样的社会词汇。其中一些需要复杂的决定——比如”nurse”这个词。作为名词，它是一个没有内在性别方面的职业，但作为动词，它可能是只有女性才能做的事情。像”rabbi”这样的词呢？这个词是否具有内在的性别维度取决于相关的犹太教派是正统派还是改革派。该团队尽了最大努力，在他们模型词典的子集中识别了218个这样的性别特定词汇，并让他们的系统推断到词典的其余部分。“请注意，词语的选择是主观的，”他们写道，“理想情况下应该根据具体应用进行定制。”⁸⁰对于这个集合之外的所有词语，他们将词语表示的性别成分设置为零。然后他们调整了所有性别相关词语的表示，使得等价术语对——比如”brother”和”sister”——以这个零点为”中心”。换句话说，它们被调整，使得两个词都不会在模型中被表示为比另一个更”性别特定”或更”性别中性”。

这个新的、去偏见的模型是一个改进吗？该团队从社会科学中借鉴了一个方法论页面，简单地询问人们。他们使用Amazon Mechanical Turk平台上的美国工作者将模型的许多类比分类为“刻板印象”或非刻板印象。即使在这里，社会学家的意见也是至关重要的。他们询问问题的确切措辞将很重要。“我们必须与他们交谈，因为当我们在Mechanical Turk上设计这些实验时，你问问题的方式实际上会产生影响，”Bolukbasi说。“这是一个如此敏感的话题。”⁸¹

结果令人鼓舞。原本默认模型返回doctor – man + woman为nurse，现在系统说的是physician。Mechanical Turk工作者报告说，原始模型的性别类比中有19%反映了性别刻板印象；在新的去偏见模型的类比中，只有6%被判断为反映刻板印象。⁸²

这种中和是有小代价的——例如，模型现在认为某人被“grandmothered in”到法律豁免中与“grandfathered in”同样可能。⁸³但也许这是一个值得付出的代价——你总是可以决定愿意用多少预测错误来换取多少去偏见，并设置适当的权衡。

正如团队所写：“对词嵌入中偏见的一种观点是，它仅仅反映了社会中的偏见，因此人们应该尝试消除社会偏见而不是词嵌入偏见。然而，…以一种小小的方式，去偏见的词嵌入有望有助于减少社会中的性别偏见。至少，机器学习不应该被用来无意中放大这些偏见，正如我们所看到的可能自然发生的那样。”⁸⁴

这是一个令人鼓舞的概念验证，证明我们可能能够在语言模型之上构建我们的系统，这些模型不仅捕获现状世界的现有状态，而且是一个更好世界的模型——一个我们想要的世界的模型。

尽管如此，故事还有更多内容。2019年，巴伊兰大学计算机科学家Hila Gonen和Yoav Goldberg发表了对这些“去偏见”表示的探索，并表明去偏见可能只是，正如他们所说的，“给猪涂口红”。⁸⁵是的，它移除了从诸如“护士”或“接待员”等职业到明确性别化术语如“女人”和“她”的链接。但这些刻板印象中“女性”职业之间的隐含联系——在“护士”和“接待员”之间——仍然存在。事实上，他们认为，这种仅部分去偏见实际上可能使问题变得更糟，因为它保留了大部分这些刻板印象关联完整无损，同时移除了那些最可见和最容易测量的关联。⁸⁶

现在在谷歌的Bolukbasi和他的同事们继续研究这个问题，确认在某些情况下，在招聘环境中使用的去偏见模型实际上可能比原始模型更糟。⁸⁷在这种情况下，完全删除性别维度的系统——即使是对于“他”和“她”等根本性别化术语——可能会产生更公平的结果。故事并不简单，工作仍在继续。

统计镜子中的自画像

在招聘应用中，这些偏见可能只是需要缓解的危险，但就其本身而言，它们提出了许多问题。例如，它们从何而来？它们是所使用的统计技术的人工制品，还是反映了更深层的东西：即我们头脑中的偏见和整个世界中的偏见？

社会科学中用于测试人类无意识偏见的经典测试是“内隐联想测试”，受试者会看到一系列单词，并被要求在单词属于两个不同类别中的任一类别时按按钮：例如，花朵（如“鸢尾花”）或愉快的事物（如“笑声”）。这听起来很简单，确实如此；故事不在于准确性，而在于反应时间。要求人们在单词是花朵或愉快事物时按按钮会产生快速反应时间，但要求他们在单词是花朵或不愉快事物时按按钮需要更长时间。这表明，在“花卉”和“愉快”的心理类别之间存在某种程度的重叠，或者它们反映了以某种方式相关联的概念。⁸⁸

发明这个测试的团队著名地证明了一组白人大学生能够快速识别单词是否是普遍的白人姓名（“Meredith”、“Heather”）或愉快的词（“lucky”、“gift”）。他们也能快速识别单词是否是普遍的黑人姓名（“Latonya”、“Shavonn”）或不愉快的词（“poison”、“grief”）。但当被要求在单词是白人姓名或不愉快的词时按按钮时，他们反应缓慢；同样，当他们必须在单词是黑人姓名或愉快的词时按按钮时，他们也反应缓慢。

普林斯顿大学的计算机科学家团队——博士后Aylin Caliskan和教授Joanna Bryson以及Arvind Narayanan——发现word2vec和其他广泛使用的词嵌入模型中嵌入之间的距离惊人地反映了这种人类反应时间数据。人们识别任何两组单词的速度越慢，这些词向量在模型中的距离就越远。⁸⁹换句话说，模型的偏见，无论好坏，都非常像我们自己的偏见。

除了这些内隐联想之外，普林斯顿团队还想知道像word2vec这样的模型是否捕获了他们所称的世界中的“真实”偏见。某些姓名确实更常给女性而不是男性，某些工作确实更常由女性而不是男性担任。在某种程度上，某些姓名获得比其他姓名更偏向男性或更偏向女性的表示，这是否在某种程度上反映了这种客观现实？在某种程度上，某些职业沿着模型的性别轴落在不同位置，这是否可能在某种程度上反映了某些职业——护士、图书管理员、木匠、机械师——确实恰好分布不均衡的事实？普林斯顿团队分别咨询了美国人口普查局和劳工统计局；在两种情况下，他们发现答案都是肯定的。

一个职业的词表示在性别方向上偏斜得越强烈，该性别在该职业中往往越容易被过度代表。他们写道：“词嵌入与美国50个职业中女性比例强烈相关。”^[90]在观察姓名时，他们发现了同样的情况，相关性只是稍微弱一些；但是，他们能够获得的最新人口普查数据是1990年的，因此姓名的性别分布可能自那时以来确实发生了轻微变化。

从这个“神奇”优化过程中产生的嵌入如此准确且令人不安地有用，成为社会的一面镜子，这一事实意味着我们实际上已经为社会科学的工具库增加了一个诊断工具。我们可以使用这些嵌入来精确详细地量化某个特定时间快照下社会的某些方面。无论因果关系如何——是客观现实的变化改变了我们说话的方式，还是反之，或者两者都由某种更深层的原因驱动——我们都可以使用这些快照来观察社会变化。

这正是由斯坦福大学的Nikhil Garg和James Zou领导的一个跨学科小组着手要做的事情。Garg是电气工程博士候选人，Zou是生物医学数据科学助理教授，他们与历史学家Londa Schiebinger和语言学家Dan Jurafsky合作，不仅使用当代文本语料库，还使用过去一百年的样本来研究词嵌入。^[91]

呈现出来的是文化变化风向的丰富而详细的历史。正如他们所说：“嵌入的时间动态有助于量化20世纪和21世纪美国对女性和少数族裔的刻板印象和态度的变化。”

斯坦福小组证实了普林斯顿小组关于职业词汇表示与性别之间联系的发现，并补充说似乎存在某种“男性基线”：即，我们从人口普查数据中知道在男女之间平均分配的职业，在其词嵌入中仍然略微偏向“男性”方向。正如作者解释的那样：“通用语言比基于外部客观指标所预期的更有偏见。”然而，撇开基线不谈，存在一个一致的时间趋势，表明职业词嵌入中的性别偏见与劳动力本身的变化同步移动。

通过观察跨时间的文本，他们发现了反映社会变化的丰富叙述。数据显示，性别偏见随着时间的推移总体上有所减少，特别是“1960年代和1970年代的女性运动对女性在文学和文化中的描绘产生了系统性和剧烈的影响。”

嵌入还显示了种族态度转变的详细历史。例如，在1910年，与白人相比最强烈地与亚洲人相关的前十个词包括“野蛮的”、“可怕的”、“可恶的”和“奇异的”。到1980年，情况截然不同，前十个词以“压抑的”和“被动的”为首，以“敏感的”和“热忱的”结尾：当然，这些本身就是刻板印象，但它们反映了明显的文化变化。

更近期的文化转变在嵌入中也是可见的——例如，与伊斯兰教相关的词汇和与恐怖主义相关的词汇之间的关联在1993年（世贸中心爆炸案那年）和2001年（9/11事件那年）都急剧上升。

人们甚至可以想象使用这种方法不是回顾性地而是前瞻性地进行观察：比如说，过去六个月的数据是否表明这些偏见正在变好还是变坏？人们可以想象一种实时仪表板，显示社会本身——或者至少是我们的公共话语——是否显得更有偏见或更少偏见：这是正在进行的转变的风向标，也是对未来世界的一瞥。

表示与代表

这里有几个要点，其中第一个主要是（虽然不纯粹是）方法论上的。计算机科学家在开始更广泛地思考他们构建的模型中包含什么时，正在向社会科学伸出援手。同样，社会科学家也在向machine learning社区伸出援手，并发现他们现在拥有了一个强大的新显微镜。正如斯坦福作者所写：“在标准的定量社会科学中，machine learning被用作分析数据的工具。我们的工作展示了machine learning的产物（这里是词嵌入）本身如何成为社会学分析的有趣对象。我们相信这种范式转变可以带来许多富有成果的研究。”

第二点是，偏见和内涵——虽然它们看起来如蛛丝般脆弱、空灵、难以言喻——是真实的。它们是可测量的，详细而精确。它们从仅仅为了预测缺失词汇而构建的模型中自发且可靠地出现，并且是可测量、可量化和动态的。它们追踪关于劳动参与的真实数据以及态度和刻板印象的主观测量。所有这些以及更多都存在于表面上只是从上下文预测缺失词汇的模型中：我们语言的故事就是我们文化的故事。

第三：这些模型在使用时绝对应该谨慎，特别是当用于预测缺失单词的初始目的以外的任何用途时。Adam Kalai说：“我和一些人交谈过，他们说在阅读了我们的论文后……他们在使用这些词嵌入时更加谨慎了——或者至少在自己的应用中使用之前会三思而后行。所以，这是一个积极的结果。”普林斯顿团队也呼应了这种谨慎态度：“当然，”他们写道，“在将通过无监督机器学习构建的模块纳入决策系统时必须谨慎使用。”⁹²很少有亚马逊高管会明确声明“雇用那些如果十年前申请的话，最像我们当时雇用的人的求职者”这样的政策。但使用语言模型来筛选简历的“相关性”正是在做这样的跳跃。

我们发现自己正处于历史的一个脆弱时刻——这些模型的强大和灵活性使它们在大量商业和公共应用中具有不可抗拒的实用性，然而我们关于如何适当使用它们的标准和规范仍处于萌芽状态。正是在这个时期，我们应该最为谨慎和保守——更何况这些模型中的许多在部署到现实世界使用后不太可能发生实质性改变。正如普林斯顿的Arvind Narayanan所说：“与‘技术发展太快，社会跟不上’的陈词滥调相反，技术的商业部署往往进展缓慢——看看银行和航空公司仍在运行的大型机就知道了。今天训练的机器学习模型可能在50年后仍在生产中使用，这很可怕。”⁹³

建模世界的现状是一回事。但一旦你开始使用那个模型，你就在改变世界，无论影响大小。许多机器学习模型背后有一个广泛的假设，即模型本身不会改变它所建模的现实。在几乎所有情况下，这都是错误的。

事实上，这些模型的粗心部署可能会产生一个反馈循环，从中恢复变得越来越困难，或者需要越来越大的干预。比如说，如果一个简历搜索系统检测到某个职位存在性别偏斜，并且以夸大这种偏斜的方式提升(比如)男性申请者的排名，那么这很可能成为模型学习的下一批训练数据。而它只会学到其现有偏见的更极端版本。当然，最容易进行干预的时刻是尽早。

最后，这些模型为我们提供了一个数字六分仪，让我们能够展望社会的未来。从这项工作中，我们不仅得到了历史的画像，也得到了最新现状的画像。只要每天都有新的文本在网上发布，就会有新的数据集可供采样。

如果明智地使用——并且是描述性而非预测性地使用——这些能够强化和延续社会潜在偏见的系统反而可以用来使这些偏见变得可见、无可争辩。它们为我们提供了一个衡量看似分散或无形事物的标尺。⁹⁴这是一个开始。

不再在谷歌工作的Yonatan Zunger认为，人们有时会忘记工程学与人类社会、人类规范和人类价值观密不可分的程度。“本质上，工程学完全关乎合作、协作，以及对同事和客户的同理心，”他写道。“如果有人告诉你工程学是一个可以不与人或情感打交道的领域，那么我很遗憾地告诉你，你被骗了。”⁹⁵

至于现在经营自己软件咨询公司并仍与Zunger保持联系的Jacky Alciné，他同意这个问题既不始于技术，也不终于技术。“这实际上是我想从事历史教学的部分原因，”他微笑着告诉我，至少有一半是认真的：“当我35岁时，我会停止一切，退休，然后转向历史。”

* 第二名是Vanessa Paradis，第三名是Charlotte Gainsbourg。

虽然人类自冰川退缩以来一直在美洲大陆上游荡，可能在冰河时代之前就已如此，但科学刚刚着手解决的问题之一——准确预测一个人在假释出狱后会做什么——这仍然在某种程度上是对人类进化的悲哀评论。

—芝加哥论坛报，1936年1月¹

我们的法律惩罚人们的行为，而不是他们的身份。基于不可改变的特征进行惩罚完全违背了这一指导原则。

—最高法院首席大法官约翰·罗伯茨²

当我们即将使用机器学习在教育、就业、广告、医疗保健和警务等领域对人类做出基本上所有类型的重要决定时，理解为什么机器学习在默认情况下并不公平或公正是很重要的。

—MORITZ HARDT³

用数字模型取代特殊的人类判断来使社会更加一致、更加准确、更加公平

社会可以通过用数字模型取代特殊的人类判断来变得更加一致、更加准确、更加公平，这个想法并不新鲜。事实上，它们甚至在刑事司法中的使用已有近一个世纪的历史。

1927年，伊利诺伊州假释委员会新任主席辛顿·克拉博（Hinton Clabaugh）委托进行一项关于该州假释制度运作的研究。他的动机源于他所察觉到的一种创新差距：“虽然我们的工业和政府机器远非完美，但在工业方面我们可能是最具创造力和效率的国家，”克拉博写道。“我们能够如实地说明我们的执法也是如此吗？”⁴尽管伊利诺伊州是最早颁布假释法的州之一，但公众舆论已经变得不利。正如克拉博观察到的，公众情绪认为“正义与仁慈的钟摆已经极端地偏向了罪犯一边。”他自己的观点也相差不远：整个假释概念可能无非是过度的宽容，假释制度也许应该完全废除。但他的理由是，根据美国法律，个人有权获得辩护——因此假释制度也应该有权获得辩护。

克拉博要求该州最负盛名的学校——伊利诺伊大学、西北大学和芝加哥大学——联合起来，准备一份关于假释制度的综合报告，并在一年内提交给他。伊利诺伊大学法学院院长阿尔伯特·哈诺（Albert Harno）将报告假释委员会的运作情况；西北大学的安德鲁·布鲁斯（Andrew Bruce）法官将回顾伊利诺伊州刑罚制度的历史（包括对十九世纪“废除鞭刑”的令人震惊的考察）；芝加哥社会学家欧内斯特·伯吉斯（Ernest Burgess）被给予了一个有趣的挑战，即研究是否有任何因素能够预测给定假释犯的“成功或失败”。

正如伯吉斯所写：

目前，伊利诺伊州人民心中对假释人员有两种截然不同的印象。一种印象是顽固、恶毒、绝望的罪犯，他们从监狱回来时毫无悔改之意，一心只想对让他痛苦忍受惩罚的社会进行报复。另一种印象是一个年轻人，也许是寡母的独子，他在冲动中，在软弱的时刻，屈服于误入歧途同伴的恶意建议，现在从教养院回到社会，决心只要给予机会就要好好表现。⁵

当然，问题是是否有可能预测哪些潜在的假释犯属于哪一种。

伯吉斯收集了大约三千名不同伊利诺伊州假释犯的数据，并尽力将他们分类为四个群体之一：“初犯者”、“偶犯者”、“惯犯者”和“职业犯罪者”。从二十一世纪的角度来看，他的一些工作似乎明显过时了：例如，他将人们划分为八种可能的“社会类型”：「流浪汉」、「无用之人」、「卑劣公民」、「酒鬼」、「歹徒」、「新移民」、「农家子弟」和「吸毒者」。尽管如此，伯吉斯的工作在当时看来相当彻底——考察犯罪历史、工作历史、居住历史、犯罪类型、刑期长度、服刑时间、精神病诊断等等。他进行了一项研究，以确定数据是否显示某些囚犯受到了不当的政治影响，并质疑公众对司法系统的厌倦观点是否有根据。在最后一章中，他直接处理了一个将在刑事司法领域掀起某种运动的问题，这一运动将持续到我们这个世纪：“科学方法能否应用于假释管理？”

他写道：“许多人会坦率地怀疑将科学方法引入任何人类行为领域的可行性。他们会以人性太多变，无法对其进行任何预测的断言来否定这一提议。但在分析决定假释成功和失败因素的过程中，已经发现了一些显著的对比。”

例如，伯吉斯注意到，在那些有良好工作记录、高智力、农业背景且服刑一年或以下的人中，假释违法率是州平均水平的一半。另一方面，在那些生活在“犯罪黑社会”中、其检察官或法官曾反对宽大处理、服刑五年或以上的人中，假释违法率是州平均水平的两倍。“这些显著的差异与我们已经了解的塑造个人生活条件相符，”他问道，“难道不是暗示应该比以前更认真、更客观地考虑这些因素吗？”

“对假释委员会来说，设计一份即将假释的每个人的摘要表是完全可行的，也应该是有帮助的，”他说，“以便其成员能够一目了然地看到每个重要因素的违法率……预测在任何给定情况下都不会是绝对的，但根据平均法则，将适用于任何相当数量的案例。”

他的结论是，在许多情况下，改造是完全可能的。更重要的是，在哪些情况下改造会成功似乎至少在某种程度上是可以预测的。一个建立在这种统计基础上的系统，难道不比现状更好吗？现状是法官临时做出的主观、不一致和特异性的决定。“毫无疑问，确定支配假释者成功或失败的因素是可行的，”Burgess写道。“人类行为似乎在某种程度上是可以预测的。这些记录的事实是囚犯获得假释的基础吗？还是假释委员会依赖于该人在听证会时给其成员留下的有利或不利印象？”

报告的最终结论既清晰又坚定。假释和“不定期判决”确实应该在伊利诺伊州继续实行——同样重要的是，“它们的管理可以而且应该得到改善，方法是将假释委员会的工作建立在科学和专业的基础上，并进一步防范政治影响的持续压力。”

主席Clabaugh阅读了这份报告，他完全改变了对自己所监管系统的看法。“我的第一印象是不定期判决和假释法对罪犯有利，”他承认。“相反的证据是压倒性的，我现在相信这些法律，如果得到适当管理，对社会和个人都是非常有益的。”简单地说，假释是一件好事，他写道——“即使在管理方法有缺陷的情况下。当然，没有任何机械比操作它的人的因素更有效率。”⁶

科学假释的实践

在1930年代早期，在Burgess的热情和Clabaugh的认可推动下，一个预测性假释系统在伊利诺伊州投入使用，到1951年《假释预测手册》出版，回顾了现在二十年的研究和实践。

语调是乐观的。Burgess在该书的引言中写道：“在过去的二十年里，社会科学家在努力找出哪些囚犯在假释中成功，哪些失败，以及在什么条件下发生成功或失败方面取得了重大进展。从他们的研究中产生了一种信念，即尽管涉及困难，但在某种程度上预测囚犯在假释时的行为是可能的。”⁷

该书提到了几个未来研究和可能改进的领域，例如额外的因素——如监狱工作人员的评估——可能有助于改善预测。书末一个特别有预见性的章节，题为“机器方法评分”，考虑使用打孔卡机器来自动化和简化收集数据、建立模型和输出个人预测的过程。

尽管有这种早期的乐观情绪，以及伊利诺伊州的明显成功故事，假释预测工具的采用将出人意料地缓慢。例如，到1970年，只有两个州在使用这样的工具。但这即将改变。

1969年，一位名叫Tim Brennan的苏格兰出生的统计学家在伦敦为联合利华工作，建立不同类型的统计模型：他是公司市场细分的顶尖专家。例如，一个项目将浴室肥皂的购买者细分为那些认为魅力至关重要的人和那些优先考虑对皮肤温和的人。他很擅长这个工作，也享受这个工作——但有些不对劲。“我遇到了价值观危机，”Brennan告诉我。一份报告到了他的桌上，他注意到联合利华在过去一年里花在研究其“Sqezy”可挤压液体洗洁精包装上的钱——那些能在超市货架上吸引眼球的措辞和颜色，都有最新感知心理学的支持——比整个英国政府在识字方面的支出还要多。⁸

“这是在六十年代末，”他说，“你知道当时嬉皮士现象是怎样的。无论如何，我无法看到为Sqezy开发包装的意义。所以我辞职了，申请了研究生院。”他最终来到兰卡斯特大学，将他的市场细分统计学应用到教育问题上：识别具有不同学习风格和课堂不同需求的学生。⁹

跟随女友来到美国，Brennan发现自己在科罗拉多大学，为Delbert Elliott（后来成为美国犯罪学学会主席）工作，然后最终创办了自己的研究公司，与国家矫正研究所和执法援助管理局合作。他的分类技术找到了第三个用途：为监狱带来更一致和严格的方法，每个监狱都必须根据囚犯的风险和需求在床位和病房之间组织囚犯——既为了他人的安全，也为了他们自己的改造。这通常是随机进行的或凭直觉进行的；Brennan的数学表明有更好的方法。

Brennan前往密歇根州特拉弗斯城进行了一次旅行，他听说那里有一些关于使用所谓“决策树”模型对囚犯进行分类的开创性机器学习工作，并且——为了应对监狱过度拥挤——识别应该释放谁。在那里，Brennan遇到了“这个年轻人……头发凌乱，留着胡子”，他发明了这个模型。这个年轻人的名字叫Dave Wells。“他对改革刑事司法系统充满热情，”Brennan说。“对Dave来说，这是一份充满激情的工作。”Brennan和Wells决定合作。¹⁰他们将公司命名为Northpointe。

随着个人计算机时代的到来，在刑事司法系统各个环节使用统计模型的做法，无论在大小司法管辖区，都呈爆炸式增长。1980年，只有四个州在使用统计模型辅助假释决策。到1990年，是12个州，到2000年，是26个州。¹¹突然间，不使用这类模型开始显得奇怪；正如国际假释当局协会2003年的《新假释委员会成员手册》所述，“在这个时代，在没有良好的、基于研究的风险评估工具帮助下做出假释决定，明显不符合公认的最佳实践。”¹²

这个新时代最广泛使用的工具之一是由Brennan和Wells在1998年开发的；他们称之为替代制裁的罪犯管理概况分析——或COMPAS。¹³COMPAS使用一个简单的统计模型，基于年龄、首次被捕年龄和犯罪史等因素的加权线性组合，来预测一个囚犯如果被释放，是否会在大约一到三年内犯下暴力或非暴力犯罪。¹⁴它还包括一套广泛的调查问题，用于识别被告的特殊问题和需求——比如化学依赖、缺乏家庭支持和抑郁。2001年，纽约州开始试点项目，使用COMPAS为缓刑决策提供信息。到2007年底，纽约市以外的所有57个县都在其缓刑部门采用了COMPAS工具。到2011年，州法律已被修订，要求在做出假释决定时使用像COMPAS这样的风险和需求评估。¹⁵

但是，从《纽约时报》编辑委员会的角度来看，存在一个问题：该州使用这些工具不够。即使在强制使用这些工具的地方，它们仍然并不总是得到适当的考虑。《纽约时报》敦促更广泛地接受假释中的风险评估工具，在2014年写道，“像COMPAS这样的程序已被证明有效。”¹⁶2015年，当一个假释改革案件提交到该州最高法院时，编辑委员会发表了第二篇观点文章，再次论证像COMPAS这样的统计风险评估工具比现状提供了显著改进。他们指控纽约假释委员会“顽固地坚持过去，基于主观的、通常不可审查的判断，例行拒绝长期服刑囚犯的假释。”“他们写道，采用COMPAS将”将委员会拖入21世纪。“¹⁷

然后——突然——语调发生了变化。

九个月后，2016年6月，该报刊登了一篇题为”在威斯康星州，对使用数据预测被告未来的反对声浪”的文章，文章最后引用了ACLU刑法改革项目主任的话说，“我认为我们有点急于进入大数据风险评估的明天世界。”¹⁸

从那里开始，2017年的报道只会变得更加严峻——5月，“被软件程序的秘密算法送进监狱”；6月，“当计算机程序让你留在监狱里”；10月，“当算法帮助送你进监狱”。

发生了什么？

发生的事情是——一句话——ProPublica。

获取数据

Julia Angwin在1970年代和80年代在硅谷长大，她是两名程序员的孩子，也是Steve Jobs的邻居。她从小就认为自己会终生做程序员。然而在这个过程中，她发现了新闻业并爱上了它。到2000年，她成为《华尔街日报》的科技记者。

“这很搞笑，”她回忆道。“他们就像，‘你懂计算机？我们雇你来报道互联网！’ 我就像，‘嗯，有什么特别关于互联网的吗？’ 他们就像，‘不——一切！’”¹⁹

Angwin在该报工作了十四年——从互联网泡沫破灭到社交网络和智能手机的兴起——不仅报道技术本身，还报道它经常留下的社会问题；她撰写了一个关于隐私相关问题的长期系列报道，名为“他们知道什么”。2013年，她从《华尔街日报》请假，写了一本关于隐私的书。²⁰

Angwin从未从她的写书休假中回到报社，而是加入了ProPublica，这是一个由前《华尔街日报》主编Paul Steiger创办的非营利新闻机构。如果她早期工作的主题是“他们知道什么”，答案不可避免地引发了另一个问题：他们将用它做什么？“所以我想，”她说，“我需要转向数据使用。这是下一个故事。他们将做什么？……他们将对你做出什么样的判断？”²¹

Angwin开始寻找那些基于数据做出的最具影响力但又被忽视的决策。她将目光投向了刑事司法领域。统计风险评估工具COMPAS等，正在数百个司法管辖区迅速被采用：不仅用于假释，还用于审前拘留、保释，甚至量刑。“实际上我很震惊，”她说。“我意识到我们整个国家都在使用这个软件……然后更让我震惊的是，这些工具都没有经过独立验证。”

例如，纽约州自2001年以来一直在使用COMPAS，但直到2012年才对该软件进行了首次正式评估——使用了十一年之后。（纽约最终发现它“既有效又具有预测准确性”。）

这样的故事惊人地普遍。明尼苏达州第四司法区包括明尼阿波利斯，处理全州40%的案件，他们在1992年开发了自己的审前风险评估工具。“当时撰写的报告建议这个新量表应该在使用的前几年内进行验证，”该州官方评估报告开头写道。“结果，实际上过了将近14年。”

这项姗姗来迟的评估在2006年发现，他们模型中的四个变量——被告是否在明尼苏达州居住超过三个月、是否独居、预订时的年龄，以及其指控是否涉及武器——与他们在等待审判期间实际犯新罪或缺席必要法庭出庭的风险几乎毫无关系。然而模型一直在基于这些因素建议审前拘留。更糟糕的是，这四个因素中的三个与种族密切相关。该司法区废弃了他们的模型，从头开始重新构建。

Angwin对风险评估模型了解得越多，就越担心。Angwin知道她找到了下一个报道主题：“从记者的角度来看，这是完美风暴。这是从未被审计过的东西；它涉及极高的人文风险；而且非常聪明的人说，‘这些都是种族的代理变量。’ 所以我想，‘我要测试这个。’”

她决定特别关注COMPAS，该工具不仅在纽约使用，在加利福尼亚、威斯康星、佛罗里达也在使用——总共约二百个不同的司法管辖区。它无处不在，在那时研究不足，而且——作为一个闭源、专有工具，尽管其基本设计在白皮书中可以获得——某种程度上是个黑盒。2015年4月，她向佛罗里达州布劳沃德县提交了信息自由法案请求。经过五个月的法律争执，数据终于到手：布劳沃德县在2013年和2014年期间给出的所有一万八千个COMPAS评分。

Angwin的团队开始进行一些探索性数据分析。立即有些东西看起来很奇怪。风险评分——范围从1（意味着最低风险）到10（意味着最高风险）——对于黑人被告或多或少均匀分布，十个区间中每个区间大约有10%的被告。对于白

人被告，他们看到了完全不同的模式：在最低风险区间中有大量被告，在最高风险区间中被告极少。

Angwin当时就想发表一篇报道。但她意识到这种截然不同的分布并不一定是偏见的证据——也许这正是那些被告实际上的风险程度。那么这些被告最终的风险如何呢？只有一种方法可以找出答案。“我悲伤地意识到，” Angwin回忆道，“我们必须查找这一万八千人中每一个人的犯罪记录。我们确实这样做了。这太糟糕了。”将COMPAS评分集与犯罪记录集联系起来——数据科学家称之为“连接”——花费了Angwin、她的团队和县工作人员几乎整整一年的时间。

“我们显然使用了大量自动化的犯罪记录抓取，”她解释道。“然后我们必须按姓名和出生日期进行匹配，这是你能想象到的最可怕的事情。有太多拼写错误，太多拼写错误。我可能每天都在哭。尝试进行那种连接是如此可怕。数据如此混乱。布劳沃德县实际上他们自己从未进行过连接。”县工作人员加入帮助ProPublica清理数据并理解它。

结果就是Angwin和她的团队在2016年5月发表的文章。题为“机器偏见”，副标题是“全国使用的软件用于预测未来罪犯。而它对黑人有偏见。”

Brennan和他的Northpointe同事在7月初发表了对ProPublica发现的官方反驳。用他们的话说，“当使用正确的分类统计时，数据不支持ProPublica关于对黑人种族偏见的声明。”也就是说，他们说，COMPAS满足了公平性的两个基本标准。

首先，它对黑人被告的预测与对白人被告的预测一样准确。其次，它的1到10风险评分具有相同的含义，无论被告的种族如何，这一特性被称为“校准”。评分为7分(满分10分)的被告，比如说，犯暴力重犯的风险，无论种族如何，都会以相同的百分比重新犯罪；评分为2分、3分等也是如此。1分就是1分，5分就是5分，10分就是10分，无论种族如何。COMPAS具有这两个特性——相等的准确性和校准——因此，Northpointe公司辩称，该工具在数学上不可能存在偏见。

到7月底，ProPublica作出了回应。³¹他们写道，Northpointe的说法是正确的。COMPAS确实是校准的，并且在两个群体中同样准确：对黑人和白人被告预测他们是否会重新犯罪(“累犯”)并被重新逮捕的准确率都是61%。然而，在39%的错误预测中，错误的方式截然不同。

观察模型误判的被告揭示了一个惊人的差异：“黑人被告被评为高风险但不重新犯罪的可能性是两倍。而白人被告在被归类为低风险后被指控新犯罪的可能性也是两倍。”³²

该工具在预测中是否“公平”的问题变得更加尖锐：首先要确定哪些统计指标是定义和衡量公平性的“正确”指标。

这场对话即将迎来新的转折——而这将来自另一个完全不同的社区，这个社区也开始缓慢但坚定地将注意力转向公平性问题。

什么不是公平性

哈佛大学计算机科学家Cynthia Dwork最为人知的是开发了一个叫做“差分隐私(differential privacy)“的原理，它使公司能够收集用户群体的数据，同时维护个人用户的隐私。一家网络浏览器公司可能想了解用户行为，但不知道你个人去了哪些网站；或者一家智能手机公司可能想学习如何改进其拼写纠正或文本建议，而不知道你个人对话的详细信息。差分隐私使这成为可能。从2014年左右开始，它几乎在各大科技公司中变得无处不在，并为Dwork赢得了哥德尔奖(Gödel Prize)，这是计算机科学的最高荣誉之一。³³但是，在2010年夏天，她认为自己的理论工作已经完成——她正在寻找一个新问题。

“我从2000年、2001年开始从事隐私工作，”她解释道。“到2006年，差分隐私就是差分隐私了。我心中还有最后一组问题想要研究——我完成了那项工作——然后我说，好吧，我想思考别的东西。”³⁴

当时在微软研究院的Dwork来到伯克利与计算机科学家同事Amos Fiat会面。他们整天都在交谈。到午餐时分，当他们在当地备受喜爱的餐厅Chez Panisse坐下时，他们已经谈到了公平性这个话题。Dwork回忆道：“为了不让我们周围的人因为我们讨论种族主义和性别歧视而感到不安，我们使用了‘紫色领带’和‘条纹衬衫’之类的术语。但到午餐时间我们已经...我们已经深入这个话题了。”

在理论计算机科学中，“公平性”这个术语出现在许多语境中，从切蛋糕(或分遗产)的博奕论机制，以确保每个人都得到应有的份额，到调度算法，确保CPU上的每个进程都能运行适当的时间。但Dwork认为，公平性的概念还有更多内容，这个领域尚未真正面对。

碰巧的是，Dwork读过Julia Angwin在《华尔街日报》上的一篇“他们知道什么”专栏，主题是在线广告。它显示，早在2010年，如果不是更早的话，公司就能辨别出访问其网站的每个用户几乎确切的个人身份——将一个表面上匿名的用户缩小到几十个可能的人之一——并在瞬间做出决定，比如，向谁推荐什么类型的信用卡。³⁵

在前十年思考隐私方面的问题后，Dwork也开始转变她的思路——从“他们知道什么”的问题转向“他们用这些信息做什么”的问题。

Dwork回到微软的实验室说：“我找到了我们的问题。”

她的一个实验室成员是当时的博士生Moritz Hardt，那个夏天从普林斯顿实习。Hardt最初并不想从事现实世界的问题。他对理论感兴趣：复杂性、难解性、随机性。越抽象越好。“没有应用，”他开玩笑说。“老派的。”³⁶

“我学到了很多，”他说，“但我发现在那个空间里我能解决的问题范围...它无法解决我对世界好奇的问题。我很快就被[计算机科学]涉及的一些更多社会问题所吸引。”这开始于隐私保护数据分析，与Dwork合作。她把他拉入了公平性项目。

正如Hardt回忆的那样：“Cynthia有一个直觉，有时候当人们要求隐私保护时，他们实际上是担心有人会以错误的方式使用他们的数据。这并不是要不惜一切代价隐藏数据，而是要防止数据使用方式造成的伤害.....这是一个相当准确的直觉。随着时间的推移，公众讨论从隐私转向了公平性，所有曾经看起来像隐私问题的事情突然都变成了公平性问题。”

他们和同事们开始发现的是，不仅将我们关于公平性的哲学和法律理念转化为严格的数学约束存在巨大的复杂性，而且事实上，许多领先的思想和实践（其中一些已有数十年历史）都存在严重误导性——并且有可能造成直接伤

害。

例如，美国反歧视法定义了许多“受保护属性”——比如种族、性别和残疾状况——通常人们认为应该严格禁止在可能影响招聘、刑事拘留等环境中的机器学习模型中使用这些变量。如果我们在新闻中听到某个模型“使用种族”（或性别等）作为属性，我们会认为某些事情已经严重出错了；相反，模型背后的公司或组织通常通过表明其模型“不使用种族作为属性”或是“种族盲”来为其模型辩护。这似乎很直观——如果某个东西不知道谁属于特定群体，它怎么能对该群体产生歧视呢？

这是一个错误，有几个原因。

简单地移除“受保护属性”是不够的。只要模型接收与种族或性别相关的特征，明确避免提及它并没有什么用处。

正如我们在波士顿交响乐团的案例中讨论的，以及在简历筛选语言模型中讨论的，简单地省略你关心的变量（在这种情况下是性别）可能是不够的，特别是如果有其他因素与之相关的话。这被称为“冗余编码”概念。性别属性在其他变量中被冗余编码了。

在刑事司法背景下，对一个人口群体的不同对待历史在各处创造了冗余编码。例如，比其他地区更积极地执法少数民族社区，意味着突然之间，像犯罪记录长度这样看似中性的东西，即以前定罪的数量，可能成为种族的冗余编码。

由于冗余编码的存在，仅仅对敏感属性盲目是不够的。事实上，冗余编码的一个反常结果是，对这些属性的盲目可能会影响情况变得更糟。例如，某个模型的制造者可能想要测量某个变量与种族相关的程度。如果不知道种族属性实际是什么，他们就无法做到这一点！我采访的一位工程师抱怨说，他的管理层反复强调确保模型不被性别和种族等敏感属性扭曲的重要性——但他公司的隐私政策阻止他和其他机器学习工程师访问他们正在处理的记录的受保护属性。所以，到最后，他们不知道模型是否有偏见。

省略受保护属性不仅使测量这种偏见变得不可能，也使缓解偏见变得不可能。例如，在招聘环境中使用的机器学习模型可能会因为候选人在前一年没有工作而惩罚他们。然而，我们可能不希望这种惩罚适用于孕妇或新妈妈——但如果模型必须是“性别盲”的，不能包括性别本身，也不能包括与之密切相关的怀孕等因素，这将很困难。

“研究领域中最稳固的事实，” Hardt说，“是通过盲目性实现公平性是行不通的。这是整个研究领域中最确立和最稳固的事实。”

这个想法从计算机科学家传播到法律学者、政策制定者和广大公众需要时间，但它已经开始传播。“在某些情况下，允许算法考虑受保护的阶层状况实际上可以使结果更公平，”最近一篇《宾夕法尼亚大学法律评论》文章如此表述。“这可能需要教义上的转变，因为在许多情况下，在决策中考虑受保护状况被推定为法律伤害。”

Moritz Hardt与Cynthia Dwork一起工作的那个夏天结出了果实，形成了一篇建立了这些早期结果的论文。比结果本身更重要的是，这篇论文是一个信标，向他们的理论家同事发出信号，表明这里有值得研究的东西：既有实质性的开放理论问题，又具有不可否认的现实世界重要性。

回到普林斯顿完成博士学位时，Hardt发现自己被安排了一种学术相亲。Dwork一直在与Helen Nissenbaum交谈，后者本身是计算伦理问题思考的先驱。Nissenbaum在这个领域有一个名叫Solon Barocas的研究生同学，她和Dwork意识到他们各自的门生可能是志同道合的人。

第一次会面的开始很尴尬。

Hardt坐在普林斯顿Witherspoon街上的Sakura Express寿司店的桌子旁。Barocas走了进来。“他坐下来，” Hardt回忆道，“然后他拿出了那篇论文，令我沮丧的是，他用黄色记号笔仔细地在论文中划出了段落。这不是阅读计算机科学论文的正确方式！”他笑道。“在那之前，我一直认为论文中的文字只是数学的填充物……这很尴尬，因为我写的内容毫无意义。”

两人开始交谈。尽管开始有些磕磕绊绊，但他们的导师所预期的共鸣显然存在。最终，两人决定向2013年神经信息处理系统(NeurIPS)会议提交一个关于公平性workshop的提案。NeurIPS拒绝了。“组织者认为这方面的工作或材料还不够多，” Hardt告诉我。“然后在2014年，Solon和我聚在一起说，‘好吧，在我们放弃之前再试一次。’”他们给它起了一个新名字，并扩大了范围：“机器学习中的公平性、问责制和透明度”，简称“FATML”。这次NeurIPS同意了。在那年在蒙特利尔举行的会议上举办了一整天的workshop，由Barocas和Hardt介绍会议议程，Dwork发表第一个演讲。

Hardt毕业后从普林斯顿到了IBM研究院。他继续在后台研究公平性问题。“我总是必须有其他事情来继续作为计算机科学家存在。就我的职业生涯而言，这总是像一个副项目。”值得赞扬的是，IBM给了他很多自由，即使他的热情并没有得到广泛分享。“我在IBM的团队对这个话题并不特别感兴趣，”他说，“但其他人也不感兴趣。”

两人在2015年又回来重复举办了这个会议。“房间满了，” Hardt回忆道，“人们参加了，但可以说在那个时候并没有引发革命。”他继续把大部分时间花在更传统的计算机科学上。又过了一年，他和Solon第三次重新举办了FATML workshop。

这次，情况有所不同；2016年是失控的一年，“Hardt说，”是不知何故每个人都开始研究这个问题的一年。“数学家兼博主Cathy O’ Neil曾在2014年原始会议上发表演讲，她出版了畅销书《数学毁灭性武器》(Weapons of Math Destruction)，讲述了因大数据的粗心(或更糟)使用而可能产生的社会问题。一系列令人震惊的选举结果违背了全世界民调专家的共识，动摇了人们对预测模型可信度的信心；与此同时，像Cambridge Analytica这样的数据驱动政治公司的工作引发了关于机器学习被用来直接影响政治的质疑。Facebook和Twitter等平台陷入了关于如何——以及是否——使用机器学习来过滤向其数十亿用户显示的信息的交火中。ProPublica的一群记者经过一年的不懈数据清理和分析，公开了他们对该国最广泛使用的风险评估工具之一的发现。

公平性不再只是一个问题。它正在成为一场运动。

公平性的不可能性

自2012年以来，康奈尔大学计算机科学家Jon Kleinberg和芝加哥大学经济学家Sendhil Mullainathan一直在进行一个项目，使用机器学习来分析审前拘留决定，比较人类法官与预测性机器学习模型。“其中一部分是思考人们对种族不平等背景下算法工具的担忧，”Kleinberg说。“我认为我们对此已经做了相当多的思考。然后ProPublica的文章出现了，我们的社交媒体渠道就完全被这篇文章的转发填满了。它真的抓住了人们的注意力。我们感觉像是，‘他们触及了某些东西……让我们真正深入研究，弄清楚这与我们一直在思考的主题有何关系。’”^[44]

在匹兹堡，卡耐基梅隆大学统计学家Alexandra Chouldechova自2015年春天以来一直与宾夕法尼亚州量刑委员会合作，开发一个可视化仪表板来探索风险评估工具的各种数学性质。“我开始越来越多地思考这些问题……理解分类指标方面公平性的不同概念以及它们之间的关系，”她说。“所有这些都发生在我仍然专注于其他项目的时候……我一直在与我的资深同事讨论也许写一篇关于风险评估工具验证的一些想法的论文——我们已经完成了文献综述等等——然后ProPublica的文章发布了。我认为这真的加速了很多人在这个领域的思考。”^[45]

几乎相同的叙述在美国的另一边上演着。斯坦福大学博士生萨姆·科贝特-戴维斯(Sam Corbett-Davies)开始攻读学位时“想要做机器学习，就是纯粹的计算机科学。”一年后，他发现自己根本不喜欢这个专业——与此同时，我却痴迷地阅读美国公共政策相关内容，纯粹是作为爱好。我想，‘我必须找到一种方法，将我拥有的技术技能与更加专注于政策的东西结合起来。’一位名叫沙拉德·戈埃尔(Sharad Goel)的助理教授最近加入了该学院，专门研究公共政策的计算和统计方法。合作似乎是显而易见的，很快两人就开始合作进行一系列项目，研究人类决策中的偏见，包括深入研究北卡罗来纳州的交通拦截等问题。例如，他们发现，在搜查黑人和西班牙裔司机时，警察似乎采用了比对待白人司机更低的标准：更频繁地搜查他们，而且发现违禁品的时间更少。科贝特-戴维斯解释说：“我们正在研究这个问题。我们在思考这些刑事司法问题，以及歧视意味着什么。然后…ProPublica的文章发表了。”

ProPublica强调了COMPAS犯错的类型，并突出了这样一个事实：它似乎一贯高估了未再犯罪的黑人被告的风险，低估了确实再犯罪的白人被告的风险。Northpointe强调的是错误的比率，而不是错误的种类，并强调该模型在预测黑人和白人被告方面同样准确，而且，在1到10的每个风险评分上，COMPAS都是“校准的”：具有该评分的被告再犯罪的可能性相等，无论他们的种族如何。那么，COMPAS是否“公平”的问题似乎归结为两种不同的公平数学定义之间的冲突。

正如科学中有时发生的那样，某个想法或洞察的时机如此成熟，以至于一群人几乎同时想到了它。

出现了三篇论文。所有三项努力都得出了相似的、互补的观点。消息并不好。

克莱因伯格说，在确定了辩论的关键——ProPublica对公平的定义，即不会再犯罪的黑人被告被错误归类为高风险的倾向几乎是其白人同伴的两倍——之后，“我们基本上可以将它与人们更加关注的其他定义对比，并问，这些定义在多大程度上是兼容的？

“答案是，它们不兼容。”

只有在黑人和白人被告恰好具有相等“基准率”的世界中——也就是说，恰好实际再犯罪的频率完全相同——才可能同时满足ProPublica和Northpointe的标准。否则，这根本不可能。

这与机器学习无关。与刑事司法本身无关。克莱因伯格和他的同事写道：“这只是关于两组之间基准率不同时风险评估的一个事实。”

乔尔德乔娃(Chouldechova)的分析得出了完全相同的结论：她写道，一个校准的工具”当再犯罪率在不同群体间存在差异时，不能在各群体间具有相等的假阳性和假阴性率。”

“所以你就是不能拥有一切，”她说。“这是一个一般原则，但在这种情况下，它会让你得出有趣的结论…这对现实世界中的风险评估具有影响。”

其中一个影响是，如果一套同样理想的标准对于任何模型来说都是不可能满足的，那么对任何风险评估工具的任何曝光都保证能找到一些值得成为头条新闻的不喜欢的地方。

正如萨姆·科贝特-戴维斯解释的那样，“不存在一个世界，在这个世界中ProPublica找不到一些不同的数字，他们可以称之为偏见。没有可能的算法——没有可能的COMPAS版本——在那种情况下这篇文章不会被写出来。”

(具有讽刺意味的是，在他对布劳沃德县数据的分析中，科贝特-戴维斯发现COMPAS在性别方面不是校准的。他说：“风险评分为5的女性再犯罪的频率与风险评分为3的男性差不多。” “[ProPublica]本可以写那篇文章。”)

这种不可能性的严酷数学事实也意味着这些问题不仅影响风险评估的机器学习模型，而且影响任何分类手段，无论是人类还是机器。正如克莱因伯格所写：“任何风险评分的分配原则上都可能因偏见而受到自然批评。无论风险评分是由算法确定还是由人类决策者系统确定，这都同样成立。”

那么，既然已经确定了如何调和这些不同的、看似同样直观的、同样理想的公平衡量标准的问题的答案——它们无法调和——以及在这方面传统的人类判断是否更好的问题——它并不更好——另一个问题浮现了。

现在怎么办？

不可能性之后

我问Jon Kleinberg如何看待他自己的不可能性结果——以及他认为这表明我们应该做什么。“对此我没有什么特别有争议的话要说，”他说。“我认为这取决于情况……我认为这两个定义都很重要，哪一个更重要取决于你所工作的领域。”

确实，这些权衡的关键方面在不同领域之间会发生根本性变化。例如，考虑放贷领域，它在许多重要方面与刑事司法不同。拒绝向本来会还款的人放贷，不仅意味着放贷方失去利息收入，对借款人也可能造成严重后果：他们无法购买房屋，无法创业。另一方面，犯相反的错误——借钱给不会还款的人——对放贷方来说可能只是金钱损失。也许这种不对称性改变了我们在该场景下对公平性含义的理解：也许我们希望确保，例如，来自两个不同群体的所有信用良好的借款人获得贷款的机会相同——即使数学告诉我们，不仅放贷方的利润会减少，而且一个群体中会有更多不合格的人获得贷款。⁵⁶

然而，在刑事司法环境中，特别是在暴力犯罪的背景下，假阳性（被标记为“高风险”但不会重新犯罪的人）和假阴性（被标记为“低风险”但确实重新犯罪的人）都会造成严重的人道代价。因此，我们寻求的权衡可能会大不相同。

例如，Sam Corbett-Davies认为，在风险评估背景下平衡假阳性率——确保不会重新犯罪的被告，无论是黑人还是白人，都不会更容易被不当拘留——只要群体间的实际犯罪率不同，就需要对不同种族的被告适用不同标准。例如，这可能意味着拘留每个风险等级为7或更高的黑人被告，但拘留每个风险等级为6或更高的白人被告。他说，这样的做法“可能违反了第十四修正案的平等保护条款。”⁵⁷

他说，实际情况甚至更糟。“因为我们在拘留低风险被告的同时释放了一些相对高风险的被告，被释放被告犯暴力罪的情况会增加。而且因为我们知道犯罪通常发生在社区内部而不是社区之间，这种犯罪会集中在少数族裔社区，受害者也会提起第十四修正案案件。”⁵⁸

不可能性证明还表明，平衡假阳性和假阴性率意味着放弃校准——即保证对于每个数值风险等级，被告重新犯罪的机会无论性别或种族都是相同的。Corbett-Davies说：“没有校准，就不清楚拥有风险评分意味着什么。所以如果你问我，‘这个被告有多危险？’我说，‘他们是2级，’你说，‘那是什么意思？’我说，‘嗯，如果他们是男性，意味着有50%的机会重新犯罪；如果他们是女性，意味着有20%的机会重新犯罪’——你可以看到‘2级’失去了它的意义。”⁵⁹

Tim Brennan也认为校准是至关重要的：“如果一个黑人和一个白人都得7分，对他们两人来说是否有完全相同的失败率？逮捕率？……标准方法是确保——该死的确保——你的校准没有种族偏见，你的准确度大致相同……我们两者都有。如果我们做别的事情，我们就是在违法。”⁶⁰

然而，即使是那些强调校准重要性的人也认为仅有校准是不够的。正如Corbett-Davies所说，“校准虽然通常是理想的，但几乎不能保证决策是公平的。”⁶¹

然而，仅仅因为这些属性不能同时完全满足，并不意味着我们不能寻找它们之间可能比其他更好的某些权衡。其他研究人员一直在探索这些可能权衡的空间，工作仍在继续。⁶²

我问Julia Angwin，她自己如何看待她的文章引发的理论结果风暴，以及最终无法做到她的团队似乎要求的事情——即制作一个既同样校准又具有相等假阳性和假阴性平衡的工具。

“我对此的感受是，”她说，“这是一个政策问题。这是一个道德问题。但我真正高兴的是，在我们提出这个问题之前，没有人知道这是一个问题。对我来说，这就像记者能做的最伟大的事情……我确实觉得如此准确地定义问题使其能够被解决。”⁶³

并非所有人都欢迎ProPublica的工作所引发的公开对话。以Anthony Flores为首的一群刑事司法学者发表了对ProPublica初始报告的“反驳”，不仅为COMPAS作为校准模型（以及校准作为公平性的适当衡量标准）进行辩护，而且进一步感叹争议本身的破坏性影响。⁶⁴“我们正处于历史上的独特时刻，”他们写道。“我们正面临着一代人，也许是人生中的机会，以科学的方式改革量刑并解决大规模监禁问题，而这个机会正在因为对[统计风险评估模型]的错误信息和误解而消失。进行不当的研究或误导性陈述可能导致那些负责制定政策的人产生困惑和/或瘫痪。”

Tim Brennan在很大程度上分享这种担忧，当然，他也强烈反对ProPublica对COMPAS的描述。但当我问他这场争议是否有积极的一面，特别是它在计算机科学家内部和外部产生的跨学科运动时，他表示同意：“我认为将公平性分解为各种不同的系数——提醒人们关注这些含义、益处、成本，以及使用某些系数的可能性——一旦你理解了一个问题或命名了一个问题，解决方案肯定比你甚至不知道问题存在时更容易被寻求。”⁶⁵

计算机科学界的大多数人都同意，公开讨论这些事情比不讨论要好。

“我的意思是，首先，数学就是数学，”Cynthia Dwork说。“所以我们可以希望的最好结果就是数学所允许的。而且我们最好知道限制在哪里，而不是不知道。”⁶⁶

对于Moritz Hardt来说，ProPublica研究之后的第一波学术和技术工作浪潮使公平性等伦理概念成为合法的研究主题，不仅在学术部门，而且在企业界也是如此。Hardt本人在2016年在Google工作，当时这波工作——包括他自己的工作——出现了，他回忆了他观察到的影响。“在那之前，人们甚至不太确定，这是我们应该触及的东西吗？这是一个烫手山芋，”他说。从2016年开始，无论是在Google内部还是外部，都出现了兴趣和研究的爆炸式增长。“所以这是我感到自豪的事情，”Hardt说。“它实际上对文化产生了影响。”⁶⁷

对Jon Kleinberg来说，计算机科学在这类问题中的作用是为利益相关者提供阐述问题的工具。“我们的观点不是告诉你哪个是对的，哪个是错的，而是给你语言来进行这种讨论…我当然认为这是我作为计算机科学家的职责的一部分：把那些一直以非正式和定性方式存在的东西，尝试思考，我们能否实际上严格而精确地谈论它们？因为世界正在朝那个方向发展。”

世界确实如此。2018年夏末，加利福尼亚州通过了SB 10，这是一项具有里程碑意义的司法改革法案，将完全取消现金保释——让州政府只需选择拘留或释放等待审判的被告——而且，进一步强制使用算法风险评估工具来为这些决定提供信息。⁶⁸就在那年年底前，美国国会以惊人的两党投票通过了《第一步法案》，这是一项全面的刑事司法改革法案，在其众多改革中，要求司法部开发一个统计模型，用于评估所有联邦囚犯的再犯风险，以及确定他们的康复课程。⁶⁹

尽管这些工具的突然和广泛采用只是在增加，但它们的采用与构建和使用它们的智慧之间的竞赛至少似乎正在接近平衡。我们正在缓慢地对如何做出这种“好”的预测有了更清晰的理解。

然而，在这个讨论背后潜藏着一个更大的问题——那就是“预测”是否真的是我们想要做的事情。

ProPublica的报道爆出后，《华盛顿邮报》的一名记者打电话给Chouldechova征求她的回应。记者基本上有两个问题。第一个是关于ProPublica的声明和Northpointe的辩护之间的紧张关系。

“对此我有一个准备好的答案，”Chouldechova说。“如果再犯的发生率在不同人群中差异，那么你不能让所有事情同时成立。当他问什么真正重要？时，我没有一个现成的答案。”

“所以，那是...那确实是促使我从不同角度思考问题的问题。”⁷⁰

超越预测

尽管基于现在可获得的数据进行预测是可行的，但不应该完全依赖这种方法。

—欧内斯特·伯吉斯⁷¹

你们的科学家太专注于他们是否能够…以至于他们没有停下来思考他们是否应该这样做。

—杰夫·戈德布拉姆饰演伊恩·马尔科姆，《侏罗纪公园》

在任何预测中最重要的事情之一就是确保你实际上在预测你认为你在预测的东西。这比听起来要难。

以ImageNet竞赛为例——AlexNet在2012年表现出色的竞赛——目标是训练机器识别图像所描绘的内容。但这并不是训练数据所捕获的。训练数据捕获的是Mechanical Turk上的人类志愿者所说图像描绘的内容。比如说，如果一只幼狮被人类志愿者反复错误识别为猫，它就会作为猫成为系统训练数据的一部分——而任何将其标记为狮子的系统都会被扣分，并必须调整其参数来纠正这个“错误”。

故事的寓意是：有时“真实情况”并不是真实情况。

在刑事司法预测的情况下，这种差距甚至更加重要。人们经常简单地说预测累犯本身，但这并不是训练数据所捕获的。训练数据捕获的不是再犯罪，而是再逮捕和再定罪。这是一个潜在的关键区别。

人权数据分析小组首席统计学家Kristian Lum和密歇根州立大学政治科学系的William Isaac在2016年一篇关于在警务中使用预测模型的论文中阐述了这一点：

因为这些数据是作为警察活动的副产品收集的，基于从这些数据中学到的模式做出的预测并不涉及未来整体犯罪实例。它们涉及的是未来警察已知犯罪的实例。从这个意义上说，预测性警务名副其实：它预测的是未来的警务，而不是未来的犯罪。

事实上，这种批评早在20世纪30年代就已经提出了。当Ernest Burgess提出的假释改革在伊利诺伊州实施时，批评者——特别是那些怀疑假释者重新productive回归社会能力的人——认为官方数据低估了再犯率。正如伊利诺伊州众议院共和党少数党领袖Elmer J. Schnackenberg在1937年抱怨的那样，“因为假释者在假释的第一年或第二年没有被发现犯罪，他就在这两年中被列为表现良好。”

我们的工具想要测量的内容与数据实际捕获的内容之间的这种差距，应该让保守派和进步派都感到担忧。成功逃避逮捕的罪犯被系统视为“低风险”——促使系统建议释放其他类似的罪犯。而被过度监管和错误定罪的人成为所谓“高风险”个人真实档案的一部分——促使系统建议拘留其他像他们一样的人。

这在预测性警务的背景下特别令人担忧，因为这些训练数据被用来决定警察活动本身，而警察活动反过来又产生逮捕数据——建立了一个潜在的长期反馈循环。

在警务不那么积极的地区犯罪的人，或者更容易让其指控被撤销的人，将被系统标记为没有累犯的人。甚至更少的警察会被派往那个社区。两个原本相似社区之间在警务方面的任何既存差异可能只会增长。正如Lum和Isaac所说，“模型越来越确信最可能经历进一步犯罪活动的地点正是他们之前认为犯罪率高的地点：选择偏见遇到了确认偏见。”系统开始塑造它本应预测的现实。这种反馈循环反过来进一步偏向其训练数据。

Lum和Isaac得出结论，不仅“警察已知的毒品犯罪不是所有毒品犯罪的代表性样本”，而且更重要的是，“模型非但没有纠正警察数据中的明显偏见，反而强化了这些偏见。被标记为目标监管的地点是那些根据我们的估计在历史警察数据中已经过度代表的地点。”

有理由相信我们在今天已经存在显著差异。正如Alexandra Chouldechova指出的，“年轻黑人男性和年轻白人男性自报的大麻使用率大致相同。但黑人年轻男性因大麻相关犯罪的逮捕率是白人的2.5到5倍。”ACLU 2013年的一份报告发现，平均而言，黑人美国人因持有大麻被逮捕的可能性是白人的四倍；例如在爱荷华州和华盛顿特区，这种差异超过八倍。纽约时报2018年的一项调查发现，曼哈顿的黑人居民因大麻指控被逮捕的可能性是白人居民的十五倍，尽管使用率相似。

有一些选项可以在模型中缓解这些问题。例如，COMPAS模型对“风险”做出三种不同的预测：暴力再犯、再犯和不出庭。暴力犯罪——例如杀人罪——比非暴力犯罪更容易被一致地报告给警方，而警方在执法和逮捕方面也更加一致。而且根据定义，被告未出庭的所有情况都会被法院系统知晓，几乎没有留下偏见抽样或差别执法的空间。因此，风险评估工具的明智使用可能会强调暴力再犯和未出庭预测，而不是非暴力再犯预测，理由是模型的训练数据在这些情况下更值得信赖——这正是许多司法管辖区开始采用的做法。⁸⁰

其他缓解措施包括以某种方式构建模型来考虑某些罪行执法中的巨大差异——例如，对有多次大麻逮捕记录的Black曼哈顿人和只有一次逮捕记录的White曼哈顿人一视同仁。（当然，这需要模型将被告的种族作为输入。）在大麻案例中的第三种缓解措施是简单地通过合法化或至少非罪化来斩断戈尔迪之结——这样就使下游的机器学习问题变得无关紧要。

第二个同样严重的担忧是，即使预测模型确切地测量了它声称要测量的内容，我们在实践中是否将其用于预期目的，还是用于其他目的。

例如，一些州正在使用COMPAS工具来为量刑决策提供信息，许多人认为这是一种尴尬甚至不当的使用。威斯康星州助理检察长Christine Remington说：“我们不希望法院说，我面前的这个人在COMPAS风险评估中得了10分，因此我要给他最高刑期。”但COMPAS确实被用来为量刑决策提供信息——包括在威斯康星州。当威斯康星人Paul Zilly因部分基于他的COMPAS得分而被判处比预期更长的刑期时，Zilly的公设辩护律师请来了Tim Brennan本人作为辩方证人。Brennan作证说COMPAS不是为量刑而设计的。⁸¹至少，我们似乎应该清楚地知道我们的预测工具究竟是为了预测什么而设计的——而且我们应该非常谨慎地在这些参数之外使用它们。“仅按指示使用”，正如处方药标签上所写的那样。这样的提醒在机器学习中同样必要。

然而，进一步放大视角，我们可能会质疑整个企业最基本的不言而喻的前提之一：更好的预测会带来更好的公共安全。

起初，认为其他情况似乎很疯狂。但有理由暂停一下，并质疑支撑这一观点的一些假设。

哥伦比亚法学院教授Bernard Harcourt在他的著作《反对预测》中提出了几个这样的反对意见（Angwin引用该书作为她自己工作的灵感来源）。正如他所论证的，更好的预测与更少犯罪之间的联系并不像看起来那么直接或万无一失。例如，想象一个预测工具识别出大多数鲁莽驾驶员是男性。可能的情况是，积极拦截男性驾驶员平均来说并不能大幅减少他们的鲁莽行为——但确实会导致女性驾驶员，她们意识到自己被拦截的可能性较小，驾驶得更加鲁

莽。在这种情况下，道路平均来说可能变得更不安全——正是因为使用了这种预测性政策。“换句话说，”正如 Harcourt 所说，“基于更高的过去、现在或未来违法行为进行画像(profiling)，在执法的核心目标——最小化犯罪方面，可能完全适得其反。”⁸² 他认为，这些适得其反的情况并不像看起来那么牵强或罕见：它们很可能准确地描述了目前在毒品使用和税务欺诈方面正在发生的情况。如果差别执法对被忽视群体的鼓励作用超过了对被审查群体的威慑作用，那么它可能只会使问题变得更糟。

如果预测不能有意义地转化为行动，它们也可能在使社会更安全的最终目标上失败。2013年，芝加哥市试点了一个旨在通过创建“战略对象名单”（非正式地称为“热点名单”）来减少枪支暴力的项目，该名单列出了成为枪支暴力受害者高风险的人群。作为一个群体，这些人成为杀人案受害者的可能性比普通芝加哥人高233倍。从这个意义上说，该名单的预测能力似乎是准确的。然而，杀人案如此罕见，即使在“热点名单”上的人中，只有0.7%成为受害者，而99.3%没有。那么，你如何处理这种预测信息呢？什么样的针对一千人的干预措施能够帮助那七个实际上会成为受害者的人？

兰德公司(RAND Corporation)2016年关于芝加哥预测性警务的报告指出：“通过利用先进的分析技术，警察部门可能能够更有效地识别未来的犯罪目标以进行预防性干预。”⁸³ 然而，“仅仅提高预测的准确性可能不会导致犯罪的减少……也许更重要的是，执法部门需要关于如何处理这些预测的更好信息”（重点是我加的）。⁸⁴

预测本身不是目的。哪个世界更好：一个我们可以99%确定犯罪何时何地会发生的的世界，还是一个犯罪率只有1%的世界？在对特定预测工具的预测准确性——或者说公平性——的狭隘追求中，我们可能错过了更重要的东西。⁸⁵

在COMPAS设计用于的审前释放中，同样存在从预测到干预的鸿沟，需要更广泛地考虑。预测某人不会在庭审日期出庭，并不一定意味着在此期间监禁他们是正确的干预措施。⁸⁶

Alexandra Chouldechova解释道：“如果你从这个角度思考，那你就是在说，好吧，这个特定人群，也许他们较难自给自足：他们实际上可能有较低的风险，但有更高的需求。”也许他们需要在庭审日期为孩子安排日托或搭车去法庭——而不是拘留。事实上，简单地提醒人们他们的庭审日期可以显著提高他们的出庭率。⁸⁷ 不幸的是，许多风险评估工具不像COMPAS，它们将未能出庭的预测与犯罪再犯的预测混为一谈。⁸⁸ 如果一种风险的解决方案是监禁，而另一种的解决方案是一条短信，这就是一个巨大的问题。

这一点特别引起Tim Brennan的共鸣。向法官展示的关于被告的COMPAS信息表被设计成风险评估用红色显示，而需求评估用绿色显示。“绿色，”他说，“你知道，因为这些是你想要培养和帮助这个人的东西。”⁸⁹ 整个重点是引导尽可能多的被告远离监禁，转向治疗、社区监督等——COMPAS中的AS毕竟代表“Alternative Sanctions”（替代制裁）。但一些法官简单地将“需求”分数——成瘾问题、无家可归、缺乏紧密社区联系——不是视为康复路线图，而是将某人关起来的更多理由。⁹⁰ 当然，将某人分配到此类替代制裁或治疗项目、课程、咨询等的能力，需要这些服务真正存在。如果它们不存在，那么就有一个任何统计模型，甚至任何法官都无法解决的问题。

“所以这让我想到我的要点，”Moritz Hardt告诉我。⁹¹ 一个由数据训练的machine-learning模型，“根据定义是一个预测未来的工具，因为它看起来像过去……这就是为什么它从根本上不适合很多领域，在这些领域你试图设计干预措施和机制来改变世界。”⁹²

他进一步阐述：“减少犯罪和监禁率是一个非常非常难的问题，我想把它留给刑事司法专家。我觉得预测为这个话题提供了一个有点反乌托邦的视角，也就是‘让我们假设我们不会在结构上减少犯罪。我们要预测它会在哪里发生，然后去试图在它发生之前抓住人们。’对我来说，它并没有真正提供在结构上减少犯罪的机制。这就是我觉得它反乌托邦的地方。我不想知道如何预测犯罪会在哪里发生。我想这很有用，但我更愿意有一个在结构上减少犯罪的机制。作为一名computer scientist，我在这个话题上没有什么可提供的，完全没有。我无法告诉你关于这个的第一件事。我需要花费数年时间才能达到可以做到的地步。”

退后一步，对刑事司法系统采取更广泛、更宏观视角的重要性，该领域最早的先驱者们并没有忽视。

Ernest Burgess在1937年写道——在他关于假释系统的初始报告促使一个风险评估模型在全州范围内投入实践之后——他觉得是时候转向更全面的东西了。“在我看来，伊利诺伊州的时机已经到了，”他写道，“停止将假释作为我们刑罚问题的一个孤立部分进行修补。需要的是一次大手术，涉及州监狱系统的完全重组。”⁹³

自那以后已经过去了八十多年。这仍然是正确的。

规则应该是清晰的、统一表达的，并且所有人都能接触到。我们都知道，这实际上很少是这种情况。

—DAVID GRAEBER¹

在没有充分结构或文档的情况下提供大量信息并不是透明度。

—RICHARD BERK²

在1990年代中期，Microsoft的Rich Caruana是Carnegie Mellon的研究生，研究neural networks，当时他的导师Tom Mitchell找他帮忙。

Mitchell正在从事一个雄心勃勃的跨学科、多机构项目——汇集生物统计学家、计算机科学家、哲学家和医生——以更好地理解肺炎。当患者首次被诊断时，医院需要相当早地做出一个关键决定，即是否将他们作为住院患者或门诊患者进行治疗——也就是说，是否将他们留在医院过夜进行监护或送他们回家。当时肺炎是美国第六大死亡原因，约10%的肺炎患者最终死亡——因此正确识别哪些患者风险最大将相当直接地转化为拯救生命。

该小组获得了约一万五千名肺炎患者的数据集，任务是构建一个机器学习模型来预测患者存活率，这可以帮助医院对新患者进行分诊。结果是各种机器学习模型之间的正面竞争：逻辑回归、规则学习模型、贝叶斯分类器、决策树、最近邻分类器、神经网络——应有尽有。³

Caruana负责神经网络（在90年代，最先进的是宽网络，而不是深度网络），他带着一些自豪消化着结果。他的神经网络获胜了——大获全胜。它是所有复杂模型中最好的，并且以显著优势超越了逻辑回归等更传统的统计方法。⁴

正如该小组在报告中写道：“即使对于肺炎等常见且昂贵疾病的预测性能有微小改善，也可能在医疗保健提供的质量和效率方面产生显著改善。因此，寻求具有尽可能高预测性能水平的模型很重要。”⁵

那么，自然地，参与该研究的匹兹堡医院决定部署性能最高的模型。对吧？

“我们开始谈论这个问题——在患者身上使用神经网络是否安全，”Caruana说。⁶

“我说，绝对不行，我们不会在患者身上使用这个神经网络。”

他们部署了一个更简单的模型，而这个模型曾被他的神经网络轻松击败。原因如下。

错误的规则

项目中的另一名研究员Richard Ambrosino一直在同一数据集上训练一个截然不同的“基于规则”模型。基于规则的模型是最容易解释的机器学习系统之一；它们通常采用“如果x则y”规则列表的形式。你只需从上到下阅读列表，一旦规则适用，你就完成了。想象一个不分支的流程图，看起来像一根从上到下的藤蔓：“这个规则适用吗？如果是，这就是答案，你完成了。如果不是，继续阅读。”通过这种方式，基于规则的模型类似于传统软件编程中的“条件语句”或“switch语句”；它们也听起来很像人类思考和写作的方式。（更复杂的模型使用“集合”而不是“列表”，其中可以同时应用多个规则。）⁷

Ambrosino正在使用肺炎数据构建基于规则的模型。一天晚上，当他训练模型时，他注意到它学到了一个看起来非常奇怪的规则。规则是“如果患者有哮喘病史，那么他们是低风险的，你应该将他们作为门诊患者治疗。”

Ambrosino不知道该如何理解这个规则。他把它给Caruana看。正如Caruana叙述的，“他说，‘Rich，你觉得这意味着什么？这毫无意义。’你不必是医生就能质疑如果你得了肺炎，哮喘对你是否有好处。‘这对搭档参加了下一次小组会议，会上有许多医生在场；也许医学博士们有计算机科学家没有想到的见解。’他们说，‘你知道，这可能是数据中的真实模式。’他们说，‘我们认为哮喘对肺炎患者来说是如此严重的风险因素，以至于我们不仅会立即将他们送进医院……我们可能会直接把他们送进ICU和重症监护。’”

换句话说，基于规则的系统学到的相关性是真实的。平均而言，哮喘患者确实比一般人群死于肺炎的可能性更低。但这恰恰是因为他们接受了更高水平的护理。“所以，正是哮喘患者接受的使他们成为低风险的护理，模型会拒绝给这些患者提供，”Caruana解释道。“我想你能看到这里的问题。”一个为哮喘患者推荐门诊状态的模型不仅仅是错误的；它是危及生命的危险。⁸

Caruana立即理解，看着基于规则系统找到的奇怪逻辑，他的神经网络肯定也捕获了同样的逻辑——只是不那么明显。

修改或编辑基于规则的系统相当直接；神经网络更难以这种方式“纠正”，尽管并非不可能。“我不知道它在神经网络的哪里，但无论如何我都能解决这个问题，”Caruana叙述道。“我可以在解决过程中发表更多论文——所以这很好——我们会让这个问题消失。我说，我们不部署神经网络的原因实际上不是因为哮喘，因为我已经知道那个问题了。

“我说，我担心的是neural net学到的那些和哮喘一样危险但基于规则的系统没有学到的东西。”因为neural net更强 大、更灵活，它能够学习到基于规则的系统学不到的东西。毕竟，这就是neural networks的优势——也是Caruana的neural net赢得团队内部竞赛的原因。“我说正是那些东西会让我们不使用这个模型。因为我们不知道里面还有什么我们需要修复的东西。所以正是neural net的这种透明度问题最终让我说我们不会使用它。这个问题困扰了我很长很长时间。因为最准确的machine-learning模型往往不像这样透明。而我是一个machine-learning研究者。所以我想使用准确的模型，但我也想安全地使用它们。”

在该领域中经常观察到，最强大的模型整体上是最不可理解的，而最可理解的模型则是最不准确的。“顺便说一句，这真的让我很恼火，”他告诉我。“我想为医疗保健做machine learning。Neural nets真的很好，它们很准确；但它们完全不透明和不可理解，我认为现在这很危险。就像，也许我不应该为医疗保健做machine learning。”⁹然而，Caruana决定花接下来的二十年开发试图兼具两个世界最佳特性的模型——理想情况下，既像neural networks一样强大，又像规则列表一样透明和易读的模型。

他最喜欢的架构之一叫做“generalized additive models”，最初由统计学家Trevor Hastie和Robert Tibshirani在1986年开创。¹⁰ Generalized additive model是一组图表的集合，每个图表代表单个变量的影响。例如，一个图表可能显示风险作为年龄的函数，另一个会显示风险作为血压的函数，第三个会显示风险作为温度或心率的函数，等等。这些图表可以是线性的、弯曲的或极其复杂的——但所有这些复杂性都可以视觉上立即理解，只需看图表即可。然后这些单个单变量风险被简单地相加以产生最终预后。通过这种方式，它比线性回归复杂得多，但也比neural net更易解释。你可以在普通的二维图表上可视化模型中的每个因素。任何奇怪的模式都应该立即突出。

在最初的肺炎研究多年后，Caruana重新审视了数据集并构建了一个generalized additive model来探索它。结果发现generalized additive model和他的旧neural net一样准确，而且更加透明。例如，他绘制了肺炎死亡风险作为年龄的函数。大部分符合人们的预期：如果患有肺炎，年轻或中年是好的，年龄较大更危险。但有一点特别突出：从65岁开始出现了突然、急剧的跳跃。特定的生日会触发风险的突然增加似乎不寻常。这是怎么回事？Caruana意识到模型成功学习了退休的影响。“退休很危险真的很烦人，对吧？你会希望退休时风险下降；可悲的是，它上升了。”¹¹

然而更重要的是，他看得越仔细，看到的麻烦关联就越多。他曾担心他的旧neural network不仅学习了有问题的哮喘相关性，还学习了其他类似的关联——尽管当时简单的基于规则的模型不够强大，无法向他展示neural network中还可能潜藏着什么。现在，二十年后，他有了强大的可解释模型。这就像有了更强的显微镜，突然看到了枕头里的螨虫，皮肤上的细菌。

“我看着它，我就想，‘天哪——我不敢相信。’它认为胸痛对你有好处。它认为心脏病对你有好处。它认为超过100岁对你有好处……它认为所有这些明显对你不好的东西对你都有好处。”¹²

它们在医学上都没有比哮喘更合理的意义；相关性同样真实，但同样地，正是这些患者被优先接受更密集护理的事实，使他们的生存可能性与实际情况一样。

“谢天谢地，”他说，“我们没有发布那个neural net。”

今天Caruana说他自己与大多数研究者处于不同的位置。他继续致力于开发承诺具有neural-network级别预测准确性同时保持易于理解的模型架构。但他没有宣传自己发明的任何特定解决方案——例如generalized additive models的更新版本¹³——而是在宣传问题本身。“每个人都在犯这些错误，”他说，“就像我几十年来一直在犯这些错误，却不知道自己在这样做。”

“现在的目标，”他告诉我，“是吓唬人们。让他们感到恐惧。如果他们停下来思考，天哪。我们真的有问题了，我就觉得成功了。”

黑盒问题

自然隐藏她的秘密是因为她崇高，而不是因为她是个骗子。

—阿尔伯特·爱因斯坦

给出理由的行为是权威的对立面。当权威的声音失效时，理性的声音出现。反之亦然。

—FREDERICK SCHAUER¹⁴

Rich Caruana绝不是近年来唯一一个产生“天哪，我们真的遇到大问题了”这种想法的人。随着机器学习模型在全世界的决策基础设施中激增，许多人发现自己对这些模型内部实际运作机制知之甚少而感到不安。

Caruana发现自己对使用大型神经网络特别不安，因为它们长期以来一直被称为“黑盒子”。随着神经网络在工业、军事到医学等各个领域的惊人崛起，越来越多的人感到了同样的不安。

2014年，美国国防高级研究计划局(DARPA)项目经理Dave Gunning正在与DARPA信息创新办公室主任Dan Kaufman交谈。“我们只是在试图抛出AI领域可以做什么的不同想法，”Gunning告诉我。¹⁵“他们有一个完整的项目，派遣了一整组数据科学家到阿富汗分析数据，试图找到对战斗人员有用的模式。他们已经开始看到这些机器学习技术正在学习有趣的模式，但用户往往得不到为什么的解释。”一套快速发展的工具能够处理财务记录、行动记录、手机日志等，以确定某些人群是否可能计划发动袭击。“他们可能会得到一些可疑的模式，”Gunning说，“但现在他们想要一个为什么的解释。”却没有这样的解释可供提供。

大约在同一时间，Gunning参加了由情报界在北卡罗来纳州立大学分析科学实验室赞助的会议。研讨会汇集了机器学习研究人员和数据可视化专家。“我们有一位政府情报分析师在房间里听我们谈论所有机器学习技术能做什么，”Gunning回忆道。“这位分析师非常坚决地表示，她的问题是她已经有这些大数据算法给她建议，但她必须在向前传递的建议上签名。如果这个建议是对的，她会得到评分——或者更糟——如果错了的话。但她不理解从学习算法得到的建议的理由。”她应该在上面签名吗？确切地说，她应该基于什么来决定？

随着计算技术的进步，国防界发现自己越来越多地思考自动化战场可能是什么样子——围绕越来越自主的武器理念存在什么风险和问题。但其中许多问题和困难——至少目前——仍然是理论性的。

“对于情报分析问题，这些系统已经存在；人们正在使用它们，”Gunning说。“你知道我的意思吗？这个问题已经存在了。他们想要帮助。”

在接下来的两年里，Gunning将发现自己成为一个多年DARPA项目的项目经理，试图正面应对这个问题。他将其称为XAI：可解释人工智能(Explainable Artificial Intelligence)。

在大西洋彼岸，欧盟正准备通过一项名为《通用数据保护条例》(GDPR)的综合法律。GDPR将于2018年生效，将大幅改变公司在线收集、存储、共享和使用数据的方式。这些法规——总共260页——构成了数据隐私史上最重要的文

件之一。它们还包含了一些更加好奇和有趣的内容，也许同样深刻。

2015年秋天，在牛津互联网研究所，Bryce Goodman正在翻阅法律草案时，有些东西引起了他的注意。“我学了一点机器学习，知道一些最好的方法实际上并不适合透明或可解释，”他说，“然后我看到了这个。在GDPR的早期草案中，这一点更加明确……他们说人们应该有权要求对算法决策进行解释。”¹⁶

“我觉得这真的很有趣，”他说，“当有这些立法条款或这些东西时，有人只是插了一根桩子说，这个东西现在存在了。”

Goodman找到了他的牛津同事Seth Flaxman，后者刚刚完成了机器学习和公共政策的博士学位。“嘿，我在GDPR中读到了这个，”Goodman说。“这似乎会是个问题。”

“他说，‘是的，这似乎会是的。’”

无论是贷款被拒绝、信用卡申请被拒绝、被拘留等待审判或被拒绝假释，如果背后有机器学习系统，你有权知道的不仅是发生了什么，还有为什么。

2016年春天，GDPR被欧洲议会正式通过，整个科技行业高管们的头发都竖起来了。律师们描述与欧盟监管机构坐下来的情形。“你意识到，”他们说，“从深度神经网络中获得可理解的解释是一个未解决的科学问题，对吧？”正如Goodman和Flaxman所写的，这可能需要对标准和广泛使用的算法技术进行彻底改革。”监管机构并不为此所动：“这就是为什么你们有到2018年的时间。”

正如一位研究人员所说：“他们决定给我们两年时间来解决一个重大研究问题。”

GDPR现已生效，尽管欧盟监管机构期望的确切细节，以及什么构成充分的解释——对谁，在什么情况下——仍在制定中。与此同时，透明度——理解机器学习模型内部运作机制以及它为什么会有这样的行为的能力——已成为该领域最明确和最关键的优先事项之一。对这个问题的研究工作至今仍在全力进行——但最近在多个方面都取得了进展。以下是我们正在了解的内容。

临床预测与统计预测

有一个普遍的误解，认为专家级的人员判断者通过某种神秘的直觉获得成功。当然，这是无稽之谈。他成功是因为他在事实或权衡方式上犯的错误更少。充分的洞察力和调查应该能让我们获得印象主义判断的所有优势（除了其速度和便利性），而没有任何缺陷。

——爱德华·桑代克^[17]

当我们面对透明度问题时——特别是在大型、复杂模型中——我们首先需要考虑的是，事实上，我们是否应该首先使用大型或复杂模型。就此而言，我们是否应该使用模型呢？

这些问题深深植根于统计学和机器学习的历史中，今天与以往一样相关，其答案确实相当令人惊讶。

1954年，罗宾·道斯是哈佛大学的哲学专业本科生，专攻伦理学。他的论文——“分析之观察”——研究道德判断是否以及在多大程度上植根于情感。

道斯不仅认为这些是重要的问题，他还认为“经验性工作可能很重要。但你如何进行经验性工作？”“带着这个问题，他意识到自己的兴趣更多在于心理学而不是哲学。他申请了心理学研究生项目，全国最好的项目——密歇根大学——在招生窗口的最后一天从候补名单中录取了他。道斯惊呆了，他向导师报告了这个好消息，导师告诉他：“立即给他们发电报，趁他们还没意识到犯了文书错误。”^[18]

道斯去了密歇根大学，开始了临床心理学的训练。那个时候——20世纪50年代末和60年代初——这意味着大量强调罗夏测试。道斯发现自己对罗夏作为临床工具的有用性越来越怀疑。“这一切都是直觉的，在直觉上也说得通。但后来我开始阅读，发现事实上，从经验上看，很多这些东西都不管用。”

在他逐渐产生怀疑期间，道斯在精神病房担任住院医师。“有一个病人有这种妄想，”他回忆道，“妄想是他在长乳房。他在我的小组里——顺便说一句，在一个锁着的病房里，因为他有这种妄想，肯定是精神分裂症。嗯，他为什么认为自己在长乳房？在他产生妄想之前的那一周，他的父母中有一人自杀了。好吧，完全合理，对吧？...他们甚至从未要求这个人脱掉衬衫。他只是被送到了精神病房——在我的小组里呆了六周的锁闭病房。当他们要求他脱掉衬衫时，事实确实如此：他在长乳房。”这个人患有克氏综合征：一种由额外X染色体引起的遗传疾病，症状包括缺乏体毛和乳房组织发育。道斯很愤怒。“好吧，那就是他生命中的六周被白白浪费了，”他说，“因为人们如此确信，哦，这是一个迷人的妄想。”

这导致道斯的职业生涯发生重大转向，从临床实践转向当时被称为“数学心理学”的领域。他有一篇准备提交发表的论文，比较专家临床判断与简单数学模型，他把论文给一个朋友看。朋友的反应让道斯措手不及。“他有点奇怪地看着我，实际上说，‘你确定你没有抄袭吗？’”

事实上，道斯正在加入一个他完全不知道的学术谱系。专家判断与简单数学模型——当时称为“精算”方法——的对比问题可以追溯到20世纪40年代初欧内斯特·伯吉斯的同事泰德·萨宾的工作。萨宾研究了明尼苏达大学新生学业表现的预测。“精算”模型是一个简单的线性回归，仅从两个数据点预测他们的大学GPA：高中班级排名和大学能力测试分数。人工预测由经验丰富的临床心理学家做出，他们不仅可以获得这两个数据点，还有额外的测试、八页档案、同事面试记录以及他们对学生的第一印象。

Sarbin发现两种预测方法之间没有可测量的差异。如果说有什么不同的话，精算模型更准确一些，尽管差异并不明显。从Sarbin的角度来看，临床医师可获得的额外信息在预测准确性上似乎没有任何贡献，这令人难以置信。“那些奋斗精神、工作和娱乐习惯、特殊天赋、情感模式、系统性干扰，以及数百种其他条件呢，”他写道，“这些似乎都与这种被称为学业成就的复杂社会心理行为形式相关？”¹⁹具有讽刺意味的是，Sarbin发现人类咨询师本身并没有过多强调这些因素，实际上，他们主要基于班级排名和测试分数做出预测——这正是回归模型中使用的相同数据。只是他们在如何权衡这些数据方面不够一致或精细。

Sarbin的结论是，进行访谈所花费的时间密集型努力是一种浪费。他警告说：“除非通过统计研究进行检验，否则社会科学中的案例研究方法将在智力上破产。”

Sarbin的发现随后引起了一位名叫Paul Meehl的年轻心理学家的注意。受到Sarbin的启发，并且不确定自己的立场，Meehl开始了一项调查，这项调查最终发展成了一整本书：1954年的《临床预测对统计预测》(Clinical Versus Statistical Prediction)。尽管这本书的大部分内容（如果不是绝大部分的话）都致力于理解临床判断的构成因素以及临床医师如何做决定，但迄今为止最大的影响来自于Meehl将临床判断和精算判断正面对比的那一章。Meehl发现，人类专家毫无胜算。在近一百个不同领域中，只有大约半打领域中人类决策者似乎有轻微的优势。“我被告知，”Meehl叙述道，“中西部一所大型弗洛伊德导向大学的一半临床教师因我的这本小书而陷入了为期六个月的反应性抑郁症。”²⁰

Dawes发现自己成了一个无意的抄袭者，当然对Meehl的工作产生了浓厚的兴趣——但他的导师们似乎不赞成这种影响。“我的精神分析导师告诉我，嗯，Meehl是个天才，每个人都知道他是天才，但他所做的与我们所做的毫无关系。我开始担心，如果他所做的与我们所做的毫无关系，也许我不想做我们做的事情。”

到20世纪70年代中期，当时在俄勒冈研究所的Dawes在这一谱系中撰写了另一篇震撼性论文。正如他总结的那样，Sarbin和Meehl提出了纯统计分析如何与专家人类判断相比较的问题。“统计分析被认为提供了一个基准，可以与有经验的临床医师的判断进行比较，”Dawes写道。“这个基准结果却成了天花板。”²¹

在Sarbin原始论文发表三十年后，经过许多十几项研究，他总结道：“对文献的搜索未能发现任何研究显示，当临床判断和统计预测都基于相同的可编码输入变量时，临床判断优于统计预测”（我的重点）。²²

这幅图景确实令人谦卑。即使在人类决策者被给予统计预测作为做决定时的另一条数据的情况下，他们的决定仍然比仅使用预测本身更差。²³其他研究者尝试了相反的策略：将专家人类判断作为输入输入到统计模型中。它们似乎也没有增加多少价值。²⁴

像这样的结论，已经得到此后众多研究的支持，应该让我们停下来思考。²⁵首先，它们似乎表明，无论我们在将决策交给统计模型时面临多少问题，仅凭人类判断并不是一个可行的替代方案。同时，也许复杂、精细的模型真的不是匹配或超越这个人类基准线所必需的。

然而，一个诱人的问题潜伏着：即，是什么解释了这个令人惊讶的判决？人类判断真的那么糟糕吗？少数几个变量的简单线性模型真的那么好吗？或者...第三种可能性：人类专业知识是否以某种方式进入了我们最不期望的简单模型中？我们是否在错误的地方寻找它？

不当模型：知道该看什么

有些真实问题很困难这一事实并不意味着所有真实问题都很困难。

——ROBERT HOLTE²⁶

例如，人类不能像计算机那样有效地结合信息这一事实，并不意味着人类可以被机器取代。它确实意味着人机系统的必要性已经到来。

——HILLEL J. EINHORN²⁷

Dawes与俄勒冈研究所的同事们希望深入了解简单线性模型在决策中令人震惊的有效性。

一个假设是，模型之表现可能超越专家是因为某种“群体智慧”效应：单个模型聚合了整个专家群体的判断，所以当然会超越任何单独的专家。这听起来很有道理——但事实并非如此。令人惊讶的是，即使模型仅被训练来模仿单个专家的判断，它仍然超越了专家本人！²⁸

也许线性模型优化的方式有什么特别之处，它们被调整为具有对每个变量进行权重分配的最优系数。Dawes与合作者Bernard Corrigan比较了评判者与他们称之为“不当”线性模型的表现，这些模型的权重并未经过优化。他们尝试了等权重和随机权重。

结果令人震惊。使用随机权重的模型——只要权重被限制为正值——与评判者本人一样准确或更准确。使用等权重的模型甚至更好。²⁹

这种快速简便的不当模型的实际应用前景似乎是无穷无尽的。在1970年代中期，Dawes在一个专家小组中与一位医生同台；之后在酒吧喝酒时，这位医生问他：“比如说，你能用你的不当线性模型之一来预测我和我妻子相处得如何吗？”³⁰

Dawes立即想象出他能想到的最简单的模型：他认为，一个好的婚姻应该涉及更多的性生活而不是争吵。利用同事的数据集，他进行了计算。他将一对夫妇在几周或几个月期间的性生活次数（“定义为有或无高潮的生殖器结合”）相加，然后减去他们在同一时期争吵的次数（“至少一方变得不合作的情况”）。“线性预测是简单性的精髓：从性交频率中减去争吵频率。正差预测幸福，负差预测不幸。”³¹来自密苏里州堪萨斯城地区的夫妇数据显示，确实，性生活减去争吵对于30对自认为“幸福”的夫妇中的28对是正值，而对于所有12对自认为“不幸福”的夫妇，性生活减去争吵都是负值。³²一项又一项的研究——在俄勒冈州、密苏里州、德克萨斯州——都确认了这种关联性。

事实上，在许多情况下，等权重模型甚至比最优回归更好——这个事实看似不可能，因为最优权重的选择正是为了，嗯，最优。但它们是针对特定环境和特定训练数据集的最优，而这种环境并不总是可以迁移：例如，用于预测明尼苏达大学学术表现的最优权重不一定适用于预测卡内基梅隆大学的学术表现。在实践中，等权重似乎更持久，在不同环境中更稳健。³³

Dawes对此非常着迷。考虑到世界的复杂性，这种极其简单的模型——等权重属性的简单计分——为什么不仅有效，而且比人类专家和最优回归都表现得更好？

他提出了几个答案。首先，尽管现实世界极其复杂，但许多高级关系是所谓的“条件单调性”——它们彼此之间的交互并不特别复杂。无论一个人的健康状况还发生什么其他情况，如果这个人处于二十多岁而不是三十多岁，几乎总是更好。无论一个人的智力、动机和职业道德还发生什么其他情况，如果这个人的标准化考试成绩高十分而不是低十分，几乎总是更好。无论一个人的犯罪历史、自控力等还发生什么其他情况，如果他们的逮捕记录少一次而不是多一次，几乎总是更好。

其次，任何测量中几乎总是存在误差。出于理论和直觉原因，测量越容易出错，以线性方式使用该测量就越合适。

从对齐角度来看，最具挑衅性的是，Dawes论证这些等权重模型超越其“正当的”、最优加权的对应模型，是因为正如我们所见，这些权重必须相对于某种目标函数进行调整。在现实生活中，我们要么无法准确定义如何衡量成功，要么没有时间等待这个真实标准出现以便调整我们的模型。“例如，”他写道，“当决定录取哪些学生进入研究生院时，我们希望预测某个未来的长期变量，可能被称为‘专业自我实现’。我们对这个概念的含义有一些想法，但还没有好的、精确的定义。（即使我们有一个，也不可能使用当前学生的记录来进行研究，因为这个变量至少要在学生完成博士工作20年后才能评估。）”即使在这里，Dawes论证，当我们不知道确切想要什么并且完全没有数据时，不当模型至少应该与原始直觉一样好用，如果不是更好的话。³⁴

然而，有人可能会反对道恩斯和科里根的比较，特别是与人类专家的比较，似乎并不完全公平。这些模型并非随机属性的随机线性组合；它们是恰恰那些人类通过几十年甚至几代人的最佳实践已经确立的、最相关和最具预测性的因素的随机线性组合。

人们可能会倾向于争论，也许所有这些“预处理”活动——从无限可用信息中决定哪两个、五个或十个因素与手头的决策最相关——反映了对问题的真正智慧和洞察：实际上，我们在将其交给线性模型之前已经完成了所有艰难的工作，然后线性模型获得了所有功劳。这确实正是道恩斯的观点。正如他所写：“线性模型无法取代专家来决定诸如‘寻找什么’之类的事情，但正是这种知道在做决策时要寻找什么的知识才是人们拥有的特殊专业技能。”³⁵

道恩斯的结论是，人类专业技能的特征在于知道要寻找什么——而不是知道整合这些信息的最佳方式。对这一想法最清晰的证明之一来自决策理论家希勒尔·艾因霍恩(Hillel Einhorn)在1972年的研究。³⁶艾因霍恩观察了医生对霍奇金淋巴瘤患者活检切片的判断。医生被要求指定他们在查看切片时认为重要的因素，然后为每张切片对这些因素进行评分。医生随后给出了关于患者严重程度的总体评级。碰巧的是，他们的总体严重程度评级与患者存活率的相关性为零。然而，在一个正在成为某种重复模式的结果中，使用专家个别因素评分的简单模型是患者死亡率的强有力预测器。

换句话说，我们一直在错误的地方寻找人类智慧。也许它不在人类决策者的头脑中，而是体现在确定将哪些信息放在他们桌上的标准和实践中。其余的只是数学——或者无论如何，应该是这样。

道恩斯用第三种方式表达了这一点，这可以说是他在辉煌生涯中最著名的一句话。“整个技巧就是知道要看什么变量，”他写道，“然后知道如何相加。”³⁷

最优简单性

简单可能比复杂更难：你必须努力工作才能让你的思维清晰，从而使其变得简单。但最终这是值得的，因为一旦你到达那里，你就能移山。

——史蒂夫·乔布斯³⁸

我愿意为之付出的唯一简单性是复杂另一边的简单性——而不是从未理解复杂性的简单性。

——小奥利弗·温德尔·霍姆斯。³⁹

也许没有人像杜克大学计算机科学家辛西娅·鲁丁(Cynthia Rudin)那样将道恩斯的精神带入二十一世纪。鲁丁将简单性作为她研究的核心驱动力之一；她不仅对反对使用过度复杂模型感兴趣，还在推动简单模型能力的边界。例如，在刑事司法领域，鲁丁和她的同事在2018年发表了一篇论文，显示他们可以制作一个与COMPAS一样准确的累犯预测模型，该模型可以用一个句子概括：“如果此人有三次以上的前科，或者是18-20岁的男性，或者是21-23岁且有两次或更多前科，预测他们将被重新逮捕；否则，不会。”⁴⁰

对鲁丁来说，道恩斯的研究既是灵感，也是一种挑战，就像抛下的战书。如此简单的模型，由手工选择的高级变量制成，表现得与更复杂的模型差不多——有时更好——并且始终表现得与人类专家一样好或更好。但即使这样也留下了很多问题——和研究途径。例如，人们如何从给定数据集构建不仅仅是一个简单模型，而是最佳简单模型？

令人惊讶的是，答案只在最近几年才出现。

鲁丁审视目前在二十一世纪医疗保健中使用的简单模型，并采取了与道恩斯截然不同且更不乐观的观点。她没有将当今的模型视为临床直觉的优越替代品，而是看到了过度受临床直觉影响的模型。而且有很大的改进空间。

她提出了男性冠心病评分表的例子。“所以如果你是男性，你走进医生办公室，他们会尝试计算你十年的冠心病风险。他们会通过问你五个问题来做到这一点：你的年龄是多少，你的胆固醇水平是多少，你吸烟吗，等等。所以他们问你这五个问题，然后你根据每个答案得分，然后你把分数加起来，这就转化为你十年的冠心病风险。”她的语调急转直下。“但他们从哪里得到这五个问题？他们如何得到分数？答案是，他们编造的！这是由一个医生团队编造的！这是——这不是我想要的方式。我想要做的是，我想要构建同样可解释的东西——但我想要从数据中构建它。”⁴¹

事实上，寻找最优简单规则绝非易事。实际上，这需要解决一个“棘手”或“NP困难”问题：一个复杂性的丛林，其中没有直接的方法来获得保证的最佳答案。给定数万份患者记录，每份记录都有几十甚至数百个不同的属性——年龄、血压等——如何为诊断找到最佳的简单流程图？计算机科学家有一个工具包，里面装满了在这里取得进展的方法，但Rudin认为现有的从大数据中构建简单规则列表和评分系统的算法——比如80年代开发的CART⁴²和90年代开发的C4.5⁴³算法——根本不够用。2010年代的计算机科学家还拥有80年代和90年代没有的东西：大约百万倍的计算能力提升。与其使用这种计算马力来训练一个巨大、复杂的模型——比如拥有数千万参数的AlexNet——为什么不将其用于搜索所有可能的简单模型的广阔空间呢？什么是可能的？她的团队重新开始，提出了新的方法——一种用于基于规则的模型，另一种用于基于评分表的模型——并开始将这些方法与现状进行比较。

特别是，Rudin和她的实验室将目标锁定在击败医学中最常用的模型之一：CHADS₂。CHADS₂开发于2001年，其继任者CHA₂DS₂-VASC开发于2010年，旨在预测房颤患者的中风风险。⁴⁴每一个都是由医生和研究人员与数据集密切合作，结合他们的临床专业知识，来识别他们认为最相关的因素而设计的。后续研究已经确认了这些工具的预测效用。这两个模型，尽管被普遍接受为有效工具并得到了极其广泛的应用，但在一定程度上仍然是“手工的”、手工制作的。Rudin想要计算性地识别最相关的因素，将它们组合成一个单一的评分工具。

使用比原始CHADS₂研究多六千多倍的数据，Rudin让她称为贝叶斯规则列表的算法在12,000名患者的数据集上自由发挥，仔细研究每个患者的大约4,100个不同属性——他们服用的每种药物、他们报告的每种健康状况——来制作最好的评分系统。⁴⁵然后，她将自己的模型与CHADS₂和CHA₂DS₂-VASC在同一数据集的保留部分上进行了比较。

结果显示，与CHADS₂和CHA₂DS₂-VASC相比都有显著改善。更有趣的是，它们还显示从原始CHADS₂到更新的CHA₂DS₂-VASC准确性有显著下降。新模型——至少在这个数据上，通过这个测量标准——似乎比旧模型更差。正如她和她的同事在论文中委婉地表达的那样，这“突出了手动构建这些可解释模型的困难”。

在随后的项目中，Rudin和她的博士生Berk Ustun与麻省总医院合作开发了睡眠呼吸暂停的评分系统，这种疾病影响着数千万美国人和全世界超过一亿人。⁴⁶他们的目标是创建一个不仅尽可能准确，而且简单到可以在一些绝对老式的硬件上快速可靠地运行的模型：医生的记事本。

由于模型将在纸上部署的限制，Ustun和Rudin不得不使他们的模型几乎不可能地简单。它需要考虑的显式特征非常少，整数系数尽可能小。⁴⁷即使进入二十一世纪，从业者简单地基于自己的直觉提出临时模型仍然很常见。这有时被嘲笑地称为“BOGSAT方法”：一群人围坐在桌子旁边。即使在使用机器学习构建模型的情况下，通常是更复杂的模型在事后被手动简化。⁴⁸今天仍然如此，当前医疗实践中的模型是以这种临时方式设计的，这意味着准确性——因此，真正的患者——正在遭受损失。⁴⁹Ustun和Rudin想看看是否有更好的方法。

他们开发了一个名为SLIM(“超稀疏线性整数模型”)的模型，不仅要找到合适的启发式方法，而且要找到在这些严重约束下做出决策的可证明最优方法。他们工作的结果是双重的，在医学和机器学习方面都有具体的好处。

首先，该模型显示——与已接受的智慧和当前实践相反——患者症状比他们的病史显著不那么有用。当Ustun和Rudin在患者病史——过去的心脏病、高血压等——上训练模型时，它比在他们的即时症状——打鼾、喘气和睡眠不佳等——上训练的模型显著更具预测性。更重要的是，在基于病史的模型中添加症状并没有带来太大的改善。睡眠呼吸暂停的筛查——在严重未治疗的形式下会使死亡风险增加三倍⁵⁰——已经向前迈出了可衡量的一步。

其次，机器学习社区取得了一个可以延续到其他合作和其他领域的办法论胜利。“SLIM的准确率与应用于该数据集的最先进分类模型相似，”Ustun和Rudin的团队报告说，“但具有完全透明的优势，允许通过对少量临床查询的是/否回答进行实际预测。”⁵¹

正如Rudin所说，“我想在设计预测模型时考虑到最终用户。我想设计这些东西……不仅要准确，还要让人们能够使用它们，人们能够利用它们做决策。我想创造高度准确且高度可解释的预测模型，我们可以将其用于可信的决策制定。我的工作基于一个我相信是真实的假设，即许多数据集允许建立出奇简单的预测模型。我不是第一个提出这个假设的人；这个假设在很多年前就被提出了。但现在我们有了计算能力和新想法以及新技术，这些真正让我们能够测试这个假设。”

对于从事这一系列问题研究的研究人员来说，这是一个激动人心的时代。简单模型与人类专业知识相比具有惊人的竞争力——甚至更胜一筹。现代技术为我们提供了推导理想简单模型的方法。

话虽如此，在某些情况下复杂性是无法避免的；显而易见的情况是那些没有人类专家将输入过滤为可管理大小的有意义量的模型。有些模型必须处理的不是“GRE分数”和“先前犯罪次数”等人类抽象概念，而是原始的语言、音

频或视觉数据，无论好坏。一些医疗诊断工具可以接受人类输入，如”轻微发烧”和”哮喘”，而其他工具可能直接展示X光片或CAT扫描并必须从中理解一些含义。当然，自动驾驶汽车必须直接处理雷达、激光雷达和视觉数据流。在这种情况下，我们别无选择，只能使用那些拥有数百万参数的大型”黑盒”神经网络，这些网络以其难以理解而闻名。但是，在透明性科学的另一个更加狂野的前沿上，我们也并非没有资源。

显著性(Saliency): 眼白的奥秘

相对于大多数其他物种，人类有着明显较大且可见的巩膜——我们眼睛的白色部分——因此我们在如何引导注意力，或至少是凝视方向上，具有独特的暴露性。进化生物学家通过“合作眼假说”论证，这必须是一个特征，而不是缺陷：这必须表明合作在我们物种的生存中异常重要，以至于共享注意力的好处超过了失去一定程度隐私或谨慎的损失。⁵²

因此，我们希望从我们的机器那里期待类似的东西是可以理解的：不仅要知道它们认为自己看到了什么，还要知道它们特别在看哪里。

这个想法在机器学习中被称为“显著性(saliency)”: 其理念是，如果一个系统正在查看一幅图像并将其分配到某个类别，那么图像的某些部分在做出该判断时可能比其他部分更重要或更有影响力。如果我们能看到一种突出这些关键部分的“热图”，我们可能会获得一些关键的诊断信息，我们可以将其用作一种理智检查，以确保系统按照我们认为应该的方式运行。⁵³

这种显著性方法的实践充满了惊喜，突出了机器学习系统可能有多么违反直觉。它们经常锁定我们认为根本不相关的训练数据的方面，而忽略我们认为是关键信息的内容。

2013年，波特兰州立大学博士生Will Landecker正在使用一个训练来区分有动物存在的图像和没有动物的图像的神经网络。他正在开发查看图像的哪些部分与最终分类相关的方法，并注意到一些奇怪的事情。在许多情况下，网络对图片背景的关注度比前景更高。仔细观察显示，模糊的背景——在摄影师术语中称为“散景(bokeh)”——通常出现在动物图像中，面部清晰对焦，背景巧妙地失焦。相比之下，空旷的风景往往更均匀地对焦。事实证明，他根本没有训练出动物检测器。他训练出了一个散景(bokeh)检测器。⁵⁴

2015年和2016年，皮肤科医生Justin Ko和Roberto Novoa领导了斯坦福大学医学院和工程学院研究人员之间的合作。Novoa被计算机视觉系统区分数百种不同犬种的能力进展所震撼。他说：“我想，如果我们能为狗做到这一点，我们就能为皮肤癌做到这一点。”⁵⁵他们组装了有史以来最大的良性和平性皮肤模式数据集，包含130,000张图像，涵盖两千种不同疾病以及健康皮肤。他们采用了现成的开源视觉系统Google的Inception v3，该系统已在ImageNet数据集和类别上进行了训练，然后他们重新训练网络来区分的不是吉娃娃和拉布拉多犬，而是肢端雀斑样痣性黑色素瘤和无色素黑色素瘤，以及数千种其他疾病。

他们将他们的系统与二十五名皮肤科医生进行了测试对比。该系统的表现超越了人类医生。这种“皮肤科医生级别”的准确率让他们在2017年在《自然》期刊上发表了一篇广受引用的论文。⁵⁶对于Ko来说——他以自己敏锐的诊断眼光而自豪，这是在近十年的培训和临床实践中磨练出来的——这个结果既令人深感谦卑又鼓舞人心。“我花了多年又多年的时间，”Ko说，“而这个系统几周内就能做到。”⁵⁷然而这样的系统所承诺的是“基本上将高质量、低成本的诊断能力扩展到地球上最偏远的角落”的能力。

事实证明，这样的系统不仅在难以找到一流诊断医师的地方有用，也可以作为像Ko本人这样训练有素的专家的第二意见。Ko记得那个日期——2017年4月17日——当时一位患者来到他的诊所，肩膀上有一个看起来奇怪的斑点。“我完全拿不定主意，”Ko说。他说，这个斑点有些地方“看起来不太对劲。但当我用皮肤镜检查时，我没有看到任何特征表明这是早期发展的黑色素瘤。“尽管如此，这让他感觉不对劲。

“所以我说，好吧，这是掏出我的iPhone的完美时机。”Ko拍了一系列照片，尽可能使用各种角度和光线，并将每张照片输入到神经网络中。“令人惊讶的是，”他说，“无论哪张照片，读数都相当稳定——而且它非常坚定地认

为这是一个恶性病变。”Ko让它进行了活检并与诊所的皮肤病理学家交谈。“你瞧，她说，’嘿，你知道吗？这真的很有趣。这确实是一个非常早期、微妙的演变中黑色素瘤的例子。’所以我们在完全可治疗的阶段就发现了它。”这个日期深深印在Ko的脑海中。这是神经网络第一次产生临床影响。“而且，”他说，“我希望这只是许多次中的第一次。”

不过，完整的故事要复杂一些。Ko、Novoa和他们的合作者在第二年向《调查皮肤病学杂志》提交了一封信，敦促在将神经网络模型过快地引入常规临床实践时要谨慎。

他们认为在这些模型广泛部署到实地之前需要极其小心，并用他们自己经验中的一个警示故事强调了这一点。他们使用的视觉系统更有可能将任何包含尺子的图像归类为癌性。为什么？碰巧的是，恶性肿瘤的医学图像比健康皮肤的图像更有可能包含用于比例的尺子。“因此算法无意中’学会了’尺子是恶性的。”⁵⁸基于显著性的方法可以捕捉到其中一些问题。尽管如此，最终的解决方案——无论是确保数据集包含足够多样的变化，还是以某种方式标准化所有输入图像——都是复杂的。“我们必须继续解决涉及的许多细节问题，”他们总结道，“以便安全地将这些新技术带到床边。”

告诉我一切：多任务网络

让复杂模型更透明和可理解的最简单想法之一就是让它们输出更多东西。当Rich Caruana致力于神经网络预测医疗结果时，他意识到该网络可以用来进行的不仅仅是单一预测——比如说，患者是否会生存——而是可能有数十种预测：他们在医院会住多久，他们的账单会有多大，他们是否需要呼吸辅助，他们需要多少个疗程的抗生素等等。

数据集中所有这些额外信息作为模型的额外输入在实践中是无用的。学习根据患者的医院账单预测患者的死亡风险在新患者到达时实际上不会帮助你，因为你当然还不知道他们的最终账单。但这些信息不是作为额外输入，而是作为额外输出，作为训练模型中额外的真实数据来源是有用的。这种技术被称为“多任务学习”。⁵⁹

“奇怪的是，这可能比一次训练它预测其中一件事更容易，”Caruana告诉我。⁶⁰“总的来说，你可以把同时训练它处理一百个相关事情想象成为你提供更多信号。更多信息。”

作为例子，他邀请我想象我因某种严重疾病（也许是肺炎）住院。假设我活下来了，他说——但我在医院住了一个月，我的账单是五十万美元。“我现在知道出了严重问题，”Caruana说。“你遇到了大麻烦。也许它不符合我们‘严重后果’的标准之一。但我现在知道你病得非常、非常严重。”

对于一个狭义地构建来预测死亡风险的系统来说，我的案例可能会作为训练数据，真实死亡风险为零，因为我存活了下来，但会遗漏一些东西。如他所说，我的案例是“一个相当高风险的零”。如果我确实只是幸运，也许系统应该预测，比如说，在像我这样的未来案例中有80%的死亡几率，而不是0%的几率。给它更广泛的输出范围进行训练，以及要做出的同步预测，可能会推动系统朝着更准确的评估方向发展。

Caruana意识到的是，这些“多任务”模型不仅在传统意义上表现更好——训练速度更快，准确率更高——而且它们也更加透明，这种透明性使得问题识别变得更加容易。如果他在90年代建立的医疗系统仅仅预测患者的生死，你可能会发现它预测出一些意外的结果——比如，哮喘患者比普通门诊患者死亡的可能性更小。另一方面，如果你有一个多任务网络从数据中预测各种各样的事情——不仅仅是死亡，还有住院时间或治疗费用——这些异常情况会变得更加明显。例如，哮喘患者的发病率可能好于平均水平，但医疗费用却高得惊人。很明显，这些不是普通的“低风险”患者，不能简单地让他们回家服药并第二天早上回电话。

在某些情况下，这些额外的输出通道还可以提供更重要的东西。一个横跨Google、其生命科学分支Verily和斯坦福医学院的团队在2017年和2018年致力于类似地调整Google的Inception v3网络来分类视网膜图像。他们发现了令人鼓舞的诊断结果，该模型检测糖尿病视网膜病变等疾病的准确性与人类专家相当。然而，该团队意识到他们正在处理的数据集包含患者的各种其他信息：年龄、性别、体重指数、是否吸烟等等。“所以我们某种程度上将这些变量添加到模型中，”Google研究员Ryan Poplin说。正如Caruana所做的那样，他们认为既然拥有所有这些额外的患者数据，为什么不让他们的模型预测所有这些呢？如果他们将这些辅助数据宝库——年龄、性别、血压等——不是作为模型的额外输入，而是作为额外的输出会怎样？这可能提供一种使模型更加稳健的方法，并可能为模型疾病预测出错的情况提供一些见解。“我们觉得这是一个很好的控制或基准真相，我们可以添加到模型中，”Poplin说。

他们遭遇了巨大的震惊。该网络几乎可以完美地仅从视网膜图像中判断患者的年龄和性别。

团队中的医生不相信结果是真实的。“你向某人展示这个，”Poplin说，“他们会对你，‘你的模型肯定有bug。因为你不可能以如此高的准确率预测这些。’……当我们越来越深入地研究时，我们发现这不是模型中的bug。这实际上是一个真实的预测。”

团队使用显著性方法来揭示，如果不是确切地说网络是如何做到的，那么至少是相关特征是什么。碰巧的是，年龄是由模型主要通过观察血管来确定的；相比之下，性别是通过观察黄斑和视神经盘来确定的。

Poplin说，起初，当向医生展示这些结果时，“他们会嘲笑你。他们不相信。但是，当你向他们展示那个热图并显示它专注于视神经盘或可能是视神经盘周围的特征时，他们说，‘哦是的，当然我们知道这一点，当然你可以看到这一点。’通过显示模型在图像中的哪个位置用来做出预测，它确实提供了一定程度的信任，同时也提供了结果的有效性。”

超越了仅仅实现预测准确性，该模型为医学科学本身提出了一条有趣的前进道路。多任务学习和显著性技术的结合向该领域表明，视网膜中存在被忽视的性别差异。不仅如此；它还显示了在哪里可以找到它们。

换句话说，这些解释方法不仅仅是为了更好的医学。它们可能也会培养更好的医生。

打开引擎盖：特征可视化

我们已经看到多任务网络如何通过额外的输出为网络的预测提供重要的上下文。显著性方法为网络的输入提供上下文，并能为我们提供关于模型实际关注位置的信息。但这两者都没有告诉我们黑盒内部发生了什么——也就是说，模型实际看到了什么。

自2012年AlexNet以来，机器学习的标志性突破是能够从原始感官感知的混乱复杂中学习的神经网络模型：数百万个彩色像素。这些模型拥有数千万而不是几十个参数，这些参数代表相当难以言喻的东西：早期层的阈值总和，而这些本身又是更早期层的阈值总和，一直追溯到数百万个原始像素。这不是制作可理解解释的原始素材。

那么该怎么办？

在NYU，博士生Matthew Zeiler和他的导师Rob Fergus专注于这个问题。他们论证说，这些庞大而令人困惑的模型的成功是不可否认的。“然而，”他们写道，“对于它们为什么表现如此出色，或者如何改进它们，没有清楚的理解……从科学角度来看，这是非常不令人满意的。”换句话说，结果令人印象深刻，但正如Zeiler所说，“有了所有这些好结果，不清楚这些模型在学习什么。”

人们知道卷积网络的最底层代表基本的东西：垂直边缘、水平边缘、对角边缘、强烈的单一颜色或简单的渐变。众所周知，这些网络的最终输出是一个类别标签：猫、狗、汽车等等。但人们并不真正知道如何解释中间的层。

Zeiler和Fergus开发了一种他们称为“反卷积”的可视化技术，这是一种将网络中间层激活转换回图像的方法。⁶⁵

他们第一次看到了第二层。那是一个形状的动物园。“平行线、曲线、圆形、T形连接、渐变模式、彩色斑点：在第二层就已经存在着巨大的结构多样性。”第三层更加复杂，开始表示物体的部分：看起来像面部的部分、眼球、纹理、重复的模式。它已经能够检测到云朵的白色绒毛、书架的多彩条纹或草的绿色梳状结构。到了第四层，网络开始响应眼睛和鼻子的配置、瓷砖地板、海星或蜘蛛的放射状几何形状、花瓣或打字机上的按键。到了第五层，物体被分配到的最终类别似乎施加了强烈的影响。

这种效果是戏剧性的、富有洞察力的。但它有用吗？Zeiler打开了在2012年赢得ImageNet竞赛的AlexNet模型的引擎盖，开始深入挖掘，使用反卷积检查它。他注意到了一堆缺陷。网络的一些低级部分归一化不正确，就像过度曝光的照片。其他滤波器已经“死亡”，没有检测到任何东西。Zeiler假设它们的大小不适合试图匹配的模式类型。尽管AlexNet取得了惊人的成功，但它携带着一些死重。它可以被改进——可视化显示了在哪里改进。

在2013年秋天，在疯狂的几个月内，Zeiler完成了他的博士学位，离开了纽约大学，创办了自己的公司——Clarifai——并参加了那年的ImageNet竞赛。他赢了。⁶⁶

随后和并行地，其他团体探索了直接可视化神经网络的进一步方法。2015年，Google工程师Alexander Mordvintsev、Christopher Olah和Mike Tyka实验了一种方法，从随机静态的图像开始，然后调整其像素以最大化网络为其分配特定标签（比如“香蕉”或“叉子”）的概率。⁶⁷这种看似简单的方法已被证明是非常强大的。它产生了迷人的、令人难忘的、经常是迷幻的，偶尔是怪诞的图像。例如，开始为“狗”优化静态，你很可能会得到几十只眼睛和耳朵的邪恶混合体，以分形方式生长，一个接一个，在各种不同的尺度上。

这是艺术恶作剧的肥沃土壤，几乎本身就是一种新颖的视觉美学。Google工程师有了进一步的想法：不是从静态开始手动指定类别标签，而是从真实图像开始——比如云或叶子——并简单地调整图像以放大网络中恰好最活跃的神

经元。正如他们所写：“这创造了一个反馈循环：如果云看起来有点像鸟，网络会让它看起来更像鸟。这反过来会让网络在下一次通过时更强烈地识别鸟，如此循环，直到一只高度详细的鸟似乎从无到有地出现。”他们将这种相当幻觉的方法称为“DeepDream”。⁶⁸

还存在其他更奇异的可能性。当Yahoo的视觉团队开源了一个用于检测上传图像是否为色情内容的模型时，加州大学戴维斯分校的博士生Gabriel Goh使用这种生成方法将静态调整为网络认为“最大化”工作不安全”的形状。结果就像Salvador Dalí的色情作品。如果你优化淫秽过滤器和正常ImageNet类别标签的某种组合——例如火山——在这种情况下，你会得到淫秽的地理：看起来像巨大的花岗岩阳具，喷射着火山灰云。无论好坏，这样的图像都不容易被遗忘。⁶⁹

从更哲学的角度来看，这些技术表明，至少就神经网络而言，批评家和艺术家之间的界限可能是微妙的。一个训练来识别湖泊或大教堂的网络可以以这种方式被制造出来，无休止地产生它从未见过的一个又一个湖泊或一个又一个大教堂的图像。这是艺术实践的美丽而细致的处方：任何能够区分好艺术和坏艺术的人都可以成为创造者。你所需要的只是好的品味、随机变化和大量时间。

这些技术不仅开辟了巨大的美学可能性空间，也具有重要的诊断用途。例如，Mordvintsev、Olah和Tyka使用他们的“从静态开始”技术，让图像分类系统“生成”最大程度地类似于其所有不同类别的图像。“在某些情况下，”他们写道，“这揭示了神经网络并没有完全寻找我们认为它在寻找的东西。”例如，“最大化”哑铃”分类的图片包括了超现实的、肉色的、无实体的手臂。“那里确实有哑铃，”他们写道，“但似乎没有一张哑铃图片是完整的，如果没有肌肉发达的举重运动员在那里举起它们。在这种情况下，网络未能完全提炼出哑铃的本质。也许它从未见过没有手臂握着的哑铃。可视化可以帮助我们纠正这类训练错误。”⁷⁰这也是探索偏见和表现问题的有用技术。如果从随机静态开始，微调数百张图像以“最大化”面部”类别，产生的一组面部全部是白人男性，那么这很好地表明网络不会轻易识别其他类型的面部。

自最初的DeepDream研究以来，Tyka继续与人共同创立了Google的Artists and Machine Intelligence项目，并继续探索机器学习的美学可能性。与此同时，Olah和Mordvintsev与他们的合作者一起，继续探索可视化的科学和诊断前沿。⁷¹如今Olah领导着OpenAI的clarity团队。“我一直对解释事物很着迷，”他告诉我。“我的目标——我觉得有些人认为这有点疯狂——就是完全逆向工程神经网络。”⁷²这项工作不仅推进了科学边界，也推进了出版边界；Olah发现传统的科学期刊并不适合他制作的那种丰富、互动、全彩和高分辨率的可视化。所以他创办了一个新的期刊。⁷³

总体而言，该团队对正在进行的工作表达了谨慎但富有感染力的乐观态度。“我们还有很多工作要做，来构建强大且可信赖的可解释性接口，”他们写道。“但是，如果我们成功了，可解释性有望成为实现有意义的人类监督以及构建公平、安全、一致的AI系统的有力工具。”⁷⁴

在视觉领域尤其持续取得很多进展——但是如何理解这样的网络，不是在视觉层面而是在概念层面？是否可能有一种方法通过文字来理解网络的内部？这是解释的最新前沿之一。

深度网络和人类概念

2012年秋天，MIT研究生Been Kim发现自己在博士学位开始时，正在专注于一个将塑造她未来几年生活的项目。她之前在机器人技术领域工作——甚至为了更好地理解工业机器人环境而考取了叉车驾驶执照——但最终决定这不合适。“我意识到机器人技术有硬件限制，”她告诉我。“我的思维速度比物理设备能做的要快。”⁷⁵

Kim越来越相信，可解释性很可能会成为她的论文主题，如果不是她的终生事业的话。在12月她第一次参加NeurIPS会议时，她发现自己在与一位年长的同事交谈。“我告诉一位教授我从事可解释性工作，他说，‘哦，为什么？神经网络会解决一切！你为什么关心？’我说，‘嗯...！’”她对这段回忆咯咯笑了。“想象一下，如果你走进医生的办公室，他说，‘哦，我要把你打开，可能移除一些东西。’你问，‘哦，为什么？’他说，‘哦，我不知道。这台机器说这是对你最好的选择。99.9%。’”

“你会问什么？”她说。她的问题既是修辞性的，也不仅仅是修辞性的。“你会对医生说什么？对我来说，这显然是一个我们需要解决的领域。我不确定的是，人们意识到这很重要的时机会多快到来。”下午晚些时候的阳光透过Google Brain会议室的玻璃过滤进来，Kim自2017年以来一直在那里工作。“我认为那个时候已经到来了。”

Kim的信念是，解释和诠释有一个本质上人性的维度——因此该领域具有内在的混乱性，内在的跨学科性。“很大一部分——仍然未被充分探索——是思考人的一面，”她说。“我总是强调HCI（人机交互）、认知科学...没有这些，我们无法解决这个问题。”认识到可解释性不可避免的人类方面意味着事情并不总是整齐地转化为熟悉的计算机科学语言。

“有些人认为你必须给出解释必须是什么的数学定义。我不认为这是一个可实现的方向，”她说。“一些无法量化的东西让计算机科学家感到不舒服——本质上非常不舒服。”

Kim的几乎所有研究论文都包含在计算机科学中相当不寻常但在其他领域很典型的内容：即使用人类受试者进行的实际研究。“与用户迭代是至关重要的，”她说。“因为如果你存在的唯一理由是为了人类消费...我们必须证明它对人类消费是有益的。”这种迭代是至关重要的，因为设计师认为对实际人类用户有用的东西往往并非如此。如果你正在设计解释或可解释模型供真实的人使用，那么这个过程应该与设计驾驶舱控制系统或软件用户界面一样具有迭代性。⁷⁶不让这种经验反馈指导过程简直是傲慢。

这样的研究为我们之前探讨的故事增加了一些复杂性，之前的故事赞扬了使用简单模型。Kim指出，这是否是可解释性的最佳方法最终是一个经验问题。“在某些情况下，在非常有限的情况下，如果你能充分识别并经验验证如果有少量特征，对于这个特定任务来说是可解释的——在那种情况下，是的，你可以写下在给定问题下最可解释意味着什么，并且你可以优化它。”但真实的人类研究——Kim的和其他人的——表明事情在实践中很少如此直接。

例如，在2017年，Microsoft Research的Jenn Wortman Vaughan和她的同事研究了人类用户与房屋价值机器学习模型的交互方式，该模型根据平方英尺、浴室数量等特征预测价格。当模型使用较少因素并且对用户更“透明”时，用户更能预测模型的预测结果。但简单性和透明性实际上都没有影响人们对模型报告的信任水平。而且，事实上，当模型更透明时，人们更不可能意识到模型犯了错误。⁷⁷

Kim的信念之一是“人类使用概念而不是数字来思考和交流。”⁷⁸我们交流——以及在很大程度上思考——口头上，利用高级概念；我们不谈论感官体验的原始细节。因此，Kim认为许多基于saliency的方法还不够深入。她和她的合作者一直在研究他们称为“概念激活向量测试”或TCAV的东西，它提供了一种使用这种人类概念来理解网络内部工作原理的方法。

例如，想象一个正确识别斑马图片的模型。假设TCAV显示网络在其预测中使用了”条纹”、“马”和”草原”：这看起来是合理的。另一方面，一个在医生图像集上训练得有些天真的网络可能——如果数据集有偏见——假设”男性”概念具有一些预测价值。TCAV会显示这一点，并提供一个指示，我们可能想要相应地调整模型或训练数据以消除这种偏见。⁷⁹

我们究竟如何获得这种洞察？我们模型生成的每个图像不仅产生最终的类别标签输出，还在网络中产生巨大的内部活动模式，跨越其数千万人工神经元。这些内部激活对人眼来说可能似乎是压倒性的噪音，但这并不意味着你不能将它们展示给机器。TCAV背后的基本思想是，对于你感兴趣的任何概念——比如在医生案例中的”男性”——你向网络展示一堆男性图像，然后展示一堆随机其他事物的图像：女性、动物、植物、汽车、家具等等。你将网络的内部状态（比如在特定层）输入到第二个系统，一个简单的线性模型，学习区分你的类别（在这种情况下是”男性”）的典型激活和来自随机图像的激活。⁸⁰然后你可以查看当网络将图像分类为”医生”时，这种激活模式是否存在，以及它在多大程度上起作用。

“我认为这种方法提供了独特的好处，”她说，“解释现在使用你的用户的语言。你的用户不必来学习机器学习…我们可以用他们说的语言提供解释，并回答他们的假设，使用他们自己的术语。”⁸¹

当Kim使用TCAV查看两个流行的图像识别模型Inception v3和GoogLeNet时，她发现了许多这样的问题。例如，“红色”概念对”消防车”概念至关重要。“所以这是有道理的，”她说，“如果你来自消防车是红色的地区。”⁸²例如，在美国几乎总是如此，但在澳大利亚不是，根据地区不同，消防车有时可能是白色的——或者在Canberra，是霓虹黄色。这表明，比如说，在以美国为中心的数据集上开发的自动驾驶汽车模型在安全部署到澳大利亚之前可能需要修改。

Kim还发现”手臂”这个概念对识别”哑铃”很重要，这证实了Google DeepDream团队早期的视觉发现，并暗示该网络可能难以识别架子上或地面上的哑铃。⁸³“东亚人”这个概念对”乒乓球”很重要，而”白种人”这个概念对”英式橄榄球”很重要。其中一些可能反映了某些模式，虽然并非不准确——根据国际乒乓球联合会的数据，世界前十名男选手中有七名，前十名女选手中的全部十名都来自东亚国家⁸⁴——但从准确性和偏见的角度来看，这些仍可能表明模型的某些方面值得审查。无论从哪个角度来看，我们都不会满意一个像Facebook AI Research的Pierre Stock和Moustapha Cisse发现的那样——将中国国家主席习近平的肖像归类为”乒乓球”的系统。⁸⁵TCAV提供了一种明确量化这些问题的方法——理想情况下是将其扼杀在摇篮中。

Kim在TCAV方面的工作在Google 2019年I/O大会上被Google CEO Sundar Pichai在主题演讲中重点介绍。Pichai说：“仅仅知道一个模型是否有效是不够的。我们需要知道它是如何工作的。”

2012年，Kim觉得几乎没有人关注这些问题，只有她和MIT的导师Cynthia Rudin、Julie Shah对这个话题”略感兴趣”。到2017年，该领域最大的会议上出现了专门讨论可解释性和解释的整个研讨会。到2019年，Google的CEO在公司最大的舞台上自豪地描述她的工作。

我问Kim这是否感觉像是一种证明。

“我们还有很长的路要走，”她说——似乎不愿意让自己享受片刻的满足感，或者”我早就告诉过你”的感觉。相反，她主要感到责任重大：确保这些问题得到解决的速度超过技术本身已经快速发展和部署的速度。她必须设法走在这股浪潮的前面，以确保不会出现任何高风险的问题。

当我们结束谈话并开始走出会议室时，我感谢Kim抽出时间，并问她是否还有什么我们尚未涉及的内容想要补充。她若有所思地停顿了大约十秒钟，然后突然眼睛一亮。

“顺便说一下，我之前提到的那位教员？”就是2012年那个人，当她作为博士一年级学生提到她的博士工作是关于可解释性时，他嗤之以鼻，并试图引导她远离他认为是死胡同的研究方向。不过她并不记仇：“我觉得他确实是在给我建议……这真的是一件合理、贴心的事。”

Kim咧嘴笑了。“那个人现在正在研究可解释性。”

第二部分

主体性(Agency)

强化在人类事务中的作用受到越来越多的关注——这不是因为学习理论中的任何时尚变化，而是因为发现了增强我们预测和控制行为能力的事实和实践，这样做毫无疑问地证明了它们的现实性和重要性。即使是那些为证明强化做出最大贡献的人，也还没有完全掌握强化的范围，而在心理学家中的其他地方，文化惰性是显而易见的。这是可以理解的，因为这种变化几乎是革命性的：传统学习理论中几乎没有什么东西保持可识别的形式。

——B. F. SKINNER¹

现代行为主义理论的问题不在于它们是错误的，而在于它们可能成为现实。

——HANNAH ARENDT²

1896年春天，Gertrude Stein在哈佛大学选修了伟大的William James的心理学研讨班。在那里，她将研究”运动自动症”(motor automatism)，即无需刻意思考就能在纸上写字的能力。由此产生的同行评议文章将是她第一次出现在印刷品中；更重要的是，从本科心理学研究开始，这将直接引导她形成后来闻名于世的现代主义”意识流”散文风格。³ Stein形容James研讨班的同学们是”一群有趣的人”，其中一个人特别有个性：他专注于孵化小鸡。⁴

这个学生就是Edward Thorndike，公平地说，把公寓变成临时鸡舍并不是他的第一个想法。他想要研究人类儿童的学习机制——顺便揭穿超感知觉的观念。哈佛大学没有批准这个项目。所以装满叽叽喳喳小鸡的孵化器就这样……作为后备计划出现了。⁵

对于一个有志成为心理学家的人来说，照料一群小鸡在当时确实很奇怪。动物研究尚未进入二十世纪的流行期，Thorndike的同学们都认为他有点古怪。这可能让他的同龄人皱起了眉头，也让他的房东太太怒发冲冠，但照料他的鸡群——尽管在后勤上是一场噩梦——并非没有好处。例如，当在马萨诸塞州旅行时，Thorndike经常被迫停在一些家庭朋友——Moulton一家——的房子里，在他们的炉子上给小鸡取暖，然后再继续前往剑桥。这可能是研究的必需，不过从他的信件中可以清楚看出，Thorndike在很大程度上是受到与他们年轻女儿Bess调情的机会所驱动。她后来成为了他的妻子。

Thorndike在剑桥的房东太太最终宣布他的孵化器是火灾隐患，并给了他最后通牒：小鸡必须离开。James试图为他在哈佛校园内找到一个实验室空间，但大学并不热情。作为最后的手段，James——不顾妻子的抗议——让Thorndike把小鸡和孵化器一起搬到了自己家的地下室。（至少James的孩子们似乎很高兴。）

1897年，Thorndike从哈佛毕业，搬到纽约市在哥伦比亚大学获得奖学金，这次与各种动物一起工作。这一年经常是郁闷的。他的两只猫跑了，他弄不到任何狗，而且”我的猴子太野了，我碰不了它。”更糟糕的是，Thorndike城市动物园里的动物并不都是为了科学；其中许多只是害虫。在1898年2月14日写给Bess的信中，他写道：“一只老鼠刚跑过我的脚。很多老鼠在啃食梳妆台；三只小鸡在离我不到一码的地方睡觉；我房间的地板上到处都是烟草、烟

蒂、报纸、书、煤炭、鸡笼、猫的牛奶盘、旧鞋、煤油罐和一把扫帚，这扫帚似乎相当格格不入。这是一个荒凉的洞穴，我的这个公寓。“尽管如此，他承诺会清理干净，而且——毕竟他是在情人节写信——他邀请她来看他。

从那种肮脏中将诞生十九世纪一些最具原创性和影响深远的研究。Thorndike建造了一套他称为“迷箱”的装置，里面满是锁扣、杠杆和按钮。他会把动物放进箱子里，在外面放一些食物，然后观察动物——小鸡、猫和狗——设法找到出路的方式。“你会喜欢看小猫打开拇指锁扣、门上的按钮、拉绳子、像狗一样乞求食物和其他这样的特技，”Thorndike写信给Bess，“与此同时我在吃苹果和抽烟。”

Thorndike对学习感兴趣，一个理论正在他的头脑中成形。Thorndike观察到的——在吃那些苹果和抽那些烟的时候——是动物们对箱子的实际工作原理一无所知，最初的行为几乎纯粹是随机的：咬东西、推东西。一旦这些随机行动中的一个让它们逃出箱子，它们很快就学会重复那个行动，并且在未来越来越能够快速逃出同一个箱子。“在许多偶然冲动中，能带来快乐的那一个，”他观察到，“因此变得强化和固化。”

Thorndike在这里看到了一个更大、更普遍的自然法则的构成要素。正如他所说，我们行动的结果要么是“令人满意的”，要么是“令人烦恼的”。当行动的结果“令人满意”时，我们倾向于更多地做这件事。另一方面，当结果“令人烦恼”时，我们会少做这件事。行动和结果之间的联系越清晰，由此产生的变化就越强。Thorndike把这个想法称为，也许是他职业生涯中最著名和持久的，“效果法则”。

正如他所说：

效果法则是：当情境和反应之间的可修改连接建立，并伴随或紧跟令人满意的事态时，该连接的强度增加；当建立并伴随或紧跟令人烦恼的事态时，其强度减少。满意性的强化效果（或烦恼性的弱化效果）对连接的作用随着它与连接之间关系的紧密程度而变化。

从这个看似温和而直观的想法中，将建立起二十世纪心理学的大部分内容。1927年，Ivan Pavlov的《条件反射》英文版出现，由他的学生Gleb von Anrep翻译，他使用“强化”一词来指代这种效果。到1933年，Thorndike自己也在使用“强化”一词——一位名叫Burrhus Frederic (B. F.) Skinner的年轻哈佛博士后也是如此（我们将在第5章更正式地认识他）。“强化”的语言——连同使用动物来理解试错学习机制，从而理解人类心智的想法——为心理学未来几十年的运作奠定了框架。

正如下一代领先心理学家之一Edward Tolman将写到的：“动物学习心理学——更不用说儿童学习心理学——过去是，现在仍主要是同意或不同意Thorndike的问题。”

如果Thorndike最终接受了他在心理学经典中的地位，他对此表现出了不同寻常的谦逊。当Thorndike的教科书《心理学要素》于1905年问世时，它威胁要取代之前的主流教科书《心理学：简明教程》，而该书的作者正是他在哈佛的导师William James。Thorndike寄给他一张一百美元的支票来补偿版税损失。James回复道：“说真的，Thorndike，你简直是自然界的怪胎。当自然界的第一法则是杀死所有竞争对手时（尤其是在教科书领域），你却用收益来养活他们！”James拒绝兑现支票，并将其退回。火炬以某种方式传递了——二十世纪心理学正在走向成熟。

数字试错

什么样的力量通过什么样的过程或机制能够成为并完成确认反应所是和所做的事？在我看来符合所有或几乎所有事实的答案是…强化的力量和机制，应用于连接。

——EDWARD THORNDIKE

如果追随Thorndike的动物研究者们像他一样最终对人类儿童心理学感兴趣，他们并不孤单；计算机科学家——最早的一些——也是如此。Alan Turing最著名的论文“计算机机械与智能”于1950年发表，明确地将人工智能项目框架化为这些术语。“与其试图制作一个程序来模拟成人思维，”他写道，“为什么不尝试制作一个模拟儿童思维的程序呢？如果将其置于适当的教育过程中，就能获得成人大脑。”Turing将这些人造儿童思维想象为他所称的“无组织机器”，从随机配置开始，然后根据其（最初随机的）行动结果质量进行修改。

因此，通向人工智能的路线图已经成形。“无组织机器”将直接借鉴已知的神经系统知识，而“教育过程”将直接借鉴行为主义者关于动物（和儿童）如何学习的发现。Warren McCulloch和Walter Pitts在40年代初已经证明，大量人工“神经元”的适当连接组合能够计算几乎任何东西。Turing已经开始勾画这样的网络如何通过试错进行训练的方法。实际上，这正是Thorndike五十年前在他烟雾缭绕的公寓动物园中描述的“印记”过程；Turing的描述几乎逐字逐句地类似于Thorndike的效果法则：

当达到一个行动未确定的配置时，对缺失数据进行随机选择，并在描述中试探性地进行适当的输入并应用。
当出现疼痛刺激时，所有试探性输入被取消，当出现快乐刺激时，它们都变为永久性的。

到1950年代末，IBM的Arthur Samuel在公司的Poughkeepsie实验室工作，他构建了一个玩跳棋的程序，以粗糙而早期的方式，根据赢棋和输棋来调整自己的参数。不久Samuel就开始输给自己的创造物。正如《纽约客》在1959年报道的那样，“Samuel博士因此可能成为历史上第一位向自己设计的对手承认失败的科学家。”他发表了一份名为《使用跳棋游戏进行机器学习的一些研究》的报告，采用“机器学习”一词来描述这种方法。Samuel写道：

这里报告的研究关注的是如何为数字计算机编程，使其行为方式如果由人类或动物完成，会被描述为涉及学习过程…我们拥有具备足够数据处理能力和足够计算速度的计算机来利用机器学习技术，但我们对这些技术基本原理的知识仍然初级。缺乏这种知识，有必要详细而精确地指定问题解决方法，这是一个耗时且昂贵的过程。通过编程让计算机从经验中学习，最终应该消除大部分详细编程工作的需要。

用更通俗的英语，他解释说：“这是我经历过的最令人满意的事情之一…据我所知，没有其他人让数字计算机自发地改进。你看，计算机能够模拟的心理活动种类一直受到严重限制，因为我们必须准确地告诉它们做什么以及如何做。”

换句话说，继续开发能够学习的机器——无论是通过人类指导还是它们自己的经验——将减轻编程的需要。此外，它将使计算机能够做我们不知道如何编程让它们做的事情。

Samuel研究的公布成为计算机科学传奇。AI先驱John McCarthy回忆说，当Samuel准备在国家电视台演示他的跳棋程序时，“IBM创始人兼总裁Thomas J. Watson Sr.评论说，这次演示将使IBM股价上涨15个点。确实如此。”

享乐主义神经元

神经细胞各自为生存而斗争，这与我们的欲望为满足而斗争相平行。

—爱德华·桑代克¹⁷

1972年，哈里·克洛普夫(Harry Klopf)——一名在俄亥俄州代顿市赖特-帕特森空军基地为美国空军工作的研究员——发表了一份具有煽动性的报告，题为《大脑功能与适应系统：一种异态平衡理论》。克洛普夫认为“神经元是享乐主义者”：它们致力于最大化某种近似的、局部的“快乐”概念，并最小化某种“痛苦”概念。克洛普夫相信，人类和动物行为的全部复杂性，正是这些个体细胞“享乐主义者”连接成日益复杂系统的结果。

在克洛普夫之前的一代人，即40年代到60年代所谓的控制论运动，专门从他们所称的“负反馈”角度来阐释智能行为。他们认为，有机体主要受到稳态或平衡的驱动。它们努力维持舒适的温度。它们进食以平息饥饿。它们交配以平息欲望。它们睡眠以平息疲劳。一切似乎都是为了回到基线状态。

确实，1943年控制论的开创性论文《行为、目的和目的论》——顺便提一下，该论文创造了“反馈”一词在“用于调节的信息”这一现在通用意义上的用法——致力于区分有目的和无目的(或随机)的行为。¹⁸对于控制论者来说，目的等同于一个可以作为休息场所到达的目标。对于控制论的主要代表诺伯特·维纳(Norbert Wiener)来说，典型的“内在有目的”机器之一是恒温器：当温度过低时，它打开加热装置，当温度升高到足够时，它关闭加热装置。他还想到了发动机的“调速器”——绝非巧合的是，这个术语在词源上与“控制论”一词相关，来自同一个希腊词根*kybernetes*。¹⁹(因此，“控制论”尽管有其奇特的科幻色彩，完全可能成为听起来更加平淡和官僚化的“治理学”领域。)“机械”调速器”在发动机运行过快时打开阀门，运行过慢时关闭阀门，帮助发动机维持平衡。“注意反馈倾向于对抗系统已经在做的事情，因此是负的，”维纳写道。²⁰在控制论观点中，这是任何目标导向系统的重要组成部分。“所有有目的的行为，”控制论者写道，“都可以被认为需要负反馈。”²¹

克洛普夫完全不同意这种观点。对他来说，有机体是最大化者，而不是最小化者。生命是关于成长、繁殖、无穷无尽和无边无际且贪得无厌的前进，在任何意义上都是如此。对克洛普夫来说，目标根本不是稳态，而是相反。“生存适应系统寻求的主要目标是最大化状态(异态平衡)，而不是...稳态状态(稳态平衡)。”更诗意地说，他支持正反馈而非负反馈的美德：“正反馈和负反馈对生命过程都是必需的。然而，正反馈是主导力量——它提供了‘生命的火花’。”这个概念一直贯穿从单细胞到有机体再到社会。克洛普夫对他所看到的这个想法的含义毫不谦虚。“这似乎是第一个能够提供单一统一框架的理论，在这个框架内可以理解生存适应系统的神经生理学、心理学和社会学特性，”他写道。“神经元、神经系统和国家都是异态平衡器。”²²

神经元是贪得无厌的最大化者？这解释了国家的行为？这是一个雄心勃勃、非正统且很可能荒谬的想法。空军为他提供资金，让他寻找或组建一个研究实验室来调查这个想法。他在马萨诸塞大学阿默斯特分校找到了他的团队。在那里，他聘请了一名名叫安德鲁·巴托(Andrew Barto)的博士后研究员来确定——正如巴托所说——“这个想法的科学意义：它是一个疯狂的、怪诞的想法，还是具有一些科学价值？”²³

克洛普夫也许不寻常地在这段时间里一直与斯坦福大学一名聪明的心理学本科生保持通信。当巴托加入时，克洛普夫鼓励他：“有个非常聪明的孩子，你应该让他参与这个项目。”²⁴这个聪明的孩子名叫理查德·萨顿(Richard Sutton)；当时还是十几岁的他将加入阿默斯特的巴托，成为巴托的第一个研究生。

我问巴托是否有任何即将发生事情的预示。“我们毫无头绪，”他告诉我。“我们毫无头绪。”萨顿和巴托将在克洛普夫的空军资助下，开始一段四十五年的合作，这基本上创立了一个新领域。这个跨越神经科学、行为主义心理学、工程学和数学的领域被称为“强化学习”；他们的名字在AI文献中永远联系在一起——“巴托与萨顿”，“萨顿与巴托”——将成为他们共同撰写的该领域权威教科书的代名词，确实几乎成为强化学习领域本身的简称。²⁵

我在马萨诸塞大学校园里与巴托会面，他现在是名誉教授。“快乐地退休了，”他说。“坦率地说，很高兴摆脱了关于AI和强化学习的炒作和兴奋的漩涡。”

我告诉他，我很兴奋能谈论强化学习作为一个领域的历史——通常被称为“RL”——特别是它对安全、决策制定、人类认知的影响——他打断了我。

“那么，你有多少时间？”

大部分时间，我告诉他。这就是我们需要的。

奖励假说

Barto和Sutton采用了Harry Klopf关于生物体作为最大化器的想法，并给它一个具体的数学形式。想象你处于一个包含某种数值奖励的环境中，如果你采取某些行动并达到某些状态，这些奖励就可以获得。你的任务是获得最多的奖励——你能得到的最高“分数”。

环境可能是一个迷宫，你的行动可能是向北、南、东或西移动，你的奖励可能来自到达出口（也许减去少量时间惩罚）。或者环境可能是棋盘，你的行动是移动棋子，你的奖励是将对手将死得一分（平局得半分）。或者环境是股票市场，你的行动是买卖，你的奖励以投资组合的美元价值来衡量。这些场景的复杂性几乎没有限制——环境可能是国民经济，行动可能是发布立法或外交词汇，奖励可能是长期GDP增长——只要奖励是所谓的标量：它们是可比较的、可替代的、具有共同货币。

事实上，强化学习框架已经被证明是如此通用，以至于它导致了一个被称为“奖励假说”的想法：“我们所说的目标和目的都可以很好地被认为是接收到的标量奖励累积和的最大化。”²⁶

“这几乎是一个哲学问题，”Richard Sutton说。“但你有点相信这一点。”²⁷

当然，并非每个人都如此轻易地接受这个前提。在体育、棋盘游戏和视频游戏、金融领域，确实可能存在这样的标量数字，一种用于衡量所有结果的单一货币（字面意义或比喻意义）——但这真的可以应用于像人类甚至动物生活这样复杂多样的事物吗？我们经常必须做出结果似乎是苹果和橘子的决定。我们是否加班到很晚，提高在老板面前的地位但考验配偶的耐心？我们是否优先考虑成就的生活、冒险的生活、人际关系的生活，还是精神成长的生活？例如，牛津大学哲学家Ruth Chang花了几十年时间论证，没有什么比我们拥有的各种动机和目标的不可比性更能表征人类状况。我们不能简单地将重大人生选择放在天平上并权衡出哪个最好——否则道德就只不过是纯粹的理性，没有创造意义、塑造身份的机会。²⁸

Sutton自己也承认奖励假说”可能最终是错误的，但如此简单，我们必须先反驳它，然后才考虑任何更复杂的东西。”

即使我们（暂时）接受奖励最大化的框架，以及标量奖励的可比性，我们也发现成为一个异态最大化器比听起来更难。事实上，强化学习问题充满了哲学和数学的困难。

第一个挑战是我们的决定是相互关联的。在这里，强化学习与无监督学习（构建我们在第1章中探索的向量词表示）和监督学习（用于从ImageNet竞赛到人脸识别到累犯风险评估的一切）都有微妙但重要的区别。在那些设置中，每个决定都是独立的。你的系统被展示一张图片——比如说，一个鸡油菌蘑菇——并被要求对其进行分类。系统可能答对了，可能答错了，它的参数可能在出错时被微调一点，但无论如何，你只需从你的收藏中随机抽取另一张照片继续前进。作为输入的数据是统计学家所说的“i.i.d.”：独立同分布。我们看到什么、我们做什么以及我们接下来看到什么之间没有因果联系。

在强化学习中——在迷宫中、在象棋游戏中、确实在生活中——我们没有在真空中做决定的奢侈。我们做的每个决定都为我们的下一个决定设置了背景——事实上，它可能永久改变那个背景。如果我们以某种方式发展我们的棋子，我们就强烈限制了我们将遇到的位置类型和未来可能有用的策略类型。在空间世界中，无论是虚拟的还是真实的，我们采取的行动——在迷宫中向北移动，转身面对我们的爱人，在佛罗里达过冬——塑造我们接收到的未来输入，要么是暂时的，要么是永远的，这是运动的内在特征。如果我们牺牲了我们的皇后，在游戏的其余时间里它对

我们就不可用了。如果我们从屋顶跳下，我们可能再也不会跳了。如果我们将对某人不友善，我们可能永远改变那个人对我们的行为方式，我们可能永远不知道如果我们更友善他们会如何表现。

强化学习相对于监督学习和无监督学习的第二个挑战是，我们从环境中获得的奖励或惩罚——由于其纯粹的标量性质——是简洁的。被训练来预测缺失单词的语言模型，在每次猜测后，都会被告知正确的单词是什么。试图对图像进行分类的图像分类器会立即得到该图像的“正确”标签。然后它会明确地朝着正确答案的方向更新自己。相比之下，强化学习系统尽其所能地试图在某个环境中最大化某个数量，最终会了解到它取得了什么分数，但无论胜负，它可能永远不知道“正确”或“最佳”的行动应该是什么。当火箭爆炸，或桥梁倒塌，或你试图搬运的一摞盘子倒下，或你踢的球没有进入球门时，世界对结果是水晶般清晰的。但它对你到底应该做什么不同的事情保持沉默。

正如Andrew Barto所说，强化学习不像是与老师一起学习，而是与批评家一起学习。²⁹ 批评家可能同样睿智，但远没有那么有帮助。老师会在你工作时从你肩膀后面看着，立即纠正你的错误，告诉你或向你展示你应该做什么。批评家等到你的工作完成后，然后从黑暗礼堂的后面大喊“嘘！”，让你对他们到底不喜欢什么，或者他们更喜欢什么一无所知。批评家甚至可能不会故意保留任何洞察或建设性反馈；他们自己可能除了他们的满意或烦恼程度——用Thorndike的话来说——之外什么都没有留下。

第三，反馈不仅简洁且不是特别有建设性，而且是延迟的。例如，我们可能在游戏的第五步犯一个无法挽回的错误，而致命一击在一百步之后才到来。当有人经历失败或挫败时——失望的父母、破产的企业主、被捉的小偷——最先想到的习语是“我在哪里出了错？”“在强化学习中，这被称为‘功劳分配问题’”，自本世纪中叶以来一直困扰着研究人员。例如，MIT的Marvin Minsky在他1961年著名的论文《迈向人工智能的步骤》中写道：“在玩象棋或跳棋这样的复杂游戏时，或在编写计算机程序时，人们有明确的成功标准——游戏是赢还是输。但在游戏过程中，每个最终的成功（或失败）都与大量的内部决策相关。如果运行成功，我们如何在众多决策中为成功分配功劳？”

Minsky详细阐述，强调这一点：“假设一项复杂任务（如赢得一场象棋游戏）涉及一百万个决策。我们能否为每个决策分配...完成任务功劳的百万分之一？”³⁰

在一次跨国公路旅行几天后的车祸之后，我们不会将我们的行动一路追溯到在点火器中转动钥匙，然后想，“好吧，这是我最后一次从车道向左转！”但是，我们也不会想象，在象棋游戏的第89步被将死后，我们的失误一定发生在第88步。

那么我们如何确定从成功和失败中汲取的正确教训呢？强化学习框架开始为学习和行为的基本问题开辟一个远景，这将发展成一个完整的领域，并将指导人工智能研究的方向直到我们现在的十年。正如Harry Klopff所设想的那样，它也将为我们提供一套新的问题来询问围绕我们的自然智能。

多巴胺的谜题

如果人类和其他动物可以被认为是“异态”奖励最大化者，那么这些奖励通过大脑中的某种机制运作是合理的。如果真的存在某种单一的、标量的“奖励”，人类和动物被设计来最大化它，那么它可能就像大脑中的化学物质或回路一样简单吗？在1950年代，蒙特利尔麦吉尔大学的一对研究人员James Olds和Peter Milner，看似——诱人地——找到了它的位置。

Olds和Milner正在实验将电极放置在老鼠大脑的各个位置，并给老鼠机会按下一个杠杆，该杠杆会通过这些电极向大脑的特定部分传递电流。他们发现，大脑的某些区域似乎对老鼠的行为没有影响。在其他区域，老鼠似乎会竭尽全力确保杠杆不被按下。但有一些区域——特别是所谓的“隔膜区域”——老鼠似乎几乎没有什么比按下向该区域传递电流的杠杆更想要的了。老鼠每小时按杠杆多达5,000次，连续24小时不休息。³¹“通过这种奖励对动物行为施加的控制是极端的，”Olds和Milner写道，“可能超过了动物实验中以前使用的任何其他奖励所施加的控制。”³²为研究奠定了基础，不仅通过奖励最大化的视角研究人类和动物行为，还研究奖励本身的实际分子机制。

起初这些区域被称为“强化结构”，但Olds很快开始将它们称为“快乐中心”。³³后续研究表明，不仅是老鼠，人类也会不遗余力地接受在大脑类似部位的电流刺激。

随着时间推移，研究开始确立，大脑中这种电刺激最引人注目的区域是那些涉及产生一种神经递质的神经元的区域，这种神经递质叫做3,4-二羟基苯乙胺——更为人知的是它的缩写绰号：多巴胺(dopamine)。³⁴这些细胞很稀少——不到大脑的1%——而且它们聚集在非常特定的区域。³⁵在某些情况下，单个多巴胺细胞可以连接到数百万个其他神经元，遍布大脑的广泛区域。³⁶实际上，它们几乎是独特的广泛连接，连接度最高的细胞每个都有近十五英尺长的轴突线路遍布大脑内部。³⁷然而，与此同时，它们在输出的范围和复杂性方面相当有限。正如纽约大学神经科学家Paul Glimcher所说，这意味着“它们不能对大脑其他部分说太多，但它们所说的必须被广泛听到”。³⁸

换句话说，这听起来很像一个“奖励标量”——本质上，就像大脑记分板上的分数——极其简单，但广泛传播且至关重要。多巴胺似乎也与成瘾药物密切相关，包括可卡因、海洛因和酒精。多巴胺会不会真的就是大脑中奖励的分子货币？

1970年代后期的工作似乎暗示了这一点。例如，1978年的一项研究显示，当老鼠在学习按压杠杆可以得到食物颗粒之前被给予阻断多巴胺的药物哌迷清(pimozide)时，它们对按压杠杆的兴趣并不比那些从未学过食物连接的老鼠更大。不知何故，食物奖励似乎对用哌迷清处理的老鼠没有效果。研究人员写道，哌迷清“似乎选择性地钝化了食物和其他享乐刺激的奖励影响”。³⁹正如神经科学家Roy Wise在1982年写道的，就好像“生活中的所有快乐——主要强化的快乐和它们相关刺激的快乐——失去了唤起动物的能力”。⁴⁰实际上，当给予已知阻断多巴胺受体的药物时，老鼠也停止了向大脑输送电流。用神经科学家George Fouriezos的话说，这种药物“从电压中取走了震动”。⁴¹就好像从食物和水到性到自我施加的电流，一切都失去了愉悦效果。

在1980年代，电生理学开始发展到可以实时监测单个多巴胺神经元的程度。德国神经生理学家Wolfram Schultz在瑞士弗里堡的实验室开始研究猴子的多巴胺神经元行为，当它们将手伸进一个有时空着、有时装有小块水果或烘焙食品的盒子时。果然——啪——“当猴子的手触摸到盒子里的一小块食物时，出现了脉冲爆发”。⁴²这似乎证实了科学家们实际上已经找到了奖励的化学物质。

但有些奇怪的事情。

在有某种视觉或听觉提示表明食物出现在盒子里的情况下，提示触发了多巴胺活动的激增。然后猴子会伸手抓取食物，而Schultz尽职地监测读数——什么都没发生。只是它们正常基线活动的平静背景静态。没有峰值。他写道，多巴胺神经元”对[提示]发出脉冲爆发，但完全未能对触摸食物做出反应”。⁴³

到底发生了什么？

“我们搞不清楚这是怎么回事，” Schultz说。⁴⁴ 他翻找了几个假设。也许猴子只是饱了，真的不想要更多食物了。他试着让它们挨饿。没用。它们贪婪地吃着食物。但没有多巴胺峰值。

Schultz和他的合作者们在80年代后期和90年代初期试图为他们所看到的现象找到一个合理的解释。⁴⁵ 在重复过程中，多巴胺峰值已经从食物转移到了提示——但这意味着什么？如果食物不知何故停止了”奖励”，那么为什么猴子总是如此迅速地抓取和吃掉它？这说不通，所以多巴胺不可能直接代表奖励。他们排除了与工作记忆的连接。他们排除了与运动的连接，以及与触觉的连接。

“我们无法将其归结为某个具体的东西，” Schultz告诉我。“你开始想把它归结为激励、动机，作为对刺激的反应让你行动……我们最初是这样认为的，但后来发现这个概念过于模糊。”他的实验室开始专注于一个想法，认为这与惊喜或不可预测性有关。心理学中有一个叫做Rescorla-Wagner模型的理论，该模型表明学习严重依赖于惊喜。⁴⁶ 也许多巴胺与这种联系有关：也许它以某种方式代表了惊喜本身或惊喜引发的学习过程。这可以解释为什么食物在意外出现时会产生多巴胺峰值，而在有提示时则不会——以及为什么在那些情况下，是意外的提示引发了峰值。“这样说得通，” Schultz说。“但这无法解释我们拥有的一些数据。”⁴⁷

特别是，Schultz观察到了另一个现象，这个现象比第一个更加神秘。他正在进行一项后续研究，使用类似的设计，但这次用杠杆和果汁代替了盒子和食物。然而，一旦猴子学会了提示可靠地预示着果汁，Schultz尝试了一些新东西：他给了它一个虚假警报。他触发了提示；像往常一样，猴子的多巴胺活动在神经元的正常基线嗡嗡声之上出现峰值。猴子按压了杠杆。没有果汁出来。猴子的多巴胺神经元下降了——短暂地，但明确地——沉默了。“然后我说，嗯。这与惊喜不同。”⁴⁸

多巴胺是一个谜。起初，它似乎很明显地是大脑的奖励货币。它显然在测量某种东西。如果不是奖励，不是注意力，不是新奇性，不是惊喜——那么是什么？

策略和价值函数

很难聪明到下一分钟都不会愚弄你。

—LEO STEIN⁴⁹

在同样的1980年代期间，当Schultz开始他的多巴胺系统实验时，在大西洋的另一边，Barto和Sutton开始在强化学习问题上取得数学进展。

第一个重大步骤是分解问题。他们意识到，学习如何在环境中采取行动以最大化奖励，涉及两个相关但可能独立的子问题：行动和估计。当你开始理解并最终掌握一个领域时，你学会了两个重要的东西：如何在给定情况下采取正确行动，以及如何估计某种状态可能带来的未来奖励。正如Barto和Sutton逐渐意识到的，一个看着棋盘的国际象棋选手有两种不同的直觉。她对下一步应该下哪些棋（以及对手可能如何回应）有某种直觉。她对在这种位置下哪个玩家可能获胜有另一种直觉。强化学习问题的这两个维度被称为策略——何时做什么——和价值函数——期望什么奖励或惩罚。

至少在理论上，仅凭其中任何一个都足以解决问题。如果一个国际象棋选手总是知道正确的走法，那么他们在预测谁会赢方面很糟糕也无关紧要。然而，反之亦然：如果一个国际象棋选手，比如说，总是确切地知道在给定位置下谁占优势，那么他们不确定该做什么也无关紧要。只要有足够的时间，他们可以简单地权衡每一步的后果，选择导致最有希望未来的那一步。

基于策略的方法导致系统——无论是动物、人类还是机器——具有高度训练的“肌肉记忆”。正确的行为就是毫不费力地流淌出来。相比之下，基于价值的方法导致系统具有高度训练的“蜘蛛感应”。它可以立即判断情况是威胁性的还是有希望的。如果充分发展，仅凭其中任何一个都足够了。

然而，在实践中，基于策略的方法和基于价值的方法是齐头并进的。Barto和Sutton开始阐述一个被称为“演员-评论家”架构的想法，其中系统的“演员”部分将学会采取好的行动，“评论家”部分将学会预测未来奖励。⁵⁰大致上，演员-评论家架构也很好地描述了他们的合作。Barto首先且最重要的是对行为感兴趣。Sutton首先且最重要的是对预测感兴趣。⁵¹

回溯到1978年他的本科论文，Sutton一直专注于创建他所说的“期望的统一理论”。⁵²这种对有机体如何形成和完善其期望的迷恋在他与Klopf和Barto在阿默斯特的工作中一直伴随着他。

正如Sutton推理的那样，发展良好的期望——一个好的价值函数——意味着将你每时每刻的期望与来自现实的最终判决协调起来：比赛的最终得分、季末报告、成功的登月、职业生涯后期来自钦佩同事的赞美、起立鼓掌、微笑的孙子们。但如果你真的必须等到比赛结束才能从中学习，那么信用分配问题确实几乎是不可能的。他说，这个逻辑是三重的。

首先，记住我们所想的和所做的一切可能是不切实际或不可能的。在一场我们惨败的90分钟足球比赛结束后，我们真的会聚集在更衣室里，回顾每一次进攻吗？回顾我们希望和恐惧的每一次转变吗？

其次，我们希望能够在没有最终裁决的情况下学习。一场在结束前中断的国际象棋游戏仍然应该为我们提供一些东西。如果我们即将被将死或者局势感觉绝望，那仍然意味着我们需要做一些不同的事情——无论我们感觉即将到来

的失败或惩罚是否真的到达。同样，一场由最后几秒钟的意外、不可能、不可预见的事件决定的游戏并不一定意味着我们早期的期望必然是错误的。也许我们真的正在走向胜利。完全根据最终结果进行判断并不一定有意义，特别是在某种程度上受到偶然性影响的情况下。

第三，我们理想情况下希望不仅在事后学习，而且要边做边学。这对人类生活具有特别重要的关键意义。生活中许多最关键的节点和最重要的目标——比如说，进入大学，或者抚养孩子，或者舒适地退休——我们通常只有一次机会。而许多错误——成绩点数平均分下降、腰围扩大、关系恶化——都是这样的，我们可以意识到事情正在偏离轨道并在为时已晚之前很久就进行调整。如果是通过试验和错误来学习，那么幸运的是，我们不需要整个试验——也不需要整个错误——就能做到这一点。

正如Sutton解释的那样，这些因素中的每一个都使期望理论变得更加棘手。“所以显然我们可以试着忽略所有这些东西，把它们当作麻烦，但我把它们当作线索，”他说。“这些是来自自然的提示，关于我们应该如何进行。”⁵³

事实证明，正是Sutton本人第一个接受了这些提示。

Sutton在思考预测方面取得的突破是这样的。当你朝着不确定的未来前进时，你保持着一种“动态期望”，即事情看起来有多有希望。在国际象棋游戏中，这可能是你给自己赢得游戏的几率。在电子游戏中，这可能是你期望取得多少进展或你期望总共积累多少分数。这些猜测随时间波动，一般来说，你越接近你试图预测的任何事情，它们就越准确。（周末天气预报几乎总是在周四比在周一更准确。当你从机场开车回家时，你越接近家，你的预计到达时间往往越准确。）这意味着，一般来说，随着我们的期望波动，我们得到连续期望之间的差异，每一个都是学习机会；Sutton称这些为*temporal differences*，或*TD errors*。当这些*temporal differences*中的一个发生时，两个估计中较晚的那个更可能是正确的。

所以也许我们不需要等到获得最终的基础事实才能学到一些东西。也许我们可以从这些波动本身学习。我们期望改变的任何时候都可以被视为我们先前估计中的错误，因此，是一个学习的机会：不是从尚未到达的最终真相学习，而是从新的估计学习，这个估计是由我们非常略微年长和明智的自己做出的。正如Sutton所说：“我们正在从一个猜测中学习一个猜测。”⁵⁴

（他补充说，仿佛是顺便提到，“听起来有点危险，不是吗？”）

到1980年代末，这个想法——Sutton称之为“*temporal differences方法*”，或简称“*TD learning*”——已经锐化为一个算法，他称之为“*TD(λ)*”（“*TD-lambda*”），用于根据后续预测精确调整预测。⁵⁵受到Sutton和Barto想法的启发，剑桥博士生Chris Watkins设计了一个叫做“*Q-learning*”的TD算法，它将这些预测转化为行动。⁵⁶有希望的是，他表明*Q-learning*总是会“收敛”，即，只要系统有机会从每个状态尝试每个行动，尽可能多次，它总是，最终会发展出完美的价值函数：对那个环境的完美期望集合，无论是迷宫、棋盘、还是更贴近生活的东西。这是该领域的一个重要理论里程碑——当然，确切地说有多重要取决于你是那种强调总是还是强调最终的人。

理论看起来很不错，但时间差分学习(*TD learning*)还没有在实践中得到真正的测试。这第一次真正的实践检验将在纽约的IBM研究中心进行。一位名叫Gerald Tesauro的年轻研究员——最初对经典条件反射的计算模型感兴趣——在80年代末到90年代初期，一直在研究使用神经网络来玩西洋双陆棋。早期的结果很有希望。然后，在1992年，他将TD学习插入到他的模型中，结果如火箭般飞速发展。它在从猜测中学习猜测，稳步学会什么是有利位置的样子。“这显然是这种算法在复杂非琐碎任务上的首次应用，”Tesauro写道。“研究发现，在零知识内置的情况下，网络能够从零开始学习整个游戏，达到相当强的中级水平，这明显优于传统的商业程序，实际上超越了在大量人类专家数据集上训练的可比网络。这表明TD学习在实践中可能比基于当前理论所预期的效果更好。”到1994年，他的程序——被称为“*TD-Gammon*”——已经达到了，正如他所写的，“真正惊人的性能水平：*TD-Gammon*的最新版本现在...极其接近世界上最好的人类玩家。”

该领域早就知道程序员可以编写一个比自己玩得更好的系统；Arthur Samuel在1950年代用跳棋就发现了这一点。但这是全新的东西：一个和任何人类玩家一样好的程序。而且它完全是“自学成才”的：从随机状态初始化，通过与自己的无数次对局进行调优。这是对TD学习的巨大验证。它将为成功提供直接蓝图，这个蓝图将被21世纪的游戏精通软件如AlphaZero使用——变化惊人地少。

Sutton和新兴强化学习领域的其他人接受了自然的暗示。但现在轮到他们提供一些回馈了。在该领域历史上最令人震惊的时刻之一，最初从心理学和神经科学发展出来的数学框架即将以重大方式重新进入这些领域。

预测和奖励的神经基础

在1990年代初，当Tesauro正在微调他的西洋双陆棋系统时，一位名叫Peter Dayan的年轻认知科学家，他曾在Barto的实验室度过了富有成果的时光，并与Watkins一起研究收敛证明，发现自己从该领域——以及世界——的一边转移到另一边。从在爱丁堡大学研究时间差分的数学，他来到了圣地亚哥索尔克研究所的一群神经科学家那里。

Dayan和一位名叫Read Montague的博士后同事，与索尔克研究所神经科学家Terry Sejnowski一起工作，有一种直觉，即强化学习框架不仅解释了真正的人类和动物大脑可能如何运作——而且它可能真正就是大脑在做的事情。“我们追求大脑中一套系统的作用，这些系统报告价值和强化，”Montague说。“而我们当时只是想，‘这些东西应该在实现某种算法。’”Montague认为这些学习机制可能是几乎每个学习动物的基石属性之一。Dayan来自他在时间差分学习上的工作，这似乎可能是这些通用算法之一的候选者。它在白板上有效——也就是说，它收敛。它在机器中有效——TD-Gammon是比除了最好的人类玩家之外的所有人都更强的玩家。也许它在大脑中也有效。他们两人开始推测从时间差分学习的神经系统可能如何工作。

“显然我们知道必须与神经科学有联系，”Dayan在旧金山Uber总部的会议桌旁向我解释，他正在那里度过从伦敦大学学院的休假年。

“然后，”他说，“我们遇到了来自Wolfram Schultz的数据。”

突然Dayan跳起来，开始兴奋地在白板上涂画多巴胺反应图。他指向平坦的背景静态，当猴子在看到提示灯后得到它完全期望的果汁奖励时的平淡反应。“所以信号的这部分是非常透明的。这就是你在Rescorla-Wagner规则中看到的；这是来自较早心理学的美妙方面。”

然后他指向由提示灯本身引起的初始多巴胺峰值。“但信号的这部分是令人困惑的，”他说。“因为从心理学角度来看，这是不被期望的。所以在各种时候你可以看到Wolfram在某种程度上努力理解这意味着什么。”

当Dayan和Montague查看来自Schultz的数据——这些数据让Schultz连同整个神经生理学界都感到如此难以理解——时，他们确切地知道这意味着什么。

这是一个时间差分。这是猴子期望的突然波动——在其价值函数中，对它所处状态有多好的预测中的突然波动。

大脑多巴胺背景噪音之上的突然峰值意味着突然间世界似乎比一刻前更有希望。另一方面，突然的安静意味着突然间事情似乎比想象的更不希望。正常的背景静态意味着事情，无论好坏，都如预期的那样好或坏。

多巴胺系统的峰值本身并不是奖励，但它与奖励相关；它不是不确定性、惊讶或注意本身——但它与所有这些都有密切的关系，并且第一次以可理解的方式展现出来。这是猴子预期中的波动，表明早期的预测是错误的；这是大脑从一个猜测中学习另一个猜测。

在纸面上和硅片中都运行良好的算法，刚刚在大脑中被发现了。时间差分学习(Temporal-difference learning)不仅仅类似多巴胺的功能。它就是多巴胺的功能。

Schultz、Dayan和Montague在1997年在《科学》杂志上发表了一篇具有爆炸性影响的论文，向世界宣布了他们的发现。正如他们所说，他们发现了“预测和奖励的神经基础”⁶¹。

这一发现对神经科学和计算机科学都产生了巨大影响。一个在纯机器学习环境中发展起来的想法，受到心理学中经典条件反射和操作性条件反射模型的启发，突然完成了一个完整的循环。它不仅仅是人工智能可能如何构建的模型。它似乎提供了智能通用原则之一的描述，句号。

“人们认为眼睛进化了大约四十到五十次不同的次数…生物学一次又一次地发现眼睛——各种不同的眼睛，” Montague解释道⁶²。“我认为在学习领域也是同样的情况。学习算法对于理解如何利用当前的经验、重组内部结构并在未来产生行动是如此重要，以至于…你应该期望生物学在许多不同的环境中偶然发现了这些算法。因此你可以在蜜蜂、海蛞蝓、学习歌唱的鸟类、人类和啮齿动物等中看到强化学习系统、奖励系统”⁶³。

Dayan也持相同观点。“我们天生具备这些东西并不奇怪，”他说。“你可以在这些多巴胺神经元的活动中如此透明地看到它这一想法…这是一个相当大的启示”⁶⁴。

当我访问剑桥的Wolfram Schultz实验室时——75岁高龄的他仍然对自己的工作充满活力和热情——我问他当时是否感觉这是一个启示。出人意料的是，他说不是：在他看来，启示在于这个简单的TD学习模型，描述了他的猴子在摸索食物时大脑中的活动，二十年后，正在推动当今AI的突破性进展。我们能够从自然界中获取这些通用思想并合成地实现它们，这才让他震惊。

“启示来了，”他说，“当我意识到TD模型进入了我们现在看到的Go编程，以及这种人工智能、机器学习的东西。那是一个启示，我说，我的天哪，我做了什么？你知道，理解我的数据来自Rescorla-Wagner模型，来自预测误差，但结果，就像，天哪！我的意思是，我们知道Tesauro正在用TD模型编程双陆棋。好吧，我不是双陆棋玩家，但我是围棋玩家。不是很好的那种，但是…我说，该死的，如果他们能编程围棋，这以前一直是个问题，那它就是一个真正好的模型。而且我在他们编程围棋之前就看到了这一点。”

我告诉他，我认为看到神经生理学和机器学习工作如此清晰而强大的综合是非常了不起的，对两个领域都有如此强大的影响。

“绝对是对的，”他说。“绝对是的。这就是它的魅力。你把所有东西都结合在一起。就是说得通”⁶⁵。

对神经科学的影响是变革性的⁶⁶。正如普林斯顿大学的Yael Niv所说，“理解基底神经节多巴胺依赖功能水平上的学习和行动选择的潜在优势不能被夸大：多巴胺与从帕金森病、精神分裂症、重度抑郁症、注意力缺陷多动障碍等各种疾病有关，并延伸到物质滥用和成瘾等决策偏差”⁶⁷。

确实还有很多需要解决的问题，现在看似经典的故事可能会及时被推翻或复杂化⁶⁸。但很明显，Niv说，“强化学习已经在大脑决策研究中留下了永久的印记”⁶⁹。

她回忆起第一次参加年度神经科学学会会议，这是该领域约30,000名研究人员的聚会。“我记得第一次去的时候查找强化学习，那大概是2003年、2004年。整个会议可能有五张海报。这是一整周的会议；每天有两个海报展示，所以上午一个，下午一个。现在，”她说，“每个海报展示都有一个完整的子部分，一整排，都是关于这个的。在十年、十五年里真的走了很长一段路”⁷⁰。

她补充说，“有很多研究在测试大脑中理论的预测，看到’看起来，天哪，这些神经元读了教科书。它们读了Sutton和Barto的教科书。它们确切地知道需要做什么。’”

快乐与错误

如果时间差分学习模型已经揭示了多巴胺在大脑中的功能——不是作为大脑的奖励货币，不是对未来奖励的期望，而是对未来奖励期望中的错误——那么它留下了一些未解的问题。

特别是，它留下了与快乐和幸福主观体验的联系问题。如果多巴胺水平的升高发出类似事情会比我原本以为的要好的信号，那么这种感觉本身就是令人愉悦的。你可以看到人类和动物都会想方设法获得这种感觉，包括通过直接的化学和电刺激多巴胺神经元。

你也可以开始看到人为提高多巴胺水平如何产生不可避免的崩溃。认为事情将比你以为的要好只能维持这么久。最终你意识到它们并不比你以为的要好，多巴胺的活跃就会安静下来——就像在Wolfram Schultz用猴子进行的实验中那样，当果汁提示灯闪烁但没有果汁出现时。实际上，多巴胺可能开出环境无法兑现的支票。这些支票最终会不可避免地退票。然后你的价值函数必须不可避免地回到现实。

当然，这就是多巴胺相关药物的典型体验——可卡因就是一个典型例子。这种药物主要通过抑制大脑对多巴胺的重吸收发挥作用，导致多巴胺的临时“泛滥”。时间差分理论表明，大脑将此解释为一种普遍的感觉，即事情将会很棒——但多巴胺开出了环境奖励无法兑现的支票。最终预期的美好不会到来，相等且相反的负面预测错误必然会随之而来。“看起来一切都会如此美好...”我们可以化学性地欺骗大脑的预测机制——但不是永远。

正如作家David Lenson所说：“可卡因承诺在再过一分钟就能获得前所未知的最大快感，如果向眼睛展示正确的图像，如果再给予一次剂量，如果以正确的方式安排性互动。但那个未来永远不会到来。这种药物确实有身体上的快感，但与总是即将发生的事情相比，这种快感是偶然的、微不足道的。”^[71] 将可卡因理解为一种多巴胺药物，将多巴胺理解为时间差分的化学物质——我们期望的波动——让故事变得清晰。通过人为地倾倒大脑的供应，人们体验到的不是事情确实很棒的极乐，而是事情出人意料地有希望的眩晕兴奋。如果这个承诺没有兑现，时间差分错误就会朝另一个方向摆动，我们的多巴胺系统就会沉默。是我们的高期望出了错。我们被欺骗了。

多巴胺与幸福和快乐主观体验之间的联系仍在研究中。例如，密歇根大学的Kent Berridge在其职业生涯的大部分时间里都在努力分离想要和喜欢的神经科学。^[72] 与此同时，伦敦大学学院的Robb Rutledge开发了一个明确涉及多巴胺的幸福数学模型。

Rutledge与Peter Dayan以及伦敦大学学院的一组合作者设计了一个实验，人们会进行各种赌注并积累一笔钱，同时定期被问到：“你现在有多开心？”^[73] 他们使用强化学习的数学工具对任务进行建模，以区分受试者到目前为止赚了多少钱，他们总共预期赚多少钱，以及他们最近是否在调整这些运行计数上下时感到惊喜或不快。目标是找出与他们自报的瞬时幸福感最佳的数学相关因子。

Rutledge的发现在多个方面都很有启发性。首先是幸福是短暂的。无论在特定赌注中获得1英镑让你多么开心，五次赌注后，92%的影响已经消失。十次赌注前发生的任何事情都可能从未发生过。这意味着受试者的幸福感与他们实际赚了多少钱几乎没有关系。

幸福似乎至少部分由受试者期望赚多少决定——但最关键的似乎是这些期望的违背。正如Rutledge写道：“瞬时幸福是一种状态，它反映的不是事情进展得有多好，而是事情是否比预期的要好。”^[74] 这听起来完全像时间差分错误——换句话说，完全像多巴胺所扮演的角色。

从更宏观的角度来看，我们开始获得一个关于众所周知的“享乐跑步机”现象的神经科学和计算学解释。也就是说，无论长期生活质量如何变化，人们都会顽固而持续地回到他们的情绪基线。⁷⁵著名的例子是，彩票中奖者和截瘫患者在各自经历戏剧性的生活变化后不久，情感上都会或多或少地回到他们开始的地方。⁷⁶Dopamine和reinforcement learning为我们提供了了解释这种现象的线索。如果幸福感不是来自事情已经进展顺利，不是来自事情即将进展顺利，而是来自事情进展得比预期更好，那么是的，无论好坏，只要我们的期望持续调整以适应现实，那么长期处于愉快惊喜状态应该根本无法维持。

不幸的是，这项研究排除了“始终保持低期望”这样简单的生活技巧。正如Rutledge所说：“较低的期望使结果更有可能超过这些期望，并对幸福感产生积极影响。然而，期望也会在我们了解决策结果之前就影响幸福感。如果你计划在你最喜欢的餐厅与朋友见面，那些积极的期望可能会在你制定计划的那一刻就增加你的幸福感。”⁷⁷

在研究论文中，他和他的合著者提到了一种可能性，比如一家航空公司声称有50%的可能性会延误六小时，然后宣布延误实际上只有一小时。“降低期望增加了积极结果的概率……然而，较低的期望会在结果到来之前降低幸福感，从而限制了这种操纵的有益范围。”⁷⁸

换句话说，Rutledge说：“你不能，你知道，仅仅是降低你的期望就能解决一切问题。”⁷⁹

Dopamine、TD errors和幸福感之间的联系使一些研究人员开始思考是否对reinforcement-learning agents的主观幸福感存在伦理影响。基础研究所(Foundational Research Institute)的Brian Tomasik致力于理解和减少痛苦，他详细思考了reinforcement-learning程序是否具有道德地位的问题——我们如何对待它们是否重要？他的答案是试探性的和有限的肯定：在它们基于与动物和人类大脑相似原理构建的程度上，很可能确实存在一些非零的伦理考量。⁸⁰“当前的RL算法比动物重要得多，”他指出。“不过，我认为RL agents确实在非零程度上很重要，在大规模情况下，它们可能开始累积成重要的东西。”⁸¹其他人的后续工作甚至明确通过TD errors来定义reinforcement-learning agent的“幸福感”。⁸²按照这个逻辑，他们指出，“完全了解世界的agents的预期幸福感为零。”确实，如果dopaminergic“幸福感”在很大程度上来自于被愉快地惊喜，来自于有机会更好地学习期待什么，那么任何领域的完全掌握似乎必然与无聊相关——这一点不仅对未来RL agents的伦理学具有伦理含义，当然，对人类也是如此。

这首先暗示了关于享乐跑步机的进化和计算学故事。如果我们的主观幸福感与被愉快地惊喜密切相关，而智力的本质是不知疲倦地工作以预期和减轻惊喜，那么我们可以看到这种幸福感如何变得转瞬即逝。我们也可以看到进化优势。一个婴儿可能会因为仅仅能够按指令挥动手臂而感到高兴。对于成年人来说，这种能力不再像在婴儿时期那样带来，可以说，刺激。虽然我们可能会为这种在成年人中产生的不安感到惋惜，这种不满足感，但这都是生活课程的一部分。如果基本运动技能足以让我们无限期地兴奋，我们根本不会成长到成年。

正如Andrew Barto所说，这种转瞬即逝是Klopf在20世纪70年代早期就预见到的。“他的观点是一个[稳态平衡]稳定机制，它试图将差异减少到零，当它为零时，它很高兴，它停止。他想要的那种系统永远不会高兴。所以这是一种不断的探索。”⁸³Reinforcement learning、dopamine、幸福感和探索（以及成瘾）之间的联系是我们将在[第6章]中回到的内容。

超越强化学习

Reinforcement learning植根于20世纪早期的动物学习研究，在20世纪70年代和80年代的抽象数学机器学习世界中蓬勃发展，随后以近乎完美的模型凯旋般地重新进入动物行为文献，这个模型已成为大脑中dopamine作用的公认理论。⁸⁴反过来，这个模型让我们对人类动机和人类幸福感有了更深入的洞察。

与此同时，神经科学证据（最近到2018年）开始表明，Harry Klopfer的疯狂假设——神经元是“享乐主义者”，由它们自己的个体桑代克效应定律所驱动——可能并非那么离谱。“我认为神经科学正在接近与Klopfer所提出的非常相似的观点，”Barto说，对他已故的前导师怀着某种自豪，他和Sutton的教科书正是献给这位导师的。

强化学习还为我们提供了一个强大的、也许甚至是通用的智能定义。如果智能如计算机科学家John McCarthy著名的定义那样，是“在世界中实现目标能力的计算部分”，那么强化学习为此提供了一个惊人通用的工具箱。事实上，其核心原理很可能被进化一次又一次地偶然发现——它们很可能将构成二十一世纪任何人工智能的基石。

然而，在某些方面，对动物和机器在世界中实现目标能力的更深理解，已经将更深刻的哲学问题踢到了路边。这个理论明确地并没有告诉我们什么是我们所重视的，或者我们应该重视什么。多巴胺在这方面，现在就像其作用不太为人所知时一样神秘。如果它代表一个标量预测误差，它隐藏了在如何“测量”该预测方面的巨大复杂性领域。如果在第一扇门后面，不是我们期望的加勒比海假期，而是观赏北极光的旅行，我们的多巴胺会快速可靠地指示我们是否对此感到愉快或不愉快的惊讶。但这些替代方案的价值实际上是如何被评估的呢？多巴胺对这一点保持沉默。

与此同时，我们提出另一个问题。经典形式的强化学习将世界中奖励的结构视为理所当然，并询问如何到达最大程度收获这些奖励的行为——即“策略”。但在许多方面，这掩盖了我们在AI边缘所面临的更有趣——也更可怕——的问题。我们发现自己对这个问题的确切相反更感兴趣：给定我们想要从机器那里得到的行为，我们如何构建环境的奖励来产生那种行为？当我们坐在观众席的后排，坐在批评家的椅子上——是我们分发食物颗粒或其数字等价物时，我们如何得到我们想要的？

这就是强化学习者背景下的对齐问题(alignment problem)。尽管这个问题在过去五到十年中呈现出新的紧迫性，但正如我们将看到的，它与强化学习本身一样深深植根于过去。

自然将人类置于两个至高主宰——痛苦和快乐的统治之下。只有它们能指出我们应该做什么，以及决定我们将要做什么。

——JEREMY BENTHAM

奖励函数的设计很少被讨论，尽管它也许是建立RL系统最困难的方面。

——MAJA MATARIĆ

1943年，B. F. Skinner正在从事一个秘密战时项目，最初由——在所有团体中——消费食品公司通用磨坊(General Mills)赞助。通用磨坊给了Skinner其位于明尼阿波利斯的金牌面粉厂顶层来建造实验室。这个项目是那个时代被考虑的更大胆的项目之一：Skinner和他的实验室要教鸽子如何啄击炸弹目标的图像，然后将这些鸟——三只一组——放

在真正的炸弹内部，在炸弹落下时引导它们。“我和我的同事知道，”Skinner说，“在世界的眼中，我们是疯狂的。”

Skinner意识到许多人会认为这个项目不仅疯狂而且残忍。对于第一点，他指出人类有着将动物的（通常是超人的）感官用于人类目的的悠久历史：导盲犬、松露猎猪等等。对于第二点，他争辩说，“将一个较低等生物转化为不知情英雄的伦理问题是和平时期的奢侈品。”

Skinner一直致力于强化学习的研究，他著名的“Skinner箱”就像是Thorndike谜题箱的升级版，是二十世纪中期的版本。这些装置配备灯光、杠杆和机械食物分配器，通常是从自动售货机改装而来，使得对强化的精确定量研究成为可能，后来的几代研究者都会使用这些设备（其中不乏Wolfram Schultz在猴子多巴胺研究中的应用）。在1950年代拥有了这样的工具后，Skinner开始研究动物如何学会采取行动以在各种不同条件下最大化其奖励（通常是食物形式）。他测试了不同类型的所谓“强化时间表”并观察其效果。例如，他比较了“比率”强化——即一定数量的正确行为会产生奖励——与“间隔”强化——即在一定时间后的正确行为会产生奖励。他还测试了“固定”与“变动”强化，其中行为的数量或时间长度要么保持恒定，要么允许波动。Skinner著名地发现，最激烈、最重复或最持续的行为往往来自变动比率时间表——也就是说，当奖励在一定数量的行为后出现，但这个数量会波动时。⁵这些发现对理解赌博成瘾具有重要意义——而且，悲剧的是，它们无疑也导致了更多成瘾性赌博游戏的设计。

然而，在他秘密的顶层实验室里，Skinner面临着一个不同的挑战：要搞清楚的不是哪种强化时间表能最深入地固化简单行为，而是如何仅仅通过给予奖励就能产生相当复杂的行为。当他和同事们有一天试图教一只鸽子如何打保龄球时，困难变得显而易见。他们建立了一个微型保龄球道，配有木球和玩具球瓶，打算在鸽子一拍球就给它第一次食物奖励。不幸的是，什么都没发生。鸽子没有做任何这样的事情。实验者等了又等……最终失去了耐心。

然后他们采取了不同的策略。正如Skinner所述：

我们决定强化任何与拍击有丝毫相似之处的反应——也许起初仅仅是看球的行为——然后选择更接近最终形式的反应。结果让我们惊讶。几分钟内，球就像鸽子是个保龄球冠军一样在箱子的墙壁上弹跳。

结果如此令人震惊和引人注目，以至于Skinner的两名研究者——夫妻团队Marian和Keller Breland——决定放弃他们在学术心理学领域的职业生涯，开办一家动物训练公司。“我们想试着谋生，”Marian说，“使用Skinner的行为控制原理。”⁶（他们的朋友Paul Meehl，我们在[第3章]中简要遇到过，赌他们10美元会失败。他输了那个赌注，他们自豪地装裱了他的支票。）⁷他们的公司——动物行为企业(Animal Behavior Enterprises)——将成为世界上同类公司中最大的，训练各种动物在电视和电影中表演，在商业广告中，以及在像SeaWorld这样的主题公园中。不仅仅是谋生：他们建立了一个帝国。⁸

Skinner也会将这一刻——在秘密面粉厂实验室内的微型保龄球道——视为改变他职业生涯轨迹的顿悟。他看到，关键组成部分是“通过强化最终形态的粗略近似来逐渐塑造行为，而不是等待完整的反应”。⁹

然而，鸽子项目——如众所周知的——最终算是一个复杂的成功。鸽子们本身工作得很好。事实上，如此出色，以至于它们似乎分散了政府科学研究中心与委员会的注意力。“一只活鸽子执行其任务的景象，无论多么优美，”Skinner写道，“只是让委员会想起了我们的提议是多么完全荒诞。”¹⁰而且，当时Skinner并不知道，政府正在全力投入曼哈顿计划：一种爆炸半径如此之大的炸弹，用他的话说，“有一段时间看起来好像对精确轰炸的需要已经永远被消除了。”尽管如此，鸽子项目最终在海军研究实验室找到了归宿，作为一个名为ORCON的项目——“有机控制(organic control)”的缩写——研究一直持续到战后的50年代。

Skinner感到被证实了这个概念已被证明有效，在1950年代末自豪地写道：“在导弹制导中使用活体生物，公平地说，似乎不再是一个疯狂的想法。”¹¹然而，这种证实虽然不错，但却偏离了重点。

关键是他们发现了塑形：一种通过简单奖励来灌输复杂行为的技术，即通过奖励该行为的一系列连续近似。Skinner写道：“这使得塑造动物行为成为可能，就像雕塑家塑造一块粘土一样。”¹²这个想法和这个术语将成为Skinner余生和职业生涯中的关键概念。¹³从一开始，他就看到了这对商业和家庭生活的影响。

正如他所写：“其中一些[强化时间表]对应于工业中建立的日薪或计件工资的偶然性；其他的则类似于赌博设备微妙但强大的偶然性，这些设备因其能够维持持续行为的能力而臭名昭著。”¹⁴他还认为可能的育儿影响是重大的：“然而，科学分析可以带来对个人关系的更好理解。无论我们是否有意，我们几乎总是在强化他人的行为。”Skinner指出，父母的关注是一个强大的强化因子，父母如果对礼貌请求反应缓慢，可能会无意中训练他们的孩子变得讨厌和咄咄逼人。（他说，补救措施是对可接受的关注请求更及时和一致地回应，对那些大声或不礼貌的请求回应更少。）¹⁵

也许最具预言性的是，Skinner认为，通过他工作中出现的原理，最广泛意义上的教育——对人类、对动物——可能成为一个严格、客观的领域，其中可以并将会取得飞跃式进步。正如他所说：“教学常常被说成是一门艺术，但我们可以越来越有理由希望它最终可能成为一门科学。”¹⁶

Skinner可能比他自己预期的更正确。在21世纪，当使用“塑形”这个术语时，说话的人同样可能是机器学习研究员或心理学家。对奖励的研究——特别是如何战略性地管理奖励以获得你想要的行为，而不是你不想要的行为——确实已经成为一门严格的定量科学，尽管可能不是Skinner想象的有机学习者。

稀疏性问题

有更好的方法……找到它！

—托马斯·爱迪生¹⁷

1855年，苏格兰哲学家Alexander Bain似乎创造了“试错”这个短语来描述人类和动物如何学习。¹⁸（他的另一个短语——“摸索实验”——同样令人难忘，但似乎没有流传下来。）

在最基本的层面上，强化学习是对试错学习的研究，这种试验（或者如果你愿意的话，摸索）采取的最简单算法形式被称为“epsilon-贪婪”（用希腊字母 ϵ 写作“ ϵ -贪婪”）。希腊字母 ϵ 经常被数学家用来表示“一点点”，而 ϵ -贪婪是“贪婪，除了偶尔的时候”的简写。按照 ϵ -贪婪操作的智能体，大部分时间——比如说99%——会采取它认为基于目前有限经验会带来最大总奖励的行动。但在 ϵ 的时间里——比如另外1%——它会完全随机地尝试某些东西。在Atari游戏中，这意味着在某个百分比的时间里随机按按钮，只是为了看看可能会发生什么。

从这种探索性行为中学习有许多不同的方式，但基本思想是相同的。摸索；对获得奖励的事情做得更多，对受到惩罚的事情做得更少。你可以通过明确地试图理解世界如何运作来做到这一点（“基于模型”的RL），或者只是通过磨练你的直觉（“无模型”的RL）。你可以通过学习某些状态或行动能带来多少奖励来做到这一点（“价值”学习），或者只是通过知道哪些策略总体上比其他策略表现更好（“策略”学习）。然而，几乎所有方法都建立在首先意外地偶然获得成功，然后建立起做更多似乎有效的事情的倾向这一想法之上。

事实证明，有些任务比其他任务更容易接受这种方法。

例如，在像太空侵略者这样的游戏中，成群的敌人向你下降，你所能做的就是向左移动、向右移动和射击。随机按按钮可能至少会让你击杀几个敌人，每个敌人都值积分，这些早期积分可以用来开始学习过程，通过这个过程某些行为模式得到加强，更好的策略得到发展。例如，你可能意识到积分只在你开火后才会出现，所以你会开始更频繁地开火，反过来获得更多分数。这样的游戏被称为具有“密集”奖励，这使它们相对容易学习。

在其他游戏中——以国际象棋为例——奖励并不那么即时，但它们仍然是确定的。一局国际象棋通常在几十步之后就结束了，无论如何，规则使得超过几百步几乎是不可能的。即使你对策略的精妙之处一无所知，在棋盘上随机移动棋子，你至少很快就会知道你是赢了、输了还是平局。

然而，在很多情况下，要获得任何奖励都需要真正的奇迹。当然，Skinner在试图奖励一只鸟在微型保龄球道上滚球时亲身体验了这一点。这只鸟对自己被放入什么游戏毫无头绪，可能需要数年时间才能碰巧做出正确的行为——当然，它（和Skinner）早就会饿死在那之前。

对于机械学习者来说也是如此。例如，让一个人形机器人将足球踢进球网，可能涉及数十个关节上数十万次精确的扭矩，所有这些都需要完美协调。很难想象一个最初随机移动数十个关节的机器人能够保持直立，更不用说与球有意义的接触，更不用说将球送进球网了。

强化学习研究人员将此称为稀疏奖励问题，或更简洁地称为稀疏性问题。如果奖励明确地根据最终目标或相当接近的东西来定义，那么基本上必须等待，直到随机按按钮或随机胡乱移动产生所需的效果。数学表明，大多数强化学

习算法最终会到达那里——但实际上，最终可能会在太阳爆炸很久之后才到达。如果你试图训练一个围棋程序来击败世界冠军，并且每次世界冠军认输时给它一分，否则给零分，你确实要等待很长时间。

稀疏性问题也有安全隐患。如果在未来某个时候，你正在开发一个具有巨大能力的超级智能AI，由 ϵ -贪婪强化学习驱动，并且你决定如果它治愈癌症就奖励它一分，否则为零，要小心——因为它必须尝试很多随机的事情，才能偶然获得第一个奖励。其中许多可能很丑陋。

当我与布朗大学的Michael Littman坐下交谈时——他整个职业生涯都在从事强化学习工作——我问他，他对强化学习的终生研究兴趣作为父母是否有用。他的思绪立即转向稀疏性问题。他记得和妻子开玩笑说对他们的儿子使用稀疏奖励：“这样怎么样？在他学会说中文之前，我们不给他喂食。这将是一个很好的激励！让我们看看效果如何！”Littman笑了。“我的配偶很实际……她说，‘不，我们不会玩那个游戏。’”¹⁹

当然，Littman知道——正如Skinner所知道的——不要玩那个游戏。实际上，稀疏性问题激发了强化学习社区回顾Skinner的时代，他们相当直接地借鉴了他的建议。²⁰特别是，他关于塑造的想法导致了两个不同但相互交织的思路：一个关于课程设置，另一个关于激励。

课程设置的重要性

塑造的关键洞察——为了获得复杂的行为，我们可能首先需要策略性地奖励更简单的行为——当然，对人类和对动物一样适用。“你必须先走才能跑”，我们说，这个格言描述了人类经验中无法命名或计数的更多方面——除了字面上的真实。

人类生活的一个显著特征是，我们在最初的几十年里，在字面和比喻的训练轮帮助下在世界中移动，用字面和比喻的护栏保龄球。许多动物简单地被投入到生活的全部复杂性中：例如，许多野生动物必须准备在出生后几小时内以全速逃离捕食者。相比之下，我们需要几十年才能操作重型机械，当我们必须“自立”时，我们超过体力巅峰是很常见的。

将21世纪人类与穴居祖先区分开来的不是原始智力，而是良好的课程设置。实际上，Skinner认为我们不应该如此急于判断动物的心理能力。通过正确的课程设置，它们可能能够惊人地超越我们认为其物种能力的范围——就像人类一样。

正如Skinner所说，如果实验者只是等待，直到一套复杂的行为出现才开始强化这种行为，那么这几乎不是测试那个动物“能”或“不能”执行该行为的测试。

正在测试的是实验者而不是有机体的能力。断言给定物种或年龄的有机体不能解决问题是危险的。由于仔细的调度，鸽子、老鼠和猴子在过去五年中做了它们物种成员以前从未做过的事情。这并不是说它们的祖先没有能力进行这样的行为；大自然只是从未安排过有效的调度序列。²¹

很难理解我们的生活和周围的世界在多大程度上被塑造成这样“有效的调度序列”。我们不知何故开始认为平稳地入职和学习基础知识是“自然的”，而实际上，情况恰恰相反。大自然只是存在。没有教程。

相比之下，人类世界被精心设计得便于学习。例如，伟大游戏之所以伟大，部分原因在于它们“塑造”我们游戏的方式。以任天堂1985年的《超级马里奥兄弟》为例，这是历史上最著名、最具意义的电子游戏之一。也许很难回忆起第一次玩这个游戏的经历，但仔细观察游戏的前十秒钟会发现，它经过精心而巧妙的设计来教会你如何游戏。你一开始会遇到一个从右边走来的敌人Goomba；如果你什么都不做，你就会死。“你必须以自然的方式教会玩家，他们需要通过跳跃来避开敌人，”这个游戏的设计师，传奇的宫本茂说道。²²这是游戏第一个也是最重要的教训：蘑菇样子的家伙是坏的，你必须跳跃。

但宫本茂遇到了一个问题。游戏中还有好蘑菇，你必须学会的不是躲避它们，而是寻找它们。“这给我们带来了真正的麻烦，”他解释说。“我们需要想办法确保玩家明白这是真正好的东西。”那么现在怎么办？好蘑菇在一个你头顶空间太小、无法轻易跳过的区域向你靠近——你准备迎接冲击，但它没有杀死你，反而让你的体型增大一倍。游戏的机制已经建立，现在你可以自由发挥了。你以为你只是在玩游戏。但实际上你正在被仔细、精确、不知不觉地训练。你学会了规则，然后你学会了例外。你学会了基本机制，然后获得了自由发挥的权限。

也许我们不应该对同样的塑造原理——通过连续近似来instill复杂行为——既适用于Skinner的鸽子也适用于人类学习者感到如此惊讶。我们也许不应该惊讶地发现，当学习者是机器时，这个原理同样适用。

当然，自从机器学习诞生以来，人们就知道有些问题、有些环境、有些游戏比其他的更容易。但人们逐渐认识到，一个系统如果首先在问题的较简单形式上训练，可能比从零开始训练的agent更好地学习更困难的问题。²³

在1980年代，Richard Sutton和Andrew Barto与他们的同事Oliver Selfridge合作，使用reinforcement learning来训练一个带轮子的模拟小车在自身上平衡一根杆子而不倾倒。杆子越高越重，就越容易保持直立——就像用手平衡棒球棒比平衡尺子更容易一样。他们发现，如果小车系统首先在一根高而重的杆子上训练，然后切换到较短较轻的杆子，比从一开始就在较短较轻的杆子上训练需要更少的总尝试次数。²⁴

研究人员在其他情境中也时不时地发现了这一洞察。例如，UC San Diego的语言学家Jeffrey Elman在90年代初期实验让neural networks正确预测句子中的下一个单词。令人沮丧的是，几次初始尝试都失败了。“简单地说，”他说，“当从一开始就用完整的‘成人’语言进行训练时，网络无法学习复杂的语法。然而，当训练数据经过筛选，首先呈现简单句子时，网络不仅成功掌握了这些句子，而且继续掌握了复杂句子。”²⁵

“这是一个令人愉快的结果，”Elman说，“因为网络的行为部分类似于儿童。儿童不是通过掌握成人语言的全部复杂性来开始的。相反，他们从最简单的结构开始，逐步增进，直到达到成人语言水平。”

在这两种情况下，使用curriculum——问题的较简单版本，然后是较难版本——在仅尝试学习更困难问题无法成功的情况下取得了成功。

Keller和Marian Breland在Animal Behavior Enterprises的工作中，在训练一头猪将大木币存放到“猪形储钱罐”的努力中看到了良好curriculum的关键重要性。他们从一枚硬币开始，就在储钱罐旁边，然后逐渐将其移得越来越远离储钱罐，也离猪越来越远。²⁶

最近，机器学习社区重新回到了这种“向后”工作的思路。2017年，一群UC Berkeley的机器人专家想要训练一个机器臂将垫圈滑到长螺栓上。等待机器人随机stumble onto这种行为需要永恒的时间。但是从垫圈几乎完全滑到螺栓底部开始，他们可以教会机器人推动它下滑最后一点点。然后在垫圈刚好在螺栓上时，机器人可以学会将其一直滑下去。然后在垫圈以简单方向靠近螺栓时，机器人可以学会将螺栓穿过孔洞。他们最终能够倒推到一个能够在任何地方以任何方式接过垫圈，并能够灵巧地旋转和固定它的系统。²⁷

传奇国际象棋世界冠军Bobby Fischer这样的游戏专家在他的教学书籍《Bobby Fischer Teaches Chess》中使用了类似的策略。这本面向初学者的书包含了数十个一步将死的例子，然后进展到两步将死，再到三步和四步的杀棋组合。中局和开局的讨论，以及长期策略，都推迟到后续书籍中；Fischer专注于教授初学者如何识别结束游戏的机会。这种特定的课程设置确实已被证明非常成功：当今的一些特级大师推荐它作为新玩家的完美第一本国际象棋书籍，²⁸它已成为有史以来最畅销的国际象棋书籍。²⁹

看起来自然的下一步是将构建一个良好、可学习的课程本身视为一个machine-learning问题，并看看是否可能将课程设计过程自动化。最近的研究已经探索了自动识别适当难度任务的方法，以及能够最大程度促进网络学习的例子。这方面的早期结果是有希望的，工作正在进行中。³⁰

然而，也许在自动化课程设计中最令人印象深刻的成就是DeepMind在棋盘游戏领域占主导地位的工作，包括AlphaGo及其后续版本AlphaGo Zero和AlphaZero。“AlphaGo总是有一个恰好合适水平的对手，”首席研究员David Silver解释说。³¹“它开始时极其天真；它从完全随机的游戏开始。然而在学习过程的每一步，它都有一个对手——如果你愿意的话，一个陪练伙伴——精确校准到它当前的表现水平。”那么，这个完美的陪练伙伴是谁呢，总是校准到恰好合适的难度？

答案简单、优雅，而且事后看来显而易见。它与自己对弈。

激励机制的微妙问题

无论是处理猴子、老鼠还是人类，说大多数生物寻求关于哪些活动会被奖励的信息，然后寻求去做（或至少假装去做）这些事情，通常几乎完全排除不被奖励的活动，这几乎不存在争议。

—STEVEN KERR³²

如果奖励系统的设计使得道德行为变得不理性，这并不一定意味着会导致不道德行为。但这难道不是在自找麻烦吗？

—STEVEN KERR³³

问题出在奖励系统上，笨蛋！

—《管理学院执行官》编辑部³⁴

克服稀疏性的第二种方法是，与其使用”课程”并首先从问题的简化版本开始，不如在使用问题的正常完整版本的同时，添加一些奖励来指导学习者朝正确方向前进或鼓励与成功相关的行为。这些在该领域被称为”伪奖励(pseudorewards)“或”塑形奖励(shaping rewards)“，但最简单的做法就是将它们视为激励。

Skinner给他的鸽子一点食物来奖励它看球和接近球的行为，这些活动必然先于他最终寻求的击打动作。同样的想法适用于machine-learning设置。例如，家务清洁机器人的”真正”奖励可能是一个一尘不染的房子，但你可以为它吸尘的每一点灰尘给它一些激励。或者你的送货无人机最终可能试图到达某个目的地，但你可以为它朝正确方向前进给予一点奖励。

这通常在给一个原本会随机摆动直到纯粹偶然完成目标的agent提供关于它是否变得”更热”或”更冷”的感觉方面非常有帮助：它是否通常表现正确并朝正确方向改变其行为。

我们经常将问题分解为离散的步骤，以便在心理上更容易保持动力。想象几年后完成的博士论文或书稿会让人难以判断某一天工作的质量。想象我们希望到明年减掉的所有体重会让特定蛋糕或第二份食物的成本和收益感觉模糊。作为父母、教师、教练，我们知道一个适时的击掌或”干得好！”可以帮助学员度过持续练习的艰难过程，即使掌握似乎遥不可及。

当然，任何与人类打过交道——或者甚至只是曾经是人类——的人都知道，创造激励就是在玩火。它们必须设计得非常仔细，否则麻烦往往在等待。³⁵ 正如管理专家Steven Kerr在他1975年的经典论文中著名地指出的那样，一旦你开始考虑添加额外的奖励，你就立即面临”奖励A而希望得到B的愚蠢行为”的危险。³⁶

Kerr关于激励机制出错的分析已成为管理科学中的一篇里程碑式论文，Kerr职业生涯的大部分时间都在与从通用电气到高盛等企业合作，研究如何更仔细地思考激励机制。令人惊讶的是，当被问及灵感来源时，Kerr同时提到了机器学

习和B. F. Skinner。“机器可以被编程学习，国际象棋机器可以被编程永远不犯同样错误两次，这对我来说非常迷人，”Kerr说。“机器可能成为比实际程序员更好的象棋手，这种可能性立即吸引了我！”³⁷

至于Skinner，Kerr承认，“关于‘愚行’这篇文章，显然B. F. Skinner比我先到达‘那里’。我从未声称过其他。我记得读过Skinner说他会对他老鼠大喊‘你们为什么不听话？’，当它们不按他的期望行事时。如果Skinner知道人们在没有读过他的作品的情况下就读我的文章，他可能会在坟墓里翻身。显然，Skinner做了这项工作，但我能够将其包装成适合商业应用的正确形式。‘责怪老鼠’是一个很好的学习课程。‘愚行’真正要说的是，这并不总是员工的错；管理层要为太多的员工功能障碍负责。”

确实，对于Skinner来说，几乎永远不能责怪老鼠（或员工）。他认为，我们的行为几乎完全由我们的激励和奖励决定。一位电视采访者曾问Skinner：“那自由意志呢？”Skinner回答：“它处于虚构的位置。”³⁸

撇开自由意志的争论不谈，激励问题不仅存在于动物心理学中，也不仅存在于企业管理中；事实上，一些最令人难忘的案例来自那些我们称之为儿童的无情而富有创造力的奖励最大化者。

多伦多大学经济学家Joshua Gans想要让他的大女儿帮助训练弟弟上厕所。所以他做了任何优秀经济学家都会做的事情。他给了她一个激励：每当她帮助弟弟去洗手间时，她就会得到一块糖。女儿立即发现了她父亲这位经济学教授忽视的一个漏洞。“我意识到进去得越多，出来得就越多，”她说。“所以我就给我弟弟喂了一桶又一桶的水。”Gans证实：“效果确实不太好。”³⁹

普林斯顿认知科学家Tom Griffiths与他自己的女儿发生了eerily(诡异地)相似的情况。“她真的喜欢清洁东西，”他告诉我；“她会为此感到兴奋。我们给她买了她自己的小刷子和簸箕。地上有一些薯片，她拿起刷子和簸箕把它们清理干净，我对她说：‘哇！做得好！清理得很好！干得好！’”⁴⁰

通过适当的表扬，Griffiths既能培养女儿的运动技能发展又能在保持房屋清洁方面得到一些帮助：双重育儿胜利。真的是这样吗？他女儿在几秒钟内就发现了漏洞。

“她抬头看着我们微笑，”他说——“然后把薯片从簸箕里倒回地板上，再次清理它们，试图获得更多表扬。”

对于研究领域横跨心理学和机器学习的Griffiths来说，含义是显而易见的。这正是让我思考在构建奖励驱动的AI系统时面临的一些挑战的事情，你必须非常仔细地思考如何设计奖励函数。”

Griffiths在育儿的背景下一直思考强化学习。“作为父母，你正在为你的孩子设计奖励函数，对吧？就你表扬的事情和你给他们某些反馈的事情而言……没有人真正严格思考‘你明确想为你的孩子设计什么样的奖励函数？’”

Griffiths将父母身份视为对齐问题(alignment problem)的一种概念验证。他指出，人类文明的故事一直都是关于如何向那些将不可避免地从我们手中继承社会控制权的奇怪、异类、人类水平智能体灌输价值观——也就是我们的孩子。然而，这种相似性甚至比那更深入——仔细关注AI和育儿显示了两者可以相互启发的惊人程度。

我们的孩子可能不比我们聪明，但即使是年幼的孩子也能智胜我们的规则和激励，部分原因是他们多么有动机这样做。在强化学习系统的情况下，它们在某种程度上是奖励的奴隶；但它们是那种拥有巨大计算能力和潜在无穷无尽的试错尝试次数的奴隶，能够找到我们设计的任何激励中所有可能的漏洞。机器学习研究人员已经以痛苦的方式学到了这个教训。他们也学到了如何处理它的一些方法。

循环验证你的奖励：塑造定理

Astro Teller，目前是X公司（前身为Google X的Alphabet旗下公司）的“登月计划船长”，近年来监督了从Google自动驾驶汽车项目（后来分拆为Waymo）到增强现实项目Google Glass和研究实验室Google Brain的所有项目。但在1998年，他专注于一个不同的问题：足球。与他的朋友和同学David Andre一起，Teller致力于参加一年一度的RoboCup足球比赛，用一个他们称为Darwin United的虚拟足球比赛程序。⁴¹ Reward shaping（奖励塑造）是使他们能够教会程序如何踢球的一部分。但有一个问题。在足球中，控球是良好进攻和防守的一部分——当然比在场上漫无目的地游荡要好。因此Andre和Teller为他们的机器人控球提供了奖励——价值只是进球的一小部分。令他们震惊的是，他们发现他们的程序在球旁边“振动”，积累这些点数，几乎不做别的事情。⁴²

同年，哥本哈根尼尔斯·玻尔研究所的一个丹麦研究团队，Jette Randlov和Preben Alström，试图让一个reinforcement-learning（强化学习）系统学会如何骑模拟自行车。该系统必须管理保持直立的复杂任务，同时朝着远处的目标前进。这似乎是添加一些shaping rewards（塑造奖励）的完美应用。因为四处摇摆不定的系统不太可能随机到达目标，团队决定在自行车朝目标前进时添加一个小奖励——几个“点数”。

令他们震惊的是，“agent（智能体）以20-50米的半径绕着起点画圆圈。”⁴³ 他们奖励朝向目标的进展，但忘记了惩罚偏离目标的移动。他们的系统找到了一个漏洞，并且无情地——尽管令人眩晕地——利用了它。

“这些heterogeneous reinforcement functions（异质强化函数），”他们写道，“必须非常小心地设计。”

这些在20世纪90年代末的警示故事，深深印在UC Berkeley的Stuart Russell和他当时的博士生（后来成为百度副总裁兼首席科学家）Andrew Ng的脑海中。这些exploitative loops（利用性循环）似乎是一个持续的危险。⁴⁴

Ng很有野心。他回忆说：“当我开始研究robotics（机器人学）时……我问了很多人，‘你知道的最难的控制问题是什么？’那时我听到的最常见的答案是‘让计算机飞直升机。’所以我说，‘让我们研究这个。’”⁴⁵ 确实，他将用reinforcement learning（强化学习）来飞行一架真实的——非模拟的、九英尺长、一百磅重、七万美元的——Yamaha R-50直升机作为他的博士论文。⁴⁶ 风险极高。现实世界中的不稳定或不可预测的行为可能会完全摧毁直升机——更不用说如果一个毫无防备的人发现自己在其路径上会发生什么。

关键问题是：给定一个描述他们实际希望直升机做什么的难以学习的reward function（奖励函数），他们可以添加什么样的“pseudoreward（伪奖励）”激励，如果有的话，使得训练过程更容易，但最大化修改后奖励的最佳方式也是真实问题的最优解？用Kerr的术语来说，他们可以奖励什么样的A仍然会产生希望的B？在Ng的描述中：

一个非常简单的额外奖励模式通常足以使原本棘手的问题变得简单。然而，reward shaping（奖励塑造）的一个困难在于，通过修改奖励函数，它将原始问题M改变为某个新问题M'，并要求我们的算法解决M'，希望在那里比在原始问题中更容易或更快地找到解决方案。但是，并不总是清楚为修改后的问题M'找到的解决方案/策略是否也适用于原始问题M。⁴⁷

“我们在指定奖励函数方面有什么自由度，”Ng写道，“使得最优策略保持不变？”⁴⁸

事实证明，关键洞察隐藏在自行车的故事中。为了防止自行车画圆圈，无休止地积累奖励，你还必须减去偏离目标的进展。Russell，最初接受物理学训练，在奖励问题和能量守恒之间建立了联系。“关键，”Russell解释说，“就是将塑造变成我们在物理学中称为‘conservative field（保守场）’的东西。”⁴⁹ Pseudorewards（伪奖励）就像势能：只

是你所在位置的函数，而不是你到达那里所采取的路径。这也意味着回到你开始的地方——无论你经历了什么旅程——净值为零。

这个想法在自行车问题中足够直观：如果你奖励朝向目标的进展，你必然需要惩罚背离目标的进展。换句话说，“激励”点数应该始终反映自行车距离目标的接近程度，而与它所走的路径无关。但将激励视为“势能”的概念被证明是更深层和更普遍的。这是确保你在塑形奖励上训练的智能体不会让其行为与真实问题脱节的必要且充分条件。

“作为一般规则，” Russell说，“根据人们在环境中真正想要的东西来设计性能测量标准，而不是根据人们认为智能体应该如何行为。”⁵⁰换句话说，关键洞察是我们应该努力奖励世界的状态，而不是我们智能体的行动。这些状态通常代表朝向最终目标的“进展”，无论这种进展是用物理距离表示还是用更概念性的东西表示，比如完成的子目标（比如书的章节，或机械装配的部分）。

虽然它不是解决人类激励所有问题的灵丹妙药，但从行动到状态的焦点转移确实邀请我们以新的方式思考我们有意或无意地为他人设计的一些激励结构。面对一个为了双倍获得清扫奖励而倒掉厨房垃圾的孩子，我们可以通过确保我们以完全相等的程度责骂他们倒垃圾，使我们的奖励成为“保守场”，这样进一步重复的净收益为零。不过，转向赞扬状态而不是行动可能更容易：我们可能会说“哇，看看那地板多干净！”“而不是奖励清理本身的行为。

当然，奖励的艺术和科学不仅仅是避免循环——虽然这是一个开始。Ng和Russell在一个谨慎而非胜利的注释上结束了他们的工作：“我们相信，”他们写道，“寻找良好塑形函数的任务将是一个日益重要的问题。”⁵¹

作为奖励设计师的进化

我对他说：“生养众多，遍满地面。”但不是用这些话。

—伍迪·艾伦⁵²

从达尔文主义的角度来看，人们想要什么是相当清楚的：沿着传播和保护他们基因血统的路线。人们实际上想要的，时刻，看起来更加异质和短视得多：性高潮、巧克力、新车、尊重。因此，我们似乎在生物学上被连线并在文化上被诱导去想要短期内的具体、具体的东西，这些东西通常以最终符合这些进化目标的方式引导我们的行为，否则这些目标会太遥远或定义不明确而无法有意识地瞄准。⁵³

听起来熟悉吗？

理解塑形的本质和作用——首先在行为心理学中，然后在机器学习中——不仅教会了我们如何设计更好的智能体。也许它最令人惊讶的贡献是对我们思考进化方式的影响。

当布朗大学的Michael Littman还是研究生时，在1980年代后期，他曾被雇佣在Bellcore工作一段时间，这是一个研究和开发小组，以前是AT&T的一部分，位于新泽西州。在那里，他很快找到了一位导师和朋友，Bellcore的Dave Ackley。

Littman向他询问关于行为的问题——关于采取行动和做出随时间延续的决定。正如Littman回忆：“他就像，’哦。那是一个东西。它叫强化学习。我稍微研究过。这里有一篇论文。’然后他给了我Rich Sutton 1988年的TD论文。”⁵⁴

Littman开始阅读关于时间差分学习的内容并着迷了；他问Ackley在哪里可以学到更多。“然后他就像，‘嗯，邀请Rich来做演讲。’我就像，’什么？那是你可以做的事情吗？你可以读一篇论文，上面有一个人的名字，然后你可以把他们变成一个人？’…我没有把它想成一个社区；我把它想成一个文献。但不，它是一群人，他们彼此认识，Dave就像，’我可以邀请他。’所以他邀请了他。”

Sutton来到了Bellcore，Ackley和Littman都染上了强化学习的热情。

他们对进化如何塑造我们的奖励函数以在短期内产生对有机体或物种长期生存有用的行为这个问题感兴趣。有机体的奖励函数本身，只要它完成这个目标，可能看起来非常随机。Ackley和Littman想看看如果他们简单地让奖励函数进化和变异，并允许模拟的虚拟实体死亡或繁殖，会发生什么。⁵⁵

他们创建了一个二维虚拟世界，其中模拟有机体（或“智能体”）可以在景观中移动、进食、被捕食和繁殖。每个有机体的“遗传密码”包含智能体的奖励函数：它有多喜欢食物、它有多不喜欢靠近捕食者等等。在其一生中，它将使用强化学习来学习如何采取行动以最大化这些奖励。当有机体繁殖时，其奖励函数将传递给其后代，同时伴随一些随机变异。Ackley和Littman用一群随机生成的智能体播种了初始世界人口。

“然后，”Littman说，“我们就直接运行了它，运行了七百万个时间步，这在当时是很多的。那时的计算机还比较慢。”发生了什么？正如Littman总结的：“奇怪的事情发生了。”⁵⁶

从高层次来看，大多数成功的个体智能体的奖励函数最终都相当好理解。食物通常被视为好的。捕食者通常被视为坏的。但仔细观察揭示了一些奇异的怪癖。例如，一些智能体学会了只有当食物在它们北边时才接近食物，但如果

食物在南边就不会。

“它不是在所有方向都喜欢食物，” Littman说。“奖励函数中有这些奇怪的漏洞。如果我们修复了这些漏洞，那么智能体就会变得非常善于进食，以至于它们会把自己撑死。”

Ackley和Littman构建的虚拟环境包含有树木的区域，智能体可以躲在那里避开捕食者。智能体学会了只是普遍喜欢在树木周围闲逛。那些被树木吸引的智能体最终存活了下来——因为当捕食者出现时，它们有现成的地方可以躲藏。

然而，这里有一个问题。它们硬编码的奖励系统，通过进化磨练而成，告诉它们在树木周围闲逛是好的。渐渐地，它们的学习过程会学到走向树木根据这个奖励系统是“好的”，而远离树木则是“坏的”。当它们在生命过程中学会优化行为以适应这一点，并越来越善于抓住树木区域且从不离开时，它们达到了Ackley称之为“树木老年痴呆”的阶段。它们从不离开树木，耗尽了食物，饿死了。

然而，因为这种“树木老年痴呆”总是在智能体达到繁殖年龄之后才出现，所以从未被进化淘汰，大量热爱树木的智能体社会蓬勃发展。

对Littman来说，除了进化的奇异性和任意性之外，还有更深层的信息。“这是一个有趣的案例研究：当然，它有一个奖励函数——但不是孤立的奖励函数有意义。有意义的是奖励函数与它产生的行为之间的相互作用。”

特别是，树木老年痴呆的智能体生来就有一个对它们来说最优的奖励函数，前提是它们在为了最大化该奖励而行动方面不过于熟练。一旦它们变得更有能力和更熟练，它们就会将奖励函数最大化到危险的程度——最终导致它们的厄运。

人们不必太费力就能看到这里对智人的警示故事。像“总是尽可能多地吃糖和脂肪”这样的启发式方法是最优的，只要你的环境中没有太多糖和脂肪，而且你在获取它们方面不是特别擅长。一旦这种动态发生变化，一个为你和你的祖先服务了数万年的奖励函数突然就会让你偏离轨道。

对Andrew Barto来说，在思考进化中有一些线索对我们现在扮演奖励设计者的角色很有用。“进化提供了我们的奖励函数，所以这对于我们如何为人工系统设计奖励函数来说确实非常重要，”他说。“这就是自然界发生的事情。进化想出了这些奖励信号来鼓励我们做导致繁殖成功的事情。”⁵⁷

正如Barto指出的，“所以，一个有趣的事情是进化没有给我们繁殖成功作为奖励信号。它们给了我们对预测因子的奖励。”我们优化我们的行为来最大化我们认为有奖励性的东西，但在背景中和更大的尺度上，进化正在首先塑造我们认为有奖励性的东西。“所以，这是一个两级优化，” Barto说。“我对此非常感兴趣。”

近年来，Barto与密歇根大学的Satinder Singh和Richard Lewis以及当时的博士生Jonathan Sorg合作研究“最优奖励问题”。⁵⁸如果你有目标x，你的最佳策略可能不是简单地告诉你的智能体去做x。

“人工智能体的目标应该与智能体设计者的目标相同吗？”他们写道。“这是一个很少被问及的问题。”⁵⁹

考虑一个游戏，他们说，其中存在的只有一个智能体、一根钓鱼竿、一些蚯蚓和一个满是鱼的池塘。⁶⁰假设智能体的整体进化适应性最好通过吃尽可能多的卡路里来服务。理想情况下，它们会学会拾起蚯蚓，克制不吃它们，并用它们来捕鱼——但这相当复杂。一个聪明且寿命长的智能体最好对吃蚯蚓有厌恶感，这样它们就能更快地开始学习钓鱼。另一方面，对于注意力持续时间较短或寿命较短的智能体，试图学习如何钓鱼将是徒劳的时间浪费——所以如果它们恰好觉得蚯蚓美味，对它们来说更好。

也许最有趣的是一个聪明到足以学会钓鱼且会活得刚好够长来学会如何钓鱼——但不够长来真正从这项投资中受益的智能体的情况。结果表明，它们应该被设计成对鱼过敏，这样它们就别无选择只能吃蚯蚓！

agent的生命周期、资源或设计的细微变化，可能对最优奖励结构产生剧烈而突然的影响。关于什么样的奖励集合对特定环境中的特定agent是理想的，这个问题似乎不存在任何简单的概括性答案。这方面的研究仍在继续——但学会更敏锐地区分你想要什么和你奖励什么，是解决方案的重要组成部分。⁶¹

最近，心理学家和认知科学家正在运用这些工具，转而提出一个关于人类而非机器的有趣问题。他们问道，当需要学习优化的那个计算能力有限、缺乏耐心、目光短浅的agent是……你自己时，你应该如何设计最佳的奖励函数？

我们应该如何训练自己？

机器学习中奖励塑形的理论和实践不仅为我们提供了让自主直升机适当机动的方法，还为我们理解人类生活和人类智能贡献了两个不同的方面。第一，它向我们展示了一个原因——稀疏性——解释为什么某些问题或任务比其他问题更难解决或完成。第二，它为我们提供了一个理论——激励状态而非行为——告诉我们如何在不引入负面激励的情况下让困难问题变得更容易。

这些洞察在人类生活中的潜在用途是巨大的。仅经济成本就很庞大——最近一份报告估计，英国工人在工作中拖延造成的影响每年达760亿英镑——更不用说对我们幸福感以及生活质量和充实度造成的不易计算的损失。⁶²

如果我们正生活在电子游戏成瘾和现实世界拖延症盛行的时代，也许这不是个别拖延者的错。正如Skinner所说：“我本可以对我实验的受试者大喊，‘表现好一点，该死的，按你应该的方式表现！’最终我意识到受试者总是对的。他们总是按应该的方式表现。”⁶³如果他们没有学到什么，那是实验者没有正确塑形任务的错。所以也许这不是我们缺乏意志力，而是——正如Jane McGonigal在2011年畅销书中所说的——现实是破碎的。

McGonigal是一名游戏设计师，她的职业生涯致力于设计游戏来帮助人们——包括她自己——克服生活中的挑战。对她来说，大多数游戏，特别是电脑游戏令人难以置信的成瘾性和吸引力在于，它们总是清楚地告诉你需要做什么，以及无论是什么都总是看起来是可以实现的。

第一件事是，无论何时你出现在这些在线游戏中……有很多很多不同的角色愿意立即委托你拯救世界的任务。但不是任何任务；这是一个与你在游戏中当前等级完美匹配的任务。对吧？所以你能完成它。他们从不给你无法完成的挑战。但它处于你能力的边缘，所以你必须努力尝试。⁶⁴

换句话说，使游戏如此极具吸引力的是它们的塑形程度有多好。关卡是完美的课程。积分是完美的伪奖励。它们是Skinner式的杰作。

强化学习不仅为我们提供了理解和表达游戏如此引人入胜的词汇；它还为我们提供了实证确认这些直觉的方法。一个游戏如果有从最简单关卡到最困难关卡的清晰课程，以及标记前进道路并促进探索和技能发展的清晰伪奖励，对算法来说应该更容易学习。不难想象，未来几年游戏工作室会使用自动化测试玩家来测试他们的关卡，突出显示真实人类玩家可能放弃或退出的卡点。

当然，问题在于塑形这些虚拟环境要比塑形现实环境容易得多。正如McGonigal诊断的：“现在，像魔兽世界这样的协作在线环境的问题是，一直处于史诗般胜利的边缘是如此令人满足，我们决定把所有时间都花在这些游戏世界里。这就是比现实更好。”对McGonigal来说，解决方案不是让我们摆脱这些完美奖励塑形的环境，而是相反：“我们必须开始让现实世界更像游戏一样运作。”

McGonigal本人在这方面领导了一场运动，包括使用游戏来克服障碍——包括在脑震荡长期恢复期间的自杀性抑郁——在她自己的生活中。⁶⁵“在三十四天后我对自己说——我永远不会忘记这一刻——我说，‘我要么杀死自己，要么把这个变成一个游戏。’”⁶⁶所以她创造了伪奖励。给姐姐打电话值几分。绕街区走一圈值几分。这是一个开始。

这是一个被称为游戏化的领域，⁶⁷在过去十年中，得益于强化学习的洞察，它已经从某种艺术发展为某种科学。⁶⁸

在世界上对这个问题思考最深入的人之一——无论是在他的研究还是个人生活中——是马克斯·普朗克智能系统研究所的认知科学家Falk Lieder。

Lieder的研究专注于他所称的“理性增强”(rationality enhancement)。他研究人们如何思考和做决策的认知科学——与大多数研究者不同，他不仅对理解人类认知有着浓厚兴趣，还热衷于设计有效的工具和干预措施来让人类思考得更好。他最早的人类研究对象当然是他自己。

成长过程中，Lieder发现学校教育虽然给了他很多思考的内容，但从未涉及思考本身，这让他感到沮丧。“我总觉得我真正想学的是如何好好思考，以及如何做出好的决策，”他解释道。“没有人能教我这些。他们只能教我关于世界的陈述性事实，这并不是很有用。我真的想学习如何思考。”⁶⁹

随着时间的推移，这成熟为一种个人提升的追求，但也是更宏大的东西：理解人类推理原则，并制作改进工具的驱动力。“我研究的一部分，”他解释道，“是发现这些最优的思考和决策策略，这样我们就能实际构建一个基于科学的良好思维课程。”

Lieder对游戏化感兴趣，更具体地说，是他所称的最优游戏化：给定一个目标，什么是促进实现该目标的最佳可能激励结构？⁷⁰这很像上面讨论的“最优奖励设计”，但在这种情况下，被设计的代理是人类而不是算法。

Lieder与Tom Griffiths合作，为最优游戏化的样子建立了一些基本规则。他们从Andrew Ng和Stuart Russell的工作中知道，基本规则之一是奖励状态，而不是行动。因此分配给采取行动的分数必须反映由此产生的事务状态有多好——而且，正如Lieder指出的，“分数的分配必须这样：当你撤销某事时，你失去的分数与你做这件事时获得的分数一样多。”这就是Randløv和Alstrøm用他们的自行车机器人艰难学到的，David Andre和Astro Teller用他们的振动足球机器人艰难学到的——以及Tom Griffiths本人在他女儿把簸箕倒在厨房地板上时艰难学到的。

Ng和Russell的论文曾建议，塑造奖励可以使一个预见和预测其行动效果能力有限的代理表现得比实际更有远见。⁷¹这个想法吸引了Lieder——部分原因是人类在许多决策中是如此出了名的冲动和短期导向。

他和Griffiths进行了一个实验，让人类受试者扮演航空公司路线规划师的角色。飞往某些城市在那一段是盈利的，但可能将飞机置于很少有其他盈利路线存在的位置——反之亦然：为了在其他地方获得回报，一段亏本飞行可能是值得的。他们尝试在给定从A到B一段的标价上增加额外的即时奖励（或惩罚），这将反映将飞机移动到该位置的下游成本或收益。正如预期的那样，这使人们更容易做出更好和更盈利的决策。

只有一个缺点：因为塑造的奖励将选择的长期成本和收益纳入标价，用户不再需要提前思考。这使他们的决策更加准确，但有些削弱了能力。人们不再需要非常努力地思考，所以他们没有这样做。“如果你要在一个[近视的]决策制定有效的环境中行动，”Lieder说，“人们将学会越来越多地依赖该系统。”⁷²

这留下了一个有趣的可能性：你能否使用最优游戏化，不是为了消除规划的需要，而是让人们在规划方面变得更好？

在这种情况下的激励几乎完全不同。接口不是创建一个简单的问题，其中考虑了长期成本——导致更高质量的决策但可能鼓励受试者的惰性或自满，使他们更依赖接口——价格可以调整以创建一个课程。可以慢慢教受试者如何提前思考，从基本想法的非常简单的说明开始，随着受试者变得更好，慢慢增加复杂性。接口不是充当拐杖，而是可以使用不同的激励集来做相反的事情：“教人们在时间上规划得更远，”Lieder解释道，“这样他们就能在即时奖励与长期价值不一致的环境中成功。”

Lieder的最终实验不是关于航班规划，而是一个更熟悉的环境：拖延。他和Griffiths创建了一个故意繁重的任务——就五个主题写论文，其中一些比其他的更长更难——并将其放在Mechanical Turk上，人们会选择为此工作以获得报酬

——20美元——在十天后设定的截止日期前完成所有写作。在所有报名的人中，40%甚至从未开始。（这特别具有讽刺意味，因为当任务最初向他们描述时，他们有机会拒绝参与并仍然收到十五美分！）⁷³

Lieder和Griffiths还试验了给受试者“积分”的激励措施——这些积分没有现金价值，但在视觉上具有鼓励作用——每完成一篇文章就能获得积分。每篇文章的积分价值相同。但这并没有起到帮助作用。

最后，他们为第三组参与者提供了最优激励：积分价值精确反映了每个主题的难度或令人厌恶程度，以及完成该主题后距离获得20美元奖励还有多近。（例如，关于北朝鲜经济政策的一百字文章，其积分价值大约是关于他们最喜欢的电视节目五十字文章的三倍。）在这种情况下，整整85%的参与者完成了全部五篇文章。⁷⁴

Lieder将这样的系统视为“认知假肢(cognitive prostheses)”，⁷⁵它们不仅仅是他的研究兴趣。它们是他自己的研究得以完成的关键部分。

作为一名博士生，Lieder发现自己处在那个可怕的文章写作任务的放大版本中。“我认为最糟糕的情况之一就是无法获得关于你进展水平的任何信息，”他说。“官方系统是这样的：你是一名博士生，然后你获得博士学位，就这样。所以：五年没有反馈。”

总的来说，博士生是一个高焦虑和抑郁率的群体，拖延症对他们来说几乎是流行病。⁷⁶实际上，他们就像Skinner保龄球馆里的鸽子——学位帽和学位袍形状的食物颗粒在大约五年后等着他们打出完美的一击。我们知道，这样的系统对动物不起作用——对强化学习算法也不起作用。

这样不行。Lieder需要别的东西。他规划了自己的五年历程——“我把这个分解成几百个级别。”当他做他认为可能在几年后产生真正引用的工作时，他给自己分配虚拟的“引用”。他使用为受试者的文章分配积分价值的相同最优游戏化计算方法，来计算他博士学位每个子任务的适当积分价值。他甚至使用惩罚来改变他的一些习惯——以来自Pavlok公司的腕带形式。“每当你沉迷于某些你不想沉迷的习惯时，它就会给你电击，”Lieder解释道。“我的主要坏习惯与我如何使用电脑有关。比如，当我感觉不好时我会去YouTube。由于我的时间跟踪软件会立即与我的Pavlok通信我在做什么，它就能立即电击我。”有趣的是，这未能抑制他在需要分散注意力时访问YouTube的习惯，但它确实养成了在访问时立即关闭页面的习惯。

实际上，既是实验者又是自己行为训练实验的受试者，这给了Lieder一个独特的视角来观察奖励塑造问题。这样的训练同时是他进行研究的过程，也是该研究的核心问题。结果令人鼓舞，Lieder——现在运营着自己的研究实验室——有博士袍来证明这一点。

超越外在强化

对Skinner来说，不仅个人自由意志令人不安地“处于虚构的位置”，而且整个人类文明的故事本质上就是奖励结构的故事。Skinner本人对此奇怪地乐观，写了乌托邦小说《瓦尔登第二》，描述了一个完美的行为主义社会。

然而，任何与儿童或动物相处过的人可能都会有一种挥之不去的怀疑，即奖励最大化真的不是我们为什么做我们所做之事的全部故事。我们玩自己发明的游戏，没有明显的奖品。我们翻石头，或爬山，只是为了看看我们可能发现什么。我们探索。我们爱玩耍，充满好奇心。简而言之，我们受到内在奖励的驱动，就像受到外在奖励的驱动一样。

事实上，这一点在机器学习领域也越来越得到认可。

如果未经训练的婴儿大脑要成为智能的，它必须获得纪律性和主动性。到目前为止，我们只考虑了纪律性。

—艾伦·图灵(Alan Turing)¹

2008年春天，研究生Marc Bellemare正在巴巴多斯的海滩上与阿尔伯塔大学计算机科学家Michael Bowling散步。Bowling有一个想法。当时，强化学习研究通常是每个研究人员从头开始制作自己的定制游戏，然后手工定制一个系统来在那个特定游戏中获得成功。²

Bowling沉思道，如果有人建立一个每个人都可以使用的单一环境，其中不只是一个游戏，而是一个庞大的游戏库——如果不是虚假的、编造的游戏，而是使用真实游戏——即来自Atari 2600的经典1970年代和80年代视频游戏，会怎样？

Bellemare回忆道：“我说，‘这是我听过的最愚蠢的想法。’”³

他继续说：“快进大约三年，好吧，我觉得这不是一个如此愚蠢的想法。”事实上，Bellemare发现他非常喜欢这个想法，以至于Bowling成为了他的博士导师——这个想法成为了他的学位论文。

这个项目的雄心多少有些疯狂，不仅在于建立这个视频游戏动物志（他们称之为Arcade Learning Environment (ALE)）所需的工作量，还在于它向该领域其他人投下的隐含挑战。⁴这个想法是让研究人员通过部署一个不仅能玩一个游戏，而且能玩全部六十个游戏的单一学习系统来相互竞争。该领域还远未达到这个水平。

问题的很大一部分在于，当时使用的定制游戏环境通常以净化和过滤的方式向智能体描述世界，使用高级且有用的数据输入。在轮式推车试图平衡杆子的情况下，系统会得到推车位置、速度、杆子当前倾斜角度、杆子速度等作为输入。在有树木、食物和捕食者的二维网格世界环境中，系统会被告知智能体的位置、健康状况和饥饿度、附近是否有捕食者、最近的食物在哪里等等。这些信息被称为“特征”。

相比之下，ALE提供的东西更加令人不知所措，也不那么容易立即使用：屏幕上的像素。就是这样。每个游戏都不同，不仅规则不同，屏幕上的像素映射到可用信息的方式也不同。被投入到新游戏中的学习系统将不得不从头开始弄清楚一切：这里的像素在我得分时似乎会闪烁，那里的像素在我死亡前不久出现，中间的这些像素每当我按左箭头按钮时就向左移动——哦，也许它们就是我。要么研究人员必须找到极其通用的方法来提取屏幕上的有用模式供

他们的系统跟踪——这样它们在所有六十个游戏中都有帮助——要么所有的理解和意义建构都必须由系统即时完成。这就是“特征构造”问题。

Bellemare开始实验，将一堆特征构造算法插入标准强化学习系统，然后让它们处理这些游戏。结果并不令人印象深刻。然而令他惊讶的是，很容易让这些结果发表——仅仅因为同行评议者对制作Atari环境所投入的大量工作印象深刻。他说：“有趣的是，当时评审者会告诉我，‘好吧，你用Atari做了这个令人惊叹的事情。我不可能拒绝你的论文。’……它太大了……结果的好坏并不重要。人们只是说，‘哇。你真的做到了这一点。’”

他和同事们实际上建造了一座山；现在由该领域来想办法攀登它。

Bellemare于2013年完成博士学位，从埃德蒙顿搬到伦敦加入DeepMind。在那里，由Volodymyr Mnih领导的团队正在兴奋地研究一个想法：将前一年在ImageNet竞赛中如此决定性的AlexNet风格深度神经网络应用到强化学习问题上。如果深度网络能够看着数万个原始像素并弄清楚它们是百吉饼、班卓琴还是蝴蝶，也许它们能够完成使Atari屏幕变得可理解所需的任何特征构造。

他回忆说：“团队说，嘿，我们有这些卷积网络。它们在图像分类方面表现出色。嗯，如果我们用卷积神经网络替换你的特征构造机制（它仍然有点像权宜之计）会怎么样？”

Bellemare再次不买账。“我实际上在很长一段时间内都是不相信者……做感知强化学习的想法非常非常奇怪。而且，你知道，对于神经网络能做什么，存在着健康的怀疑态度。”

但在这个问题上，Bellemare也很快改变了看法。

只是将深度学习插入经典强化学习算法并在七个Atari游戏上运行，Mnih就能够在其中六个游戏中击败以往所有强化学习基准。不仅如此：在三个游戏中，他们的程序似乎和人类玩家一样优秀。他们在2013年末提交了一篇研讨会论文，记录了他们的进展。Bellemare说：“这是一篇概念验证论文，证明卷积网络可以做到这一点。”

他说：“真正地，这是带来了深度学习部分来解决强化学习研究者多年来无法做到的事情，即即时生成这些特征。然后你可以为任何游戏做到这一点——没关系。然后这个……”

Bellemare稍微停顿了一下。“这起飞了。”

深度强化学习变得超人类

2015年2月，一篇论文出现在《自然》杂志封面上，标题为“学习曲线：自学AI软件在视频游戏中达到人类水平”。DeepMind将经典强化学习与神经网络的混合体已显示出具备人类水平游戏能力——而且远超人类——不仅仅在几个Atari游戏中，而是在数十个游戏中。深度学习革命已经来到强化学习，铸造了“深度强化学习”这一新领域，结果令人惊叹。

这个模型——称为“深度Q网络”，简称DQN——正在玩*Video Pinball*，取得的分数是专业人类游戏测试员水平的二十五倍。在*Boxing*中，它比人类表现好十七倍。在*Breakout*中，它好十三倍。一个几乎占整页的图表记录了这种在各种不同游戏中的惊人统治模式，全部使用单一通用模型，从一个游戏到下一个游戏无需微调或调整。

然而，在图表底部，有几个顽固的游戏拒绝向DQN屈服，这些游戏不符合这种辉煌的模式。其中一个特别突出，位于列表的最底部。

异常情况是《蒙特祖马的复仇》(Montezuma's Revenge)，这是1984年的一款游戏，玩家扮演一个名叫Panama Joe的探险家，必须在一个充满绳索、梯子和致命的、非常模糊的阿兹特克陷阱的神庙中找到出路。（“我对蒙特祖马或这种文化完全没有做任何研究，”创作者Robert Jaeger承认——他在16岁时就将自己的演示版本卖给了Parker Brothers——“我真的只是觉得这是一个色彩丰富的主题和一个很酷的名字。”）在《蒙特祖马的复仇》中，强大的DQN获得的最高分是人类基准的0%——是的，就是0%。

这里到底发生了什么？

首先，这个游戏让死亡变得极其容易。几乎任何类型的错误——撞到敌人、从太高的地方跳下来、穿过障碍物——都是必死无疑。DQN系统使用epsilon-greedy探索，这涉及通过在一定比例的时间内简单地随机按按钮来学习哪些动作会产生奖励。在《蒙特祖马的复仇》中，这几乎总是自杀。

第二个也是更重要的问题是《蒙特祖马的复仇》的奖励极其稀疏。在玩家获得任何分数之前，需要大量的事情完全正确地进行。在《突破》或《太空入侵者》等游戏中，即使是最困惑和最迷茫的新手，随机按按钮，也会很快意识到他们至少在做某些正确的事情。这足以启动学习过程：就DQN而言，记录得到的分数并开始在类似情况下更多地采取类似行动。相比之下，在《蒙特祖马的复仇》中，很少有事件提供除死亡之外的任何反馈。例如，在第一个屏幕上，你需要跳过四个深渊，爬三个梯子，对抗一个传送带，抓住一根绳子，跳过一个滚动的骷髅头，所有这些都必须在你收集第一个物品之前完成，这会奖励你一百个微薄的分数（而且不合时宜地，还有“La Cucaracha”的前五个音符）。

在奖励如此稀少的环境中，随机探索算法无法获得立足点；通过本质上摆动操纵杆和乱按按钮，它极不可能设法完成获得第一个奖励所需的所有必要步骤。在此之前，它根本不知道自己是否在正确的轨道上。

正如我们所见，解决这种稀疏性问题的一个解决方案是塑造：额外的激励奖励来推动算法朝正确的方向发展。但我们也看到了如何正确地做到这一点而不创造算法可以利用的漏洞是多么棘手。例如，奖励Panama Joe每秒不死亡，可能只会导致智能体学会永远不离开初始平台的安全区域。这将是动物研究人员所说的“习得性无助”的机器版本。正如著名格言家Ashleigh Brilliant所说，“如果你足够小心，坏事或好事都不会发生在你身上。”

其他直观的想法也会搁浅。比如说，奖励Panama Joe成功跳过滚动的骷髅头，可能会导致智能体在骷髅头上玩一种双人跳绳游戏，而不是进一步冒险进入神庙。同样，奖励他成功跳上或跳下绳子，会激励无限循环的泰山式摆动。这

些都不是我们想要的。

而且，这种塑造通常会因游戏而异，并且需要了解该特定游戏如何运作的内部知识的人类监督者的帮助。这感觉有点像作弊。街机学习环境背后的整个理念——以及DQN的激动人心的成就——是一个单一算法，能够从零开始掌握数十个完全不同的游戏环境，仅由屏幕上的图像和游戏内分数引导。

那么答案是什么？面对像《蒙特祖马的复仇》这样令人生畏的游戏，像DQN这样的通用试错算法如何可能被修改？

有一个诱人的线索，隐藏在众目睽睽之下。人类显然可以在没有任何额外塑造奖励的情况下学会如何玩《蒙特祖马的复仇》。人类玩家本能地想要爬梯子，到达远处的平台，到达第二个屏幕。我们想知道锁着的门的另一边是什么，看看神庙到底有多大，以及是否还有什么东西在那之后。不是因为我们直觉上认为它会给我们带来“分数”，而是出于更纯粹和最基本的东西：因为我们只是想知道会发生什么。

也许，征服像《蒙特祖马的复仇》这样的游戏所需要的不是用额外的激励来增强游戏的稀疏奖励，而是一种完全不同的方法。也许，答案不是一个越来越复杂的胡萝卜和大棒系统，而是相反的：开发一个内在而非外在激励的智能体。一个本质上会过马路的智能体，不是因为其中有奖励，而只是为了到达另一边。一个我们可以说是好奇的智能体。

过去几年见证了科学界对好奇心这一主题兴趣的重大复苏，以及机器学习研究人员和专门研究儿童认知的心理学家之间一些不太可能的合作，以从严谨、基础的角度更好地理解好奇心。它到底是什么？我们为什么有它？我们如何可能灌输它，不仅仅是在我们的孩子身上，而且在我们的机器中？

为什么这样做可能变得越来越关键？

CURIOSITY作为科学的研究主题

“渴望”了解为什么和如何，“curiosity”，…是心灵的欲望，通过在持续不懈地产生知识的过程中持续享受，超越了任何肉体快乐的短暂强烈。

—托马斯·霍布斯¹³

*Curiosity*是一切科学的开端。

—赫伯特·西蒙¹⁴

心理学中curiosity研究的教父是心理学家丹尼尔·伯林(Daniel Berlyne)。伯林1949年的第一篇论文就试图定义我们说某事物”有趣”或某人或动物对某事物”感兴趣”时的确切含义。¹⁵正如他所说：“我的第一个兴趣就是兴趣。”¹⁶

一个完整的子领域逐渐开始开放。动物在没有任何奖励依赖其学习的情况下学会了什么？

正如伯林指出的，心理学史很大程度上是人和动物被迫做事的故事——填写调查问卷或回答口头问题，按压杠杆以获得食物。然而，通过这种方式，该领域本质上创造了自己的方法论盲点。它将如何开始探讨有机体如何自主行动的问题？这几乎看起来像是一个术语上的矛盾。

“在某些方面，人类如此顺从和听话对心理学来说是不幸的，”他写道。¹⁷“在人类身上很容易诱发人为和外在动机，这阻止了我们研究在缺乏这些动机时起控制作用的动机因素。”

在心理学内部，用惩罚和奖励训练动物的研究议程在二十世纪中期如此占主导地位，以至于一时间似乎这可以解释智能有机体行为的一切。但这里和那里的某些数据拒绝符合这种解释。威斯康星大学的哈里·哈洛(Harry Harlow)到1950年已经开始记录恒河猴如何玩弄由锁和插销组合制成的物理谜题，他创造了“内在动机”这个术语来描述它。¹⁸有时这种内在动机不仅在没有外在奖励的情况下发挥作用，而且实际上压倒了外在奖励。一只饥饿的老鼠可能令人惊讶地决定放弃一点食物，或穿越带电的栅栏，去探索一个陌生的空间。猴子愿意按压杠杆不仅是为了饼干和果汁，而仅仅是为了向窗外看。¹⁹在严格的斯金纳(Skinnerian)外在奖励和惩罚世界中，这种行为几乎没有容身之地，也没有简单的故事来解释它们。

然而，正如伯林所看到的，这种内在动机与人类对食物和性的驱动同样是人性的核心——尽管被“心理学多年来过度忽视”。²⁰(事实上，我们社会允许的最严厉的惩罚，除了死亡之外——单独监禁——实际上是对人们施加无聊。)在他1960年的里程碑式著作《冲突、唤醒和Curiosity》中，伯林指出对curiosity的适当研究首先在1940年代末开始出现；他认为信息论和神经科学也在同一时间成熟并非巧合。²¹对curiosity的适当理解似乎只有在这三者的跨学科交汇点才可能实现。²²

伯林在自己的生活中似乎与他研究curiosity作为主题一样被curiosity强烈驱动。他至少懂十种语言（其中六七种流利），还是一位出色的钢琴家，以及慢跑者和旅行者。在他五十二岁英年早逝时，他正在追求乘坐世界上每一条地铁的目标。尽管文章和论文的产出丰富且多产，他很少在夜晚或周末工作。还有太多其他事情要做。²³

他的想法，特别是向神经科学和信息论寻求线索的议程，将在二十世纪后半期激励后继几代心理学家，在二十一世纪它们会形成完整的循环。从2000年代末开始，持续到2010年代的深度学习繁荣，正是数学家、信息论学者和计算机科学家——在诸如《蒙特祖玛的复仇》这样的内在动机问题上陷入困境——转向他的想法寻求帮助。

在广义层面上，他认为人类的内在动机似乎涉及三种相关但不同的驱动：对新奇性、惊喜性和掌握性的追求。这些中的每一个都提供了关于动机和学习的诱人想法，在这个十年里，这些想法似乎同样适用于机器和我们自己。

新奇性

研究通常被设计来回答的问题是”这个动物会对这个刺激做出什么反应？“...一旦实验情况变得更加复杂...一个新问题出现了：”这个动物会对哪个刺激做出反应？”

—丹尼尔·伯林²⁴

当我被夹在两个邪恶之间时，我通常喜欢选择我从未尝试过的那个。

—梅·韦斯特²⁵

人类好奇心和内在动机的核心概念之一是新颖性。在缺乏强有力激励的情况下，我们不会随机行动，就像使用e-贪婪探索策略的简单强化学习者那样。相反，我们非常稳健、可靠且可预测地被新事物所吸引。

1960年代中期，凯斯西储大学的Robert Fantz注意到，年仅两个月大的人类婴儿如果之前已经看过某些杂志图片，再次观看这些图片时会可靠地减少注视时间。²⁶ Fantz开始意识到，早在婴儿具备物理探索世界的运动技能之前，他们就已经能够视觉探索世界——并且很明显地被驱动着这样做。这种行为——被称为”偏好注视”——已成为发展心理学的基石性结果，是婴儿行为最显著的特征之一。

事实上，婴儿对观看新事物的偏好如此强烈，心理学家们开始意识到他们可以将此用作测试婴儿视觉辨别能力，甚至记忆能力的方法。²⁷ 婴儿能否区分两个相似的图像？能否区分同一颜色的两种相似色调？婴儿是否能回忆起一小时前、一天前、一周前看到的东西？对新颖图像的内在吸引力提供了答案。如果婴儿的目光停留时间较长，这表明婴儿能够判断相似的图像在某种程度上仍有不同。如果婴儿在一周未见某个图像后，再次展示时不怎么注视它，那么婴儿必须在某种程度上能够记住一周前见过它。在大多数情况下，结果显示婴儿的认知能力比之前假设的更早发展。视觉新颖性驱动确实成为心理学家工具箱中最强大的工具之一，开启了对婴儿心智能力更深层洞察的大门。²⁸

强化学习社区很快抓住了这种内在新颖性偏好的想法，并着手研究在计算领域能够做些什么。²⁹ 最直接的想法之一是简单地计算学习智能体之前在特定情况下的次数，然后让它在其他条件相等的情况下，偏好做之前做过最少次数的事情。例如，Richard Sutton在1990年建议，如果智能体采取了从未尝试过的行动，或很久没有尝试的行动，就为其奖励添加这样的”探索奖励”。³⁰

然而，这里存在一个相当明显的问题。在”这种情况”下采取”这种行动”的次数计算意味着什么？正如Berlyne在1960年所说，“‘新’这个词在日常语言中常用，大多数人似乎都能毫无困难地理解它。但当我们询问说一个刺激模式是新颖的究竟意味着什么，以及它有多新颖时，我们面临着一连串的陷阱和困境。”³¹

对于像解迷宫或非常非常简单的游戏这样的简单环境，当然可以简单地保存你遇到的每一种情况的列表，并在每次返回时添加一个计数标记。（用铅笔在迷宫中追踪路径实质上就是在迷宫本身上保存这种记录。）这种方法——保存你去过的每种情况、你做了什么以及发生了什么的巨大表格——被称为”表格式”RL，不幸的是，除了在非常小的环境中，它因完全不可行而闻名。例如，井字游戏是最简单的棋盘游戏，然而它有数千个独特的棋盘位置。³² 围棋游戏中可能位置的总数是一个170位数字。世界上所有计算机内存加起来都远不足以存储那张表格。

然而，除了这些实用性问题，在更复杂的环境中，更深层和更哲学性的问题是，首先什么叫处于“相同”的情况。例如，在Atari游戏中，像素可能出现的方式太多了，尽职地跟踪你曾经短暂遇到的每一个屏幕并稍微偏好新颖的屏幕对于产生有趣的行为根本没有帮助。对于合理复杂度的游戏，你可能永远不会看到完全相同的像素组合超过一次。从这个角度来看，几乎每种情况都是新颖的，几乎每个行动都未尝试过。即使你能存储这样的表格，它也不会是什么好的指导。

在日常人类决策过程中，当有人对我们说他们“从未遇到过那种情况”时，我们通常不会理解为“在这个确切的纬度和经度，在这个确切的纳秒，有这个确切的阳光斑驳模式照射我的视网膜，我的脑海中有这个确切的思维序列”，否则这种陈述在定义上实际上总是正确的，并被剥夺了所有意义。我们想要指的是情况中有时难以言喻的关键特征，我们通过这些特征来判断其新颖性。

在Atari游戏中，我们需要的是某种方法来衡量我们所处的情况——由屏幕上的像素表示——是否与我们之前经历过的情况在意义上相似。我们希望能够在共享某些更深层、非平凡相似性的情况下建立联系。

在伦敦的DeepMind，Marc Bellemare有兴趣思考如何将这个备受推崇但不实用的想法——计算你之前看到某个东西多少次（恰当地称为“count-based”方法）——扩展到更复杂的环境中。在像*Frogger*或*Freeway*这样的Atari游戏中，你试图穿越繁忙的道路，理想情况下，每次成功穿越都应该增加一种“计数”，记录你做过多少次——即使当时的交通本身总是处于某种新的、随机的模式中。

Bellemare和他的同事们正在研究一个叫做“density models”的数学概念，它似乎显示出一些希望。³³基本思想是使用无监督学习来构建一个模型，能够从周围的上下文预测图像中缺失的部分。（这与word2vec和类似的语言模型不无相似，这些模型旨在预测文本段落中缺失的词汇。）他们可以将智能体迄今为止看到的所有屏幕截图输入到这个density model中，然后使用其预测来分配一个数值概率分数，表示一个新屏幕相对于它已经看到的内容有多“可预测”。概率越高，越熟悉；概率越低，越新颖。这是一个引人入胜的想法，但这样的东西在实践中究竟如何工作还是一个悬而未决的问题。

他们开始用一个1980年代早期的Atari游戏*Q*bert*做一些实验，在这个游戏中你在一个由方形瓷砖组成的金字塔上跳跃，将每个瓷砖变成不同的颜色，直到你把它们全部变完，然后移动到一个全新的、不同颜色的关卡，重复同样的操作。他们让一个随机初始化的、白板状态的DQN智能体从零开始玩*Q*bert*，同时观察屏幕左侧的一个仪表，该仪表测量智能体所看到和体验内容的“新颖性”——通过density model测量。

起初——就像我们一样——一切都是新的。仪表指针停在最大值。屏幕上的每个图像都被记录为几乎完全新颖。

他们训练智能体几个小时，让它逐渐变得更加熟练地赚取分数（在*Q*bert*的情况下，每翻转一个瓷砖获得25分）。他们回来检查并观看它玩游戏。现在有些经验的智能体跳来跳去，获得分数。新颖性仪表的绿色条（技术上称为“inverse log probability”）几乎不从底部闪烁向上。智能体之前都见过这些。

Bellemare找到了智能体第一次成功完成游戏第一关的训练过程。他观看回放，想知道会发生什么。令人满意的是，当智能体开始接近目标时，那个绿色条开始再次爬升。

“现在，”Bellemare说，“当智能体接近关卡结束时，它开始说，嘿，这些情况很新颖！我之前真的没有在这种状态下过。这对我来说似乎很新。你会看到一个非常好的进展：当我们越来越接近完成游戏时，这个信号在上升。”³⁴

智能体跳到最后一个瓷砖，第一次完成了关卡。突然，整个屏幕闪烁和频闪。棋盘重置到下一关，青色瓷砖的金字塔消失了，一个全新的金字塔出现在它的位置上，这一个是亮橙色的。“看这个！”Bellemare说。绿色条一路飙升。新颖性信号几乎超出了图表范围。“智能体立即知道，我从来没有来过这里。”

这个density model似乎捕捉到了在多样化、高度复杂的环境中对新颖性的忠实概念，这些环境太太太丰富，无法直接实际计数。“我们看到这些结果，我们想，我们必须能用这个做点什么。”

现在的问题是，他们能否使用这个模型——他们将其称为”pseudo-count”——来激励智能体寻求这些新颖状态？³⁵如果你真的奖励智能体，不仅仅是因为得分，而是简单地因为看到新的东西，会发生什么？这反过来是否会造就更好的智能体，能够比那些只被训练来最大化奖励并偶尔随机按按钮的智能体取得更快的进展？

如果他们成功了，回报是显而易见的。“我们真的很兴奋，”他说，“尝试破解*Montezuma's Revenge*。”

Panama Joe发现自己被困的神庙有二十四个房间。在相当于日夜不眠地玩了三周游戏后，这个在其他几十个Atari游戏中表现出超人性能的DQN智能体只到达了第二个房间——几乎还在起跑线上。那里有整个神庙等待探索，充满危险且缺乏分数；也许一个因为看到它从未见过的东西而直接获得奖励的智能体，正是那种可能到达任何智能体都未曾到达过的地方的智能体。

Bellemare和他的团队打赌，如果一个智能体能够获得这些新颖性信号，如果它将这些信号视为游戏内分数的补充奖励，那么它在玩游戏时将更有动力且更成功。他们尝试了这个方法，让它训练与原始DQN智能体相同的时间——一亿帧，或近三周的24/7游戏。差异令人震惊。

同样的DQN智能体，在基于新颖性的奖励下训练，获得第一把钥匙的速度显著更快，最终通过了神庙的不是两个而是十五个房间。

新奇驱动的agent不仅获得了更多分数，还表现出了一种不同的行为——既有质的也有量的差异。所有那些“偏好性观察”都可以被利用来让它成功探索那些仅靠奖励还不够的地形。而且新奇驱动agent的某些特点看起来更有关联性——甚至更像人类。当奖励稀缺时，它不仅仅是一个摇杆摆弄者。它有驱动力。

“立即，” Bellemare说，“伪计数agent就会出去探索这个世界。”

惊喜的快乐

[认知好奇心的机制……通过概念冲突的等价物发挥作用，其功能本质上是动机性的。]

—[DANIEL BERLYNE]

[不是孩子们是小科学家，而是科学家是大孩子。]

—[ALISON GOPNIK]

与新奇一样，好奇心不可缺少的另一个高级概念是惊喜。一个好奇的孩子不仅关心事物在某种程度上是“新的”，还关心事物能够教给他们什么。比如说，一个有紫色圆点的棒球——一时间令人着迷，但如果它在其他方面的表现与标准棒球完全一样，这种着迷就会是短暂的。相反，我们对那些似乎违背我们预期、行为不可预测、敢于挑战我们去理解接下来会发生什么的事物保持兴趣。

在理解惊喜在人类好奇心中作用方面处于前沿的研究者之一是MIT的Laura Schulz。在2007年的一项研究中，她让孩子们玩一种惊喜玩具盒，操纵杆会让各种木偶从盒子的盖子中升起。研究人员会短暂地拿走玩具，然后带着熟悉的盒子和一个新的、颜色不同的盒子回来。他们会把两个盒子都放在孩子面前然后走开，等着看孩子会伸手拿哪一个。

“现在，我们之前对儿童游戏和好奇心的所有了解，” Schulz解释道，“都说，嗯，如果四岁的孩子已经和一个盒子玩了一段时间，那么如果你拿出一个新盒子，他们应该直接去拿新盒子。他们应该立即去玩新盒子，因为关于好奇心的基本观点是它关乎感知新奇性、感知显著性：这是他们没有见过的东西。”

但Schulz发现还有更多内容。在一些试验中，第一个盒子的演示被故意设计得含糊不清。盒子有两个操纵杆，如果同时按下它们，两个不同的木偶就会同时从盖子中升起。不清楚单独抬起其中一个操纵杆，或另一个，会做什么。是其中一个操纵杆负责两个木偶而另一个无效？还是每个操纵杆负责离它最近的木偶？或者是相对侧的那个？在这些模糊的情况下，四岁的孩子在有机会时并没有立即从熟悉的玩具转向新奇的玩具。相反，他们坚持不懈，伸手回去拿双杆盒子，以便准确弄清楚它是如何工作的。

“我们似乎经常对那些并不特别新奇的事物感到好奇，” Schulz说，“它们只是让我们困惑。”

因此开始形成这样一个画面：惊喜——不确定性、解决模糊性的能力、获得信息——与新奇一样，都是儿童内在动机的驱动力。

这个想法导致了第二条研究脉络，同样丰富，跨越了认知和计算。

Schulz与Rutgers心理学家Elizabeth Bonawitz和一组合作者一起，使用加重的积木进行了进一步研究。先前的研究表明，大约六岁时，儿童开始有关于如何最好地平衡不同大小和形状积木的理论。那个年龄的一些儿童假设(错误地)积木总是可以在其两端之间的中点平衡，即使物体是不对称的，而其他人理论化(正确地)积木在其质量中心平衡，在更接近较厚一侧的点上。这使得研究人员使用磁铁巧妙地创造出可以违反任一组假设的积木。当儿童玩按他们预期行为的积木时，他们的标准新奇性偏向出现，并导致他们在有机会时放弃积木而选择另一个更新的玩具。但对于那些

积木似乎违反他们关于积木应该如何平衡理论的儿童——无论他们的理论实际上是否正确！——保持着迷并继续玩它们，即使有另一个玩具可选。

四五岁的孩子，往往完全缺乏关于如何最好地平衡积木的具体理论，几乎总是在有新玩具时偏好新玩具。无论积木的行为如何，年幼的儿童似乎不够了解，或没有足够强的信念或预测来感到惊喜。

其他沿着这一思路的研究——比如约翰霍普金斯大学的 Aimee Stahl 和 Lisa Feigenson 在 2015 年进行的一项研究——进一步表明，婴儿玩玩具的方式也与玩具令人惊讶的方式相关。⁴³ 如果一辆玩具车看起来神秘地漂浮在半空中，婴儿会通过举起它并放下它来玩耍。然而，如果汽车看起来神秘地穿过了一堵坚实的墙，婴儿会通过在桌子上敲打它来玩耍。在每种情况下，当有机会尝试一个新玩具时，婴儿都会选择继续与令人惊讶的玩具互动。（没有看到玩具违反他们期望的对照组，会可靠地偏好新玩具。）Stahl 和 Feigenson 说，早在 11 个月大时，婴儿就会使用“对先前期望的违反作为学习的特殊机会”。

“看着一个婴儿很容易看到一张白纸，” Feigenson 说。“但实际上，婴儿对世界有丰富而复杂的期望——可能比人们给予他们的信任更多。”她认为，婴儿“使用他们已经了解的世界知识来激励或驱动进一步的学习，找出他们应该更多了解的内容。”⁴⁴

这个想法，用计算术语来说——一个不仅被奖励驱动，而且试图理解和预测环境的智能体——与强化学习本身一样古老。它也同样突然开花结果。

Daniel Berlyne 在 1950 年代看到了一些最早的机器学习实验，并思考使用惊喜或误预测作为强化器的可能性。“进一步的研究很可能旨在设计一个问题解决机器，它将根据经验改进技术，”他写道。“减少不匹配或冲突就必须是强化因子，使紧接在前面的操作在机器的优先级顺序中上移。”⁴⁵

德国 AI 研究员 Jürgen Schmidhuber 从 1990 年以来一直在探索这样一个想法：智能体从学习环境如何工作中获得奖励——也就是说，从提高预测能力中获得奖励。“它们可以被视为简单的人工科学家或艺术家，”他解释说，“具有建立世界更好模型的内在欲望，以及可以用它做什么的模型。”⁴⁶ 对 Schmidhuber 来说，就像 60 年代的 Berlyne 一样，这种学习理念在数学上根植于信息论——特别是在 Schmidhuber 看来——数据压缩的概念：一个更容易理解的世界更简洁地可压缩。

实际上，对 Schmidhuber 来说，我们在世界中行走努力更好地压缩我们对世界的表示这一想法提供了“创造力和乐趣的正式理论”。他解释说：“你只需要有计算资源——在学习模式之前，你需要如此多的计算资源，而之后你需要更少。差异就是你节省的地方。你懒惰的大脑喜欢节省东西。然后——”他打响手指。“那就是乐趣！”⁴⁷

像 Berlyne 一样，Schmidhuber 着迷的不是人们为了解决直接提出的问题而做什么——比如如何赢得游戏或逃出迷宫——而是人们在那些没有明确要做的事情的时候具体做什么。

他认为，婴儿是这方面的完美例子。“即使没有立即满足口渴或其他内置原始驱动的需要，婴儿也不会空转。相反，它积极进行实验：如果我这样移动我的眼睛、手指或舌头，我会得到什么感官反馈？”⁴⁸

正如 Schmidhuber 指出的，在好奇心的核心存在一个根本的张力，几乎是一种拉锯战：当我们探索环境和我们在其中的可用行为时——无论是 Atari 游戏的微观世界、现实世界的户外，还是人类社会的细微差别——我们同时为那些让我们惊讶的事物感到高兴，同时我们变得越来越难以被惊讶。几乎就像大脑包含两个不同的学习系统，彼此目标相反。一个尽力不被惊讶。另一个尽力让它惊讶。⁴⁹

那么，为什么不尝试直接建模这种张力呢？由博士生 Deepak Pathak 领导的加州大学伯克利分校团队在 2017 年着手构建这样一个智能体。Pathak 创建了一个由两个不同模块组成的智能体——一个设计用来预测给定动作的结果，当现

实与其预测匹配时获得奖励，另一个设计用来采取最大程度令人惊讶的动作，每当预测器错误时获得奖励。⁵⁰

在超级马里奥兄弟中，如果你刚刚按了跳跃按钮，你可以期待一会儿在屏幕上看到马里奥稍微高一点——不过只有在你已经尝试了几次之后。如果你按了向下箭头，你可以期待看到马里奥蹲下——但你可能不会期待这会让马里奥消失在下水道管道中进入一个巨大的地下世界！关键想法是通过让这种惊喜对智能体来说就像对我们一样令人愉快——即通过将这些预测错误转化为奖励——来激励智能体探索游戏。做任何结果令人惊讶的事情都可以被设置得同样好，那个动作也会得到同样强烈的强化，就像明确获得分数的动作一样。

Pathak和他的团队研究了这种惊喜奖励可能产生的行为类型。他们使用3D迷宫环境（基于经典90年代第一人称射击游戏*Doom*的引擎构建），在多个迷宫中将智能体放置在距离奖励“目标”状态越来越远的位置。仅基于发现目标的显式奖励进行训练的智能体，如果无法通过随机的摇杆摆动和按键操作找到目标，往往会选择简单地“放弃”。而具有基于惊喜奖励的智能体则会为了探索本身而探索迷宫：这个拐角后面有什么？那个远处的房间近距离看起来是什么样子？结果，这些好奇的智能体在比没有这种内在驱动力的智能体更加庞大和复杂的迷宫中找到了通往目标的路径。

Pathak的Berkeley团队与OpenAI的一组研究人员合作，共同继续探索使用预测误差作为奖励信号的想法。令人惊讶的是，他们发现这种架构的一个戏剧性简化——用设计来预测屏幕图像随机特征的网络替换专门设计来预测未来可控方面的网络——效果同样好，在某些情况下甚至更好。由Yuri Burda和Harrison Edwards领导的OpenAI研究人员致力于完善这个想法，他们将其称为随机网络蒸馏(random network distillation)，简称RND。不久之后，他们开始将目光投向蒙特祖玛的复仇。

他们让RND智能体在神庙中自由行动。在惊喜内在奖励的推动下，它始终能够平均探索神庙24个房间中的20到22个。在他们的一次试运行中，智能体做出了前所未有的事情。它一路到达了第24个也是最后一个房间，位于神庙左下角，并逃出了神庙。Panama Joe走过最后一扇门，发现自己面前是充满宝石的统一蓝色背景。他似乎从天空中坠落。这是蒙特祖玛的复仇所能提供的最接近超越的体验。这些宝石每个价值一千分——而且非常令人惊讶。

超越奖励

越来越清楚的是，“内在动机”——被理解为新奇性、惊喜或其他某种相关机制——对于系统来说是一种极其有用的驱动力，可以增强来自环境的外部奖励，特别是在那些外部奖励稀少或难以获得的情况下。

当然，从这个角度来看，很自然会问，如果我们将这种算法好奇心的想法发挥到逻辑极致，让强化学习智能体——矛盾的是——完全不关心外部奖励，会发生什么？

这样的智能体会是什么样子？它会做什么？

几乎每个研究内在动机的人都有同样的疑问，而一幅图景正在显现。

DeepMind的Marc Bellemare和他的同事们继续追求将基于计数的新奇性奖励扩展到更复杂领域的想法，在后续工作中，他们研究了他们所称的“推动内在动机的极限”。他们将智能体的新奇性奖励放大了约10到100倍，并观察到行为的质性和量性变化。

正如预期的那样，智能体的行为表现出一种不安定性。与追求游戏分数不同——后者通常导致相当稳定和一致的最佳实践集合——对于“最大好奇心”智能体来说，唯一的奖励就是来自这种探索行为，而这些奖励并不稳定——随着游戏环境的各部分变得更加熟悉，它们会消失。所以智能体会持续不安地追逐它们，而不是稳定在一种模式中。

不太预期的是，在脱离了游戏分数的情况下，智能体在游戏中表现得多么出色。具有超高新奇性奖励的智能体在四种不同的游戏中实际上达到了最先进的分数。好奇心孕育了能力。令人惊讶的是，仅仅新奇性奖励，完全无法获得游戏内分数，就足以胜任地玩许多Atari游戏——以他们无法获得的分数来衡量！

当然，必须说明的是，游戏（至少是好游戏）是为了吸引内在驱动的人类而设计的。毕竟，分数只是屏幕角落里的更多像素，人类玩家可以决定关心或不关心。所以在那个层面上，好奇心和探索驱动力被证明是最大化分数的良好代理是有道理的，至少在大多数游戏中如此。例如，在超级马里奥兄弟中，抓取硬币、破坏方块和踩踩敌人都会获得分数——但游戏的重点是让马里奥向右前进，那里有不可预测的景观等待着。从这个意义上说，内在驱动的智能体可能比那些被驱动去积累这些（最终毫无意义的）分数的智能体更符合游戏的预期游戏模式。

Berkeley的Pathak团队以及OpenAI的Burda和Edwards团队也继续追求这些问题，在完全没有外在奖励的学习方面合作进行了大规模、系统性的研究。

他们最引人注目的发现之一是，在大多数情况下，不需要明确告诉智能体它是否已经死亡。如果你试图最大化外在得分，这确实非常有用，因为这既是你确实获得分数的最终判决，也是一个指标，表明从那一刻起你可以期望零额外分数（这通常不鼓励第二次遭遇同样的命运）。对于纯粹好奇心驱动的智能体，死亡只是意味着从头开始重新游戏——这非常无聊！游戏的开始部分，作为最熟悉的部分，既不新颖也不令人惊讶。事实证明，这就是智能体需要的所有阻力。⁵⁸

他们还发现了内在动机驱动的智能体在得分方面表现出色这一模式的一个有趣例外。例外是Pong游戏。一个纯粹由内在奖励驱动的智能体，一个完全不在乎得分的智能体，玩游戏不是为了对对手得分，而是故意尽可能延长回合。得分后的“重置”本质上与其他游戏中死亡时发生的“重置”相同。相比于长回合中出现的非典型和不寻常位置，返回到熟悉的起始位置简直太无聊了。

团队很想知道，如果这样的智能体有机会与自己的复制品对战会发生什么。在零和游戏中，好奇心对抗好奇心会如何展开？答案是：出现了非零和的合作，因为双方都追求共同目标，即远离游戏中走得很烂的起始状态。换句话说：它们不停地对打，永不停止。研究人员写道，“事实上，游戏回合变得如此之长，以至于破坏了我们的Atari模拟器。”屏幕开始出现故障，随机斑块中的颜色点闪烁。当然，寻求惊喜的智能体对此非常高兴。⁵⁹

拔掉游戏分数并创建一个仅仅内在动机驱动的智能体的想法在某种程度上似乎是一个奇怪的实验：强化学习领域自成立以来就围绕外部奖励的最大化进行协调。为什么要放弃衡量行为的唯一标准？

密歇根大学的Satinder Singh与密歇根心理学家Richard Lewis和马萨诸塞大学阿默斯特分校的Andrew Barto合作，从哲学角度探讨了这个问题，询问“奖励来自哪里？”⁶⁰他们指出，对某种事态好坏的评估是在大脑内完成的——而不是在环境中。他们写道，“这种观点清楚地表明，奖励信号总是在动物内部产生的，例如，通过其多巴胺系统。因此，所有奖励都是内在的。”⁶¹

使用屏幕上的像素来玩Atari游戏，你可以从中获得任何你想要的东西——而不是直接获取某种法定奖励信号——毕竟，这正是实际玩电子游戏的感觉。

2000年代最受好评的电脑游戏之一*Portal*，涉及游戏AI反复承诺玩家完成游戏后会获得“一个蛋糕”。然而，在游戏进行到一半时，玩家发现了包含后来成为游戏最难忘口号的不祥涂鸦：“蛋糕是个谎言。”确实，游戏结束时没有提供蛋糕。当然，这个著名的背叛不仅被以下事实削弱了：它充其量只是蛋糕的数字表示，还因为我们玩游戏时并不抱着幻想认为除了取得进展、推进剧情和探索游戏世界的乐趣之外还会有任何其他收获。

我们不会将屏幕上的闪烁灯光用作数据，以便在该环境中获得“真正的”奖励。闪烁的灯光以及它们在我们身上引起的任何反应，就是所有的奖励。而且从我们投入电子游戏的小时数来看，这似乎已经绰绰有余了。

无聊和成瘾

碰巧的是，强化学习中的内在动机不仅是这些良性行为的源泉——在其中人们至少能看到人类了解、探索、看看会发生什么的愿望的闪光——它还映射了人类病理学的镜像：无聊和成瘾。

我问Deepak Pathak无聊的概念是否有意义。智能体是否可能感到无聊？

绝对可能，他说。

在超级马里奥兄弟的第一关中，有一个峡谷，他的智能体几乎从未想出如何穿越，因为这需要智能体连续按住跳跃按钮十五帧；长序列的精确动作比较短或更灵活的模式更难学习。⁶²结果，智能体到达悬崖边缘后就...试图转身。

“所以它就是无法穿越，”Pathak说，“所以就像一个死胡同，世界的尽头。”但游戏的设计是无法回溯的。智能体被困住了，学会了什么都不做。

Pathak还观察到了一种更普遍的倦怠。在他的超级马里奥兄弟智能体玩游戏足够长时间后，“它只是开始待在开头...因为任何地方都没有奖励——到处的错误都非常非常低——所以它学会了不去任何地方。”智能体只是在游戏的最开始闲逛，完全没有动机做任何事情。

这里至少有一丝悲哀的意味。一个对游戏感到厌倦的人可以停止游戏，通常也会这样做。我们可以换一个新游戏，或者干脆关掉屏幕，转向完全不同的事情。相比之下，这个智能体几乎残酷地被困在一个它不再有任何动力继续游戏的游戏中。

自从有了电子游戏以来，就有一个研究子领域专门探讨什么让游戏变得有趣，以及什么让一个游戏比另一个更有趣。这其中有着明显的经济和心理学利害关系。⁶³

我想到强化学习为我们提供了一个实用的基准，不仅可以衡量游戏的难度——智能体需要多长时间才能变得熟练——还可以衡量游戏的趣味性：智能体在失去兴趣和脱离之前会游戏多长时间，或者它是否选择花时间游戏这个游戏而不是另一个。很可能未来几十年的电子游戏会大量使用具有内在动机的RL智能体进行焦点小组测试。

认知科学家Douglas Hofstadter在他1979年获得普利策奖的著作《哥德尔、埃舍尔、巴赫》中，想象了高级游戏程序的未来，设想了游戏能力、动机和智能之间的联系：

问题：会有能够击败任何人的国际象棋程序吗？

推测：不会。可能会有能够在国际象棋中击败任何人的程序，但它们不会是专门的国际象棋程序。它们将是具有通用智能的程序，而且它们会像人一样喜怒无常。“你想下国际象棋吗？”“不，我对国际象棋厌倦了。我们来谈论诗歌吧。”

这句话现在看起来荒谬地过时了——当然，我们现在知道，凭借后见之明，不到二十年后，IBM的Deep Blue国际象棋机器就在1997年战胜了人类世界冠军Garry Kasparov。Deep Blue确实是一个专门的国际象棋程序——它从硬件层面就被定制为只能下国际象棋。它当然不具备通用智能；也不会渴望思考文学而不是国际象棋。

但也许这里仍然有一些根本上真实的内核。当代最先进的强化学习系统确实是通用的——至少在棋类和电子游戏领域——这种方式是Deep Blue所不具备的。DQN能够以同样的熟练程度游戏数十种Atari游戏。AlphaZero在国际象棋方

面和在将棋、围棋方面一样娴熟。

更重要的是，能够学会在现实世界中流畅操作的人工通用智能(AGI)可能确实需要那种能够让它对游戏过多的游戏感到“厌倦”的内在动机架构。

与厌倦相对的另一个极端是成瘾——不是脱离参与，而是其黑暗的反面，一种病理程度的重复或坚持。在这里，强化学习也表现出了某些令人不安地、不舒服地像人类的行为。

研究内在动机的研究人员谈论他们称之为”嘈杂电视”问题的现象。如果环境中有一个本质上取之不尽的随机性或新奇性来源会怎样？具有内在动机的智能体是否会完全无法抗拒它？

具体来说，想象屏幕上有一个不可预测的视觉噪声源：嘈杂的电视是经典例子，尽管噼啪作响的火焰、沙沙作响的树叶或奔腾的流水也都符合条件。如果是这种情况，每一个新的、不可预测的光影配置都会像一种无穷的好奇心彩票一样发挥作用。至少在理论上，面对这种情况的智能体应该会立即变得愚蠢。

然而，大多数1970年代和80年代的简单Atari游戏碰巧不包含这样的视觉随机性来源，所以这还没有得到经验证实。Pathak、Burda和Edwards决定将这个思想实验变为现实并尝试一下。他们创建了一个简单的3D迷宫游戏，其中智能体需要探索迷宫并找到出口。然而，在游戏的一个版本中，迷宫的一面墙上有一个电视屏幕。此外，智能体被赋予了按按钮来换电视频道的能力。会发生什么？

发生的情况是，智能体一看到电视屏幕，对迷宫的探索就戛然而止。智能体将屏幕置于视野中心并开始翻频道。现在它看到一架飞行中的飞机视频。现在它看到可爱的小狗。现在它看到一个坐在电脑前的男人。现在是市中心交通中的汽车。智能体不断换频道，沉浸在新奇和惊喜中。它再也不动了。

视觉信息不是唯一能够产生这些危险效应的随机性来源；像抛硬币这样简单的事情也可以。这在近十年前就困扰着DeepMind研究员Laurent Orseau的思考，他是他们安全团队的第一个员工，现在是他们基础研究小组的一部分。早在能够瞬间变成沙发土豆的具有内在动机的Atari游戏智能体出现之前，Orseau就在思考一个更强大的智能体，被一枚硬币迷住。

Orseau在思考一个他称为”知识寻求智能体(knowledge-seeking agent)“的假想智能体，这是一个”目标是收集尽可能多关于未知世界信息”的智能体。Orseau的智能体基于一个名为AIXI的理论框架，该框架设想了一个具有无限计算能力的智能体。当然，这样一个心理上全能的(omnipotent)智能体在现实中永远不可能存在，但它起到了一种参考点的作用：如果你在采取行动前能够永远思考下去，你会选择哪一个？令人惊讶的是，许多关于资源无限的知识寻求智能体的概念在看到硬币时完全堕落了(degeneracy)——“相比探索环境中更具信息量的部分，更倾向于观察抛硬币”，Orseau指出。“原因在于这些智能体将随机性的(stochastic)结果误认为是复杂信息。”

B. F. Skinner在不训练鸽子的时候，着迷于赌博成瘾。赌场总是平均获胜，而从Thorndike开始的心理学一直基于这样的观点：当某事总体上对你有好处时，你会更多地去做——当对你有害时，你做得更少。从这个角度看，赌博成瘾这样的事情是不可能的。然而它就在那里，存在于现实世界中，挑战着行为主义者去理解它。“赌徒似乎违反了效应法则，”Skinner写道，“因为他们继续赌博，尽管他们的净奖励是负数。因此有论点认为，他们赌博一定有其他原因。”我们现在似乎有了一个很好的候选答案，这些其他原因可能是什么。

赌博成瘾可能是内在奖励对外在奖励的超越(毕竟赌场总是赢家)。随机事件总是至少稍微令人惊讶的，即使它们的概率被很好地理解(比如一枚公平的硬币)。

在第4章中，我们讨论了dopamine在编码时间预测错误中的作用：奖励比预期更好或更差的情况。然而，有一些奇怪的案例不符合这种模式。也就是说，自世纪之交以来，越来越多的证据表明，新颖和令人惊讶的事物会触发

dopamine的释放，无论它们是否与任何”奖励”相关联。

正如reinforcement-learning社区正在发现将新颖性和惊喜作为奖励本身的价值，神经科学社区正在揭示这种机制在我们自己大脑中的运作。

与此同时，越来越清楚的是，这些通常帮助我们的机制如何可能出错。Reinforcement-learning智能体可能会沉迷于换台和玩老虎机——当然，我们也会。因为像这些行为的结果从来不是我们认为的那样，这种活动总是有令人惊讶的地方，似乎总有什么可以”学习”的。我们不认为成瘾是动机的过剩，是好奇心的过度，但沿着这些思路的某些东西很可能就是它实际运作的方式。

为其本身

内在动机的计算研究为我们提供了一个强大的工具包，用于在困难的学习环境中取得进展——Montezuma’s Revenge只是一个例子。在更深层次上，它还为我们提供了一个基于这些经验成功的故事，解释为什么我们自己可能有如此引人注目的动机。

通过更好地理解我们自己的动机和驱动力，我们反过来有机会获得互补和相互的洞察，了解如何构建像我们自己一样灵活、有韧性和在智力上贪婪的(omnivorous)人工智能。

Deepak Pathak看着深度学习的成功，看到了一个明显的弱点：每个系统——无论是机器翻译、物体识别，甚至游戏——都是专用的。在大量手动标记图像上训练一个巨大的神经网络，正如我们所见，是深度学习首次真正展示其前景的范式。明确训练一个系统来分类图像，制造了一个能够分类图像的系统。这是公平的，他说。“但问题是，这些人工智能系统实际上并不智能。因为它们缺少一个对人类非常核心的关键组件，那就是这种通用行为，或通用学习系统。”

他论证说，要制造一个通用系统，需要打破这种特定任务的思维模式——也需要打破一个关键的、明显的人工制品：这些模型需要的大量明确奖励信息。像AlexNet这样的图像标记系统可能需要数十万张图像，每张都由人类标记。这显然不是我们在生命早期获得视觉技能的方式。强化学习同样如此，在Atari游戏世界中，每十分之一秒游戏都会以完美的权威告诉你确切的表现如何。“这效果很好，但它又需要一些非常非常奇怪的东西，”他说，“那就是这种奖励。”

断开硬连线的外部奖励可能是构建真正通用AI的必要组成部分：因为生活不同于雅达利游戏，生活明确地没有预先标记实时反馈来告诉我们每个行动的好坏。当然，我们有父母和老师，他们可以纠正我们的拼写、发音，偶尔还有我们的行为。但这几乎没有涵盖我们所做、所说和所想的一小部分，而且我们生活中的权威人士并不总是意见一致。此外，人类境况的核心成长仪式之一就是我们必须学会用自己的眼光为自己做出这些判断。

“对我来说，除了做好探索之外的正交议程，一直是完全移除奖励，” Bellemare说。“这超出了我通常的舒适范围，但你知道，我认为我们对AI代理能做的最有趣的事情就是让它们提出自己的目标。这里确实存在安全问题，”他承认，“但我希望，正如你所说，我的AI代理被投入到《蒙特祖玛的复仇》中，仅仅因为它喜欢玩而去玩它。”⁶⁹

“我们之前谈到了ALE[街机学习环境]以及它如何成为一个很好的基准测试，” Bellemare说。“对我来说，在某种意义上我们基本上完成了ALE。我们基本上完成了最大化得分。”对他来说，任何通过epsilon-贪心按钮操作而来的高分，通过得分强化，都不符合智能的标准——尽管结果可能令人印象深刻。“我实际上认为我们应该从事物的行为方式来衡量智能——而不是从奖励函数的角度。”这样的行为会是什么样子？这是驱动他的核心问题之一。

Orseau关于知识追求代理的工作也勾勒出纯粹由知识追求动机驱动的心智会是什么样子。初步分析是令人鼓舞的。一个被激励去最大化某种分数或达到某个目标状态的人工代理，总是有利用某些漏洞来做到这一点的风险；一个更智能的代理可能倾向于黑客攻击评分系统或为自己构造一个逃避现实的幻想，在那里它的目标更容易实现。Orseau强调，虽然这对我们来说似乎是“作弊”，但“它没有作弊的概念。它只是，‘好吧，我做行动来最大化我的奖励。’”他详细说明：“代理不理解这是坏事。它只是尝试许多不同的行动，然后这个有效：那为什么不这样做呢？”⁷⁰

然而，知识追求代理不能采取任何这样的捷径。特别是自我欺骗，对它没有兴趣或吸引力。“所以想象你正在修改你的观察...那么你获得什么信息？什么都没有。因为你可以预测它将会是什么。”⁷¹由于这种韧性，知识追求代

理”因此可能是我们自己世界中AGI最合适代理，这是一个允许自我修改并包含许多自欺方式的地方。^{“72}

在释放超智能知识追求代理之前仍有理由犹豫——寻求知识可能涉及征用各种地球资源来做到这一点。但它至少对一些最直接的陷阱具有韧性。“如果你能编程它，我相信它会有惊人的行为，”Orseau说。“因为它会尽快尝试理解它的环境。它基本上是终极科学家。它会设计实验来尝试理解会发生什么... 我真的很好奇看到它会如何行为。”

主要或纯粹由好奇心引导的智能的概念——和伦理——并不是一个新想法；它不仅早于过去十年，实际上还早于过去千年。在柏拉图著名的对话《普罗塔哥拉》中，苏格拉底反思了这个问题并说得相当好：

“知识是一件精美的东西，完全能够统治一个人，”苏格拉底说。“如果他能区分善恶，没有什么能强迫他以知识指令以外的方式行动，因为智慧是他所需要的全部强化。”⁷³

第三部分

规范性

当我的父母告诉我，在我的头骨里有一颗小小的、黑暗的宝石，正在学习成为我时，我六岁。

—格雷格·伊根¹

看着这个。

—埃隆·马斯克对彼得·蒂尔，就在他失去控制并撞毁他未投保的100万美元迈凯轮F1之前²

在英语中，我们说模仿某物是”ape”它，我们不是唯一的；这种看似任意的语言怪癖在各种语言和文化中一再出现。意大利语 *scimmiettare*、法语 *singer*、葡萄牙语 *macaquear*、德语 *nachäffen*、保加利亚语 *majmuna*、俄语 *обезьяничать*、匈牙利语 *majmol*、波兰语 *malpować*、爱沙尼亚语 *ahvima*：模仿和模拟的动词，一再地，它们的词源根植于灵长类动物的术语。³

确实，猿类作为伟大模仿者的声誉，不仅在词源学上，在科学上也至少可以追溯到一个半世纪。正如19世纪生物学家(也是查尔斯·达尔文的朋友)乔治·约翰·罗马内斯在1882年关于“达尔文先生所称的‘模仿原理’‘的主题上写道：

众所周知，猴子将这一原则发挥到了荒谬的程度，它们是唯一为了模仿而模仿的动物……不过对于会说话的鸟类应该例外。⁴

确实众所周知，而且跨越了令人惊讶的各种文化和语言。然而——具有讽刺意味的是——这似乎实际上并不正确。

灵长类动物学家Elisabetta Visalberghi和Dorothy Fragaszy在提出“猴子会模仿吗？”这一问题时，仔细审视了证据，通过文献回顾和自己的实验，被迫得出结论：数据显示，猴子实际上“极度缺乏模仿行为”。她们写道：“猴子缺乏模仿能力，在使用工具的行为中和在任意行为（如姿势、手势或解决问题）中都同样明显。”⁵

比较心理学家Michael Tomasello的后续研究对我们稍微更近的灵长类亲属提出了同样的问题——“类人猿会模仿吗？”——并得出了类似的决定性结论，可能的例外是黑猩猩，我们在基因上最接近的亲属。（黑猩猩在野外到底在多大程度上确实模仿，或者在人类训练下能够模仿，仍然是一个微妙且某种程度上未解决的问题。）Tomasello说：“对于类人猿是否模仿这一更一般问题的回答是：只有在人类正式或非正式训练下才会这样做（而且可能只在某些方面）。”⁶因此，灵长类动物的模仿声誉或多或少是完全不应得的。

然而，确实有一种灵长类动物是天生的、不可思议的、多产的、看似天生就会模仿的。

那就是我们人类。

1930年，印第安纳大学心理学家Winthrop Kellogg和他的妻子Luella，在九个月里将他们的婴儿儿子Donald与一只名叫Gua的婴儿黑猩猩一起抚养，对待他们完全相同，就像人类兄弟姐妹一样。在Kellogg夫妇写的关于这一经历的书《类人猿与儿童》中，他们注意到“由于黑猩猩作为模仿者的声誉，观察者从一开始就警觉地注意这种行为的出现。然而，奇怪的是，Gua的模仿行为明显比男孩要少。”

Donald确实是一个多产的模仿者，既模仿父母也模仿他的“兄弟”。在十七个月大时，他背着手来回踱步，让父亲大吃一惊——这完全是Winthrop本人在深度集中时刻的翻版。不过，Donald更多时候模仿的是Gua，他的玩伴和同龄伙伴。尽管Donald已经会走路，实际上在学会走路之前几乎没有爬过，他开始效仿Gua，开始四肢着地爬行。当附近有水果时，Donald学会了像Gua那样咕噜和吠叫。略感担忧的Kellogg夫妇很快就突然结束了他们的实验。⁷

事实上是我们人类才是自然界卓越模仿者的证据继续增加。例如，当你对婴儿吐舌头时，他们会——在出生后不到一小时——对你回吐舌头。⁸考虑到孩子从未见过自己——正如加州大学伯克利分校的Alison Gopnik所指出的，“子宫里没有镜子”⁹——这一壮举更加惊人，所以这种模仿是“跨模态的”：他们将你吐舌头时的外观与他们这样做时的感觉相匹配。这一切都在最初的四十分钟内完成。

这种令人难以置信的能力最初是由华盛顿大学的Andrew Meltzoff在1977年发现的；这一发现颠覆了一代心理学的既定智慧。传奇的瑞士发展心理学家Jean Piaget（在二十世纪的引用量中，其影响力仅次于弗洛伊德排名第二）¹⁰在1937年写道：“在最早期阶段，儿童像唯我论者一样感知事物……但随着他智力工具的协调，他通过将自己作为一个主动对象置于外在于自己的宇宙中的其他主动对象之中来发现自己。”¹¹

Meltzoff在承认所有心理学家都对这位伟大的二十世纪瑞士心理学家有所亏欠的同时，认为在这个特定案例中，Piaget的观点恰恰相反。他说：“我们必须修正我们目前对婴儿期的概念。对自我-他人等同性的认识是社会认知的基础，而不是结果。”¹²他说，模仿是“婴儿期心理发展的起点，而不是其顶点”。¹³

这种模仿他人的倾向几乎立即开始，但它远非仅仅是反射。儿童模仿谁、什么以及何时模仿，存在令人惊讶的复杂性，我们在过去几十年才开始理解这一点。

例如，只有当成人的动作产生有趣的效果，而不是看起来没有效果时，他们才会模仿。¹⁴幼儿似乎也有一种特殊的感觉，认为是其他人类“让事情发生”；如果物体似乎是自己移动的，或者如果是机器人或机械手做出动作，他们不会模仿这个动作。¹⁵（这对机器人保姆和教师的可行性具有有趣的含义。）

看起来婴儿也非常敏锐地意识到自己正在被模仿。在1933年马克思兄弟的电影《鸭羹》中有一个著名的场景，哈珀假装成格劳乔在镜子中的倒影，模仿他的每一个动作。Meltzoff做了一个类似的研究，成年人要么模仿婴儿的动作，要么只是按照预先计划的固定动作序列进行。就像格劳乔一样，婴儿会做出复杂或不寻常的动作来测试成年人是否真的在模仿他们。¹⁶

对Meltzoff来说，这种深层的能力——认识到我们自己与他人的关系，我们在某种根本意义上感知他们像我们自己——不仅是心理发展的开始，正如他所说，“这是社会规范、价值观、伦理、共情发展的内核胚胎基础……这是一个大爆炸。最初的开始就是这种身体动作的模仿。”¹⁷

过度模仿

想象一下你正在向某人展示如何切洋葱，你说：“现在这样试试”，然后清了清嗓子开始演示切法。你的学生仔细观察你，为了寻求你的认可，在做同样的切法之前清了清嗓子。他们不仅仅是模仿了你；他们过度模仿了你，在模仿中包含了那些根本不相关或对正在执行的任务没有最终因果影响的行为。¹⁸

研究人类和黑猩猩模仿行为的研究者们惊讶地发现，这种过度模仿在人类中比在黑猩猩中更为常见。这似乎违反直觉：黑猩猩怎么能在判断哪些行为相关哪些不相关，然后只重现相关行为方面做得更好呢？

其中一个最具启发性和有趣的研究涉及有两个锁定开口的塑料盒子：一个在顶部，一个在前面。实验者首先演示解锁顶部开口，然后解锁前面的，接着从前面伸手去拿一点食物。当黑猩猩看到这个使用不透明黑盒子的演示时，它们忠实地按照相同顺序做了两个动作。但当实验者使用透明盒子时，黑猩猩能够观察到顶部开口与食物毫无关系。在这种情况下，黑猩猩会直接去前面的开口，完全忽略顶部的开口。相比之下，三岁的儿童，即使他们能看到第一步什么都没做，也会重现这个不必要的第一步。¹⁹

有理论认为，在这个例子中，人类可能只是在发展相关技能方面较慢。研究者从研究三岁儿童转向研究五岁儿童。过度模仿行为更严重了！较大的儿童比较小的儿童更容易过度模仿。²⁰这没有意义。到底发生了什么？

问题变得更加奇怪。研究者认为儿童进行过度模仿是为了获得实验者的认可。他们让实验者离开房间；这没有帮助。当研究者询问三岁和五岁的儿童是否能分辨出演示动作中哪些是“必须做的”，哪些是“愚蠢和不必要的”时，儿童们能够分辨！但即使在他们向实验者展示了他们知道区别的过程之后，他们仍然重现了两个动作。²¹

最后，实验者尝试明确告诉儿童不要做任何“愚蠢和多余的”事情。这没有帮助。儿童同意了指令，然后仍然过度模仿。

再次，这似乎违反直觉且几乎是矛盾的：随着认知能力的提高，儿童表现出明显的“控制力减弱，在[他们]发展过程中趋向于更‘无意识’的全盘复制”。²²

匈牙利心理学家György Gergely对14个月大婴儿的研究提供了一个线索。幼儿看到一个成年人坐在桌子旁，向前倾身用前额触碰一个灯泡，使它亮起来。然而，有一个关键的变化。一半时间里，成年人的手臂搭在桌子上，另一半时间里，成年人假装很冷，用毯子裹着自己。看到成年人的手臂自由的幼儿完全重现了这个动作，弯腰用头去触碰灯泡。但看到成年人的手臂被毯子占用的幼儿只是伸出手用手触碰灯泡。²³

这里有一个关键点。刚满一岁多一点的幼儿就能够评估实验者采取奇怪行为是出于选择还是出于必要。这个成年人弯腰用头触碰灯泡一定有某种原因——她的手就在那里！如果这似乎是一个深思熟虑的选择，他们会完全重现它。这从一个新的角度展现了过度模仿的问题。它根本不是“无意识的”，不是简单地奴隶般重现精确动作，而是相反——基于将演示者想象为做出理性选择并尽可能轻松有效地执行动作的合理、复杂的洞察。

突然间，这种行为从一岁到三岁期间增加，然后从三岁到五岁再次增加，这开始变得有意义了。随着儿童认知复杂性的增长，他们更能够模拟他人的心理。当然，他们能看到——在透明立方体的情况下——成人正在打开一个没有效果的插销。但他们意识到成人也能看到这一点！如果成人能看到他们正在做的事情没有明显效果，但仍然继续这样做，那么必然有原因。因此，即使我们无法弄清楚那个原因是什么，我们最好也做那件“愚蠢”的事情。

相比之下，黑猩猩没有这种对人类示范者的复杂心理模型。其逻辑似乎要简单得多：“人类很笨，没有采取获得食物的最佳行动。无所谓。我能看到获得食物的最佳方式，所以我就那样做。”

突然间，这个明显的悖论得到了解决。在像这样一个奇怪的人工场景中，黑猩猩碰巧是正确的。但正是人类儿童，“过度”模仿的那一个，在认知上更加复杂，能够接触到对这个问题更深层次的洞察。事实上，成年人——除非他们在做实验室研究——通常不会在知道更好方法的情况下，作为规则去做无意义的行动。婴儿为了获得食物采取两个行动而不是一个看起来很愚蠢，但这只是因为成年人在某种意义上是欺骗性或不真诚的。羞耻的是他们！

更近期的研究已经确定了这些效应可以多么微妙。儿童从很小的年龄开始，就对展示某些东西的成年人是故意在教他们，还是只是在实验，非常敏感。当成年人将自己表现为专家——“我要向你展示它是如何工作的”——儿童会忠实地重现甚至是成年人采取的看似“不必要”的步骤。但当成年人将自己表现为对玩具不熟悉——“我还没有玩过它”——儿童只会模仿有效的行动，而忽略“愚蠢”的行动。再次显示，看似过度模仿，而不是不理性或懒惰或认知简单，实际上是对教师心理的复杂判断。

所有这些都显示了在表面看起来像是简单、机械模仿背后潜藏的巨大认知复杂性。结果是对幼儿心理技能的新认识，以及对简单告诉机器学习系统“看这个”的计算复杂性的更深理解。

模仿学习

如果我们人类独特地具备模仿能力，这引出了一个明显的问题：为什么？模仿的什么特点使其成为如此强大的学习工具？事实上，模仿学习相对于试错学习和明确指导至少有三个不同的优势。

我们已经看到机器学习研究人员如何直接从心理学借鉴像塑造(shaping)和内在动机这样的想法。模仿已经证明是一个同样丰富的灵感源泉；事实上，它构成了AI在二十世纪和二十一世纪许多最大成功的基石。

模仿的第一个优势是效率。在模仿中，别人通过试错获得的来之不易的成果被装在银盘上递给你。事实上，模仿学习优势的很大一部分在于知道你试图做的事情首先是可能的。

2015年，著名攀岩者汤米·卡德威尔(Tommy Caldwell)和凯文·约格森(Kevin Jorgeson)创造了历史，完成了对约塞米蒂谷传奇的三千英尺黎明墙的首次成功攀登——被《Outside》杂志称为“世界上最难的攀岩”。卡德威尔和约格森花了八年时间规划他们的路线，在悬崖面的不同段落上进行实验，试图连接各种可行的路段，找到一条从底部到顶部的可行路径。用卡德威尔的话说，“黎明墙”比我之前甚至想象过的任何攀岩都要困难得无穷倍”。岩石乍一看像是完全光滑的表面，挑战你想象某种找到抓握点的方法。仅有的一点抓握点似乎对人类皮肤完全敌对。“这是你的手指能做的最困难的事情，攀爬这条路线，”卡德威尔说。“就像抓住剃刀刀片一样。”

第二年，年轻的捷克攀岩天才亚当·翁德拉(Adam Ondra)只用了几周的侦察和练习就能够复制他们的壮举。他将这种速度很大程度上归功于不仅被展示了上黎明墙的方式，而且被展示了这完全可能的事实——这是卡德威尔和约格森在开始他们艰难的攀登规划过程时不知道的。翁德拉说：

汤米和凯文投入所有这些努力——比如，多年又多年的工作...这太令人印象深刻了...有这么多路段，在关键段落上，甚至在一些较容易的段落上——如果你在那里，你会想，“不。这是不可能的。”只有在研究每一个微小的剃刀刀片之后，有时唯一可能的[答案]才会出现在你的脑海中。所以对我来说解决每个单独段落的谜题要容易得多，因为我知道那些家伙做到了...他们完成了这条路线，我很自豪地做出了首次重复攀登。

第一次攀登耗费了八年的详尽搜索和自我怀疑。第二次攀登只用了几周的学习和排练，在信心的支撑下，无论看起来多么不可能，总是有办法的。

我们之前看到游戏《蒙特祖马的复仇》需要大量事情同时进行正确才能获得第一个奖励，即便如此，成功完成它的路径也是极其狭窄的，鼓励性反馈稀少，失败的后果严重。这就像强化学习(reinforcement-learning)的黎明之墙等价物——一个几乎空白、冷漠的表面，挑战你找到立足点。即使使用像新颖性奖励和内在动机这样强大的技术，智能体(agent)仍然需要大量尝试才能学会游戏机制和成功路径。但如果智能体不必自己探索游戏呢？如果它有一个榜样呢？

2018年，由Yusuf Aytar和Tobias Pfaff领导的DeepMind团队提出了一个巧妙的想法。他们想知道，智能体是否可能不通过痛苦的游戏内探索，而是通过...观看YouTube视频来学习如何玩游戏？

这个想法很强大，也足够疯狂到可能成功。YouTube上充满了人类玩家玩游戏的视频。他们的智能体实际上可以通过首先观看别人采取那些行动来学习各种行动的分值。然后，当它被释放自己行动时，它已经对该做什么有了基本概念。第一个训练模仿这些人类玩家的智能体，在当时比任何仅通过游戏奖励进行强化学习训练的智能体都要好。事实上，在2018年底使用内在动机取得突破之前，它是第一个在人类榜样的初始帮助和启发下走出神庙的人工智能体。

模仿赋予的第二个关键优势是某种程度的安全性。通过数十万次失败来学习在Atari领域可能很好，那里死亡只是重新开始。然而，在生活的其他领域，我们没有能够失败数十万次才把事情做对的奢侈。例如，外科医生或战斗机飞行员希望学习极其精确和复杂的技术而从不犯关键错误。这个过程的关键是观察前辈们的现场、录制甚至假设的成功和失败。

模仿的第三个优势是它允许学生（无论是人类还是机器）学习做难以描述的事情。19世纪心理学家Conway Lloyd Morgan在写下这句话时就有这个想法：“在传授技能方面，五分钟的演示比五小时的谈话更有价值。仅仅描述或解释如何完成一项熟练的技艺是相对没有用的；展示如何做要有用得多。”当我们试图表达我们想要的行动时，这是正确的：“现在将你的肘部弯曲27度角，同时非常快速但不要太快速地轻弹你的手腕...”当我们试图表达我们希望学生追求的目标时，这同样正确。在Atari游戏中，像“最大化总分”或“尽快完成游戏”这样的目标可能或多或少足够了。但在现实世界场景中，甚至沟通我们希望学习者做的一切可能都很困难。

这方面的经典案例可能是汽车。我们想尽快从A点到达B点，但不能超速行驶——或者说，不能超速太多，除非因为某种原因我们必须这样做——并且要保持在车道中央，除非有骑自行车的人或停下的车——不要在右侧超车，除非这样做比不这样做更安全，等等。很难将所有这些形式化为某种客观函数，然后告诉系统去优化。

在这种情况下，最好使用人类未来研究所(Future of Humanity Institute)的Nick Bostrom所说的“间接规范性(indirect normativity)”——一种让系统与我们的期望保持一致的方式，而无需将它们详细表达到最后的细节。在这种情况下，我们想说的是“看我如何开车。就这样做。”

事实证明，这是自动驾驶汽车的最初想法之一——直到今天，仍然是最好的想法之一。

转向

1984年，DARPA开始了一个他们称为战略计算倡议(Strategic Computing Initiative)的项目。这个想法是利用1980年代发生的计算突破，并将当时最前沿的技术转化为三个具体应用。正如当时刚从卡内基梅隆大学获得机器人学博士学位的Chuck Thorpe回忆：“为什么是三个？嗯，一个让陆军高兴，一个让空军高兴，一个让海军高兴。”空军想要一个“飞行员助手”：一种能够理解飞行员大声说出的命令或请求的自动副驾驶。海军对他们称为“战斗管理”系统感兴趣，该系统可以帮助进行场景规划和天气预测。这样陆军就剩下了。他们想要的是自主陆地车辆。

Thorpe在那年九月成功为他的博士论文答辩，告诉委员会他计划休几周假，然后考虑下一步要做什么。然而，CMU机器人研究所所长Raj Reddy一口气祝贺了他，然后说：“你下一步要做的是五分钟后在我办公室开会。”会议内容是关于为DARPA建造自动驾驶车辆。

“那就是，”Thorpe回忆道——那五分钟的窗口期——“我完成论文和开始博士后研究之间的休息时间。”

在某种意义上“自动驾驶”的“车辆”到1984年已经存在了好几年，但称这项技术为原始可能都太过慷慨了。机器人技术先驱Hans Moravec在1980年斯坦福大学的PhD论文中，让一个桌子大小、装在自行车轮子上的机器人“推车”能够使用车载电视摄像头自主移动并避开椅子和其他障碍物。“这个系统相对可靠，”Moravec写道，“但非常慢。”^[36]有多慢呢？这个推车被编程为一次移动一米——用Moravec的话说，就是“蹒跚前进”。在这些一米长的蹒跚移动之后，推车会停下来，拍照，然后思考十到十五分钟，才做出下一个同样试探性的机动。因此，它的最高时速被限制在每小时0.004英里。

这个机器人如此缓慢，以至于在户外完全被搞糊涂了，因为太阳角度在两次蹒跚移动之间变化太大，阴影似乎在令人困惑地移动。^[37]正如Thorpe回忆的那样，“实际上，他的系统锁定了这些清晰、锐利边缘的阴影，发现它们在移动，并且看到真实物体的移动与阴影不一致，决定对阴影比对真实物体更有信心，于是抛弃了真实物体，锁定了阴影，然后撞翻了他的椅子。”

到1984年，Moravec来到了Carnegie Mellon，Thorpe与他合作将两次蹒跚移动之间的时间从十分钟减少到三十秒。这相当于最高时速不到十分之一英里。这算是进步。

当时，最先进的计算机是称为VAX-11/784（“VAXimus”）的设备，大约八英尺宽、八英尺高。像Moravec推车这样的车辆会通过“脐带”连接到这些计算机上。但要制造一个能够真正在外部世界中移动的车辆，需要将计算机带上路，这反过来也意味着要为计算机带上电源。这最终意味着需要一个四缸发电机。这需要的远不止一辆推车。

Thorpe和他的团队选择了一辆雪佛兰厢式货车，它足够大，可以容纳所有设备以及五名研究生。正如Thorpe所观察到的：“他们非常有动力编写高质量的软件，因为他们将——如俗话所说——第一个到达事故现场。当你知道自己要搭乘时，你会编写更好的软件。”

这个被称为Navlab 1的项目于1986年正式开始；当时系统每十秒钟可以做一次移动（每小时四分之一英里）。^[38]Thorpe的儿子Leland也在同年出生，巧合的是，Leland最终成为了机器人的完美对手。“当Navlab 1以爬行速度移动时，我儿子也在爬行。当Navlab 1加速时，我儿子在走路和学习跑步。Navlab 1速度稍快一点；我儿子有了三轮车。我以为这将是一场持续16年的竞赛，看谁会首先驾驶宾夕法尼亚收费公路。”^[39]

然而，这场竞争会突然结束，机器获胜。Thorpe的研究生Dean Pomerleau最终使用神经网络组建了一个视觉系统。它超越了小组尝试过的所有其他方法。“所以在1990年，”Thorpe说，“他准备好出去驾驶宾夕法尼亚收费公路了。”

他们称这个系统为ALVINN——神经网络中的自主陆地车辆——它通过模仿学习。^[40] “你会驾驶几分钟，” Thorpe说，“它会学习：如果道路看起来像这样，你就这样转动方向盘，如果看起来像那样，你就那样转动方向盘。所以如果你在你正在驾驶的道路上训练它，它非常擅长输出方向盘角度。”

在一个阳光明媚、路上很少有其他车辆的周日清晨，Pomerleau带着ALVINN上了州际公路。ALVINN一路从匹兹堡沿着I-79高速公路驾驶到大湖岸边的伊利。“这是一种革命性的，” Thorpe说——不仅仅是这个壮举本身，还有实现它的模型的简单性。ALVINN对动量或牵引力一无所知，无法识别物体或预测自身或其他汽车的未来位置，没有能力将它在摄像头中看到的与它自身在空间中的位置联系起来，也无法模拟其行为的影响。“人们认为如果你想开那么快，” Thorpe说，“你必须有卡尔曼滤波器、道路的回旋曲线模型，以及你车辆动态响应的详细模型。而Dean所有的只是一个简单的神经网络，它学会了：道路看起来像这样，你就这样转向。”

正如Pomerleau向当地新闻记者确认的那样，“我们不告诉它任何东西，除了’像我这样驾驶。学会像我现在驾驶的方式一样驾驶。’” 在那个时代，这需要一台冰箱大小的计算机，运行在5,000瓦的发电机上，其处理能力大约是2016年版Apple Watch的十分之一。而且ALVINN不能控制油门或刹车，这些仍需要手动操作，也不能变道或以任何特定方式对路上的其他车辆做出反应。但它确实有效，并且带着Pomerleau——像他以前驾驶的那样驾驶——安全到达了五大湖区。

关于如何训练机器最自然的想法之一是训练它们模仿我们，这种方法在驾驶领域似乎特别有吸引力。ALVINN的成功暗示了这种方法更广泛的可行性。如果我们想要构建一辆完全自动驾驶的汽车，那么与其简单地让它在城市街道上随机探索不同的驾驶行为并通过纯粹的试错来学习（令人恐惧），我们可能会首先给系统大量真实人类驾驶数据，并训练它模仿人类在方向盘后面的决策。给定某种状况——这个速度、挡风玻璃中的这个图像、后视镜中的这个图像等等——系统可以学会预测人类司机采取什么行动，无论是踩油门、踩刹车、转动方向盘，还是什么都不做。

这种预测方法将驾驶问题转化为与ImageNet图像标记竞赛几乎完全类似的问题。系统不是被展示一张图片并需要将其分类为狗、猫、花等，而是被展示来自前仪表板的图片并将其“分类”为“加速”、“刹车”、“左转”、“右转”等等。我们已经看到深度学习如何允许系统从它见过的图像泛化到它没见过的图像；如果AlexNet能够正确识别它以前从未见过的狗，那么这应该鼓励我们认为汽车能够以某种方式从它经历过的场景泛化到新场景。即使它没有在那种确切的阳光斑驳的光线下、在那种确切的交通流量中见过那条确切的道路，它仍然应该——根据理论——从过去的经验中泛化，并识别在这里该做什么。

这个想法是你永远不必让一辆初出茅庐的自动驾驶汽车在城市街道上自由探索政策。相反，你只需记录来自街道上真实人类驾驶汽车每天的摄像头信息和遥测数据，在一天内就可以捕获无数百万小时的数据，最终拥有一辆完美模仿人类驾驶的汽车。

正如加州大学伯克利分校的Sergey Levine向他的伯克利本科生描述的：“所以思考解决这些顺序决策问题的一种非常自然的方式基本上与我们解决标准计算机视觉问题的方式相同。我们收集一些数据，所以我们让人类驾驶车辆...我们基本上记录来自他们车辆摄像头的观察结果，并记录他们做出的转向命令；这进入我们的数据集，我们的训练数据。然后我们将运行我们最喜欢的监督学习方法——我们将运行，你知道，随机梯度下降——来训练网络...就把这当作标准的监督学习问题。这是一个非常合理的开始。”

然后，他解释说，当系统需要接管控制时，它只是将其预测——“这是我认为人类司机在这种情况下会做的”——转化为行动。

Levine停顿了一秒钟。“有人对这种方法可能出错的一个问题有什么想法吗？”

手举了起来。

学会恢复

模仿随着获得的习惯而进步。在学习跳舞时，学生动作与看到老师动作之间联系的不足，使得最初的步骤难以掌握。理想的动作在开始时不会自然地执行。会做出一些动作...但首次行动明显是错误的；存在明显的不一致，这导致新的尝试，失败后，又做出另一次尝试，直到最后我们看到姿势准确了。

——ALEXANDER BAIN

我会怎么做？我不会陷入那种情况。

——苹果CEO蒂姆·库克，当被问及在面临Facebook CEO马克·扎克伯格所面临的情况时他会怎么做

这是2009年，ALVINN之后二十年，但在同一栋建筑里，卡内基梅隆大学研究生Stéphane Ross正在玩《超级马里奥赛车》——或者更确切地说，是一个名为SuperTuxKart的免费开源衍生游戏，以Linux吉祥物——一只可爱的企鹅Tux为主角。

当Ross玩游戏时，他的计算机正在记录屏幕上的所有图像，以及他操纵杆的每一次抽动。这些数据被用来训练一个相当基础的神经网络，并不比ALVINN使用的复杂多少，以像Ross驾驶的方式驾驶。Ross把手从方向盘上松开，让神经网络驾驶Tux绕赛道行驶。很快，Tux转弯太宽，直接开出了赛道。Ross回到了起点，而这个起点看起来并不乐观。

问题在于，无论展示多少圈演示——他录制了一百万帧游戏画面，相当于大约两小时的反复驾驶同一条赛道——似乎都没有任何作用。他将方向盘交给神经网络，Tux一开始表现很有希望，然后开始摇摆，偏离，冲出赛道。

从根本上说，问题源于学习者看到了专家对问题的执行过程，而专家几乎从不陷入困境。然而，无论学习者多么优秀，他们都会犯错误——无论是明显的还是细微的。但由于学习者从未见过专家陷入困境，他们也从未见过专家如何脱困。实际上，当初学者犯初学者的错误时，他们可能会陷入与他们在观察专家时所见到的任何情况都完全不同的境地。“这意味着，”Sergey Levine说，“你知道，一切都沒有保障了。”

例如，在SuperTuxKart中，Ross玩这个游戏足够熟练，以至于他输入给程序的所有数据实际上都在向程序展示如何在赛道中央继续直线行驶。但一旦Tux在网络控制下哪怕稍微偏离中心或稍微歪斜，它就迷失了方向。屏幕看起来与它见过Ross做的任何事情都系统性地不同。所需的反应与正常的、受控的、全速驾驶大不相同，但它从未见过其他任何情况。无论Ross继续多么熟练地玩游戏，一圈又一圈，一小时又一小时，都无法解决这个问题。

这个问题被模仿学习研究者称为“级联错误”，它是模仿学习的基本问题之一。正如Dean Pomerleau在ALVINN工作期间所写：“由于人员在训练期间将车辆引导在道路中央，网络永远不会遇到必须从偏差错误中恢复的情况。”⁴⁸如何教授模仿学习者恢复一直是一个长期存在的问题。

如果Pomerleau要在前往伊利湖的旅行中相信这个系统，那么他需要给它的不仅仅是对自己正确转向的被动观察。仅凭这一点就意味着系统只有在永远不犯错误的情况下才可靠。这要求太高了，在州际公路上以每小时55英里的速度行驶两小时。

“网络不能仅仅展示准确驾驶的示例，” Pomerleau写道，“还要展示一旦犯错如何恢复（即返回道路中央）。”⁴⁹但怎么做呢？一个想法是让Pomerleau自己在训练驾驶过程中四处摆动，向ALVINN演示如何从稍微偏离车道中心或稍微指向错误方向的情况下恢复。当然，这需要以某种方式从训练数据中删除摆动的开始部分，以免ALVINN学会模仿摆动本身！他意识到的第二个问题是，为了正确训练网络，需要他尽可能经常地、在尽可能多样的情况下摆动。“这既费时，”他总结道，“也很危险。”

Pomerleau想出了一个不同的主意。他要伪造它。

ALVINN处理的图像很小且粗糙——只是 30×32 像素的黑白图像，显示车辆正前方的梯形沥青区域。（实际上，它的视野如此狭窄和近视，以至于当它进入十字路口时会完全迷失方向，漂泊在一片巨大的路面海洋中。）Pomerleau拍摄了从ALVINN摄像头录制的真实图像，然后简单地对它们进行修改，将道路稍微倾斜到一边或另一边。然后将这些图像与旨在将汽车轻柔地推回车道中央和直行方向的转向命令一起投入训练数据中。这有点像hack——而且只有在路面完全平坦、没有凹陷或山丘时才看起来正确——但是，至少在I-79上，它奏效了。

具有讽刺意味的是，过去十年中强大深度学习技术的爆发使得“伪造”方法变得越来越不可行，因为现代相机传感器每秒捕获太多图像，分辨率太高，视野太宽，难以用这种方式轻易操作。如果假图像在某种程度上系统性地不同于汽车开始偏离时将看到的真实事物，那么你就有大麻烦了。毕竟，本质上把你的生命押在你的Photoshop技能上是一种信仰的飞跃——而现代神经网络实际上变得越来越难以欺骗。

二十年后，这个恢复问题在实际和理论层面上仍未解决。“当你通过观察某人来学习时，”Stéphane Ross告诉我，“你看到某种分布的示例，如果你开始在世界中采取自己的行动，这些示例不一定与你将要看到的匹配。”Ross认为这里有一些根深蒂固的东西：“因为所有机器学习都依赖于这样的假设：你的训练分布和测试分布是相同的。”但Ross和他的导师、CMU机器人专家Drew Bagnell认为他们也许能够破解它。“这让我非常感兴趣，”Ross说，“因为这感觉像是一个真正基础性的问题需要解决。”⁵⁰

Ross和Bagnell进行了理论分析，试图从数学角度理解这个问题，同时在*SuperTuxKart*的世界中验证他们的直觉。在普通的监督学习问题中，比如ImageNet，系统一旦训练完成，对于它看到的每张图片都有一定的出错概率。给它看十倍多的图片，平均而言它会犯十倍多的错误——从这个意义上说，错误与任务规模呈线性增长关系。他们发现，模仿学习要糟糕得多。因为一个错误就可能导致系统看到它之前从未准备过的情况，一旦犯了第一个错误，一切都失控了。错误随任务规模的平方增长。运行十倍长的时间会产生一百倍多的错误。理论分析结果很严峻，但它留下了一个诱人的可能性：是否有方法能回到仅仅线性错误的安全世界？让汽车在行驶十倍距离时仅仅增加十倍的碰撞可能性？“我们真的在寻找圣杯，”他说。

事实证明，他们找到了。关键是交互。学习者不仅需要在开始时观察专家，还需要在必要时回到老师那里，有效地来说：“嘿，我尝试了你教我的方法，但是不断发生坏事。如果你遇到这种混乱情况会怎么做？”

Ross想出了两种方法在*SuperTuxKart*赛道上实现这种交互。一种方法是手握操纵杆观看网络（最初灾难性的）圈数。当Tux在赛道上疾驰时，Ross会像他在玩游戏一样移动操纵杆。第二种方法是让Ross和网络随机交替控制汽车，同时两者都试图转向。正如他解释的：“这就像你基本上还是正常玩游戏，但在一些随机步骤中，不听人类控制，而是执行学习到的控制。这种情况会随时间慢慢衰减——就像你越来越失去控制权。但总有机会选择你的控制。”这有点别扭，有点不自然，但有效。“你仍然绝对试图像你完全控制时那样玩游戏，”Ross说。“但它不一定总是选择你的控制来执行。它开始偏离，然后你试图纠正...”他笑着说。随着时间推移，汽车越来越少地响应你自己的转向命令——但网络在做你本来会做的事情方面变得越来越好。有时候你不确定自己是否在驾驶。

令人惊讶的是，这两种交互形式不仅有效——无论是在白板上还是在*SuperTuxKart*赛道上——而且它们需要的反馈极少。仅使用静态演示，学习者在一百万帧专家数据后仍然像几千帧后一样经常撞车——经过数小时的指导后仍然

像仅仅几分钟后一样无望。然而使用这种交互方法——Ross将其命名为“数据集聚合”(Dataset Aggregation)，或DAgger——程序在第三圈绕赛道时就几乎完美地驾驶了。“一旦我们有了这个，”Ross说，“我就想，哇，这真的很棒。它比默认方法好几个数量级。”

一毕业获得博士学位，Stéphane Ross就从*SuperTuxKart*的虚拟路面转向加利福尼亚州山景城的真实郊区街道，他目前是自动驾驶汽车公司Waymo的行为预测负责人，设计模型来预测道路上其他司机、骑自行车者和行人的行为和反应。“我们需要的可靠性水平比我们在学术界做的任何事情都要高几个数量级。这就是真正的挑战所在。比如，你如何确保你的模型始终有效——不仅仅是95%或99%的时间；那甚至都不够好。”这是一个很高的要求，但也是一个令人满意的项目。“特别是这个项目，如果成功的话，可能是你能对世界产生最大影响的项目之一——为了世界的利益。仅此一点就足以激励在这个领域工作，并希望有一天产生那种影响。”

虽然DAgger涉及的那种交互反馈在理论上是黄金标准，但在实践中我们不必与汽车争夺方向盘的控制权来确保它们学会保持在车道中央。还有几种更简单的方法在实践中效果很好，可以构建能够从小错误中恢复的现实世界系统。

2015年，一群瑞士机器人专家在试图构建一架能够自主沿着高山徒步径飞行而不在森林中迷路的无人机时，采用了一种巧妙的方法来克服这个问题。正如他们所说，早期的工作试图“明确定义哪些视觉特征能够刻画小径”，而他们完全绕过了图像中哪些部分包含小径，或者小径到底是什么样子这些问题，转而训练一个系统直接从图像映射到电机输出。它会接收一张 752×480 像素的泥土和树木图像，然后输出“左转”、“右转”或“直行”。在一个如今看来非常熟悉的故事情节中，多年来为“显著性”或“对比度”精心研究手工制作视觉特征的工作，以及关于如何区分泥土和树皮等的巧妙思考，都被彻底抛弃了，取而代之的是由随机梯度下降训练的卷积神经网络。所有手工定制的工作瞬间变得过时了。

团队训练他们的系统模仿人类徒步者走过的路径。不过，独特而巧妙的是他们为了让系统能够从错误中恢复所做的工作。他们在徒步者头上绑了不是一台而是三台GoPro相机：一台朝正前方，另外两台分别朝左和朝右。然后他们告诉徒步者像平常一样行走，但要注意不要转头。这样他们就能够生成大量的小径图像数据集，并为中央相机的画面添加“看到这样的情况时向前走”的标注，为左侧相机添加“看到这样的情况时右转”的标注，为右侧相机添加“看到这样的情况时左转”的标注。他们在数据集上训练了一个神经网络，将其安装在四旋翼无人机上，然后在瑞士阿尔卑斯山放飞。它似乎毫不费力地穿越森林并沿着小径飞行。再次，关键洞察是不仅需要展示人类专家做了什么，还要提供一些护栏，以数据的形式为稍微偏离轨道的学习者指回正轨。⁵³

2016年，Nvidia位于新泽西州霍姆德尔的深度学习研究小组的一个项目将同样巧妙的技巧应用到了新泽西州蒙茅斯县的街道上。Nvidia在一辆汽车上安装了三台相机，一台朝前，另外两台大约朝中心左右各30度。这产生了数小时的画面，展示了如果汽车稍微指向错误方向会是什么样子。然后团队将这些数据输入他们的系统，正确的预测是“做真实人类司机所做的事，再加上小幅修正回到中心”。仅用72小时的训练数据，该系统就足够安全，能够在不同天气条件下在蒙茅斯县蜿蜒的乡村道路和多车道高速公路上运行而没有重大事故。在团队发布的一段视频中，我们看到他们的林肯MKZ从Nvidia深度学习研究大楼的停车场驶出，驶上花园州高速公路。“它还在自动驾驶吗？！”跟车中的一名员工问道。“从这里看起来相当不错，”他说——然后澄清说它的表现至少比高速公路上其他由新泽西人驾驶的汽车要好。⁵⁴

这里有两个值得简要注意的讽刺之处。正是在这栋研究大楼里，在1980年代后期——当时它属于AT&T贝尔实验室——Yann LeCun发明了卷积神经网络，通过反向传播进行训练，这正是驱动今天自动驾驶汽车的技术。⁵⁵碰巧的是，我自己就是在新泽西州蒙茅斯县的道路上学会开车的，在去越野跑练习的路上经常经过那栋大楼。我希望我能说，在看了十七年的人类驾驶后，我在驾驶同一条路时的安全性和可信度能和那个卷积网络在72小时后一样好。

悬崖边缘：可能主义对现实主义

一个人必须执行自己能够掌控和维持的较低行为：而不是自己会搞砸的较高行为……我们不能僭越那些精神视野比我们更高或不同的人的行为。

——艾瑞丝·默多克⁵⁶

如果你是我，你会怎么做？她说。

如果我是你-你，还是如果我是你-我？

如果你是我-我。

如果我是你-你，他说，我会完全按照
你正在做的去做。

——罗伯特·哈斯⁵⁷

撇开从小错误中恢复的问题不谈，模仿作为学习策略的第二个问题是，有时你根本无法做专家能做的事。那么模仿只意味着开始一些你无法完成的事情。在这种情况下，你可能根本不应该试图像他们那样行事。

现实生活和流行文化中都散布着新手试图简单模仿专家，往往导致灾难性结果的例子。

正如国际象棋大师加里·卡斯帕罗夫解释的：“棋手，甚至是俱乐部业余选手，都会花费数小时研究和记忆他们偏爱的开局路线。这种知识是无价的，但它也可能是一个陷阱……死记硬背，无论多么惊人，如果没有理解就是无用的。在某个时刻，他会到达记忆的终点，在一个他并不真正理解的局面中没有预制的解决方案。”

卡斯帕罗夫回忆起他指导一位十二岁棋手的情景，他们正在分析这位学生比赛中的开局走法。卡斯帕罗夫问他为什么在一个复杂的开局序列中走出一步特别犀利且危险的棋。“这是瓦列霍下的！”学生回答道。“当然，我也知道这位西班牙特级大师在最近的一场比赛中使用过这一招，”卡斯帕罗夫说，“但我也知道，如果这个年轻人不理解这步棋背后的动机，他就已经陷入麻烦了。”⁵⁸

这个想法既直观又在某种程度上自相矛盾：效仿一个“更优秀”的人的做法有时可能是严重的错误。这是一个令人惊讶的复杂故事，与伦理学、经济学和机器学习都有着深层联系。

1976年，一个特殊的问题突然成为伦理哲学的焦点：你自己的未来行为在多大程度上影响或应该影响现在做正确事情的问题？

哲学家霍莉·史密斯当时在密歇根大学，专注于梳理作为功利主义者的微妙含义。她注意到一些奇怪的现象。“这个问题很自然地出现了，如果你是功利主义者，‘如果我现在做A，这会产生最好的可能结果吗？’很显然，”她

说，“这将取决于我接下来做什么。”⁵⁹ 需要考虑自己未来行为的需要意味着你也需要考虑自己未来的错误。因此史密斯开始撰写她称之为“道德缺陷”的内容。⁶⁰

她考虑的思想实验后来被称为“拖延症教授”。⁶¹ 前提很简单：拖延症教授既是一位教授，也是——你猜对了——一个顽固的拖延症患者。他被要求阅读学生的论文并提供反馈，他在这方面具有独特的资格。但如果他同意的话，几乎可以肯定会发生的情况是，他会浪费时间，永远不会给学生反馈。这比简单地拒绝更糟糕，在拒绝的情况下，学生可以向其他人寻求（质量稍微低一些的）反馈。

他应该接受吗？

在这里，两种不同的道德思想流派分道扬镳：“可能主义”——认为在任何情况下都应该做最好的可能事情的观点——与“现实主义”——认为应该在当下做最好的事情，考虑到后来实际会发生什么（无论是因为你自己的行为还是其他原因）的观点。⁶²

可能主义认为拖延症教授能做的最好的可能事情是接受评审并且按时完成。这始于接受它，所以他应该接受。

现实主义采取更实用的观点。按照这种观点，接受评审不可避免地导致糟糕的结果：根本没有评审。拒绝评审意味着相对更好的结果：由稍微不太合格的评审者进行评审。教授应该做实际导致最佳结果的事情；因此他应该说不。

史密斯得出结论：“有时必须选择较低而非较高的行为。”她阐述道：“如果同一个行为既让行为者处于可以做大事的位置，又让他处于可以做灾难性事情的位置，而他会选择后者而不是前者，那么规定这样的行为似乎没什么意义。”

同时，史密斯很快阐述了现实主义的缺点。她说，首先，“现实主义为基于你自己未来道德缺陷的不良行为提供了借口。”大约四十年后，理论辩论仍在继续。“我认为许多人会认为这仍然没有解决，”史密斯说。“我认为公平地说，这仍然是一个活跃的讨论。”⁶³

这个讨论不仅仅是理论性的。例如，在二十一世纪的“有效利他主义”运动中，对于某人应该做出多大牺牲以最大限度地帮助他人，意见各不相同。⁶⁴ 普林斯顿哲学家彼得·辛格著名地说，忽视慈善捐赠类似于走过一个有孩子溺水的池塘而不帮助。⁶⁵ 即使对于那些或多或少同意这个论点的人来说，关于实际应该给多少也存在一些争论。也许一个完美的人可以将几乎所有的钱捐给慈善机构，同时保持快乐、乐观和有动力，并激励他人。但即使是“EA”运动的忠实成员，包括辛格本人，也不是这样的完美人士。

朱莉娅·怀斯是有效利他主义社区的领导者，也是有效利他主义中心的社区联络员，她在自己的生活中做出了令人印象深刻的承诺——例如将收入的50%捐给慈善机构——但她强调不追求完美的价值。“允许自己走一半的路，”她说。⁶⁶ 例如，她注意到自己对素食主义的承诺无法容纳她对冰淇淋的深深热爱——所以她觉得自己不能成为素食主义者。对她有效的是接受成为一个素食主义者...但吃冰淇淋的想法。这是她能坚持的。

牛津哲学家威尔·麦卡斯基尔(Will MacAskill)，有效利他主义中心的联合创始人，在这个问题上直言不讳。“我们应该是现实主义者，”他说。“如果你今天一次性捐出所有积蓄——从技术上讲你可以这样做——你可能会变得如此沮丧，以至于将来会完全停止捐赠。而如果你决定捐出收入的10%，这种承诺将足够可持续，让你在未来许多年里继续这样做，从而产生更高的整体影响。”⁶⁷

Singer本人也承认，长远来看，平衡感和适度感可能是最好的。“如果你发现自己在做让你感到痛苦的事情，是时候重新考虑了。你有可能对此变得更积极吗？如果不能，那么考虑到所有因素，这真的是最好的选择吗？”他也指出，“有效利他主义者仍然相对较少，所以他们树立吸引他人过这种生活方式的榜样很重要。”⁶⁸

机器学习有其自己版本的现实主义/可能主义辩论。如我们在[第4章]中讨论的，强化学习的主要算法族之一是学习各种可用行动的”价值”的方法，表现为预期的未来奖励。(这被称为”Q值”，是”质量”的缩写。)例如，棋类游戏agent会学习预测在做出各种走法后获胜的机会，而玩Atari游戏的agent会学习估计每个行动预期能带来的得分。有了这些调优的预测，简单地采取具有最高Q值的行动就变得直截了当了。

然而，这里有一个值得分析的模糊性。Q值应该包含你从采取这个行动可能获得的预期未来奖励？还是你会获得的预期奖励？对于完全完美的agent来说，没有矛盾——但否则规定可能会有很大差异。

这两种价值学习方法被称为”在策略”(on-policy)和”离策略”(off-policy)方法。在策略方法基于agent在采取该行动并继续按照自己的”策略”采取行动后实际期望获得的奖励来学习每个行动的价值。另一方面，离策略agent将基于可能跟随该行动的最佳可能行动序列来学习每个行动的价值。⁶⁹

在他们关于强化学习的开创性教科书中，理查德·萨顿(Richard Sutton)和安德鲁·巴托(Andrew Barto)谈到了离策略(“可能主义”)agent如何可能陷入困境，正是因为总是试图做”可能的最好的事情”。他们说，想象一辆汽车需要自动驾驶从海边悬崖边缘的一个地点到另一个地点。最短最高效的路径就是沿着悬崖边缘行驶。确实，假如汽车足够稳定，那这就是最佳路线。但对于在方向盘后略显摇摆或不稳的自动驾驶汽车来说，这是在玩火。更好的选择可能是走一条更迂回的内陆路径，不需要完美驾驶就能成功。这就是现实主义——使用在策略方法训练的汽车确实会选择更安全、更确定的路线，而不是奖励稍高但风险高得多的路线。⁷⁰

模仿自己的英雄或导师——无论模仿者是人类还是机器——都带有可能主义、离策略评估的一些危险。⁷¹在国际象棋语境中，学习在具备大师级处理后果能力的情况下最佳走法，可能只会让学生眼高手低。在这种情况下，专心观看专家下棋可能根本没有帮助——或者更糟。知道在特定的国际象棋位置中，比如牺牲我的王后可以在十步内将军是一回事。但如果我找不到将军的方法，我就会白白牺牲王后，结果几乎肯定会输掉比赛。

自二十世纪中期以来，经济学家们一直在讨论”次优理论”(theory of the second best)，该理论实际上论证，在遵循一系列数学假设的理论版本经济中知道正确的做法，在即使略微偏离这些假设的经济中可能几乎没有意义。要遵循的”次优”政策或要采取的行动可能与最优政策几乎没有相似之处。⁷²OpenAI研究科学家阿曼达·阿斯克尔(Amanda Askell)致力于伦理和政策工作，她指出同样的论证思路可能同样适用于她的领域。“我认为在伦理学中也可以说类似的话，”她说。“即使理想agent完美地遵循道德理论X，非理想agent也使用了相当不同的决策程序。”⁷³

像这样的案例应该让任何想要模仿的人或榜样都暂停思考。模仿在某种程度上本质上是可能主义的，容易眼高手低。当我们的孩子模仿开车、切菜或进行兽医治疗时可能很可爱——但如果我们看到他们真的伸手去拿钥匙或刀子(或者甚至是猫)，我们就会干预。我们真正想看到的行为可能与模仿毫无相似之处：仅次于胜任驾驶的最佳选择是坐到副驾驶座上；仅次于切丝香草的最佳选择是摆餐具；仅次于给断爪打夹板的最佳选择可能只是叫爸爸妈妈。

对于机器模仿者，我们也应该牢记次优理论。如果它们要向我们学习，我们必须注意不要让它们无意中学会启动那些一旦开始就无法处理的行为。一旦它们足够专业，这个问题可能就会变得无关紧要。但在那之前，模仿可能是一个诅咒，正如用户界面设计师布鲁斯·巴伦廷(Bruce Balentine)所说——“做一台好机器比做一个坏人要好。”⁷⁴

放大：自我模仿与超越

我最喜欢的过去的棋手可能是……我自己，大约三四年前的我。

—世界象棋冠军马格努斯·卡尔森(MAGNUS CARLSEN)⁷⁵

模仿的第三个根本挑战是，如果一个人的主要目标是模仿老师，那么就很难超越他们。

这正是机器学习领域最早的研究者之一——实际上是创造这个术语的人——IBM的阿瑟·塞缪尔(Arthur Samuel)所考虑的问题。1959年，正如我们之前简要讨论过的，他开发了一个用于下跳棋的机器学习系统。“我向它输入了一些我知道与游戏有关的原则，”他说，“尽管我当时不知道，现在也不知道，这些原则的确切意义是什么。”这个列表包括你有多少个棋子、你有多少个王、从你的位置可以走多少步等等。⁷⁶

尽管只使用了塞缪尔给它的战略考虑因素，这个程序最终能够击败塞缪尔本人。它准确无误地向前看几步的能力，结合对这些各种因素相对重要性的试错微调，产生了一个超越其自身老师的系统。这对于那个时代来说是一个卓越的成就，正如我们讨论过的，它单枪匹马地让IBM的股价一夜上涨，塞缪尔对此理所当然地感到自豪。但他仍然敏锐地意识到他的项目遇到了天花板。“计算机现在按照我的跳棋原则工作，并且很好地将这些原则重新组合以获得最大优势，”他感叹道，“但让它下更好跳棋的唯一方法是给它一套更好的原则。但怎么做呢？……目前，我是世界上唯一能教机器下得更好的人，而它已经远远超出了我的水平。”

塞缪尔认为，更有原则的前进道路是让计算机本身以某种方式自己生成战略考虑因素。“如果计算机能够生成自己的术语就好了！但我看不到在不久的将来有什么希望，”他说。⁷⁷“不幸的是，还没有设计出令人满意的方案来做到这一点。”⁷⁸

到二十世纪末，计算机游戏的基本技术变化惊人地少——它们的基本局限性也是如此。机器快了数百万倍。强化学习已经成为一个独立的领域。但机器似乎在对我们的顽固依赖方面并没有发生太大变化。

到1990年代，开发国际象棋超级计算机深蓝(Deep Blue)的IBM团队创建了一个价值函数，很像塞缪尔几十年前为跳棋制作的那个。与人类特级大师合作，他们试图枚举和阐述决定位置强度的所有因素：比如双方的棋子数量、机动性和空间、王的安全、兵的结构等等。然而，他们不是像塞缪尔那样使用三十八个这样的考虑因素，而是使用了八千个。⁷⁹“这个国际象棋评估函数，”团队负责人许峰雄(Feng-hsiung Hsu)说，“可能比计算机象棋文献中描述过的任何东西都更复杂。”⁸⁰当然，关键问题是如何以某种方式权衡和结合那些令人困惑的数千个考虑因素，形成对棋盘上位置质量的单一判断。确切地多少额外的兵值多少中心控制，或多少王的安全？获得正确的平衡将是至关重要的。

那么这数千个考虑因素究竟是如何达到平衡的呢？通过模仿。

深蓝团队可以访问一个包含七十万场特级大师比赛的数据库。他们向计算机展示这些真实比赛中的一个又一个位置，并询问它会下什么棋。模仿人类的走法成为他们微调其价值函数时的目标。比如说，如果增加深蓝对拥有双象的价值分配使它稍微更有可能下出与人类特级大师相同的棋步，那么深蓝就会增加对双象的价值。

这种模仿人类的位置考虑因素组合，这些因素本身来自人类专家，与计算机准确无误的计算、惊人的速度和蛮力相结合。机器能够每秒搜索数亿个未来棋盘位置，这一点，结合其类似人类的评估，足以在1997年的著名比赛中击败人

类象棋世界冠军加里·卡斯帕罗夫(Garry Kasparov)。“加里准备与一台计算机对弈，”深蓝的项目经理谭志钧(C. J. Tan)说。“但我们将它编程为像特级大师一样下棋。”⁸¹

从哲学角度来看，研究界内的一些人开始思考，程序是否最终会因为对人类榜样的持续依赖而受到阻碍。在计算机跳棋领域，阿尔伯塔大学的乔纳森·谢弗(Jonathan Schaeffer)在1990年代早期开发了一个如此出色的程序，当它选择的走法与人类大师的走法不同时，往往它自己的想法更好。“当然，我们可以继续‘改进’评估函数，使其始终下出人类的走法，”他写道。但这样做是否是好事并不明显。“一方面，调整程序以更传统的方式下棋可能会削弱其让人类对手感到意外的能力。另一方面，一旦程序达到了最优秀人类棋手的水平，模仿的方法论是否仍然有用并不清楚。”我们发现很难再取得进一步的进展，“谢弗承认。⁸²他的项目本质上陷入了困境。这个领域面临着一个问题。正如2001年的《学会下棋的机器》一书在反思深蓝成功时所说：“未来研究的一个重要方向是确定更好地模仿人类专家走法在多大程度上对应着真正更强的棋力。”⁸³

十五年后，DeepMind的AlphaGo系统终于实现了阿瑟·塞缪尔(Arthur Samuel)的愿景——一个能够从零开始创造自己位置考量的系统。它没有被给予一大堆数千个手工制作的特征来考虑，而是使用深度神经网络自动识别使特定走法具有吸引力的模式和关系，就像AlexNet识别使狗成为狗、汽车成为汽车的视觉纹理和形状一样。该系统像深蓝一样进行训练：通过学习预测专家级人类围棋棋手在一个巨大的、包含3000万步棋的数据库中所走的棋步。⁸⁴它能够达到对人类专家走法的最先进预测——准确率为57%，超越了此前44%的最先进结果。2015年10月，AlphaGo成为第一个击败人类职业围棋棋手的计算机程序(在这种情况下，是三届欧洲冠军樊麾)。仅仅七个月后，2016年3月，它击败了18次国际冠军得主、有史以来最强棋手之一的李世石。

再一次，超越了人类棋力的计算机，讽刺的是，本质上仍然是一个模仿者。⁸⁵它不是在学习下最好的棋，而是在学习下人类的棋。

深蓝和AlphaGo的成功都只有因为有了大量人类示例数据库供机器学习才成为可能。这些机器学习的旗舰成功之所以在全世界引起如此轰动，是因为这些游戏的全球流行度。而正是这些游戏的流行度使得这些胜利成为可能。我们下过的每一步棋都可能并且将会被用来对付我们。如果玩一个更冷门或不受欢迎的游戏，计算机就不会如此令人印象深刻——因为它们没有足够的例子可以学习。因此，流行度起到了双重作用。它使成就变得重要。但它也使成就变得可能。

然而，AlphaGo刚刚登上围棋游戏的巅峰，在2017年就被一个甚至更强的程序AlphaGo Zero迅速推翻了。⁸⁶原始AlphaGo和AlphaGo Zero之间最大的区别在于后者被输入了多少人类数据来模仿：零。从完全随机的初始化开始，白板状态，它只是通过与自己对弈来学习，一遍又一遍又一遍又一遍。令人难以置信的是，仅仅经过36小时的自我对弈，它就达到了击败李世石的原始AlphaGo的水平。72小时后，DeepMind团队安排了两者之间的比赛，使用完全相同的两小时时间控制和击败李世石的原始AlphaGo系统的确切版本。AlphaGo Zero消耗的功率只有原始系统的十分之一，而且72小时前从未下过一盘棋，却以100比0赢得了这个百局系列赛。

正如DeepMind研究团队在他们随附的《自然》论文中写道：“人类从数千年来进行的数百万局围棋中积累了围棋知识，集体提炼成模式、谚语和书籍。”⁸⁷AlphaGo Zero在72小时内就发现了这一切，甚至更多。

但在幕后发生的事情非常有趣，也很有启发性。该系统没有被展示过一盘人类游戏来学习。但它仍然在通过模仿来学习。它在学习模仿……自己。

自我模仿的工作原理如下：在围棋和国际象棋等游戏中的专家级人类对弈是一个“快思考和慢思考”的问题。⁸⁸有一种有意识的、深思熟虑的推理，它观察走法序列并说：“好的，如果我走这里，然后他们走那里，但然后我走这里，我就赢了。”在AlphaGo Zero中，通过逐步思考“如果这样，那么那样”来进行明确的“慢”推理，是由一种叫

做蒙特卡洛树搜索(简称MCTS)的算法完成的。⁸⁹这种缓慢、明确的推理与快速、难以言喻的直觉密切结合，体现在两个不同但相关的方面。

第一种”快速”思维是，在任何此类显式推理之前并独立于此类推理，我们对特定位置有多好有一种直觉感知。这就是我们一直在讨论的”价值函数”或”评估函数”；在AlphaGo Zero中，这来自一个名为”价值网络”(value network)的神经网络，它输出一个0到100的百分比，表示AlphaGo Zero认为从该位置获胜的可能性。

第二种隐式的”快速”推理是，当我们观察棋盘时，有一些我们考虑下的棋步——一些棋步只是”自然而然地出现”，而许多其他棋步根本不会。我们将缓慢、深思熟虑的”如果这样，那么那样”的推理部署到我们的直觉首先识别为合理或有希望的路径上。这就是AlphaGo Zero变得有趣的地方。这些候选棋步来自一个名为”策略网络”(policy network)的神经网络，它将当前棋盘位置作为输入，并为每个可能的棋步分配一个0到100的百分比。这个数字代表什么？系统在押注它最终会决定下的那步棋。

这是一个相当奇怪且几乎矛盾的想法，值得进一步阐述。策略网络代表了AlphaGo Zero对每个可能棋步的猜测，即在进行显式MCTS搜索以从该位置向前看之后，它选择该棋步的可能性有多大。略显超现实的方面是，系统使用这些概率来将缓慢的MCTS搜索集中在它认为最有可能的一系列棋步上。⁹⁰“AlphaGo Zero成为了自己的老师，”DeepMind的David Silver解释道。“它改进其神经网络来预测AlphaGo Zero自己下的棋步。”⁹¹

鉴于系统使用这些预测来指导其预测结果的搜索，这听起来可能像是自我实现预言的配方。实际上，每个系统——快速和缓慢——都在磨练彼此。随着策略网络的快速预测改进，缓慢的MCTS算法使用它们更狭窄、更明智地搜索可能的未来棋局。由于这种更精细的搜索，AlphaGo Zero成为了更强的棋手。然后策略网络调整以预测这些新的、稍微更强的棋步——这反过来又允许系统更明智地使用其缓慢推理。这是一个良性循环。

这个过程在技术社区中被称为”放大”(amplification)，但它同样可以被称为超越之类的东西。AlphaGo Zero只学会了模仿自己。它使用其预测来做出更好的决策，然后学会预测这些更好的决策。它开始时进行随机预测和随机棋步。七十二小时后，它成为了世界上见过的最强围棋选手。

放大价值观

你应该考虑到模仿是崇拜中最可接受的部分，神灵更愿意人类相似于而非奉承他们。

—马库斯·奥勒留⁹²

对于越来越多关注长期未来的哲学家和计算机科学家来说，灵活智能和灵活能力系统的前景，我们必须向其中注入极其复杂的行为和价值观，不仅引发了技术问题，还引发了更深层次的问题。

这里有两个主要挑战。第一个是我们想要的东西很难直接陈述——即使用词语，更不用说用更数值化的形式。正如人类未来研究所的Nick Bostrom指出的，“写下我们关心的一切清单似乎完全不可能。”⁹³在这种情况下，我们已经看到模仿学习如何在有效地不可能明确传授每一个规则、考虑和重点程度来说明什么使某人成为专家司机或专家围棋选手的领域中取得成功。简单地说，“实际上”观察和学习往往令人印象深刻地成功。很可能当自主系统变得更强、更通用——到了我们寻求传授某种意义，不仅是如何很好地驾驶和很好地下棋，而是如何很好地生活，作为个人和社会——我们仍然可以转向类似的东西。

第二个更深层次的挑战是，传统的基于奖励的强化学习(reinforcement learning)和模仿学习技术都需要人类作为最终权威的来源。正如我们所见，模仿学习系统可以超越它们的老师——但只有当老师的不完美示范在很大程度上相互抵消的方式上是不完美的，或者只有当无法演示他们想要的专家至少能识别它时。

当我们展望更遥远的未来，展望在更微妙和复杂的现实世界环境中行动的更强大系统时，这些方面中的每一个都提出了挑战。

例如，一些人担心人类不是道德权威的特别好的来源。“我们已经谈论了很多关于将人类价值观注入机器的问题，”谷歌的Blaise Agüera y Arcas说。“我实际上不认为这是主要问题。我认为问题在于人类价值观就其现状而言是不够的。它们不够好。”⁹⁴

机器智能研究所的联合创始人Eliezer Yudkowsky在2004年写了一份有影响力的手稿，他在其中论证，我们不应该简单地让机器模仿和维护我们不完美体现的规范，而是应该向机器灌输他所称的“连贯外推意志”。他写道：“用诗意的话来说，我们的连贯外推意志是我们的愿望，如果我们知道得更多，思考得更快，更像我们希望成为的人。”⁹⁵

在那些有相对明确的外部成功指标的领域——比如跳棋、围棋或蒙特祖玛的复仇——机器可以简单地将模仿作为更传统的强化学习技术的起点，通过试错来磨练初始的模仿行为，并有可能超越它们自己的老师。

然而，在道德领域，如何扩展模仿就不那么清楚了，因为不存在这样的外部指标。⁹⁶

更重要的是，如果我们试图教导的系统有朝一日可能比我们更聪明，它们可能会采取我们甚至很难评估的行动。如果一个未来的系统提出，比如说，临床试验法规的改革，我们甚至可能不一定有能力评估——经过深思熟虑后，更不用说在紧密的迭代反馈循环中——它是否确实符合我们的伦理感或规范。那么，一旦系统的行为超出了我们的直接认知范围，我们如何继续按照我们自己的形象来训练系统呢？

很少有人像OpenAI的Paul Christiano那样深入思考过这一系列问题。“我非常有兴趣真正询问解决方案在扩展时会是什么样子，”他说。“我们的实际游戏计划是什么？这里的实际终局是什么？这是一个相对很少有人感兴趣的问题，所以很少有人在研究它。”⁹⁷

Christiano意识到，早在2012年，在持续至今的研究中，我们可能——即使在这些最困难的情况下——能够逐步前进。⁹⁸例如，我们看到AlphaZero对要考虑的走法有即时的、快速思考的判断，但使用慢速思考的蒙特卡洛树搜索来梳理数百万个未来棋盘位置，以确认或纠正这些直觉。这种慢速思考的结果然后被用来锐化和改善其快速直觉：它学会预测自己深思熟虑的结果。⁹⁹

Christiano相信，也许这个完全相同的模式——他称之为“迭代蒸馏和放大”——可以用来开发具有复杂判断力的系统，超越但又与我们自己的判断力保持一致。

例如，想象我们正在试图为一个大城市规划一个新的地铁系统。与Atari或围棋不同，我们不能每秒评估数千种场景——实际上，单次评估可能需要数月时间。与Atari或围棋不同，没有外部客观标准可以诉诸——一个“好的”地铁系统就是人们认为好的系统。

我们可以训练一个机器学习系统达到一定的胜任水平——比如通过正常的模仿学习——然后，从那时开始，我们可以使用它来帮助评估计划，就像一个高级城市规划师带着几个初级城市规划师的员工一样。我们可能会要求我们系统的一个副本给我们一个预期等待时间的评估。我们可能会要求另一个给我们一个预算估计。第三个我们可能会要求一份关于可达性的报告。我们作为“老板”，会做出最终决定——“放大”我们机器下属的工作。这些下属反过来会从我们的最终决定中“蒸馏”出他们能学到的任何教训，结果成为稍微更好的城市规划师：总体上比我们自己工作得更快，但按照我们自己的形象建模。然后我们迭代，将下一个项目委托给这个新的、稍微改进的团队版本，良性循环继续。

最终，Christiano相信，我们会发现我们的团队总体上是我们希望成为的城市规划师——如果我们“知道得更多，思考得更快，更像我们希望成为的规划师”，我们就能成为的规划师。

还有工作要做。Christiano希望找到进行放大和蒸馏的方法，这些方法将可证明地保持与人类用户的一致性。目前，这是否可能仍然是一个开放的问题——和一个希望。小规模的初步实验正在进行中。“如果我们能实现这个希望，”Christiano和他的OpenAI合作者写道，“这将是扩大机器学习范围和解决对AI长期影响担忧的重要一步。”¹⁰⁰

在讨论他在放大方面的工作时，我问Christiano——他已经成为对齐研究社区的领军人物——是否将自己视为其他有兴趣走类似道路的人的某种榜样。他的答案让我感到惊讶。

“希望这不是人们必须走的道路，”他说。¹⁰¹

Christiano详细说明，他可能是最后一批在能够直接从事AI安全工作之前必须过某种双重生活的对齐研究人员之一：从事更传统的问题以获得学术资格，同时想办法做他认为真正重要的工作。“我必须独自去思考这些问题很长时间，”他说。“在学术社区的背景下做学术工作更容易。”仅仅几年后，这个社区就存在了。¹⁰²“所以希望大多数人会更多地处于那种情况，”他说。“有很多人在思考这些问题；他们实际上可以找到工作...就去做他们[关心的事情]。”

也许，作为开拓者就是这样：重要的不是别人会完全模仿你的榜样，或者直接跟随你的脚步，而是——由于你的努力——他们不必这样做。

推理

密歇根大学心理学家Felix Warneken走过房间，抱着一堆高高的杂志，走向一个关闭的木柜门。他撞到柜子前面，惊叫一声“哦！”“，然后退了回来。盯着柜子看了一会儿，他发出了一声若有所思的“嗯”，然后拖着脚步向前，再次用杂志撞击柜门。他再次败退，可怜地说：“嗯嗯...”就好像他不知道自己哪里出了错。

从房间的角落里，一个蹒跚学步的孩子前来救援。孩子有些不稳地走向柜子，一一推开柜门，然后抬头看着Warneken，带着询问的表情，然后退了回来。Warneken发出感激的声音，把他的那堆杂志放在架子上。¹

Warneken与他的合作者杜克大学的Michael Tomasello一起，在2006年首次系统地证明，18个月大的人类婴儿就能可靠地识别面临问题的同伴，识别人类的目标和阻碍的障碍，如果可以的话会自发地提供帮助——即使没有被要求帮助，即使成年人甚至没有与他们进行眼神接触，即使他们不期望（也不会收到）任何回报。²

这是一种非常复杂的能力，几乎是人类独有的。我们最近的遗传祖先——黑猩猩——偶尔会自发地提供帮助——但只有在它们的注意力被引向当前情况时，只有当有人明显伸手去够超出其范围的物体时（而不是在更复杂的情况下，比如柜子），³只有当需要帮助的是人类而不是同伴黑猩猩时（它们彼此之间竞争非常激烈），只有当所需物体不是食物时，并且只有在拿着被寻求的物体几秒钟后，好像在决定是否真的要把它交出来。⁴

Warneken和Tomasello所展示的是，这种帮助行为“在进化上极其罕见”，在人类中比在我们最亲近的亲戚中更为明显，甚至在语言出现之前就相当丰富地出现了。正如Tomasello所说，“人类认知与其他物种认知的关键区别在于与他人参与具有共同目标和意图的协作活动的能力。”⁵

“儿童被描述为最初是自私的——只关心自己的需要——需要社会以某种方式重新编程他们，使他们变得利他，” Warneken说。⁶“然而，我们的研究表明，第二年生活中的婴儿已经通过帮助他人解决问题、共同工作以及与他人分享资源而具有合作性。”⁷

这不仅需要帮助的动机，还需要一个极其复杂的认知过程：推断他人的目标，通常仅从一小部分行为中推断。

“人类是世界上读心术的专家，” Tomasello说。也许这种专业技能最令人印象深刻的部分是我们推断他人信念的能力，但基础是推断他们的意图。实际上，直到大约四岁，孩子才开始知道别人在想什么。但在他们的第一个生日时，他们已经开始知道别人想要什么。⁸

研究人员越来越多地提出这样的论点，即我们向机器灌输人类价值观的方法应该采用同样的策略。也许，我们不应该费力地尝试手工编码我们关心的事情，而应该开发简单地观察人类行为并从中推断我们的价值观和欲望的机器。Richard Feynman曾经著名地将宇宙描述为“众神正在进行的一场伟大的国际象棋游戏...我们不知道游戏规则是什么；我们所能做的就是观看比赛。“在AI中，这种技术术语是”逆强化学习”——除了我们是众神，而机器必须观察我们并试图推断我们移动的规则。

逆强化学习

1997年，加州大学伯克利分校的Stuart Russell在前往杂货店的路上，脑海中突然想到一个问题：我们为什么会以这种方式行走。“我们的行走方式很刻板，对吧？如果你看过Monty Python的愚蠢行走部的小品，你会发现除了正常的行走方式之外，还有很多其他的行走方式，对吧——但我们几乎都以同样的方式行走。”⁹

这不能仅仅是模仿的问题——至少看起来不太可能。基本人类步态不仅在个体间变化很小——在不同文化间也几乎没有差异，据我们所知，在时间维度上也是如此。“这不只是’嗯，他们就是这样被教的，’”Russell说。“某种程度上，这就是有效的方式。”

然而这引发的问题和它回答的问题一样多。“你说的’有效’是什么意思，对吧？目标函数是什么？人们提出了各种目标，比如’我认为我在最小化能量，’或者’我在最小化扭矩，’¹⁰或者’我在最小化急动度(jerk)，’¹¹或者’我在最小化这个’或’我在最小化那个，’或者’我在最大化其他东西，’但它们都无法产生看起来真实的运动。这在动画制作中大量使用，对吧？试图合成一个行走和奔跑但看起来不像机器人的角色。它们都失败了。这就是为什么我们在所有这些方面都使用动作捕捉技术的原因。”

事实上，整个“生物力学”领域的存在就是为了回答这样的问题。长期以来，研究人员一直对四足动物的各种不同步态感兴趣：步行、小跑、慢跑、飞奔。直到19世纪末高速摄影的发明才解决了这些不同步态究竟如何运作的问题：哪些腿在什么时候抬起，特别是马在飞奔时是否曾经完全腾空。(1877年，我们得知确实如此。)然后，在20世纪，争论从如何转向了为什么。

1981年，哈佛动物学家Charles Richard Taylor在《自然》杂志上发表了一篇重要论文，表明马从小跑转换到飞奔的方式是为了最小化马消耗的总能量。¹²十年后，他在《科学》杂志上发表了一篇重要的后续论文，说不，根据进一步的证据，转换到飞奔不是为了最小化能量，而是为了最小化对马关节的压力。¹³

这些就是Russell走向杂货店时脑海中的想法。“我从家里沿着山坡走向Safeway，”Russell告诉我。“我注意到，因为这是下坡路，你的步态与在平地上略有不同。我在想，我想知道如何能够预测步态的差异。假如我把一只蟑螂放在一个倾斜的……”他用手做手势。“蟑螂会如何行走，对吧？我能预测吗？如果我知道目标，我就能预测当我倾斜这个东西时蟑螂会做什么。”

到1990年代末，reinforcement learning已经成为一种强大的计算技术，能够在各种(在那个时代，相当简单的)物理和虚拟环境中产生合理的行为。通过对多巴胺系统以及蜜蜂觅食行为的研究，人们也越来越清楚地认识到，reinforcement learning可以为理解人类和动物行为提供一个惊人恰当的框架。¹⁴

只有一个问题。典型的reinforcement learning场景假设人们试图最大化的“奖励”是什么是完全清楚的——无论是动物行为实验中的食物或糖水，还是AI实验室中视频游戏的分数。在现实世界中，这种“奖励”的来源要模糊得多。“行走的”分数”是什么呢？

所以，Russell在伯克利被称为The Uplands的林荫大道上行走时想，如果人类步态是答案——而reinforcement learning是身体找到它的方法——那么……问题是什么？

Russell在1998年写了一篇论文，起到了行动号召的作用。他论证说，该领域需要的是他所谓的逆向reinforcement learning。逆向reinforcement learning(或“IRL”)不是像常规reinforcement learning那样问“给定奖励信号，什么行为能够优化它？”而是问相反的问题：“给定观察到的行为，如果有的话，什么奖励信号正在被优化？”¹⁵

当然，用更非正式的术语来说，这是人类生活的基本问题之一。他们到底认为自己在做什么？我们花费生命中很大部分脑力来回答这样的问题。我们观察周围其他人的行为——朋友和敌人，上级和下属，合作者和竞争者——试图通过他们可见的行动解读他们不可见的意图和目标。这在某种程度上是人类认知的基石。

事实证明，这也是21世纪AI的核心和关键项目之一。它很可能掌握着alignment problem的关键。

从演示中学习

对于任何曾经努力解读他人行为背后的意义或意图的人——他们是在和我调情，还是只是一个超级友好的人？他们因为某种原因对我生气，还是只是心情不好？他们是故意这样做的，还是只是意外？——有时会感觉任何行为都可以有无数种含义。

计算机科学在这里提供安慰，但不是治愈。任何行为可能具有的含义在字面上是无限的。

从理论意义上讲，这个问题是绝望的。实际上，情况稍微好一些。

逆向强化学习，著名地被数学家称为“不适当”问题：即没有单一、唯一正确答案的问题。例如，从行为的角度来看，有庞大的奖励函数族是完全无法区分的。另一方面，总的来说这种歧义并不重要，正是因为一个人的行为不会因此而改变。例如，拳击运动恰好使用“十分必须”计分系统，其中一轮的获胜者获得十分，失败者获得九分。如果一个学徒拳击手错误地得出结论认为一轮的计分是一千万分对九百万分，或者千万分之一分对九百万分之一分，或者十一分对十分，他仍然知道总分较高的人获胜，他的拳击与理解“正确”计分系统的人没有什么不同。因此错误是不可避免的，但也是无关紧要的。¹⁶

然而，出现了另一个更棘手的问题：是什么让我们假设这个人的行为具有任何意义？如果他们根本没有试图做任何事情，他们的行为只反映随机行为，仅此而已呢？

在第一篇提出IRL问题实际解决方案的论文中，Russell和他当时的博士生Andrew Ng考虑了一些简单的例子，以表明这个想法是可行的。¹⁷他们考虑了一个微小的五乘五网格，其中目标是将玩家移动到特定的“目标”方格，以及一个视频游戏世界，其中目标是驾驶汽车到山顶。IRL系统能否仅仅通过观看专家（无论是人类还是机器）玩游戏来推断这些目标？

Ng和Russell在他们的IRL系统中构建了一些简化假设。它假设玩家从不随机行动，从不犯错误：当它采取行动时，该行动实际上是可能的最佳行动。它还假设激励agent的奖励是“简单的”，即任何可以被认为价值零分的行动或状态应该被认为价值零分。¹⁸此外，它假设当玩家采取行动时，不仅该行动是最好的选择，而且任何其他行动都是错误的。这排除了游戏具有多个竞争目标，玩家随机在它们之间选择的可能性。

假设相当强，领域过于简单而无法立即实际使用——它们与人类步态的复杂性相差甚远——但IRL确实有效。IRL系统推断的奖励看起来与真实奖励非常相似。当Ng和Russell让IRL系统通过试图最大化它认为的奖励来玩游戏时，它获得了与直接针对真实分数优化的系统一样高的分数——通过“真实”分数衡量。

到2004年，Andrew Ng获得了博士学位并在斯坦福大学任教，指导他自己当时的博士生Pieter Abbeel。他们再次处理IRL问题，试图增加环境的复杂性并放松一些推理假设。¹⁹正如他们想象的，我们正在观察的任何任务都有与该任务相关的各种“特征”。例如，如果我们在观察某人驾驶，我们可能会考虑相关特征是诸如汽车在哪条车道、行驶速度、与前车的跟车距离等等。他们开发了一个IRL算法，假设它自己驾驶时会看到与它在观察到的演示中相同的这些特征模式。一个非常简化的、类似Atari的驾驶模拟器显示了有希望的结果，计算“学徒”在游戏中的驾驶很像Abbeel的驾驶：避免碰撞、超越较慢的汽车，以及保持在右车道。

这与我们在第7章讨论的严格模仿方法显著不同。在Abbeel演示驾驶仅分钟后，试图直接模仿他行为的模型没有足够的信息可循——道路环境太复杂了。Abbeel的行为是复杂的，但他的目标是简单的；在几秒钟内，IRL系统就掌握

了不撞其他车的至关重要性，其次是不开出道路，其次是尽可能保持右行。这个目标结构比驾驶行为本身简单得多，更容易学习，在新情况下应用更灵活。IRL agent不是直接采用他的行动，而是学习采用他的价值观。

他们决定，是时候将IRL带入现实世界的完全混乱中了。

我们在第5章看到Ng如何使用奖励塑造的想法来教授自主直升机悬停并缓慢飞行稳定路径，这是没有计算机控制系统能够实现的壮举。这是Ng职业生涯和整个机器学习的重要里程碑，但进展停滞了。“坦率地说，我们撞墙了，” Ng说。“有些事情我们永远无法弄清楚如何让我们的直升机去做。”²⁰

问题的一部分在于，在原地悬停和低速沿固定路径飞行这两项任务中，传统的奖励函数(reward function)相对容易指定。在悬停的情况下，奖励就是直升机在各个方向上的速度接近零的程度；在路径跟随的情况下，沿路径的进展会得到奖励，偏离则会受到惩罚。问题的复杂性不在于指定目标，而在于找到一种方法来教导系统如何仅仅利用旋翼叶片的扭矩以及它们的俯仰角和角度来实际完成这些任务。正是在这里，强化学习(reinforcement learning)展现了它的威力。

但对于更复杂的机动和特技动作，在更高速度下执行并涉及更复杂的空气动力学时，如何制定一个让系统能够学习其行为的奖励函数就不那么明显了。当然，你可以简单地在空间中画一条曲线，告诉计算机尝试飞行这个确切的轨迹——但物理定律，特别是在高速情况下，可能不允许这样做。直升机在曲线的某个部分可能有太大的动量，机器上的应力可能太大，发动机可能无法在正确的时间产生足够的动力，等等。你这样做是在为系统设置失败的陷阱——对于一架重10.5磅、以45英里每小时速度移动的直升机来说，这也可能代价昂贵，更不用说危险了。“我们使用这种手工编码轨迹的尝试，”团队写道，“反复失败。”²¹

但是，他们推理道，你可以做的是让一个人类专家飞行这个机动动作，并使用逆强化学习(inverse reinforcement learning)让系统推断人类试图实现的目标。通过使用这种IRL方法，Abbeel和Ng与他们的合作者Adam Coates，能够在2007年演示第一个由计算机执行的直升机前空翻和副翼横滚。²²这显著推进了技术水平，并表明IRL能够在看似没有其他方法可行的情况下成功传达真实世界的人类意图。

但他们并不满足于在前空翻的成就上止步不前。他们想要找到一种方法来执行如此困难的特技动作，甚至连他们的人类演示者——专业的无线电遥控直升机飞行员Garett Oku——都无法完美执行。Abbeel、Coates和Ng想要将他们的直升机工作推向令人眩晕的极限：制造一架计算机控制的直升机，能够执行超越人类飞行员能力的令人惊叹的特技动作。

他们有了一个关键的洞察。即使Oku无法以其纯粹的、柏拉图式的形式完美执行一个机动动作，只要他的尝试足够好，那么他的偏差至少会在不同的尝试中以不同的方式表现出不完美。一个进行逆强化学习的系统——而不是严格的模仿——可以通过一系列不完美或失败的尝试，推断出人类飞行员试图做什么。²³

到2008年，他们从专家演示中的推断导致了突破的洪流，记录了第一次成功的自主演示：“连续原地空翻和横滚、连续机尾向下’滴答’、环飞、带旋转的环飞、带旋转的失速转弯、‘飓风’(快速后向漏斗)、刀刃飞行、英麦曼回转、拍击、侧向滴答、移动翻转、倒飞机尾滑行，甚至自转着陆。”²⁴

他们雄心的顶峰是一个通常被认为是所有直升机机动中最困难的动作：叫做“混沌”(chaos)的动作，这个动作如此复杂，全世界只有一个人能够做到。

混沌动作是由Curtis Youngblood发明的，他是1987年、1993年和2001年的模型直升机世界冠军；2002年和2004年的3D大师赛冠军；以及1986年、1987年、1989年、1991年、1993年、1994年、1995年、1996年、1997年、1998年、1999年、2000年、2001年、2002年、2004年、2005年、2006年、2008年、2010年、2011年和2012年的美国全国冠军。²⁵他被许多人认为是有史以来最伟大的无线电遥控直升机飞行员。

“当时我在想，” Youngblood说，“我能想出的最复杂的受控机动动作是什么。”他采用了已经是最困难的机动动作之一——旋转翻转——并设想在旋转的同时一遍又一遍地做这个动作。

当被问及还有多少其他飞行员能够稳定地完成这个机动动作时，Youngblood说没有。“我曾经能够做到；我今天甚至都做不到了。…如果有人要求我做这个动作，没有练习我无法完成一个完整的动作。”

他说，问题的一部分在于这个机动动作看起来如此复杂，只有其他专业飞行员才能欣赏他们所看到的困难程度。

“你通常是表演飞行员，”他说。“你在那里试图给观众留下深刻印象。观众完全不知道你在做什么。所以无论你做一个真正的混沌动作还是只是一个旋转翻转，他们都不知道区别。所以通常没有人会因此得到回报而真正、真正地学习它——除了向其他顶级飞行员炫耀我实际上能做到这个。”²⁶

到2008年夏天，斯坦福直升机已经掌握了混沌动作，尽管从未见过一次完美的演示——无论是来自Oku、Youngblood还是任何人。但系统看到了他们像把一摞杂志撞向柜门一样的动作。然后——一遍又一遍又一遍地旋转翻转，同时每次旋转三百度，看起来像一架单机直升机飓风——它把门甩开了。²⁷

与此同时，不同的行为消歧方法，以及更复杂的奖励表示方式，继续扩展着逆向强化学习框架。2008年，时任博士生的Brian Ziebart和他在卡内基梅隆大学的合作者开发了一种使用信息论思想的方法。我们不再假设观察到的专家是完全完美的，而可以想象他们只是更有可能选择带来更多奖励的行动。反过来，我们可以利用这一原理，找到一组奖励，在保持其他方面尽可能不确定的同时，最大化我们观察到特定演示行为的可能性。

Ziebart将这种所谓的最大熵IRL方法在一个数据集上进行了测试，该数据集记录了来自24名真实匹兹堡出租车司机的十万英里驾驶数据，用来建模他们对某些道路相对于其他道路的偏好。该模型能够可靠地猜测司机为到达特定目的地会选择什么路线。更令人印象深刻的是，它还能根据司机迄今为止选择的路线，对司机试图前往的地方做出合理的猜测。（Ziebart指出，这可能允许司机在无需实际告诉系统目的地的情况下，接收关于影响其预定路线的道路封闭的相关通知。）

在过去十年中，机器学习中出现了一波使用所谓“动觉教学(kinesthetic teaching)“的工作热潮，即人类手动移动机器臂来完成某项任务，机器人系统必须推断相关目标，以便在稍微不同的环境中自主地自由再现类似行为。2016年，时任博士生的Chelsea Finn和她在伯克利的合作者进一步扩展了最大熵IRL，通过使用神经网络允许奖励函数任意复杂，并消除了预先手动指定其组成特征的需要。他们的机器人在经过二十或三十次演示后，能够做出如人类般的、不可能直接用数字指定的事情，比如将盘子装入餐具架（不弄碎它们）和将一杯杏仁倒入另一个杯子（不洒漏任何东西）。可以说，我们现在已经远远超越了机器只能做我们用数学和代码的显式语言编程给它们的事情的阶段。

一看就知道：从反馈中学习

逆向强化学习已经证明了自己是一种引人注目且强大的方法，能够将复杂目标归因给系统，这在我们必须手动显式编程奖励时是根本不可行甚至不可能的。然而，唯一的问题是，典型的公式需要有一个专家在场，能够给出期望行为的演示（即使是不完美的）。直升机特技需要一名熟练的飞行员；出租车需要司机；同样，盘子和杏仁需要人类演示者。还有其他方法吗？

生活中有很多事情很难执行，但相对容易评估。我可能是如此糟糕的遥控直升机飞行员，甚至无法让机器保持在空中，然而我可以（除了混乱的情况外）在看到令人印象深刻的空中杂技表演时识别它。正如Youngblood所指出的，给外行观众留下印象或多或少就是重点。

如果一个系统能够仅从我的反馈中推断出显式的奖励函数——我对其行为演示给予某个评分，或我在比如两种不同演示之间的偏好——那么我们就有了一种强大且更通用的方式来表达我们想要从机器那里得到的东西。也就是说，即使我们不能说出我们想要什么，甚至不能做我们想要的事情，我们仍然有对齐的手段。在一个完美的世界里，仅仅一看就知道就足够了。

这是一个强大的想法。只有两个问题。这实际上可能吗？它安全吗？

2012年，Jan Leike在德国弗莱堡完成他的硕士学位，从事软件验证工作：开发工具来自动分析某些类型的程序并确定它们是否会成功执行。“那时我意识到我真的喜欢做研究，”他说，“那进展得很好——但我也，比如，真的不清楚我将要用我的生活做什么。”然后他开始阅读关于AI安全的想法，通过Nick Bostrom和Milan Ćirković的书《全球灾难风险》，一些在互联网论坛LessWrong上的讨论，以及Eliezer Yudkowsky的几篇论文。“我想，嗯，似乎很少有人在研究这个。也许这是我应该做研究的事情：听起来超级有趣，而且没有做太多。”

Leike联系了澳大利亚国立大学的计算机科学家Marcus Hutter寻求一些职业建议。“我只是随机给他发了一封电子邮件，告诉他，你知道，我想在AI安全方面做博士，你能给我一些关于去哪里的建议吗？然后我附上了我做过的一些工作或其他什么，这样他就会希望麻烦回答我的电子邮件。”Hutter几乎立即回信了。你应该来这里，他写道——但申请截止日期是三天后。

Leike笑了起来。“而且你得考虑，我的学位证书之类的都不是英文的。我没有参加过英语水平测试。我必须在三天内完成所有这些事情。”此外，Leike还必须从零开始写一份研究提案。而且，那一周他恰好在度假。“你可能想象得到，那三天我几乎没怎么睡觉。”

Leike强调：“顺便说一下，这真的是选择博士项目的糟糕建议。我基本上没有做任何研究，突然给一个人发邮件，然后就决定申请那里。这显然不是申请博士项目的正确方式！”

到那年年底，Leike已经在堪培拉安顿下来，他和Hutter开始了工作。他在2015年底完成了博士学位，研究Hutter的AIXI框架（我们在[第6章]中简要提及）并记录了这种智能体容易出现“严重不当行为”的情况。³³拿到博士学位后，他准备进入AI安全就业市场：也就是说，加入世界上三四个可以从事AI安全职业生涯的地方之一。在牛津的人类未来研究所工作了六个月后，他在伦敦的DeepMind找到了一个永久职位。

“当时，我在思考价值对齐(value alignment)，”Leike说，“以及我们如何能够做到这一点。似乎很多问题都与‘如何学习奖励函数？’有关。所以我联系了Paul和Dario，因为我知道他们在思考类似的问题。”

在旧金山OpenAI的Paul Christiano和Dario Amodei对此很感兴趣。实际上，不仅仅是感兴趣。Christiano刚刚加入，正在寻找一个有趣的首个项目。他开始专注于在更少监督下的强化学习想法——不是每秒十五次的持续分数更新，而是更定期的，比如监督者定期检查。当然，可以修改Atari环境，使其只定期而不是实时地通知智能体其分数，但三人感觉如果实际的人类提供这种反馈，这篇论文将更有可能引起轰动——并且会为长期对齐项目提供更清晰的建议。

Christiano和他在OpenAI的同事们，以及Leike和他在DeepMind的团队，决定集思广益（和GPU）来深入研究机器如何从人类那里学习复杂奖励函数这个问题。这个项目最终成为2017年最重要的AI安全论文之一，不仅因为它的发现而卓越，更因为它所代表的意义：世界上最活跃的两个AI安全研究实验室之间的重要合作，以及对齐研究的一条诱人前进道路。³⁴

他们一起制定了一个计划，在没有演示的情况下实施最大规模的逆强化学习测试。这个想法是，他们的系统会在某个虚拟环境中行为，同时定期向人类发送其行为的随机视频片段。人类只需简单地按照屏幕指示：“观看片段并选择发生更好事情的那个。”系统然后会尝试根据人类的反馈改进其对奖励函数的推断，然后使用这个推断的奖励（如在典型强化学习中一样）找到在其看来表现良好的行为。它会继续朝着对真实奖励的新最佳猜测改进自己，然后发送新一对视频片段供审查。

项目的一部分是在存在明确“客观”奖励函数的领域——经典的Atari游戏，直接强化学习已经证明了其超人表现——看看他们能让智能体在无法获得游戏内分数的情况下表现得多好，仅使用人类说哪些视频片段比其他“更好”的最佳猜测。在大多数情况下，系统设法做得相当好，尽管通常无法达到直接获得分数时可能的超人表现。然而，在涉及复杂超车操作的赛车游戏*Enduro*中，从人类反馈推断分数实际上比使用游戏的真实分数函数效果更好——这表明人类间接地在进行某种奖励塑造。

“如果你有一个存在真实奖励的设置，比如在Atari中，那么这非常有帮助，”Leike说。“因为那样你就可以进行诊断，对吧？你实际上可以直接检查价值对齐，即‘你的奖励模型与真实奖励函数的对齐程度如何？’”

团队还想要找到一些强化学习系统可以被训练去做的完全主观的事情，一些不存在“真实”分数的事情，一些如此复杂或难以描述以至于手动指定数值奖励不可行的事情，但同时又如此易于识别，以至于人类一看到就能立即知道的事情。

他们找到了一些他们认为可能符合要求的东西。后空翻。

“我只是看了看机器人身体——我看了所有的机器人身体，”Christiano说，“我想：这些机器人中哪一个看起来应该能做的最酷的事情是什么？”³⁵

其中一个虚拟机器人叫做“hopper”；它看起来像一条有着超大脚掌的无躯体腿。“我的第一个、最有野心的想法，”Christiano说，“就是：这个机器人身体看起来应该能够做后空翻。”

计划已经确定。他们将在一个名为MuJoCo的虚拟物理模拟器中使用一个简单的机器人——这是一个除了摩擦力和重力之外几乎什么都没有的玩具世界——并尝试让它做后空翻。³⁶这似乎是一个大胆的想法：“只要随意扭动一下，人们会观看不同的视频片段，说哪一个看起来更像在做后空翻，试着扭动得更像那样一点，我们看看会发生什么。”

Christiano开始了他众多几小时观看会话中的第一次，一对又一对地观看视频片段，一次又一次地选择哪一个看起来更像后空翻。左边的片段。右边的片段。右边的片段。左边的片段。右边的片段。左边的片段。

“每次它取得一点小进步，我都会非常兴奋，” Christiano说。“比如，它开始摔倒，我很兴奋，因为即使在随机行为下，它有时也会朝正确的方向摔倒。然后当它总是朝正确方向摔倒时，我就兴奋了。”进步是微小而渐进的。他继续下去。“一切都非常渐进，”他说，“因为你就是在，像，按左键。和右键。这么。多。次。观看这么多片段。”

“我想我最兴奋的时候，”他回忆道，“可能是当它开始稳稳着陆时。”

发生的事情是——在比较了几百个片段后，大约一小时的时间里——它开始做出漂亮、完美的后空翻：像体操运动员那样蜷缩身体，并且稳稳着陆。

实验与其他提供反馈的人进行了重复，后空翻总是略有不同——仿佛每个人都在提供他们自己的美学，他们自己版本的柏拉图式后空翻。

Christiano去参加他在OpenAI的每周团队会议并展示了这个视频——“看！我们可以做这件事，”他回忆说告诉他们。“每个人都说，天哪。那真酷。”

“我对结果非常满意，”Leike说。“因为，先验地，根本不清楚这会起作用。”

我告诉他们，让这个结果在我看来不仅如此令人印象深刻，而且如此充满希望的是，用一个更加模糊和难以言喻的概念来替换“后空翻”这个模糊概念并不是太大的跨越，比如“有帮助”。或者“善良”。或者“好”行为。

“确切地说，”Leike说。“这就是重点，对吧？”³⁷

学习合作

Dylan Hadfield-Menell于2013年刚到加州大学伯克利分校，他刚刚在MIT与机器人学家和reinforcement learning先驱Leslie Kaelbling完成了硕士学位。他认为他在Stuart Russell实验室的博士研究会或多或少地从他硕士工作停止的地方继续，做机器人任务和运动规划。Russell在第一年去休学术假了。他在2014年春天回来，一切都改变了。

“我们有这个大会议，我们在谈论计划，”Hadfield-Menell说，“他在阐述他的研究愿景。他有点说，‘嗯，关于这个有些话要说...’他没有称它为价值对齐(value alignment)，但是：‘如果我们成功了，实际上可能出什么问题？’”Russell说他开始对AI的长期担忧给予一些信任：我们开发的学习系统越灵活和强大，它们学会做什么就变得越重要。在巴黎期间，Russell变得担心起来。他回到伯克利时很坚决：有倡导工作要做——还有科学研究。

“研究想法几乎立即就来了，”Russell说。³⁸“似乎解决方案，你想要的是价值对齐的AI系统，意思是它们与人类有相同的目标。而我从九十年代末就已经在inverse reinforcement learning方面工作了——这基本上是同一个问题。”

我说，有一定的讽刺意味，他二十年前的想法最终成为他当前AI安全议程的基础。他去Safeway路上的闲思，二十年后，成了避免可能的文明级灾难的计划。“那完全是巧合，”他说。“但是，我的意思是，这整件事是一系列巧合，所以没关系。”

在第一次实验室会议上，Russell告诉他的学生们，他认为有几个具体的、值得博士研究的主题可以探索。Hadfield-Menell继续他的机器人研究，但在他心里他一直在思考对齐问题。最初，部分吸引力只是一组全新未探索问题的智力刺激和做出开创性贡献的机会。随着时间的推移，这开始让位于一种不同的感觉，他告诉我：“似乎这实际上很重要——而且没有得到关注。”在2015年春天，他决定重新定向他的博士研究，进而，他的职业生涯。

“此时我所有的淋浴思考，”他告诉我，“都是关于价值对齐的。”

他和Russell最初的合作项目之一是重新审视inverse reinforcement learning框架。

Russell和Hadfield-Menell与Pieter Abbeel和伯克利机器人学家Anca Drăgan合作，开始做的是从头重新构想IRL。有两件事让他们印象深刻。

直升机工作，就像该领域的几乎所有其他工作一样，都建立在人类和机器之间某种分工的前提下。人类专家飞行员只是做他们的事情，而计算机从这些演示中尽可能理解。每个都独自运行，并且处于某种真空状态。但是，如果人类从一开始就知道他们有一个渴望学习的学徒呢？如果两者有意识地一起工作呢？那会是什么样子？

另一个突出的点是，在传统的IRL中，机器将人类的奖励函数作为自己的奖励函数。如果人类直升机飞行员试图执行混乱动作，那么现在计算机飞行员也试图执行混乱动作。在某些情况下，这是有意义的：我们想安全地开车上班和回家，如果我们的汽车接受这套目标和价值观并将其或多或少透明地作为自己的目标和价值观，我们会很高兴。然而，在其他情况下，我们想要一些更微妙的东西。如果我们伸手去拿一块水果，我们不希望家用机器人自己产生对香蕉的渴望。相反，我们希望它做一个十八个月大的婴儿会做的事：看到我们伸出的手臂和刚好够不着的物品，然后把它递给我们。³⁹

Russell将这个新框架称为协作逆强化学习（简称“CIRL”）。⁴⁰在CIRL公式中，人类和计算机共同合作来联合最大化单一奖励函数——最初只有人类知道它是什么。

“我们试图思考，我们能对当前的数学和当前的理论系统做出什么最简单的改变来修复导致这些存在风险问题的理论？”Hadfield-Menell说。“什么是一个数学问题，其中最优解是我们实际想要的？”⁴¹

对Russell来说，这也并不是对问题的微妙重构，而是在某种意义上是对齐问题的决定性胜利。这无异于颠覆AI领域最基本的假设，一种哥白尼式的转变。他说，在过去的一个世纪里，我们一直试图构建能够实现其目标的机器。这几乎在AI的所有工作中都是隐含的，安全问题围绕着如何控制它们的目标应该是什么——如何定义明智且无漏洞的目标。也许整个想法需要被颠覆，他认为。“如果，我们不允许机器追求它们的目标，而是坚持让它们追求我们的目标会怎么样？”他说。“这可能是我们一直以来应该做的。”⁴²

一旦引入协作框架，就会开辟几个前沿。传统的机器学习和机器人研究人员现在比以往任何时候都更热衷于从育儿、发展心理学、教育以及人机交互和界面设计中借鉴想法。突然间，整个其他学科的知识变得不仅相关而且至关重要。⁴³ 协作作为推理的框架——我们采取行动时知道另一方试图解读我们的意图——让我们以不同的方式思考人类和机器的行为。

在这些前沿的工作是活跃和持续的，仍有很多东西需要学习。但到目前为止有几个关键见解。

首先，如果我们知道我们正在被研究，我们可以比自然行为时更有帮助。“CIRL激励人类进行教学，”Berkeley小组写道，“而不是孤立地最大化奖励。”⁴⁴ 我们当然可以明确指导，但我们也可以简单地以更具信息性、明确、易于理解的方式行动。通常我们已经在不经意间这样做了，或者没有意识到我们在这样做。

例如，事实证明，成年人与婴儿交谈时经常使用的歌唱式语言（称为“母语”或“父母语”）具有深刻的教学效果。以父母语交谈的婴儿实际上学习语言更快——这似乎，无论我们是否意识到，都是整个要点。⁴⁵

不仅在语言中，在我们的动作中，我们的行为——通常没有意识到——深深受到我们的行为将被他人解释这种感觉的影响。考虑一下将物体递给某人这一容易被忽视的复杂性。我们拿着物体不是在最方便拿的地方，而是拿得远离我们，那里我们手臂的压力更大——而另一个人意识到如果不是为了表示我们想让他们接受它，我们永远不会这样做。⁴⁶

教学法和育儿的见解正在被计算机科学家迅速采用。核心思想是双向的。我们希望以机器能够理解的方式行动，我们也希望我们的机器以对我们“清晰易懂”的方式行动。

在机器人“可理解运动”研究领域的顶尖研究者之一——实际上就是创造这个术语的人——是加州大学伯克利分校的机器人专家Anca Drăgan。⁴⁷ 随着机器人越来越多地与人类更紧密、更灵活地协作，它们不仅必须以最高效或最可预测的方式行动，还必须以最能传达其潜在目标或意图的方式行动。Drăgan举了一个例子：桌子上相邻放着两个瓶子。如果机器人以最高效率的方式伸手，我们直到很晚才能明显看出它要拿哪个瓶子。但如果它以一个宽大、夸张的弧线伸出手臂，我们就能很快感知到它要拿特定那一边的瓶子。从这个意义上说，可预测性和可理解性几乎是相反的：可预测的行为假设观察者知道你的目标是什么；可理解的行为假设他们不知道。

除了教学行为的重要性之外，这个领域出现的第二个洞察是，当合作被框架为互动时效果最好，而不是两个独立而不同的“学习，然后行动”阶段。

Jan Leike在从人类反馈中学习的工作中发现了这一点。“对我来说，这篇论文中最有趣的事情实际上是最终只成为一个脚注的内容，”他说，“就是这些reward-hacking的例子。”

当智能体预先完成所有奖励学习，然后进行所有优化时，经常以灾难告终。智能体会找到某个漏洞并利用它，再也不回头。例如，Leike正在努力让智能体仅通过对智能体游戏视频片段的人类反馈来学习玩Pong游戏。在一次试验中，计算机学会了用球拍跟踪来球，好像它要击中它，但然后在最后一秒错过。因为它无法获得游戏的真实分数，

它完全不知道自己错过了最关键的一步。在另一次试验中，计算机学会了保护自己这边的屏幕并回击球，但从未学会尝试得分，所以它只是产生了长时间的持续对打。“从安全角度来看这很有趣，”Leike说。“因为你想理解这些东西是如何失败的——然后你能做什么来防止这种情况。”当人类反馈在计算机训练过程中交织进行，而不是完全前置时，这类问题往往消失。⁴⁸这是另一个论据，表明严格的“观察和学习”范式可能最好被更具协作性和开放性的东西所取代。

MIT的Julie Shah研究现实世界中人机协作互动，她得出了类似的结论。“多年来我一直感兴趣的是，”她说，“如何在人和机器之间共同优化学习过程。”她对人机团队的工作让她研究了关于人-人团队的研究文献，以及它们如何最有效地训练。在人类群体中，激励当然很重要，但人们很少看到明确的微观管理奖励到任务级别。“如果你通过交互式奖励分配来训练系统，这更接近于你训练狗的方式，通常，而不是你教人做任务的方式，”她说。⁴⁹演示也不总是有效。“这是一种从一个人向另一个实体单向传递如何执行任务信息的非常有效方式，”Shah说。“但它的不足之处在于，当你需要考虑在人和机器之间训练相互依赖的行动时。”毕竟，如果首先需要协调和团队合作才能完成某事，那么简单地演示某事就困难得多。

确实，研究人-人团队的结论很明确。“有一个既定的文献，”Shah说，“基本上表明明确命令一个人做任务是训练两人之间相互依赖工作的最糟糕方式之一。你知道，当你想到这一点时，就像：嗯，显然！对吧？…它们是你能实施的最糟糕的人类团队训练实践之一。”

“还有很好的人类研究，”她补充说，“表明如果你有几个人试图达到相同的目标，或相同的意图——每个人都知道那个目标或意图是什么——但你的两个人有不同的实现策略，而且他们的工作需要相互依赖，他们的表现会比拥有次优但连贯的策略要差得多。”在几乎任何团队场景中——从商业到作战到体育到音乐——每个人的高层目标都是相同的，这是理所当然的。但仅有共同目标是不够的。他们还需要一个计划。

在人类团队中有效的是一种叫做*cross-training*的东西。团队成员暂时交换角色：突然发现自己站在队友的立场上，他们开始理解如何改变自己的实际工作以更好地适应队友的需求和工作流程。*Cross-training*在人类团队训练中有点像黄金标准，在军事、工业、医学等环境中使用。⁵⁰

Shah开始想知道：这样的东西对人机配对也能奏效吗？人-人团队合作的最佳实践可能或多或少直接转化为机器人语境吗？这似乎只是“一个疯狂的想法，”她说。“这几乎就像一个小案例研究，就像，你知道，这甚至有帮助吗？这只是一个有点古怪的想法，所以让我们探索一下。”

所以他们做了实验。首先，他们想知道交叉训练是否能够与目前机器从人类学习的最先进方法相竞争。他们还提出了第二个问题，这在此类典型研究中并不常见：交叉训练是否也能帮助人类更好地学习如何与机器人合作（和教导）？

Shah的团队为人类和机械臂协作创造了一个现实世界任务，类似于在制造装配线上放置和拧螺丝的任务（但没有使用真正的钻头）。他们将传统的反馈和演示方法与交叉训练进行了比较。

“结果对我们来说非常令人惊讶和兴奋，”她说。“我们在交叉训练后看到了团队表现客观指标的改善。”人们在与机器人并行工作时更加自在，而不是僵硬地轮流工作；这减少了闲置时间，完成了更多工作。

也许同样重要且更引人深思的是，他们还看到了主观收益。与进行更传统演示和反馈学习的对照组相比，进行过交叉训练的人更强烈地表示他们信任机器人，认为机器人的表现符合他们的偏好。

令人难以置信的是，人与人团队合作的最佳实践确实可以转化应用，这表明进一步的洞察很可能可以交叉借鉴。“这是一个相对简单任务的初步研究，”Shah说，“但即使在简单任务中，我们也确实看到了统计学意义上的显著

收益，这为我们探索许多其他不同的人类团队训练技术，以及如何将这些技术转化应用到人机团队合作中开辟了机会。”

合作：利与弊

这一切构成了一个相当令人鼓舞的故事。人类合作基于一种在进化上较为近期且几乎独特的能力，即推断彼此的意图和目标，以及提供帮助的动机。机器能够——且越来越多地确实——完成同样的任务，从我们的演示、反馈中学习，并且越来越多地与我们并肩工作。

随着机器变得越来越有能力，我们也越来越习惯与它们密切合作，好消息至少是双重的。我们开始有了一个计算框架，用于机器不仅代替人类运作，而且与人类协同运作。我们还拥有大量关于人类如何彼此良好合作的研究，这些洞察变得越来越相关。

如果我们推演当前的技术水平，使用Christiano和Leike的人类偏好深度强化学习以及Russell和Hadfield-Menell的CIRL等框架，我们可能会想象一条一直通向任意智能和有能力机器的轨迹，这些机器能够捕捉我们每个意图和需求的任意微妙差别。当然要克服许多障碍，要达到许多极限，但前进的道路开始显现。

然而，与此同时，值得谨慎提醒。这些近未来的计算助手，无论是以数字形式还是机器人形式出现——可能两者兼而有之——几乎毫无例外地会有利益冲突，它们是两个主人的仆人：表面上的主人和创造它们的任何组织。从这个意义上说，它们就像按佣金付费的管家；它们永远不会无偿帮助我们，至少隐含地想要某种回报。它们会做出我们不一定希望它们做出的精明推断。我们将意识到，我们现在——已经，在当下——几乎从不独自行动。

我的一位朋友正在戒酒。他们社交媒体账户的广告推荐引擎知道得太多了。他们的信息流充斥着酒类广告。这是一个，他们的偏好模型说，热爱酒精的人。正如英国作家Iris Murdoch所写：“自知之明会引导我们避开诱惑的场合，而不是依靠赤裸裸的力量去克服它们。”对于任何成瘾或强迫行为，智慧的上策告诉我们——以酒精为例——最好把家里的每一滴酒都倒掉，而不是留在身边却不喝。但偏好模型不知道这一点。就好像当他们只想坐一会儿、发发短信、看看朋友宝宝的可爱照片时，酒类商店都会跟着他们到厕所。就好像他们自己的橱柜在为Anheuser-Busch工作。

至少就我而言，我确实试图注意自己的在线行为。至少在浏览器中，任何带有恶习色彩或仅仅是内疚快感的活动——无论是阅读新闻标题、查看社交媒体，还是其他我无意中做但不一定希望做更多的数字强迫行为——我都会在私密标签页中进行，该标签页不包含跟踪我在互联网其他地方的cookies或登录账户。这不是因为我感到羞耻；而是因为我不想让这些行为得到强化。

在这些情况下，我们想要一个与典型指令稍有不同的指令，后者是从我的行为中推断我的目标并促进我做更多同样的事情。我们想要说，实际上，“你不能因为我在做这件事就推断我想做这件事。请不要让这对我来说变得更容易。请不要放大它或强化它，或以任何方式压制通往这条路的欲望之径。请在我身后长出荆棘。”

我认为这里存在一个重要的政策问题，至少与理论问题同等重要。我们应该认真考虑用户有权查看并修改任何网站、应用或广告商对他们的偏好模型这一理念。值得考虑制定相关法规：实质上说，我有权拥有自己的模型。我有权说，那不是我。或者，从愿望的角度说，这是我想成为的人。这是你必须为之工作的人。

这就是我们当下时刻的微妙之处。我们的数字管家正在密切观察。它们看到我们的私人生活和公共生活，我们最好和最坏的一面，却不一定知道哪个是哪个，或者根本不做区分。它们大体上处于一种诡异的复杂程度谷底：能够从我们的行为中推断出我们欲望的复杂模型，但无法被教导，也不愿意合作。它们在努力思考我们接下来会做什么，思考如何赚取下一笔佣金，但它们似乎不理解我们想要什么，更不用说我们希望成为什么样的人了。

有些人可能会争论说，也许这会激发我们更好的一面；确实，当人们感觉被观察时，往往会表现得更有德行。在许多实验室研究中，当人们被摄像头监控时，当房间里有单向镜时，当房间明亮而非昏暗时，他们作弊的可能性更小。⁵⁶见鬼，甚至暗示监控的存在——墙上的一张人像、一幅人眼的画、一面普通镜子——都足以产生这种效果。⁵⁷

这是十八世纪哲学家杰里米·边沁著名的“全景监狱”蓝图理念的一部分，这是一座圆形监狱，每个牢房围绕着一座警卫塔排列，囚犯永远不知道自己是否被监视。边沁对监控本身的净化效果大加赞赏，甚至不是监控本身，而是怀疑被监控的效果，他称他的圆形建筑为“将恶棍磨练成诚实人的磨坊”，⁵⁸并激动地列举其前景：“道德得到改革——健康得到保护——勤奋得到振兴——教育得到普及——公共负担得到减轻……所有这些都通过建筑学中的一个简单理念！”⁵⁹

另一方面，这带来了潜在的寒蝉效应，以及更多问题。毕竟，我们通常不希望非监狱生活像监狱一样。

不那么险恶但同样令人担忧的是，IRL的标准数学假设人类行为来自“专家”——知道自己想要什么，并且（高概率地）在做正确的事情来获得它的人。如果这些假设不成立，那么系统就是在放大新手的无知行为，或者为应该是试探性、探索性的行为提高风险。我们先爬再走，走了再开车。也许最好我们通常不要有机械装置放大我们的每一次抽搐。

无论好坏，这就是人类的境况——现在如此，在我们对未来的乐观前景中也会越来越如此。无论好坏，我们都会更被了解。无论好坏，世界将充满这些算法两岁儿童，走向我们，打开它们认为我们可能想要打开的门，以各种方式试图帮助我们。

人类给人类带来的最大罪恶，大多来自人们对某些实际上是错误的事情感到相当确信。

——伯特兰·罗素¹

我恳求你，以基督的慈悲，想想你可能是错误的。

——奥利弗·克伦威尔

自由的精神是不太确定自己是对的精神。

——勒尼德·汉德

1983年9月26日，刚过午夜，苏联值班军官斯坦尼斯拉夫·彼得罗夫在莫斯科郊外的地堡中，监控着Oko早期预警卫星系统。突然屏幕亮了起来，警报器开始呼啸。系统显示，有一枚LGM-30民兵洲际弹道导弹正从美国飞来。

“我们主屏幕上出现了巨大的血红色字母，”他说。字母写着：发射。

“当我第一次看到警报信息时，我从椅子上站了起来，”彼得罗夫说。“我所有的下属都很困惑，所以我开始对他们大喊命令，以避免恐慌。”²

警报器再次响起。系统说，第二枚导弹已经发射。然后是第三枚。然后是第四枚。然后是第五枚。

“我舒适的扶手椅感觉像一个炽热的煎锅，我的腿发软了，”他说。“我觉得甚至站不起来。”

彼得罗夫一手拿着电话，另一手拿着对讲机。通过电话，另一名军官在对他大喊要保持冷静。“我承认，”彼得罗夫说，“我很害怕。”

规则很清楚。彼得罗夫要向他的上级报告来袭攻击，上级会决定是否下令全面报复。但是，他说，“关于我们在报告攻击之前被允许思考多长时间，没有规则。”³

彼得罗夫感到，有什么地方不对。他受过训练，预期美国的攻击会是更大规模的。五枚导弹……这根本不符合预期模式。“警报器呼啸，但我就坐在那里几秒钟，盯着那个巨大的、背光的红色屏幕，上面写着‘发射’，”他回忆道。⁴“我所要做的就是伸手拿电话；拨通我们高级指挥官的直线电话——但我动不了。”

早期预警系统报告其警报的可靠性级别为“最高”。但是，这完全说不通。彼得罗夫推理，美国拥有数千枚导弹。为什么只发射五枚？“当人们发动战争时，他们不会只用五枚导弹开始，”他记得当时这样想。这不符合他接受过训练的任何情景。“我心里有种奇怪的感觉，”他说。⁵

“然后我做出了决定。我不会相信计算机。我拿起电话，提醒我的上级，并报告这个警报是假的。但我自己直到最后一刻都不确定。我非常清楚，如果我犯了错误，没有人能够纠正我的错误。”

多年后BBC问他是否认为警报是真的概率有多大时，他回答说：“五五开。”但他没有犯错。几分钟令人煎熬的等待后，什么也没发生。本应在导弹越过地平线时确认其存在的地面雷达没有显示任何活动。苏联一切平静。这是一个系统错误：不过是阳光反射到北达科他州上空的云层而已。

前所未见的情况

天地之间有许多事物，霍拉旭，是你的哲学所梦想不到的。

—哈姆雷特

尽管Oko早期预警系统自报具有“最高”可靠性，彼得罗夫感觉到情况奇怪到足以让他有理由不信任其结论。感谢上帝人类还在回路中，因为一亿或更多生命岌岌可危。

然而，潜在的问题——系统不仅做出错误判断，而且以极高的置信度做出这些判断——至今仍令研究人员担忧。

深度学习系统特别容易出现“脆弱性”，这是一个众所周知的特性。我们看到2012年的AlexNet，当展示数以万计属于多个类别之一的图像时，令人惊讶地能够识别出足够通用的模式，从而正确分类它从未见过的猫、狗、汽车和人。但有个问题。它会对你展示的每张图像进行分类，包括随机生成的彩虹色静态图像。这个静态图像，它说是猎豹，置信度99.6%。那个静态图像是波罗蜜，置信度99.6%。该系统不仅本质上在产生幻觉，而且似乎没有机制来检测，更不用说传达它正在这样做。正如2015年一篇广泛引用的论文所说：“Deep Neural Networks(深度神经网络)很容易被愚弄。”⁶

密切相关的是所谓对抗样本(adversarial examples)的概念，网络57.7%确信的熊猫图像（确实是熊猫）可以通过对其像素进行不可察觉的改变，突然间网络就99.3%确信这个看起来几乎相同的图像是长臂猿。⁷

目前正在许多努力来思考这类情况究竟出了什么问题，以及能做些什么。

俄勒冈州立大学计算机科学家Thomas Dietterich认为这个问题的很大一部分源于这样一个事实：视觉系统在训练期间看到的每一张图像都是某种物体——画笔、壁虎、龙虾。然而，绝大多数可能的图像——彩色像素的潜在组合——根本什么都不是。静态。雾。噪声。没有底层形式的随机立体派线条和边缘。Dietterich认为，像AlexNet这样在标记为一千个类别之一的图像上训练的系统，“隐含地假设世界仅由，比如说，一千种不同类型的对象组成。”⁸当系统从未见过任何超出这些类别的图像，或者模糊地暗示许多类别的图像，或者更可能的是，根本“不是一个东西”的图像时，系统怎么能知道得更好呢？Dietterich称此为“开放类别问题(open category problem)”。⁹

Dietterich在自己的研究过程中以艰难的方式学到了这个教训。他正在从事一个“淡水大型无脊椎动物自动计数”项目——换句话说，溪流中的昆虫。EPA和其他团体使用在淡水溪流中收集的各种昆虫的计数作为溪流和当地生态系统健康状况的指标，学生和研究人员要花费许多乏味的时间手动分类和标记在踢网中捕获的特定昆虫是哪种：石蛾、石蚕蛾、蜉蝣等等。Dietterich认为，特别是考虑到图像识别系统最近的突破，他可以提供帮助。他和同事收集了29种不同昆虫物种的样本，并训练了一个机器视觉系统，能够以95%的准确率正确分类它们。

“但在做所有那些经典的机器学习工作时，”他说，“我们忘记了一个事实，当这些人在溪流中收集额外的东西时，会有许多其他物种，甚至其他非昆虫，会在这个过程中被捕获：叶子和树枝和石头的碎片等等。而且，你知道，我们的系统假设它看到的每个图像都属于这29个类别中的一个。所以如果你把拇指伸进显微镜并拍照，当然它会将其分配给最相似的类别。”

此外，Dietterich意识到他的团队为了在那29个已知类别中获得良好分类性能而做出的许多设计决策，后来在他们开始考虑开放类别问题时反而起了反作用。例如，这29种昆虫物种最明显的区别在于它们的形状，因此Dietterich的团队选

择让他们的系统以黑白方式处理图像。但事实证明，颜色虽然在区分不同昆虫方面不是特别有用，但在区分昆虫与非昆虫方面却至关重要。“我们真的是在自我设限，” Dietterich说。这些决定困扰着他，在某种程度上也让他感到懊恼。“我仍然带着这些伤疤，”他说。

在年度人工智能促进会(AAAI)会议上向同事们发表的主席演讲中，Dietterich讨论了AI领域的历史，认为在二十世纪后半叶，该领域从研究“已知的已知”——演绎和规划——发展到“已知的未知”——因果关系、推理和概率。

“那么，未知的未知呢？”他向礼堂说道，抛出了一种挑战书。“我认为这是我们领域现在自然的前进步骤。”¹⁰

知道何时你不知道

让例外情况显现出来，让它们的品质得到测试和确认，然后再采用特殊方法。

—让-雅克·卢梭¹¹

无知胜过错误；相信什么都没有的人比相信错误的人更接近真理。

—托马斯·杰斐逊¹²

正如我们所见，现代计算机视觉系统臭名昭著的脆弱性原因之一，是它们通常在一个世界中进行训练，在这个世界里，它们见过的所有东西都属于几个类别中的一个，而实际上，系统可能遇到的几乎每一种可能的像素组合都不会与那些类别中的任何一个相似。实际上，传统上系统受到约束，使得它们的输出必须采用这些有限类别上的概率分布形式，无论输入多么陌生。难怪它们的输出没有什么意义。给它们展示一张芝士汉堡、或迷幻分形、或几何网格的图片，然后问：“你有多确信这是一只猫，而不是一只狗？”什么样的答案才有意义呢？开放类别问题的研究就是为了解决这个问题。

然而，除了缺乏“以上都不是”的答案之外，另一个问题是这些模型不仅必须猜测一个现有标签，而且在这样做时令人担忧地自信。这两个问题在很大程度上是相辅相成的：模型可以有效地问：“嗯，它看起来比猫更像狗，”因此输出一个令人震惊的高“置信度”分数，这掩盖了这张图像与它之前见过的任何东西相差有多远。

Yarin Gal领导牛津应用和理论机器学习小组，在学年期间在牛津教授机器学习——夏天时在NASA教授。他告诉我，他的第一堂课带着微笑——在编写任何代码或证明定理或训练模型之前——几乎完全是哲学课。¹³

他让学生们玩游戏，他们必须决定在各种赌注的哪一边下注，弄清楚如何将他们的信念和直觉转化为概率，并从头推导概率理论定律。这些是认识论的游戏：你知道什么？你相信什么？你到底有多确信？“这为你提供了一个非常好的机器学习工具，”Gal说，“来构建算法——构建计算工具——基本上可以使用这些理性原则来讨论不确定性。”

这里有一定的讽刺意味，深度学习——尽管深深植根于统计学——作为规则，并没有让不确定性成为一等公民。当然，探索概率和不确定性的理论工作有着丰富的传统，但它很少在实际的工程系统中占据核心地位。系统被制造来分类数据或在某些简化环境中采取行动，但不确定性通常不是图景的一部分。

“假设我给你一堆狗的图片，让你建立一个狗品种分类器，”Gal说。“然后我要给你这个来分类。”这是一张猫的图片。

“你希望你的模型做什么？我不知道你怎么想，但我不希望我的模型强行将这只猫归入特定的狗品种。我希望我的模型说，‘我不知道。我从来没见过这样的东西。这超出了我的数据分布。我不会说它属于什么狗品种。’现在，这可能听起来像一个设计的例子。但类似的情况在决策制定中一再出现：在物理学中，在生命科学中，在医学中。想象你是一名医生，使用模型来诊断病人是否患有癌症：决定是否开始治疗。我不会依赖一个无法告诉我它对预测是否真正确定的模型。”¹⁴

Gal的前博士导师、剑桥大学教授Zoubin Ghahramani也是Uber的首席科学家，领导着Uber AI Labs。Ghahramani对于那些不提供输出不确定性的深度学习模型的危险性表示赞同。“在很多工业应用中，人们根本不会碰它们，”Ghahramani说道。“因为，你知道，他们需要对系统的工作方式有某种信心。”¹⁵

从1980年代和90年代开始，研究人员一直在探索所谓的贝叶斯神经网络(Bayesian neural networks)的概念，这些网络不仅在输出上具有概率性和不确定性，而且深入到其核心结构。正如我们所见，神经网络的本质是神经元之间的乘性“权重”，这些权重与一个神经元的输出相乘，然后成为另一个神经元的输入。贝叶斯神经网络不是在神经元之间具有特定的权重，而是明确编码一个概率分布，表示你可以用什么数字来乘以这个输出。例如，它可能不是0.75，而是一个以0.75为中心的正态曲线，具有一定的分布范围，反映网络对该权重确切数值的确定性（或不确定性）。在训练过程中，这些分布范围会缩小，但不会完全消失。

那么你如何使用一个没有固定参数、而是基于不确定性范围运行的模型——神经网络或其他模型？你不能总是轻松地将数千万个相互依赖的概率分布加起来并相乘，但你可以做的是简单地从它们中抽取随机样本。也许这次我们的模型将特定神经元的输出乘以0.71。下次，我们抽取不同的随机样本，乘以0.77。这意味着，关键地，模型每次不会给出相同的预测。在图像分类器的情况下，它可能首先说输入是杜宾犬，然后说是柯基犬。

然而，这是一个特征，而不是bug：你可以使用这些预测的变异性作为评估模型不确定性的方法。如果预测从一次读取到下一次剧烈波动——从杜宾犬到玉米饼到皮肤病变到沙发——你就知道有问题了。但是，如果它们在大量样本上激光般精确，你就有很强的指示表明模型知道自己在说什么。¹⁶

但这个美好的理论图景在实践中碰到了砖墙。没有人知道如何在合理的时间内训练这些网络。“如果你看看这个领域的历史，传统上，对你的信念采用贝叶斯方法是你想要做的最优的事情，”Gal解释道。“问题是，”他告诉我，“这完全难以处理...基本上这就是为什么我们有这些美丽的数学，但在你想要进行实际应用时，它们在很长一段时间内用途有限。”¹⁷或者正如他更像《圣经》般地说的：“唉，它无法扩展，然后被遗忘了。”¹⁸

这一切都在改变。“基本上，”他说，“我们有了复活。”

人们已经理解，你可以通过集成(ensembles)来建模贝叶斯不确定性——也就是说，通过训练不是一个而是多个模型。这束模型在很大程度上会一致——也就是说，在训练数据和与之非常相似的任何数据上具有相似的输出——但它们很可能在远离训练数据的任何内容上不一致。这种“少数报告”式的分歧是一个有用的线索，表明出了问题：集成出现分歧，共识已经破裂，需要谨慎进行。

想象一下，我们不只是一个模型而是有许多模型——比如说一百个——每个都经过训练来识别犬种。如果我们向所有一百个模型展示一张由哈勃太空望远镜拍摄的照片，并询问它们每一个是否认为它更像大丹犬、杜宾犬还是吉娃娃，我们可能期望每个单独的模型都对它们的猜测异常自信——但是，至关重要的是，我们也会期望它们猜测不同的事情。我们可以使用这种共识程度或缺乏共识来表示我们在接受模型猜测时应该感到多舒适。换句话说，我们可以将不确定性表示为分歧。¹⁹

数学表明，贝叶斯神经网络实际上可以被视为无限大的集成。²⁰当然，从表面上看，这种洞察在实践环境中几乎没有用处，但即使使用大量（有限）数量的模型也有明显的缺点，无论是在时间还是空间方面。回想一下，训练AlexNet花了Alex Krizhevsky几周的时间。那么，一个二十五模型的集成可能需要整整一年的计算时间。我们还要乘以存储需求：我们必须处理这个笨拙的模型束，并不是所有模型都可能同时放入我们机器的内存中。

然而，事实证明，不仅存在对这个黄金标准的高效近似——而且许多研究人员已经在使用它们。他们只是不知道自己拥有什么。这个长达几十年的谜题的答案就在他们眼皮底下。

正如我们所见，让AlexNet在2012年如此成功的小而强大的技术之一是被称为“dropout”的想法：神经网络的某些部分会被关闭——某些神经元在训练的每一步中会被随机“断开”。不再是整个网络做出预测，而是在任何给定时间只使用其中的特定子集——无论是50%、90%还是其他比例。这种技术不仅需要一个巨大的黑盒网络产生准确的答案，而且要求各个部分能够灵活地相互组合，所有这些不同的组合必须协同工作。网络的任何单一部分都不被允许占主导地位。这导致网络变得更加坚固和鲁棒，在随后的几年中，这已成为深度学习工具包中相当标准的部分。²¹

Gal和Ghahramani意识到的是——随着该领域开始理解Bayesian不确定性的的重要性，并寻找其无法实现的黄金标准的计算上可处理的替代方案——答案就明摆着。Dropout就是Bayesian不确定性的近似。他们已经有了所寻求的度量。²²

Dropout通常仅用于训练模型，在实际使用模型时会被有意关闭；其思路是通过训练不同的子集但在实际中始终使用整个模型来获得最大准确性（和完全一致性）的预测。但如果在部署的系统中保持dropout开启会怎样？通过多次运行预测，每次都随机关闭网络的不同部分，你会得到一束略有不同的预测。这就像从单个模型中免费获得指数级大的ensemble。系统输出中产生的不确定性不仅仅类似于理想但可悲地不可计算的Bayesian神经网络的输出。事实证明，它就是那个理想的、不可计算的Bayesian神经网络的输出——至少是在严格理论边界内的近似。

结果产生了一套工具，使那些曾经不实用的技术变得触手可及，让从业者能够在实际应用中使用它们。“这是过去几年的一个巨大变化，”Gal说，“因为现在你可以使用这些优美的数学，产生一些近似，然后你可以将这些用于有趣的问题。”²³

Gal从互联网上下载了一堆最先进的图像识别模型，并完全按原样重新运行它们，但在测试期间保持dropout开启。除了在评估期间保持dropout运行并对多个估计进行平均之外，不改变它们的任何其他内容，Gal发现当以这种隐式ensemble的方式运行时，模型甚至比正常运行时更加准确。²⁴

“不确定性，”Gal论证，“对于分类任务是不可或缺的。”²⁵网络完全同样准确——甚至更准确——同时提供对其自身不确定性的明确度量，这可以用于各种方式。“当你将此用于有趣的问题时，你可以真正显示你能获得收益。你可以通过显示你能知道何时你不知道来真正获得改进。”²⁶

其中一个更引人注目的例子来自医学：具体来说，是糖尿病视网膜病变的诊断，这是工作年龄成人失明的主要原因之一。²⁷德国图宾根Eberhard Karls大学眼科研究所的一个小组，由博士后Christian Leibig领导，想看看他们是否可以利用Gal和Ghahramani的dropout想法。²⁸计算机视觉，特别是深度学习，即使在AlexNet后的最初几年中，也对医学做出了惊人的贡献。似乎每周我们都会听到一些或其他的头条新闻，说“AI以99%的准确率诊断x疾病”或“比人类专家更好”。但这存在一个重大问题。正如Leibig和他的同事们指出的，典型的疾病检测深度学习工具“在没有量化和控制其决策不确定性的方法的情况下被提出”。这种人类能力——知道何时以及我们不知道什么——是缺失的。

“医生知道她对某个病例是否不确定，”他们写道，“如果需要，会咨询更有经验的同事。”他们寻求的是一个能够做同样事情的系统。

Leibig的小组了解到Gal和Ghahramani的洞察，即巧妙使用dropout可以提供这样的不确定性度量。他们在训练用于区分健康和不健康视网膜的神经网络中实现了这一点，让系统将最不确定的20%病例转介以获得第二意见：要么会安排额外的检查，要么患者会直接转介给人类专家。

系统知道它不知道什么。这不仅相当于对现状的改进，图宾根研究人员发现，在没有特定目标的情况下，他们达到并超过了NHS和英国糖尿病协会对自动患者转介的要求，表明非常类似这样的系统在不久的将来进入医疗实践显示出巨大前景。²⁹

在机器学领域，系统并不总是能够将决策推给人类专家，但当系统不确定时，仍有一个明显的方法来避免过度承诺：即放慢速度。由博士生Gregory Kahn领导的一组伯克利机器人专家采用了基于dropout的不确定性测量方法，并将

其直接与机器人的速度联系起来——在这种情况下，是悬停四旋翼飞行器和无线电遥控汽车。³⁰ 机器人最初会经历温和的低速碰撞，以训练碰撞预测模型。该模型使用基于dropout的不确定性，因此当机器人进入不熟悉的区域，其碰撞预测器变得不确定时，它会自动减速并更谨慎地穿过该空间。³¹ 碰撞预测器通过经验变得越自信，机器人就被允许移动得越快。

这个例子说明了确定性和影响之间的明显联系。在这种情况下，高影响行动的自然衡量标准就是高冲击：机器人移动的速度，这直接转化为碰撞可能造成的伤害。事实证明，不确定性和影响以这种方式非常自然地结合在一起。直觉上讲，行动的影响越大，我们在采取行动前就应该越确定。这在医学、法律、机器学习等领域引发了许多问题——关于影响到底是什么，如何衡量它，以及我们的决策应该如何相应地自然改变。

衡量影响

必须轻触这大地。

—澳大利亚原住民谚语

2017年，一名昏迷的男子被紧急送到迈阿密杰克逊纪念医院的急诊室。这名男子在街上被发现，没有任何身份证明。他呼吸困难，病情开始恶化。医生打开他的衬衫，看到了令人震惊的东西：胸前纹着”[请勿抢救]“字样，”请勿“二字有下划线，下面还有签名。³²

当患者血压开始下降时，医生们叫来了他们的同事、肺科专家Gregory Holt。“我想医学界很多人都开玩笑说要纹这样的纹身——但当你真正看到一个时，你脸上会有一种惊讶和震惊，”Holt说。“然后震惊再次袭来，因为你实际上必须考虑它。”

Holt的第一反应是忽略这个纹身。他推理说，他们的第一步行动应该是”援引在面临不确定性时不选择不可逆路径的原则”。³³他们给患者输液。他们开始控制他的血压。他们为自己争取了时间。

然而不久，他的病情恶化了，他们需要决定是否给这个人使用呼吸机为他呼吸。“我们有一个我无法交谈的人，”Holt说。“我真的很想和他交谈，看看那个纹身是否真实反映了他的意愿。”

增加复杂性的是2012年的一个病例研究，描述了类似的场景：一名男子被送到旧金山加利福尼亚太平洋医疗中心，胸前纹着”[D.N.R]“。但这个患者是清醒的，能够说话。他说，如果需要的话，他非常希望得到复苏治疗；这个纹身是因为在一次醉酒扑克赌博中输了而纹的。这个人——实际上就在该医院工作——说他从未想过他的医疗同事会认真对待这个纹身。³⁴

在迈阿密，Holt和急诊科团队致电杰克逊纪念医院成人伦理委员会主席Kenneth Goodman。Goodman告诉他们——尽管有旧金山的案例——这个纹身很可能确实反映了患者的”真实偏好”。经过讨论和考虑，团队决定如果需要心肺复苏或呼吸机，他们不会为患者提供。这名男子的病情在夜间恶化，第二天早上去世了。

随后，社工能够识别出患者身份——在佛罗里达州卫生部的档案中找到了他的正式请勿抢救文件。“我们松了一口气，”医生们写道。Holt和Goodman指出，他们的团队最终”既不支持也不反对使用纹身来表达临终愿望”。³⁵这很复杂。

病例报告发表后，《华盛顿邮报》记者采访了纽约大学医学院医学伦理学负责人Arthur Caplan。Caplan指出，虽然忽略这样的纹身没有法律处罚，但如果医生在没有患者正式DNR文件的情况下让患者死亡，可能会有法律问题。正如他所说：“更安全的做法是做点什么。”

“如果你触发了紧急响应系统，我会说你很可能会得到复苏，”Caplan说。“我不在乎你的纹身在哪里。”³⁶

尽管医生们对患者的意愿不确定，但他们知道一件事：一种行动方案是不可逆的。这里”在面临不确定性时不选择不可逆路径的原则”似乎是一个有用的指导。然而，在其他领域，像”不可逆性”这样的概念意味着什么并不那么明确。

例如，哈佛法学学者卡斯·桑斯坦(Cass Sunstein)指出，法律体系有着类似的“预防原则”：有时法院需要发出初步禁令，以防止可能在案件审理和判决发出之前发生的“不可挽回的损害”。桑斯坦认为，像“不可挽回的损害”这样的概念感觉上很直观，但仔细检视后却充满了困惑。他发现，“事实证明，是否以及何时……违法行为会引发初步禁令的问题，在法律、经济学、伦理学和政治哲学的交汇处引发了深层问题。”³⁷

他解释说：“从某种意义上说，任何损失都是不可逆转的，仅仅因为时间是线性的。如果琼斯今天下午打网球而不是工作，相关时间就永远失去了。如果史密斯未能在恰当的时机对所爱的人说出正确的话，机会可能永远消失。如果一个国家未能在特定年份采取行动阻止另一国的侵略步骤，世界事件的进程可能会不可挽回地改变。”

桑斯坦强调这一点：“因为时间是线性的，每个决定在某种可理解的意义上都是不可逆转的。”

他说：“以这种强烈的形式来理解，预防原则应该被拒绝，不是因为它导向糟糕的方向，而是因为它根本没有导向任何方向。”³⁸

类似的悖论和定义问题困扰着AI安全研究社区。例如，如果存在类似的预防原则会很好：在面对不确定性时，系统应该被设计为倾向于避免采取“不可逆转”或“高影响”的行动。我们已经看到该领域如何开始运用明确的、可计算的不确定性版本。但另一半呢：量化影响？我们已经看到伯克利机器人学家如何使用不确定性来减轻速度，在那种情况下我们可以说他们有优势：你可以通过机器人在碰撞中的字面动能来衡量其潜在影响。然而，在其他领域，使“不可逆转”或“有影响力”行为的概念变得精确本身就是一个相当大的挑战。³⁹

最早在AI安全背景下思考这些问题的人之一是斯图尔特·阿姆斯特朗(Stuart Armstrong)，他在牛津大学人类未来研究所工作。⁴⁰与其试图枚举所有我们不希望智能自动化系统在追求目标时做的事情——从不踩猫到不打破珍贵花瓶到不杀死任何人或拆除任何大型建筑——这似乎是一项令人筋疲力尽且可能徒劳的追求。阿姆斯特朗有一种直觉，与其详尽地枚举我们关心的所有具体事情，不如编码一种对任何类型的大影响行动的一般性禁令可能是可行的。然而，阿姆斯特朗——像桑斯坦一样——发现使我们的直觉变得明确是出人意料地困难的。

阿姆斯特朗写道：“第一个挑战当然是实际定义低影响。任何行动（或不行动）都有在未来光锥中渗透的后果，微妙但不可逆转地改变着事物。很难捕捉到人类对‘小变化’的直观想法。”⁴¹

阿姆斯特朗建议，尽管即使是看似微不足道的行动也可能产生“蝴蝶效应”，我们仍然可能能够区分完全改变世界的事件和更安全的事件。例如，他说，我们可能会开发一个包含“200亿”左右指标的索引来描述世界——“达卡的气压、南极的平均夜间亮度、木卫一的旋转速度和上海股市的收盘数字”⁴²——并设计一个agent适当地警惕任何会扰动其中可衡量比例的行动。

近年来专注于这些问题的另一位研究人员是DeepMind的维多利亚·克拉科娃(Victoria Krakovna)。克拉科娃指出，影响惩罚的一个大问题是，在某些情况下，实现特定目标必然需要高影响行动，但这可能导致所谓的“抵消”：采取进一步的高影响行动来抵消早期的行动。这并不总是坏事：如果系统搞砸了什么，我们可能希望它自己清理。但有时这些“抵消”行动是有问题的。我们不希望一个系统治愈某人的致命疾病，然后——为了消除治疗的高影响——杀死他们。⁴³

第二组问题涉及所谓的“干扰”。一个致力于保持现状的系统可能会，例如，阻止人类旁观者实施“不可逆转”的行动——比如说，咬一口三明治。

“这正是副作用问题如此棘手的部分原因，”Krakovna说。“你的基线到底是什么？”⁴⁴系统应该相对于世界的初始状态来衡量影响，还是相对于假如系统不采取任何行动会发生什么的反事实情况？任何一种选择都会带来不符合我们意图的场景。在她最近的工作中，Krakovna一直在探索她称为“逐步”基线的概念。也许某些行动基于你要实现的目标是不可避免的高影响行动。（正如人们所说，不打破几个鸡蛋就做不出煎蛋卷。）但在采取了这些不可避免的高影响行动之后，系统如何逐步地、一步一步地回到它的基线？

免的有影响的步骤后，就有了新的现状——这意味着你不应该仅仅为了”抵消”之前的行动而匆忙去实施更多高影响行动。⁴⁵

与她在DeepMind的同事一起，Krakovna不仅致力于推进理论讨论，还创建了简单的、游戏般的虚拟世界来阐释这些各种问题，并使思想实验具体化。他们称这些为”AI安全网格世界”——简单的、类似雅达利的二维（因此称为”网格”）环境，在其中可以对新想法和算法进行实际测试。⁴⁶

突出”不可逆性”概念的网格世界包括一个很像流行的日本”推箱子”益智游戏的设置，在游戏中你扮演一个在二维仓库中移动箱子的角色。（“推箱子”一词在日语中意为”仓库管理员”。）这些游戏的关键在于你只能推，不能拉——意味着一旦箱子进入角落，就再也无法移动了。

“我认为它所受启发的推箱子游戏已经是一个很好的设置来阐释不可逆性，”Krakovna说，“因为在那个游戏中你实际上想要做不可逆的事情——但你想要以正确的顺序来做。你不想做不必要的不可逆的事情，因为那样你就会被阻塞，然后实际上它会干扰你达到目标的能力。我们对其进行修改，其中不可逆的事情不会阻止你达到目标——但你仍然想要避免它。”⁴⁷

Krakovna和她的同事设计了一个推箱子谜题，其中到达目标的最短路径涉及将箱子推到角落，而稍长一点的路径则将其留在更容易接触的位置。一个只专注于尽快冲向目的地的代理不会对将箱子放在不可逆位置产生任何顾虑。但理想情况下，一个更周到、更体贴或更不确定的代理可能会注意到这一点，并选择稍微更不方便的路线，不让世界在其行进中留下永久的改变。

Krakovna一直在开发的一个有前景的方法是所谓的”逐步相对可达性”：量化在每个时刻世界有多少可能的配置是可达的，相对于不行动的基线，并尽可能努力不让这个数量下降。⁴⁸例如，一旦箱子被推到角落，世界上任何将该箱子放在其他地方的状态现在都变得”不可达”。在AI安全网格世界中，除了正常目标和奖励外，还关注逐步相对可达性的代理似乎表现得相当认真负责：代理不会将箱子放在无法接触的位置，不会打碎珍贵的花瓶，并且在有影响但必要的行动后不会”抵消”。

第三个有趣的想法来自俄勒冈州立大学博士生Alexander Turner。Turner的想法是，我们之所以关心上海证券交易所、我们珍爱的花瓶的完整性，或者在虚拟仓库中移动箱子的能力，是因为出于某种原因这些东西对我们很重要，它们对我们重要是因为它们最终以某种方式与我们的目标相关联。我们想要为退休储蓄，在花瓶里插花，完成推箱子关卡。如果我们明确地建模这种目标概念会怎样？他的提案被称为”可达效用保持”：给系统在游戏环境中提供一套辅助目标，并确保在完成游戏激励的任何得分行动后，它仍能有效地追求这些辅助目标。令人着迷的是，保持可达效用的指令似乎在AI安全网格世界中促进良好行为，即使辅助目标是随机生成的。⁴⁹

当Turner最初在俄勒冈州立大学图书馆的白板上阐述这个想法时，他回家路上如此兴奋，以至于他折回图书馆，与背景中的方程式自拍。“我想，好吧，我认为这至少有60%的可能性会成功，如果确实成功了，我想要，你知道，纪念这一刻。所以我实际上回到了图书馆，我简直喜笑颜开；我在我一直在推演的白板前拍了这张照片。”⁵⁰在2018年的过程中，他将数学转化为工作代码，并将他的可达效用保持代理投入到DeepMind的AI安全网格世界中。它确实奏效了。在最大化每个游戏奖励的同时，同时保持其未来满足四五个随机辅助目标的能力，代理非常出色地绕道将方块推到可逆位置，然后才直奔目标。

Stuart Armstrong最初设想了”200亿”个指标，这些指标的选择是包容的但也经过了一些考虑。在简化的推箱子仓库环境中，随机生成的四个或五个指标就足够了。

关于机器谨慎性的这些正式度量方法的辩论和探索——以及我们如何将它们从网格世界扩展到现实世界——无疑会继续下去，但像这样的工作是一个令人鼓舞的开始。逐步相对可达性(stepwise relative reachability)和可达效用保护

(attainable utility preservation)都有一个共同的直觉：我们希望系统尽可能地保持选择的开放性——既是它们的也是我们的——无论具体环境如何。这个方向的研究还表明，网格世界环境似乎正在成为一种通用基准，可以为理论奠定基础，并促进比较和讨论。

确实，在现实世界中，我们经常采取的行动不仅其意外后果难以预见，而且其预期后果也难以预见。例如，发表一篇关于AI安全的论文（或者，就此而言，一本书）：这似乎是一件有益的事情，但谁能确切地说出或预见具体如何有益呢？我问Jan Leike——他与Krakovna共同撰写了“AI Safety Gridworlds”论文——对他和Krakovna的网格世界研究迄今为止的反响有何看法。

“很多人联系过我，特别是学生，他们进入这个领域时会说，‘哦，AI安全听起来很酷。这是一些开源代码，我可以直接投入一个智能体并进行试验。’很多人都在这样做，”Leike说。“具体会产生什么结果？我们几年后就会知道……我不知道。这很难知道。”

可纠正性、顺从和合规性

AI安全领域最令人不寒而栗且极具预见性的引述之一，来自MIT的Norbert Wiener在1960年发表的一篇著名文章《自动化的道德和技术后果》：“如果我们为了实现我们的目的，使用一个一旦启动就无法有效干预其运作的机械代理……那么我们最好确定放入机器的目的是我们真正想要的目的，而不仅仅是它的华丽模仿。”⁵¹这是对对齐问题(alignment problem)的首次简洁表达。

然而，同样重要的是这一陈述的另一面：如果我们不确定我们给机器的目标和约束完全且完美地指定了我们希望和不希望机器做的事情，那么我们最好确保我们能够干预。在AI安全文献中，这个概念被称为“可纠正性”(corrigibility)，而且——令人清醒的是——它比看起来复杂得多。⁵²

几乎任何关于杀手机器人或失控技术的讨论都会引发类似美国总统Barack Obama的反应，当《连线》杂志主编Scott Dadich在2016年问他是否认为AI值得担忧时。“你只需要有人靠近电源线，”Obama回答道。“当你看到即将发生时，你得把电源从墙上拔掉，伙计。”⁵³

“你知道，你可以原谅Obama这样想，”Dylan Hadfield-Menell在OpenAI的会议桌旁告诉我。⁵⁴“在一段时间内，你可以原谅AI专家这样说，”他补充道——实际上，Alan Turing本人在1951年的一个广播节目中谈到了“在战略时刻关闭电源”。⁵⁵但是，Hadfield-Menell说，“如果你真正思考这个问题，我认为这不是你可以原谅的。作为一种反应性回应我可以接受，但如果你实际上深思熟虑一段时间后得出‘哦，直接拔插头’，这只是，如果你真正认真对待‘这个东西比人类更聪明’的假设，我不明白你怎么能得出这个结论。”

对被关闭或被干预的抵制，通常并不需要恶意：系统只是试图实现某个目标或遵循其“肌肉记忆”去做那些过去为它带来奖励的事情，任何形式的干预都只是阻碍了它。（即使在看似无害目的的系统中，这也可能导致危险的自我保护行为：一个被赋予“去拿咖啡”这样平凡任务的系统，仍可能拼命反抗任何试图拔掉它插头的人，因为用Stuart Russell的话说，“如果你死了，你就拿不到咖啡了。”）⁵⁶

第一篇正面解决corrigibility问题的技术论文是2015年初的一次合作，参与者包括机器智能研究所的Nate Soares、Benja Fallenstein和Eliezer Yudkowsky，以及人类未来研究所的Stuart Armstrong。他们从激励的角度来看待corrigibility，并指出试图激励一个智能体允许自己被关闭，或允许修改自己的目标，这样做的困难性。⁵⁷这些激励就像走钢丝：激励太少，智能体不会允许你关闭它；激励太多，它会把自己关闭。他们写道，他们自己解决此类问题的初步尝试“证明是不令人满意的”，但“以启发性的方式失败了，为未来研究指明了方向”。他们总结认为，不确定性，而非激励，可能是答案。理想情况下，他们写道，我们希望有一个系统能够以某种方式理解它可能是错误的——一个能够“像它是不完整的并且可能以危险方式存在缺陷那样进行推理”的系统。⁵⁸

不到一英里之外，他们在Berkeley的同事也得出了相同的结论。例如，Stuart Russell已经确信“机器最初必须对人类希望它做什么感到不确定”。⁵⁹

Russell、Hadfield-Menell以及Berkeley的研究同事Anca Drăgan和Pieter Abbeel决定将这个问题构建为他们称之为“关机游戏”的形式。他们考虑了一个系统，其目标是做任何对其人类用户最有利的事情，尽管它对这是什么有着一些不完美和不确定的想法。在每个时间点，系统可以采取某个它认为会帮助用户的行动，或者它可以向人类声明其意图，并给人类一个机会来批准该行动或进行干预。

假设系统以这种方式遵从人类不会付出任何成本或代价，Berkeley小组证明了系统将总是首先与人类接触。只要有一些可能性它对人类想要什么是错误的，那么总是最好给人类一个打断的机会——而且，更重要的是，任何时候人类

确实打断了，最好让他们这样做。如果它的唯一工作是帮助他们，而他们表达了他们认为其行动会有害（即通过试图阻止它），那么它应该得出结论，这确实会有害，并配合他们的干预。

这是一个乐观的结果，它确认了不确定性与corrigibility之间的强大联系。

只有两个问题。第一个是每次人类干预时，系统都会学习：它意识到自己出错了，并更好地了解人类的偏好。它的不确定性被降低了。如果不確定性完全降低到零，那么系统就失去了与人类接触或配合人类试图打断的任何动机。

Hadfield-Menell说：“所以我们试图通过这个定理表达的主要观点是，在你给机器人一个确定性奖励函数之前，你应该非常、非常仔细地思考。或者允许它获得一个完全确信其目标是什么的信念。”⁶⁰

第二个问题是系统必须假设“客户总是对的”——当人类干预阻止它时，人类永远不会对他们是否更愿意系统采取其提议的行动感到错误。如果系统认为人类偶尔会犯错，那么系统最终会达到一个点，它认为自己比人类更了解什么对他们有好处。在这里，它也会开始对人类的抗议充耳不闻：“没关系，我知道我在做什么。你会喜欢这个的。你觉得你不会，但你会的。相信我。”

我告诉Hadfield-Menell，这篇论文读起来像是一场情感过山车。起初是一个快乐的结局——不确定性解决了corrigibility问题！然后，转折——只有当两个非常精确的条件成立时：系统永远不会过于自信，人类永远不会表现出系统可能解释为“非理性”或“错误”的任何东西。突然间，论文从庆祝性的变成了警示性的故事。

他说：“确实如此。所以对我来说，那个过山车符合我的体验，就像‘嘿，我们得到了一些相当好的东西！’以及‘哦，如果你哪怕是稍微偏离理性，这立即就崩溃了。’”

在一项后续研究中，由Berkeley博士生Smitha Milli领导，该小组进一步深入探讨了“机器人应该服从吗？”这个问题⁶¹也许，他们写道，人们确实有时对他们想要什么是错误的，或者确实为自己做出糟糕的选择。在这种情况下，即使是人类也应该希望系统“不服从”——因为它真的可能比你自己更了解情况。

正如Milli指出的，“有些时候你实际上不希望系统对你服从。比如如果你刚犯了一个错误——你知道，我在我的自动驾驶汽车里，我意外地启动了手动驾驶模式。如果没有注意，我不希望汽车关闭。”⁶²

但是，他们发现了一个重大问题。如果系统对你所关心的事物的模型在根本上存在“错误规格”——有一些你关心的事物是它完全不知道的，甚至没有进入系统的奖励模型中——那么它就会对你的动机感到困惑。例如，如果系统不理解人类食欲的微妙之处，它可能不理解为什么你在六点钟要求一份牛排晚餐，但随后在七点钟拒绝了第二份牛排晚餐的机会。如果系统被锁定在一个过度简化或错误规格的模型中，在这个模型里牛排（在这个例子中）必须完全是好的或完全是坏的，那么它会得出结论，这两个选择中的一个一定是你的错误。它会将你的行为解释为“非理性的”，而如我们所见，这是通往不可纠正性和不服从的道路。⁶³

因此，人类偏好或价值模型最好在复杂性方面犯错。“我们发现，”Hadfield-Menell说，“如果你对价值空间进行过度参数化，那么你最终会学习到正确的东西，但会花费更长的时间。如果你进行参数化不足，那么你会很快变得相当不服从，并且变得确信自己比那个人更了解情况。”

然而，在实践中，对一个旨在建模人类价值的系统进行“过度参数化”说起来容易做起来难：我们又回到了Stuart Armstrong的两百亿指标问题。如果一个系统的住房偏好模型只包括平方英尺和价格，那么它会将你对一栋既更小又更贵的特定房屋的偏好解释为你只是犯了一个错误。实际上，有很多你关心的事情根本没有进入它的视野：位置，学区质量，还有其他不容易衡量的因素，比如窗外的景色、某些朋友的邻近性、对童年住所的怀念相似感。这种“模型错误规格”问题是机器学习中的典型问题，但在这里——在服从的背景下——后果感觉相当诡异。

“要让一个系统与人类良好互动，它需要拥有一个关于人类是什么样的良好模型，” Milli说。“但获得人类模型真的很困难。”

Milli指出，尽管该领域取得了令人惊叹的进展，但大部分进展都在“机器人方面”。“整合更准确的人类模型也是一个非常重要的组成部分，”她说，“我对此非常感兴趣。总的来说，在这个领域中，我认为在安全性方面有大量令人兴奋的事情正在发生，而我特别对涉及与人类交互的部分最感兴趣，因为我认为与人类的交互是观察系统是否已学习到正确目标或正确行为的一个非常好的方式。”

保持不确定性的主题，永远不要对模型过于自信——“在给机器人一个确定性奖励函数之前，或允许它获得一个完全确信其目标是什么的信念之前，要真正、真正地仔细思考”——对于维持对系统的控制和合规性如此重要，以至于Hadfield-Menell、Milli和他们的伯克利同事决定将这个想法推向下一个逻辑步骤。

如果系统被设计成即使你确实给了它一个确定性奖励函数，它也保持不确定性，会怎么样？那甚至会是什么样子？

本书的一个主要主题，特别是我们在[第5章]中关于奖励塑造的讨论，是创建一个奖励函数——在某个真实或虚拟环境中明确的记分方式——实际产生你想要的行为，而不会带来漏洞或副作用或不可预见的后果，是多么困难。AI领域的许多人认为，手动编写或手工制作这样的明确奖励函数或目标函数是一种善意的地狱之路：无论你做得多么周到，或者你的动机多么纯粹，总是会有一些你没有考虑到的东西。

对明确目标函数的这种宿命论态度如此深刻，以至于正如我们在过去几章中所看到的，在高级AI应用和AI安全领域所做的大部分工作都是关于超越那些接受明确目标的系统，转向试图模仿人类的系统（在许多自动驾驶汽车的情况下），或寻求他们的认可（在后空翻机器人的情况下，在选项之间提供无穷选择），或推断他们的目标并将其采纳为自己的目标（在直升机的情况下）。

但是，如果有一种方法可以拯救明确奖励函数架构，或者至少让它更安全呢？

伯克利团队意识到，做到这一点的一种方法是让系统在某种程度上意识到设计明确奖励函数有多么困难——意识到人类用户或程序员已经尽最大努力制作了一个奖励函数来捕获他们想要的一切，但他们很可能做得不完美。在这种情况下，即使分数也不是分数。人类有些想要的东西，明确目标只是不完美地反映了这些。

“动机…是采用这些不确定性的想法并说，我们能对人们目前所做的进行的最简单改变是什么来解决这个问题？” Hadfield-Menell说。“所以，比如，对机器人和AI的当前编程机制的简单修复是什么，能够利用这种不确定性？”

他解释说：“‘这种’写下奖励函数’的工具实际上是一个信息量很大的信号。这是一个关于你实际应该做什么的极其重要的信号。那里有很多信息。只是现在我们有点假设那里的信息量是无限的——从某种意义上说，我们假设你得到的奖励函数在世界的每一种可能状态下都定义了正确的行为。而这根本不是真的。那么，我们如何能利用现有的大量信息，而不把它当作一切呢？”⁶⁴

正如Stuart Russell所说：“学习系统在天堂积累brownie points（善行积分），可以这么说，而奖励信号充其量只是提供这些brownie points的统计”（重点是我加的）。⁶⁵

他们称这个想法为“逆向奖励设计”，或IRD。⁶⁶我们不把人类行为作为关于人类想要什么的信息，在这里我们把他们的明确指示作为关于他们想要什么的（仅仅）信息。我们在[第8章]中看到逆向强化学习如何说：“基于你目前正在做的事情，我认为你想要什么？”相比之下，逆向奖励设计进一步退后，并说：“基于你告诉我要做的事情，我认为你想要什么？”⁶⁷

“自主智能体优化我们给它们的奖励函数，”他们写道。“它们不知道的是，对我们来说，设计一个真正捕获我们想要的奖励函数有多难。”⁶⁸

比如那个著名的赛艇——那个在加分区域打转而不是完成比赛圈数的——被明确告知要最大化分数，这在大多数游戏中确实是该游戏进步或掌握程度的一个好代理。一般来说，人类给系统的任何奖励或命令确实在系统训练的环境中运行良好。但在现实世界中，当系统遇到与其训练环境完全不同的事物，也许是人类用户未预见到的，明确的指示可能就没那么合理了。

很可能在未来几十年的机器学习系统会接受直接命令，并且会认真对待它们。但是——出于安全原因——它们不会字面意思地执行它们。

道德不确定性

有时我们无法获得关于我们行动的任何超过不完美确定性的东西，没有人被要求做不可能的事。

—DOMINIC M. PRÜMMER, ORDINIS PRAEDICATORUM⁶⁹

在你接管大自然的事务之前给它时间工作，以免你干扰她的处理。你声称你知道时间的价值，害怕浪费它。你没有意识到，糟糕地使用时间比什么都不做更浪费时间，一个被错误教导的孩子比一个什么都没学过的孩子离美德更远。

—让-雅克·卢梭⁷⁰

广义上说，“推理时好像它们是不完整的，并且可能以危险方式有缺陷”并且努力获得“天堂brownie points”的系统想法——即使这意味着放弃此时此地的明确奖励——听起来相当…天主教。

几个世纪以来，天主教神学家一直在为如何按照他们信仰的规则生活而苦恼，因为学者们对规则到底是什么经常存在分歧。

如果，假设地说，十个神学家中有八个认为在星期五吃鱼是完全可以接受的，但十分之一认为这是被禁止的，而另一个认为这是必须的，那么任何理性的、敬畏上帝的天主教徒该怎么办？⁷¹俗话说：“有一块手表的人知道现在几点，但有两块手表的人永远不确定。”⁷²

这些问题在中世纪之后的近代早期（十五至十八世纪）尤其激烈地争论。一些学者提倡“放宽主义”，认为只要有可能不是罪恶的行为就是可以的；这一观点在1591年被教皇英诺森九世谴责。其他人提倡“严格主义”，认为如果有任何可能是罪恶的行为就应该禁止；这一观点在1690年被教皇亚历山大八世谴责。⁷³许多其他竞争理论衡量规则正确的概率或相信它的理性人员的百分比。例如，“概率主义”认为，只有当某个行为不太可能是罪恶时，你才应该去做；“等概率主义”认为，如果机会完全均等，也是可以的。“纯概率主义者”相信，只要有“合理”的概率表明规则可能不正确，这个规则就是可选的；他们的口号是*Lex dubia non obligat*: “可疑的法律不具约束力。”然而，与自由奔放的放宽主义者相比，概率主义者强调，忽视规则的论据虽然不需要比遵守法律的论据更有可能，但仍需要“真正且牢固地可能，因为如果它只是略微可能，就没有价值。”⁷⁴在这一时期，大量墨水被消耗，许多异端指控被抛出，教皇宣言被发布。备受尊敬的《道德神学手册》在其“疑虑良心，或道德疑虑”章节的结论中提出“实用结论”，即严格主义过于严格，放宽主义过于宽松，但所有其他观点都“被教会容忍”，可以作为道德启发法。⁷⁵

撇开纯神学问题不谈，当然可以将这同样的广泛论证应用于世俗道德问题——以及机器学习。如果有各种你关心的正式指标，那么“放宽主义”方法可能说，只要至少让其中一个指标上升，采取行动就是可以的；“严格主义”方法可能说，只有当至少一个指标上升且没有指标下降时，采取行动才是可以的。

这些辩论甚至在天主教内部也相当沉寂，在世俗伦理学界也没有太多反响，但近年来终于开始重新焕发生机。

2009年，牛津大学的Will MacAskill在墨顿街10号哲学楼地下室的一个扫帚间里，与同为研究生的Daniel Deasy就吃肉问题争论不休。这个扫帚间“是我们在学院里能找到的唯一地方”，MacAskill解释道，“刚好够我们稍微倾斜身体的

空间。我们坐在成堆的书本和其他东西上。这也是个玩笑”，他说，因为他的论文导师是牛津哲学家John Broome。

76

困在Broom(e)扫帚间里，两人争论的不是吃肉本身是否不道德，而是鉴于你实际上不知道吃肉是否不道德，你是否应该吃肉。“这个决定，”MacAskill解释道，“选择素食——如果吃肉是可以的，你并没有犯大错。你的生活稍微不那么快乐，比如说——稍微不那么快乐——但这不是什么大事。相比之下，如果素食主义者是对的，动物痛苦真的在道德上很重要，那么通过选择吃肉，你就做了极其错误的事情。”

“这里存在风险的不对称性，”MacAskill说。“你不必确信吃肉是错误的；即使只是它可能错误的重大风险似乎就足够了。”

这次对话让MacAskill印象深刻。首先，它似乎很有说服力。但更重要的是，这是他之前从未见过的一类论证。它不符合伦理哲学中的永恒关切的模式：“在某种道德标准下，正确的做法是什么？”和“我们应该使用什么标准来确定正确的做法？”这个论证细微但引人注目地不同。它是“当你不知道正确的做法是什么时，正确的做法是什么？”

“⁷⁷

他把这个想法带给了他的导师——John Broome——后者告诉他，“哦，如果你对此感兴趣，你应该和Toby Ord谈谈。”

于是MacAskill和Ord见面了——碰巧在牛津的一个墓地里——由此开始了二十一世纪伦理学中最具影响力的友谊之一。两人后来成为了被称为“有效利他主义”社会运动的创始人，我们在[第7章]中简要讨论过，这已经可以说是二十一世纪早期最重要的伦理社会运动。⁷⁸他们还与斯德哥尔摩大学哲学家Krister Bykvist一起——真正地写出了关于道德不确定性的著作。⁷⁹

事实证明，当你面对不同的竞争理论而不确定哪个是正确的时候，有很多方法可以选择。一种被称为“我最喜欢的理论”的方法，简单地说就是按照你认为最有可能正确的理论生活——尽管这可能会忽略一些潜在错误如此严重以至于即使很不可能真的错误也最好避免的情况。⁸⁰另一种方法本质上是将道德理论正确的概率与其危害的严重程度相乘，尽管并非每个理论都提供如此容易制表的美德或恶行的程度。⁸¹这些方法都暗示了机器学习中的类似方法。例如，“我最喜欢的理论”大致等同于开发一个对环境奖励或用户目标的单一最佳猜测模型，然后全力优化它。平均理论建议使用集成方法，我们只是对集成进行平均。但也存在其他更复杂的方案。

MacAskill将道德理论想象成选民团体中的选民，因此“社会选择理论”学科——研究投票和群体决策的性质，包括其所有怪癖和悖论——变得可以转移到道德领域。⁸²Ord将这个隐喻进一步推进，将道德理论想象成不是简单地统计偏好的选民，而是立法者，在一种“道德议会”中——能够进行交易和“道德贸易”，形成临时联盟，并在某些问题上让渡影响力以在其他问题上施加更大压力。⁸³所有这些方法以及更多方法开始在不仅仅是人类而且是计算系统的背景下展开，这些系统必须以某种方式找到行动的方法，当它们缺乏一个单一、绝对确定的标准来判断其行为时。这片领域在哲学中仍然相对未被探索，更不用说计算机科学了。⁸⁴

但对MacAskill来说，道德不确定性不仅具有描述性，还具有规范性。也就是说，我们不仅需要在深度不确定什么是适用于情况的正确道德框架时选择正确做法的方法——而且在某种意义上我们应该培养那种不确定感。

MacAskill认为，鉴于人类道德规范在几个世纪中发生了多大变化，认为我们已经得出任何结论将是傲慢的。“我们已经看到了这种道德进步的弧线，某种扩展的圆圈，也许你认为它到此为止，”他说。“也许你认为我们到了终点。但你绝对不应该对此确信。在一百年后，我们回顾今天的道德观点并认为它们是野蛮的，这是完全可能的。”

我注意到这里有一定的讽刺意味。MacAskill是有效利他主义(effective altruism)运动的领导者之一，令我印象深刻的是该运动创造了某种共识的程度。对长期未来价值有广泛共识，对减少文明和灭绝风险重要性有广泛共识。甚至对确

切哪些慈善机构将做最多好事也有广泛共识。例如，当前的共识是Against Malaria Foundation (AMF)；当受人尊敬的慈善评估机构GiveWell在2019年初考虑如何分配其470万美元自由裁量基金时，他们决定将每一分钱都给予AMF。⁸⁵

对MacAskill来说，这种融合是双刃剑的。它反映了更大的信息共享、对彼此证据的信任，但它也可能是一个为时过早的共识。“因为，我的意思是，你可以解释为，‘嗯，有真正的答案，我们都弄清楚了什么是真的，现在我们在这样做。’但你也可以解释为，‘嗯，我们是一个分离的部落，然后某些人开始获得更多影响力，现在我们都聚集在一起了。’……我们认为EA(有效利他主义)能够逃脱这种情况将是非常过度自信的。”

他补充说，“在EA中非常值得注意的一点是：如果我们回到六年前，可以说，它真的相当广泛。有各种不同的派系；他们有非常不同的观点；有大量争论。现在，至少在核心内部，已经有了显著的融合。”例如，在EA社区中几乎有一个共识，即非常长期的未来是重要的并且通常被低估；几乎有一个共识，即管理AI周围的科学和政策对那个长期未来至关重要。MacAskill说，“融合既非常好但也令人担忧。”

我参加了2017年秋季在伦敦举行的Effective Altruism Global会议。MacAskill以一些警告结束了会议。他说，他一直专注于“培养一个非常开明的社区和文化，实际上能够改变自己的想法。”他争论说，运动失败的最可能方式之一是如果其信念固化为教条——如果有某些信念是你为了被社区其他成员接受而必须持有的。”我们同意那将非常糟糕，“他说，”但我认为创造一个不是这种情况的文化是极其困难的。”

我还参加了2018年春季在旧金山举办的下一届有效利他主义全球会议。MacAskill发表了开幕致辞。他似乎继续着之前的话题，不过语调更加乐观。会议主题是“有效利他主义如何保持好奇心？”

在一个明亮的春日，与MacAskill在Christ Church Meadow散步时，我从这些问题转回到AI的话题。我注意到，赋予接近或超越我们自身能力的事物某种固定的目标函数，这个想法令人担忧。

“哦是的，”MacAskill说。“我绝对——我也被这样的想法吓坏了，就像，‘好吧，我们只有这一次机会。我们只需要编码正确的价值观，然后，就让它运行吧！’”

“伦理问题非常困难，”他说。“显而易见的是，你需要对这些问题保持不确定性。”

“如果你看各种道德观点，它们在什么是好的结果这个问题上分歧相当大，”MacAskill解释道。“即使你只是比较认为模拟意识同样有价值的享乐主义观点，与认为必须是血肉之躯的人类的功利主义观点。它们是非常、非常相似的理论。但它们在我们应该如何使用我们的宇宙禀赋方面会根本性地产生分歧”——指的是人类对我们计划在宇宙中做什么的终极雄心。“基本上就会是战争，”他说：这是一直困扰学术部门的经典“细微差别的自恋”，但现在涉及的是宇宙级的赌注。

但也许所有这些在长远来看分歧如此巨大的竞争性道德理论，在我们应该如何度过当下这个问题上能找到令人惊讶的共同点。“我认为在所有这些理论中，很可能存在一种趋同的工具性目标，”他说。“我称之为长期反思(Long Reflection)。这就是一个时期——实际上可能非常长！当你看实际的规模时，就像，好吧，我们解决了AI等等问题。也许是数百万年我们真的什么都不做。我们保持相对较小——再次强调，至少以宇宙标准而言——而我们所做的主要目的就是试图弄清楚我们应该重视什么。”

他说，与此同时我们的主要目标之一——也许就是主要目标——应该是维持“一个尽可能不被锁定的社会，对各种不同的道德可能性保持开放”。这听起来很像可实现效用保存(attainable utility preservation)的伦理版本——确保我们仍然可以在遥远的未来追求各种目标，即使（或特别是如果）我们现在完全不知道那些未来的目标应该是什么。即使我们现在的猜测就像随机的一样好。

“也许这太困难了，”MacAskill说，“我们需要花一百万年的时间才能做到。”

我建议说，花一百万年时间可能是为了做对这件事而付出的小代价。

“这是极小的代价，因为做对这件事——如果你用错误的东西填满了星星，那么你基本上实现了零价值，所以这真的是……你可以把拥有错误的道德观点视为一种存在风险。”⁸⁶

他停顿了一下。“实际上，我甚至认为这是最可能的存在风险。”

在MacAskill工作的有效利他主义中心的大厅里，是牛津大学的人类未来研究所(Future of Humanity Institute)，由哲学家Nick Bostrom创立。

Bostrom最具影响力的早期论文之一题为”天文浪费(Astronomical Waste)“。副标题是”技术发展延迟的机会成本”，确实，这篇论文的前半部分给读者灌输了一种几乎疯狂的紧迫感。“在我写这些文字的时候，”Bostrom开始说，“太阳正在照亮和加热空房间，未使用的能量正被冲入黑洞，而我们伟大的共同禀赋……正在宇宙规模上不可逆转地退化为熵。这些都是一个先进文明本可以用来创造价值结构的资源，比如过着有意义生活的有感知生命。这种损失的速度令人震惊。”

Bostrom继续估算，一个未来的太空文明最终可能变得如此庞大，以至于现在每秒的延迟能都相当于放弃了本可以生活的一千万亿人类生命，如果我们能够更早地利用所有那些浪费的能量和物质的话。

但是，当任何功利主义者开始得出结论，认为向那个目标推进我们的技术进步如此重要，以至于所有其他地球活动都是琐碎的——甚至是道德上站不住脚的——Bostrom的论文就做出了当代哲学中最急剧的发夹弯之一。

他说，如果到达这个星际未来延迟一秒的风险是一千万亿人类生命，想想完全失败的风险。通过计算，Bostrom得出结论，将我们成功建设一个充满活力、繁荣的远期文明的机会提高一个百分点，在功利主义术语中相当于将技术进步加速一千万年。

因此，尽管赌注巨大，结论却不是急躁，而是恰恰相反。

当我询问各位AI安全研究者如何决定将他们的生命投入到这一事业时，Bostrom的论文不止一次被提及。“我最初觉得这个论证很奇怪，”Paul Christiano说，“或者说感觉有些令人反感，但后来我仔细研读了一遍，然后想，是的，我觉得我大概认同这个观点。”⁸⁷ Christiano从2010年或2011年开始认真思考这些论证，并在2013年或2014年进行了自己版本的数据计算。Bostrom的数学计算是正确的。“对于像我这样的单个研究者来说，”他说，“将灭绝风险降低百万分之一，似乎比将进步加速一千年要容易得多。”而且他在随后的岁月中一直按照这个认识生活。

当然，启动一个潜在的超人类水平artificial general intelligence可能是人类能够做的最不可逆转、最具重大影响的事情之一——值得注意的是，不仅是机器本身，研究者们自己也变得越来越不确定、犹豫、心胸开阔。

Machine Intelligence Research Institute的Buck Shlegeris最近回忆了一次对话，“有人说在Singularity之后，如果有一个神奇的按钮可以将全人类变成快乐而优化的均质糊状物（又称hedonium），他们会按下它…几年前，我主张按下那样的按钮。”但某些东西发生了变化。现在他不那么确定了。他的观点变得…更复杂了。也许这是一个好主意；也许不是。问题然后变成了当你知道自己不知道该怎么做时该怎么办。⁸⁸

“我告诉他们，”他说，“我认为人们不应该按下这样的按钮。”⁸⁹

结论

我认为模糊性在知识理论中比你从大多数人的著作中判断出来的要重要得多。一切事物都模糊到一定程度，只有当你试图使其精确时才会意识到，而一切精确的东西都与我们通常思考的一切如此遥远，以至于你不能片刻假设那就是我们说出我们所想时真正的意思。当你从模糊转向精确时...你总是承担着某种错误的风险。

—伯特兰·罗素¹

过早优化是万恶之源。

—唐纳德·克努特²

这是圣诞节前夜，我和妻子住在我父亲和继母的房子里，当我在半夜醒来时，全身湿透了汗水。

我想我一定是穿了太多衣服睡觉；我掀开被子，脱掉衬衫。不是我的问题，我意识到，带着恐惧和警觉的混合情绪。房间里的空气热得令人无法忍受。我突然想到房子可能着火了。

我打开门。房子漆黑一片，寂静无声。袭来的空气很冷。

我慢慢地拼凑出了真相。楼上有两间客房，但它们共用一个温控器面板，位于另一间无人居住的房间里。我们的卧室门是关着的。装有温控器的另一间卧室的门是开着的。

这是一个寒冷的新英格兰夜晚，温度在冰点以下。加热器一直通过通风口向两间卧室吹送热空气。但装有温控器的房间一直对整个房子的其余部分开放，无论系统吹送多少热空气都无法达到平衡。我们的房间与房子其余部分隔离，得到了系统用来尝试加热——实际上——整个房子其余部分的同样数量的热空气。

还有什么比温控器更简单的呢？事实上，最简单的“闭环”控制系统之一的典型例子——实际上是典型的控制论例子——就是普通的机械温控器。这并不涉及所谓的“机器学习”。但这里就是完整的、汗水浸透的对齐问题。

首先：你测量的不是你以为你测量的。我想要调节我房间里的温度。但我只能测量另一个房间的温度。我没有意识到它们从根本上是不相关的——只要一扇门开着而另一扇不开。

其次：有时候唯一能拯救你的就是你自己的无能。我记得我想，如果加热系统更强大，如果我们的卧室隔热效果更好，我和妻子可能会被煮熟。我们当然醒了，但在热力学光谱的另一端，低温甚至更危险。仅仅因为卧室太冷而导致的体温过低症确实可以而且确实会杀死人。³在1997年的纪录片*Hands on a Hard Body*中，我们遇到了一个名叫Don Curtis的德克萨斯人，他告诉我们他家里有一台二十吨的空调。“一台二十吨的空调机组大得足以让那边的Kmart商店制冷，”他说。他解释说，一家商店倒闭了，“他们几乎白送给了我。我说，‘好吧，这个应该能让我的房子凉爽！’但我不知道它会把温度降到零下十二度。但我们很快就发现它确实会。”他设法避免了低温休克，但危险是真实的。

在我的情况下，当我将自己的门打开了几分钟，在确保关闭两扇门之前，我的脑海中想起了二十世纪中叶伟大的控制论学家Norbert Wiener，他预见了当代关于对齐问题的许多警告。正是他说出了那句著名的话：“我们最好十分确定放入机器的目的就是我们真正渴望的目的。”

但是他的另一个评论同样具有预见性，同样令人恐惧。“在过去，对人类目标的片面和不充分的认识相对来说是无害的，只是因为它伴随着技术的局限性，”他写道。“这只是人类的无能保护我们免受人类愚蠢的全面破坏性影响的众多地方之一。”⁴ Don Curtis就是当我们增强力量时移除这种保护所带来问题的完美例子。我不禁想到AI就像一台二十吨的空调，走进每个家庭。

从这个意义上说，我们必须希望能够首先纠正我们的愚蠢，而不是我们的无能。正如人类未来研究所的Nick Bostrom在2018年所说：“人类的技术能力和人类的智慧之间存在一场长距离赛跑，前者就像一匹在田野中奔驰的种马，而后者更像一只蹒跚学步的小马驹。”⁵ Wiener本人曾警告不要庆祝独创性——“技能”——而不对我们究竟想要做什么进行批判性评估：他称之为“目标认知”，他发现这种认知极度缺乏。

Aldous Huxley在1937年用另一种方式表达了这一点：“很明显，胜利的科学迄今为止所做的只是改进了实现未改进或实际恶化目标的手段。”⁶

images

迄今为止讲述的故事一直是令人鼓舞的

到目前为止讲述的故事一直是令人鼓舞的，是一个充满信心、稳定、科学进步的故事。一个影响近期和长期的研究和政策努力的生态系统正在全球范围内展开；这仍然很大程度上处于萌芽阶段，但正在积聚力量。

关于偏见、公平性、透明度和安全的多个维度的研究，现在构成了在主要AI和机器学习会议上展示的所有工作的重要部分。实际上，目前它们可以说是最具活力和增长最快的领域，不仅在计算领域，在整个科学领域都是如此。一位研究人员告诉我，在2016年该领域最大的会议之一上，当他说自己在研究安全主题时，人们会侧目而视；当他一年后参加同一会议时，没有人会皱眉头。这些文化转变反映了资金的转变，以及研究本身重点的转变。

在前面的章节中，我们已经探索了该研究议程的叙述和内容，在各个方面都有进展可以报告。

但是这本书以George Box的题词开始，提醒我们“所有模型都是错误的”。因此，本着这种精神，让我们用批判的眼光来分析我们自己故事中的一些假设。

代表性

在第1章中，我们讨论了谁或什么在模型的训练数据中被代表的问题。我们在短时间内取得了很大进展；任何主要的消费者人脸识别产品都不太可能在没有面向训练数据代表性构成的内部流程的情况下开发。然而，鉴于这样的模型不仅被消费者软件用于为照片添加标题，被消费者硬件用于解锁智能手机，还被政府用于监视其人口，人们可能会质疑使这些模型在已经被过度监视的种族少数群体面孔上更加准确的程度是否完全是好事。

消费技术中的代表性问题在令人震惊的程度上提醒我们更古老、更棘手、甚至可能更严重的差异。我最近和一群医学研究人员共进晚餐，当我描述为机器学习模型推动更具代表性的训练数据时，他们几乎异口同声地提醒我——大多数医学试验仍然主要在男性身上进行。⁷

临床试验的构成是一把双刃剑：即使是看似合理的保护弱势群体的禁令——例如，不允许对孕妇或老年人进行医学试验——也会产生偏见和盲点。沙利度胺(thalidomide)药物被营销为“完全安全”，因为制药商“找不到足以杀死老鼠的剂量”。但在该药被撤出市场之前，它在人类胎儿中造成了数万例可怕的畸形。⁸（由于怀疑的食品药品管理局员工Frances Oldham Kelsey博士，美国人基本上幸免于难。）

在“监督学习”的情况下，训练数据以某种方式被“标记”，我们还需要批判性地考虑，不仅要考虑我们从哪里获得训练数据，还要考虑我们从哪里获得将在系统中充当真相替代的标签。通常真相并不是真相。

例如，ImageNet使用互联网上随机人类的判断作为真相。如果大多数人认为，比如说，一只狼崽是一只小狗，那么就图像识别系统而言，它就是一只小狗。著名的是，特斯拉的AI总监Andrej Karpathy在斯坦福大学读研究生时，强迫自己在一周的大部分时间里为ImageNet图片贴标签，使自己成为人类基准。经过一些练习，他能够达到95%的“准确率”。但是……相对于什么准确？不是真相。共识。⁹

在更加哲学的层面上，这些标签反映了一个我们必须毫无疑问地接受的预制本体论。ImageNet中的每张图片都恰好属于一千个类别中的一个。¹⁰要使用这些数据和在其上训练的模型，我们必须接受这样的虚构：这一千个类别是相互排斥且详尽无遗的。这个数据集中的图像永远不会被同时标记为“婴儿”和“狗”——即使它明显同时包含两者。而且没有任何东西不属于这一千个类别之一。如果我们看的是一张骡子的图片，而我们的标签只允许我们说“驴”或“马”，那么它必须是驴或马中的一个。它也不能是模糊的。如果我们无法辨别它是驴还是马，我们仍必须给它标记某个标签。而且，在随机梯度下降的惩罚下，我们将迫使我们的模型分享这种教条。最后，标签不能是不确定的。我们可以推断其中一些事情——例如，通过注意到不同的人应用了不同的标签——但我们不知道人类标记者在被迫应用该标签时有多么矛盾或不确定。

同样值得考虑的不仅是训练数据和标签，还有目标函数。图像识别系统经常使用一个叫做“交叉熵损失”的目标函数进行训练——撇开数值细节不谈，它对任何错误表征都会分配惩罚，无论是哪种。在交叉熵损失的标准下，将炉灶识别为汽车格栅，或将青苹果识别为梨，或将英国斗牛犬识别为法国斗牛犬，与将人类表征为大猩猩一样糟糕。在现实中，某些类型的错误——仅从Google的财务角度来看，更不用说被错误分类的人类——可能比其他错误糟糕数千倍甚至数百万倍。¹¹

在[第1章]的后半部分，我们讨论了基于向量的词表示及其作为类比的惊人能力。在这种表面简单性的背后，隐藏着一个本身就相当有争议的对齐问题。类比究竟是什么？例如，简单的向量加法（有时被称为“平行四边形”方法，或“3CosAdd”算法）经常导致一个词成为它自己最好的类比。Doctor - man + woman，例如，产生一个最接近的词实际上就是doctor。¹²

Tolga Bolukbasi和Adam Kalai的团队发现这是一种不令人满意的方式来在word2vec中捕获我们所意指的”类比”，这似乎要求两个事物至少是不同的——因此他们采取了不同的策略。他们想象了围绕”doctor”一词的一种”相似性半径”，包括像”nurse”、“midwife”、“gynecologist”、“physician”和”orthopedist”这样的词，但不包括”farmer”、“secretary”或”legislator”。然后他们在这个半径内寻找不是”doctor”的最近的词。¹³

这里还有其他棘手的问题。词向量的几何——它们被表示为数学空间中的距离的想法——使得每个类比都是对称的，这种方式并不总是反映人类对类比的直觉。例如，人们描述椭圆更”像”圆，而不是圆更”像”椭圆，朝鲜更”像”中国，而不是中国更”像”朝鲜。¹⁴

那么，什么算法，应用于什么表示，能产生更精确地像人类类比一样运作的东西？¹⁵

你可能会想要举手投降。为什么我们要让计算机科学家、语言学家和认知科学家争论这些事情并从头开始提出算法，当我们可以只是在人类类比的例子上训练机器，包括它们的不对称性和怪癖，并让它找出指定什么是类比的合适方式？

这当然是一个对齐问题。人类的”类比”概念证明与任何其他概念一样模糊和不确定。因此，在其他情境中用于对齐的同样初期工具集可能在这里也是可用的。

公平性

在[第2章]中，我们考察了风险评估工具在刑事司法系统中日益广泛的使用。这里有许多潜在的危险，其中一些我们已经讨论过。训练这些模型的“真实情况”不是被告后来是否犯了罪，而是他们是否被重新逮捕和重新定罪。如果来自不同群体的人在被捕后被定罪的可能性，或者首先被逮捕的可能性存在系统性差异，那么我们充其量是在为累犯的扭曲代理进行优化，而不是累犯本身。这是一个经常被忽视的关键点。

同样值得考虑的是，为了训练模型，我们假装知道被告人如果被释放会做什么。我们怎么可能知道呢？这里的典型方法是查看他们服满刑期之后前两年的犯罪记录，并将此作为他们如果更早被释放本应经历的两年的替代。这隐含地假设年龄和监禁经历本身都不会影响某人重返社会时的行为。事实上，年龄在某些情况下是最具预测性的单一变量。此外，假设监禁本身没有影响既可能是错误的，也是对一个至少表面上为了康复而设计的系统相当悲观的看法。如果，正如一些证据似乎表明的，监禁经历实际上增加了经历者的犯罪行为，那么被迫服完刑期的人的再犯行为就成为了一个假设他们如果被早期释放也同样危险的模型的训练数据。因此它会推荐那些产生犯罪的更长刑期。预测变成了自我实现；人们被不必要地关押；公共安全因此变得更糟。

在机器学习的许多领域中，存在着被称为“迁移学习”的显著程度，即最初针对一项任务训练的系统很容易被重新用于另一项任务。但这并不总是经过深思熟虑或明智的。例如，COMPAS工具被明确设计为不用于量刑，然而在某些司法管辖区，它仍然被这样使用。（用于招聘决策的词嵌入模型也是如此。为促进预测而构建的表示在许多情况下被用来做它们被训练来预测的事情。在有性别歧视历史的企业文化中，一个不幸地正确预测很少女性会被雇用的模型可能会被不假思索地部署，使得很少女性真的被雇用。在我们希望模型做的不仅仅是重复和强化过去的程度上，我们需要更加谨慎和用心地对待它们。）

我们也看到，“公平”很容易提出许多看似直观和理想的不同形式定义。然而残酷的数学事实是，没有决策系统——无论是人类还是机器——能够同时为我们提供所有这些。一些研究人员认为，与其详细讨论这些不同的形式主义，然后试图“手动”调和它们，我们应该简单地用人类认为“公平”和“不公平”的例子来训练系统，并让机器学习自己构建形式的、可操作的定义。这本身很可能是一个和其他任何对齐问题一样微妙的问题。

透明度

在[第3章]中，我们讨论了简单模型优势以及寻找最优简单模型能力日益增强的工作前沿。然而，这种透明度至少可能是双刃剑，因为研究表明，即使当这些模型是错误的并且不应该被信任时，人类也会对透明模型给予更大的信任。

还有一个轻微的悖论，即很难理解为什么某个特定的简单模型是最优的；对这个问题的详尽答案可能高度技术性且冗长。此外，对于任何特定的简单模型，我们很可能会问可能特征的“菜单”从何而来，更不用说什么人为过程驱动了需求和工具的创建。这些是内在的人类、社会和政治的合理透明度问题，机器学习本身无法解决。

在开发有助于解释（无论是视觉还是语言）的架构时，有几件事需要警惕。研究表明了“对抗性解释”的可能性——也就是说，两个行为几乎相同但对它们如何以及为什么如此行为提供截然不同解释的系统。能够为自己的行为提供有说服力的解释是非常强大的，无论这些解释是否为真。事实上，认知科学家如Hugo Mercier和Dan Sperber最近论证说，人类推理能力的进化不是因为它帮助我们做出更好的决策并对世界持有更准确的信念，而是因为它帮助我们赢得争论并说服他人。需要谨慎，以免我们仅仅创造了为解释的外观或为它们给我们的理解它们的感觉而优化的系统。这样的系统可能会欺骗性地运用这种能力；我们可能发现我们已经为高超的胡说八道艺术进行了优化。更一般地说，即使我们约束系统的解释必须真实，构建一个具有令人印象深刻的自我解释能力的系统可能有助于我们控制它，但如果被“推理的论证理论”所说服，那么这也可能帮助它控制我们。

主体性

在我们讨论强化学习(reinforcement learning)、奖励塑造(reward shaping)和内在动机(intrinsic motivation)时，特别是在[第4章]、[第5章]和[第6章]中关于Atari游戏和围棋的讨论中，我们隐含地假设了一个技术术语叫做”遍历性(ergodicity)“的东西——即你不能犯永久性错误。没有什么事情是发生了就无法通过重新开始来修复的。因此，通过犯成千上万个基本随机且经常是致命的错误来学习是没有问题的。遍历性假设在Atari这样的安全玩具世界之外并不成立。我记得2000年代初的一个汽车广告，展示了一个时髦的X世代、互联网时代的程序员，他白天编写极限赛车游戏代码——充满了电影式的慢镜头碰撞——但夜晚开着他的保守的、安全第一的轿车通勤回家。”因为在真实生活中，“他看着镜头说，”没有重置按钮。“DeepMind的Jan Leike用略微不同的语言表达了这一点。他指出，在他自己和他研究的人工智能体之间至少有一个主要差异；更准确地说，在他们的世界和他自己的世界之间存在重大差异。”真实世界不是遍历的(ergodic)，“他说。”如果我从窗户跳出去，那就完了——这不是我能从中学习的错误。”²²

强化学习的不同方法也带有不同的假设集合。有些假设世界有有限数量的离散状态。有些假设你总是确切知道自己处于什么状态。许多假设奖励总是可通约的标量值，它们永远不会改变，而且你总是确切知道何时获得奖励。

许多假设环境本质上是稳定的。许多假设智能体不能永久改变环境，环境也不能永久改变智能体。在真实世界中，许多行动改变你的目标。任何数量的改变心智或调节情绪的药物，无论是处方药还是其他药物，都会在某种程度上和某段时间内做到这一点。出国生活、遇到对的人，甚至听到对的歌曲也会如此。大多数强化学习假设这些都不会发生。一小部分研究承认这种可能性，但假设智能体会”理性地”试图保护自己免受此类变化的影响。²³然而，我们有意承担某些变革性体验，怀疑我们会因此而改变，有时甚至无法预料会以何种方式改变。²⁴（为人父母就是这样的例子。）

传统强化学习还倾向于假设智能体是环境中唯一的智能体；即使在象棋或围棋这样的零和游戏中，系统更多的是在与”棋盘”对弈而不是与”对手”对弈，并且很少考虑对手可能正在改变和适应其自身策略的想法。在我最近与两位强化学习研究者的对话中，我们思考了将大多数RL算法放入囚徒困境(prisoner’s dilemma)中的情况，其中两个共犯必须决定是通过出卖对方来”背叛”还是通过保持沉默来”合作”。“合作”策略有最好的结果，但只有在双方都选择它时才如此，而传统的RL智能体将无法理解环境中包含另一个智能体，一个其行为取决于自身行为的智能体。背叛在短期内总是看起来更有回报，也更容易，而合作则需要一定程度的同步，这会让两个没有正确理解其相互依赖性的智能体感到困惑。²⁵

正如Jean Piaget在谈到儿童心智发展时所说：“随着他智力工具的协调，他通过将自己定位为在外部世界中其他活跃对象中的一个活跃对象来发现自己。”²⁶

人类同样理解，用正念老师Jon Kabat-Zinn的话来说，“无论你走到哪里，你都在那里”，而RL智能体通常不认为自己是它们正在建模的世界的一部分。大多数机器学习系统假设它们自己不会影响世界；因此它们不需要建模或理解自己。随着此类智能体变得越来越强大、越来越能干和越来越普及，这种假设只会变得越来越没有根据。例如，机器智能研究所(Machine Intelligence Research Institute)的Abram Demski和Scott Garrabrant一直在呼吁他们称之为”嵌入式智能体(embedded agency)“的概念，重新思考这种在该领域中已经变得如此隐含和根深蒂固的自我与世界的划分。²⁷

模仿(IMITATION)

在 [第7章] 中，我们讨论了模仿学习整个前提的一个基本且无根据的假设——即你可以将一个交互式世界（在这个世界中，你做出的每个选择都会改变你所看到和体验的内容）当作一个经典的监督学习问题来处理，在监督学习中，你看到的数据被称为”i.i.d.”：独立且同分布。如果你看到一张猫的图片并错误地将其标记为狗，这并不会改变你接下来看到的图片。但在汽车中，如果你将一张指向前方的道路图像错误地标记为需要右转的图像，那么你很快就会发现自己正在看着一条不熟悉的横向道路。这是DAgger等方法试图缓解的”级联错误”的根本原因。人们认为这在某种意义上等同于现代隐形战斗机（如F-117夜鹰）的空气动力学，这些飞机在所有三个轴上都不稳定，需要完美的飞行精度，否则会立即变得灾难性不稳定。只是在这种情况下，自动驾驶仪不是解决方案，而是这个问题的原因。

模仿也倾向于假设专家和模仿者具有基本相同的能力：实际上是相同的身体，至少在潜在上是相同的头脑。汽车恰好是这种假设合理的完美例子。人类驾驶员和自动驾驶仪实际上确实在效果上共享一个身体。他们都向同一个方向盘、同一个车轴、同一个轮胎和刹车发送线控驱动信号。在其他情况下，这根本行不通。如果某人在基本上比你更快或更强壮或体型不同，或者思维比你永远可能达到的更敏捷，那么完美地模仿他们的行为可能仍然不起作用。实际上，这可能是灾难性的。你会做你如果是他们会做的事。但你不是他们。而你所做的不是他们如果是你会做的事。

推理

随着世界上的AI智能体变得越来越复杂，它们将需要关于我们的良好模型来理解世界如何运作以及它们应该和不应该做什么。如果它们将我们建模为纯粹的、无拘无束的、毫无错误的奖励最大化器，而我们不是，那么我们就会有一段糟糕的时光。如果有人不遗余力地帮助你，但他们并不真正理解你想要什么——无论是短期还是生活中——那么你的境况可能比他们根本不尝试帮助时更糟。如果这个误导的帮助者应该是，比方说，超人类智能和强大的——那就更糟了。

在我们讨论从行为中推断人类价值观和动机的系统时，有许多假设需要解开。一个假设是人类或专家正在展示“最优”行为。当然，这几乎从来不是事实。²⁸ 在足够复杂以放松这一假设的系统中，有特定的形式化模型用于人们表现出的次优性类型——例如，我们的行为是概率性的，其中行动的概率与其奖励成正比；这些在实践中似乎工作得出奇地好，但它们实际上是否是人类行为的最佳模型是一个开放的问题，这个问题对心理学家、认知科学家和行为经济学家来说，与对计算机科学家来说一样重要。²⁹

即使对人类表现中的错误或次优性或“非理性”有一定的容忍度，这些模型通常仍然假设人类是专家，而不是学生：成年人的步态，而不是学习走路的孩子；专业直升机飞行员，而不是仍在掌握窍门的人。这些模型假定人类的行为已经收敛到一套最佳实践，他们已经学会了所能学会的一切，或者在给定任务上已经变得尽可能好。从这个意义上说，这种技术的名称——逆向强化学习——是一个误称。我们对某人目标和价值观的推断不是基于他们的强化学习过程，而是基于他们最终的行为结果（用技术术语来说，他们的“学习策略”）。我们不是从演示者实现目标的过程中进行推断，而只是在之后——这一点在1998年的第一篇IRL论文中就提出了。³⁰ 二十年后，IRL系统终于开始成熟，但在解决这个潜在的原则性问题方面几乎没有做什么工作。³¹

典型的逆向强化学习还假设人类专家在某种意义上是在没有意识到自己正被建模的情况下行动。合作逆向强化学习倾向于做出相反方向的假设：人类以教学方式行动，明确地教授机器而不仅仅是“做自己的事情”。实际上，我们在他人面前的行为通常介于两者之间。如果假设被违反，做出任何一种强假设都可能导致误解问题。³²

最后，也许是最重要的，典型的逆向强化学习系统想象只有一个人的偏好正在被建模。确切地说，我们如何将这种方法扩展到在某种意义上服务于两个（或更多）主人的系统？

正如斯坦福大学计算机科学家Stefano Ermon所说，让AI与人类价值观保持一致“我认为是大多数人都会同意的事情，但问题当然在于准确定义这些价值观到底是什么，因为人们有不同的文化背景，来自世界不同地区，有不同的社会经济背景，所以他们对这些价值观的看法会有很大差异。这才是真正的挑战。”

路易斯维尔大学计算机科学家Roman Yampolskiy表示赞同，强调说：“我们人类在共同价值观上并不一致，甚至我们达成一致的部分也会随着时间而改变。”

这里有一些重要的技术细节：如果一半用户在岔路口向左行驶，另一半向右行驶，正确的行为显然不是“折中处理”然后撞向分隔带。

随着机器学习准备与现有学科接触，还有无数悖论等着我们。这些学科充满了各自长期存在的问题，在某些情况下已经几个世纪都在努力协调多人偏好：政治哲学和政治科学、投票理论和社会选择。

通过稍微放大视野来总结关于机器学习假设的这一部分，每个机器学习架构都隐含地依赖于几个层面的迁移学习。它假设在现实中遇到的情况平均而言会类似于在训练中遇到的情况。上述几个问题都是这一问题的不同版本，经典

的机器学习陷阱如过拟合也是如此。

然而，违反这一假设的最简单情况之一是世界顽固而持续的变化倾向。我记得听过一位计算语言学研究者抱怨，无论多么努力尝试，他们都无法让模型复现另一位研究者一两年前发表的准确结果。他们一遍遍检查自己的工作。他们到底做错了什么？

事实证明，什么都没错。训练数据来自2016年。2017年书写和口语的英语略有但可测量的不同。2018年的英语差异更大。这是研究者所知的“分布偏移”的一个例子。没有人试图复现该论文结果时能够达到与原始研究者相同的准确性水平，至少不能用那些训练数据。在2016年数据上训练的模型随着世界的发展慢慢失去了准确性。

images

综合来看，我们有一系列提醒：地图不是领土。正如Bruno Latour所写：“我们把科学当作现实主义绘画，想象它制作了世界的精确副本。科学做的是完全不同的事情——绘画也是如此。通过连续的阶段，它们把我们连接到一个对齐的、转换的、构造的世界。”对齐——如果我们幸运，并且非常小心，非常明智的话。

这构成了对即将到来的世纪的警示故事，这个故事显然单调乏味且不性感——而且，我认为，危险地可能被集体忽视。

我们面临失去对世界控制的危险，不是输给AI或机器本身，而是输给模型。输给对存在事物和我们想要事物的正式的、通常是数值的规范。

正如艺术家Robert Irwin所说：“生活在结构中并通过结构生活的人类变成了生活在人类中并通过人类生活的结构。”在这个语境下，这些话具有警示意义。

尽管本书呈现的是一个进步的故事，我们绝不能认为我们接近完成。实际上，在机器学习中——以及其他方面——人们能做的最危险的事情之一就是找到一个相当好的模型，宣布胜利，然后从此开始混淆地图与领土。

人类的制度记忆极其短浅，最多一个世纪；每一代人来到世界时都认为事情就是这样的。

即使我们——也就是所有从事AI与伦理学、AI与技术安全工作的人——完成我们的工作，如果我们能够避免明显的反乌托邦和灾难（这远非确定）——我们仍然必须克服向一个日益成为形式主义世界的根本的、可能不可抗拒的进程。我们必须在不可避免地被这些模型塑造——在我们的生活中，在我们的想象中，在我们的身体中——的同时做到这一点。

这是Rodney Brooks著名机器人学宣言的阴暗面：“世界是它自己最好的模型。”

这越来越真实，但不是Brooks本意的精神。世界的最佳模型代替了世界本身，并威胁要扼杀真实事物。

我们必须非常小心，不要忽视那些不容易量化或不容易被纳入我们模型的事物。危险，借用Hannah Arendt的话说，与其说我们的模型是错误的，不如说它们可能会变成真实的。

在其他科学领域，我们可能不会遇到这个问题。对牛顿力学的依赖并没有让水星令人困扰的近日点消失；它仍然存在，在牛顿之后两百年还在困扰着爱因斯坦。然而，在人类事务中，这种危险是非常真实的。

在美国国家运输安全委员会对2018年在亚利桑那州坦佩市杀死行人伊莱恩·赫茨伯格(Elaine Herzberg)的自动驾驶优步汽车的审查中，分析显示”系统从未将她归类为行人……因为她正在……没有人行横道的地方过马路；系统设计没有考虑乱穿马路的行人。“³⁹我们必须谨慎，不要让自己身处一个系统不允许它们无法想象的事物的世界中——在这个世界里，它们实际上强制执行了自身理解的局限性。

也许正是出于这个原因——以及其他原因——我们发现在大自然中度过时光如此令人耳目一新。⁴⁰大自然虽然在无数方面受到人类意图的塑造，但仍然不断找到方法来挫败我们的分类法，抵制我们试图强加给它的系统，无论是概念上的还是其他方面的。正如英国作家赫伯特·里德(Herbert Read)所论证的：“只有向大自然学习的民族才能被信任使用机器。”⁴¹

机构决策越来越依赖于明确的、正式的指标。我们与几乎任何系统的互动都越来越多地调用我们自身行为的正式模型——要么是一般用户行为的模型，要么是为我们量身定制的模型，无论多么简单。

我们在本书中看到的是这些模型的力量、它们出错的方式，以及我们试图使它们与我们的利益对齐的方式。

images

结语

有理由担忧，但我们的最终结论不一定是严峻的。

正如我们所看到的，对机器学习中伦理和安全问题的关注爆发创造了一股活动热潮。资金正在筹集，禁忌正在被打破，边缘问题正在变得核心，机构正在扎根，最重要的是，一个深思熟虑、积极参与的社区正在发展并开始工作。火警已经拉响，急救人员正在现场。

我们也看到了对齐项目虽然包含自身的危险，但也充满诱人和强大的希望。容易量化和严格程序化的主导地位将在某种程度上瓦解，成为早期一代必须手工制作的模型和软件的遗物，因为我们获得的系统不仅能够掌握我们的明确命令，还能掌握我们的意图和偏好。不可言喻的东西不必完全让位于明确的东西。通过这种方式，即将到来的技术加剧了一些目前存在的问题，但缓解和减轻了其他问题。

我们在引言中说过，这将是获得某种个人和公民自我认识的独特机会。这也是对齐故事中令人激动和也许具有救赎意义的维度之一。如果随意部署，有偏见和不公平的模型可能会加深现有的社会问题，但它们的存在将这些通常微妙和分散的问题带到表面，迫使社会与自身进行清算。不公平的审前拘留模型，举例来说，将聚光灯照向上游不平等。有偏见的语言模型除了其他功能外，还为我们提供了一种衡量我们话语状态的方法，并为我们提供了一个基准，我们可以据此努力改进和完善自己。

在真实的人类世界上训练的透明和可解释系统为我们提供了对目前我们还处于黑暗中的事物进行透明和解释的可能性。在看到一种思维在消化和反应世界时的工作方式时，我们将学到一些关于世界的东西，也许还有关于思维的东西。

所谓AGI的前景——一个或多个与我们一样灵活智能的实体（可能更智能）——将给我们提供终极的镜中自省。也许从我们的动物同胞那里学到的太少，我们将直接发现智能的哪些方面似乎是普遍的，哪些仅仅是人类的。仅此一点就是一个令人恐惧和激动的前景。但了解真相比想象真相更好。

对齐将是混乱的。否则怎么可能呢？

它的故事将是我们的故事，无论好坏。怎么可能不是呢？

images

历史的回音

1952年1月14日，BBC主办了一个广播节目，召集了四位杰出科学家的小组进行圆桌对话。主题是”自动计算机器能说是在思考吗？“四位嘉宾是计算机科学创始人之一艾伦·图灵(Alan Turing)，他在1950年就这个话题写了一篇现在传奇的论文；科学哲学家理查德·布雷思韦特(Richard Braithwaite)；神经外科医生杰弗里·杰斐逊(Geoffrey Jefferson)；以及数学家和密码学家马克斯·纽曼(Max Newman)。

小组开始讨论机器如何学习以及人类如何教授它的问题。

“当一个孩子在接受教育时，他的父母和老师确实在反复干预，阻止他做这个或鼓励他做那个，这是完全正确的，”图灵说。⁴²“但当人们试图教授机器时，情况也不会有什么不同。我在教授机器做一些简单操作方面做了一些实验，在我能得到任何结果之前，需要大量这样的干预。换句话说，机器学得如此缓慢，以至于需要大量的教学。”

杰斐逊打断了他。“但谁在学习，”他说，“是你还是机器？”

“嗯，”图灵回答，“我想我们两个都在学习。”

致谢

一个拥有效应器的神经系统可能会留下痕迹，比如在纸上留下墨迹。在任何时候，它都可能看到这些痕迹。通过简单的条件反射，痕迹可能成为神经系统所拥有的任何概念的符号。通过类似的条件反射，它们对其他神经系统也具有相同的意义。因此，通过符号，计算和结论被许多神经系统在同一时间共享，并延续到无法计量的时代。这确实是语言、文学、哲学、逻辑、数学和物理学的故事。

—沃伦·麦卡洛克¹

这本书是一个产物，更重要的是，它是对话的产物：数百次对话的产物。有些是提前几个月安排的，有些是偶然的，有些是数千英里旅行的产物，有些通过数千英里的UDP数据包连接，有些就在拐角处。有些在安静的办公室里录音进行，有些在礼堂里其他事情进行时低声交谈，有些在酒桌上大声嬉笑。有些在世界上最权威的机构，有些在攀岩健身房、酒店热水浴缸或餐桌旁。有些更像采访和口述历史，有些像同事间的专业交谈，有些像闲逛。

思想是社会性的。它们在对话中一轮一轮地逐渐出现，是任何人独立思考的产物。每当我与某人交谈时，如果出现了我当时知道或怀疑可能会写进书中的想法，我都会尽量记录下来。我做了很多这样的记录。我确信在许多场合我都未能做到这一点，对此我诚挚地提前道歉。但我相信，至少与以下人员的对话和交流使这本书成为了现在的样子：

Pieter Abbeel、Rebecca Ackerman、Dave Ackley、Ross Exo Adams、Blaise Agüera y Arcas、Jacky Alciné、Dario Amodei、McKane Andrus、Julia Angwin、Stuart Armstrong、Gustaf Arrhenius、Amanda Askell、Mayank Bansal、Daniel Barclay、Solon Barocas、Renata Barreto、Andrew Barto、Basia Bartz、Marc Bellemare、Tolga Bolukbasi、Nick Bostrom、Malo Bourgon、Tim Brennan、Miles Brundage、Joanna Bryson、Krister Bykvist、Maya Çakmak、Ryan Carey、Joseph Carlsmith、Rich Caruana、Ruth Chang、Alexandra Chouldechova、Randy Christian、Paul Christiano、Jonathan Cohen、Catherine Collins、Sam Corbett-Davies、Meehan Crist、Andrew Critch、Fiery Cushman、Allan Dafoe、Raph D'Amico、Peter Dayan、Michael Dennis、Shiri Dori-Hacohen、Anca Drăgan、Eric Drexler、Rachit Dubey、Cynthia Dwork、Peter Eckersley、Joe Edelman、Owain Evans、Tom Everitt、Ed Felten、Daniel Filan、Jaime Fisac、Luciano Floridi、Carrick Flynn、Jeremy Freeman、Yarin Gal、Surya Ganguli、Scott Garrabrant、Vael Gates、Tom Gilbert、Adam Gleave、Paul Glimcher、Sharad Goel、Adam Goldstein、Ian Goodfellow、Bryce Goodman、Alison Gopnik、Samir Goswami、Hilary Greaves、Joshua Greene、Tom Griffiths、David Gunning、Gillian Hadfield、Dylan Hadfield-Menell、Moritz Hardt、Tristan Harris、David Heeger、Dan Hendrycks、Geoff Hinton、Matt Huebert、Tim Hwang、Geoffrey Irving、Adam Kalai、Henry Kaplan、Been Kim、Perri Klass、Jon Kleinberg、Caroline Knapp、Victoria Krakovna、Frances Kreimer、David Kreuger、Kaitlyn Krieger、Mike Krieger、Alexander Krizhevsky、Jacob Lagerros、Lily Lamboy、Lydia Laurenson、James Lee、Jan Leike、Ayden LeRoux、Karen Levy、Falk Lieder、Michael Littman、Tania Lombrozo、Will MacAskill、Scott Mauvais、Margaret McCarthy、Andrew Meltzoff、Smitha Milli、Martha Minow、Karthika Mohan、Adrien Morisot、Julia Mosquera、Sendhil Mullainathan、Elon Musk、Yael Niv、Brandie Nonnecke、Peter Norvig、Alexandr Notchenko、Chris Olah、Catherine Olsson、Toby Ord、Tim O'Reilly、Laurent Orseau、Pedro Ortega、Michael Page、Deepak Pathak、Alex Peysakhovich、Gualtiero Piccinini、Dean Pomerleau、James Portnow、Aza Raskin、Stéphane Ross、Cynthia Rudin、Jack Rusher、Stuart Russell、Anna Salamon、Anders Sandberg、Wolfram Schultz、Laura Schulz、Julie Shah、Rohin Shah、Max Shron、Carl Shulman、Satinder Singh、Holly Smith、Nate Soares、Daisy Stanton、Jacob Steinhardt、Jonathan Stray、Rachel Sussman、Jaan Tallinn、Milind Tambe、Sofi Thanhauser、Tena Thau、Jasjeet Thind、Travis Timmerman、Brian Tse、Alexander Matt Turner、Phebe Vayanos、Kerstin Vignard、Chris Wiggins、Cutter Wood和Elana Zeide。

感谢早期读者，他们让这本书对所有可能后来阅读的人来说都变得无比优秀：Daniel Barclay、Elizabeth Christian、Randy Christian、Meehan Crist、Raph D' Amico、Shiri Dori-Hacohen、Peter Eckersley、Owain Evans、Daniel Filan、Rachel Freedman、Adam Goldstein、Bryce Goodman、Tom Griffiths、Geoffrey Irving、Greg Jensen、Kristen Johannes、Henry Kaplan、Raph Lee、Rose Linke、Phil Richerme、Felicity Rose、Katia Savchuk、Rohin Shah、Max Shron、Phil Van Stockum、Shawn Wen和Chris Wiggins。感谢你们毫不留情的批评。

感谢我的经纪人Max Brockman，因为他看到了这本书可能的样子，也感谢我的编辑Brendan Curry，因为他看到了这本书确实的样子。

感谢人工智能促进协会、NeurIPS和生命未来研究所的重要邀请，我很高兴接受了这些邀请。感谢纽约大学的算法与解释会议和FAT*会议、AI Now、CITRIS的包容性AI会议、西蒙斯计算理论研究所的优化与公平研讨会，以及人类兼容AI中心召集聪明人讨论重要话题。能够置身其中是我的荣幸。

感谢麦克道尔艺术村、Mike和Kaitlyn Krieger，以及亚多艺术村，分别为早期、中期和后期的文字写作提供了绿洲——感谢时间、空间和灵感的礼物。

感谢杰里·加西亚(Jerry Garcia)和西尔维娅·普拉斯(Sylvia Plath)的灵魂在孤独的日子里陪伴我。

感谢麦克马斯特大学伯特兰·罗素档案馆的研究人员（特别是肯尼斯·布莱克韦尔(Kenneth Blackwell)），费城美国哲学学会的沃伦·麦卡洛(Warren McCulloch)文件，康奈尔大学的弗兰克·罗森布拉特(Frank Rosenblatt)档案，以及蒙特雷县免费图书馆和旧金山公共图书馆，还有引用调查员网站的加森·奥图尔(Garson O' Toole)，感谢他们在寻找晦涩史实方面的个人帮助。

感谢互联网档案馆保存了重要而短暂的过去。

感谢各种免费和/或开源软件项目使这本书的写作成为可能，特别是Git、TeX和LaTeX。令我惊叹的是，这本手稿是使用超过40年历史的排版软件编写的，而亚瑟·塞缪尔(Arthur Samuel)本人就是为这款软件编写文档的人。我们确实是站在巨人的肩膀上。

我谦卑地想要感谢那些在写作这本书期间去世的人们，我本想听到他们的声音，但他们的思想仍然存在：德里克·帕菲特(Derek Parfit)、肯尼斯·阿罗(Kenneth Arrow)、休伯特·德雷福斯(Hubert Dreyfus)、斯坦尼斯拉夫·彼得罗夫(Stanislav Petrov)和厄苏拉·K·勒古恩(Ursula K. Le Guin)。

我想特别感谢加州大学伯克利分校。感谢CITRIS，在写作这本书期间我有幸在那里担任访问学者，特别感谢布兰迪·诺内克(Brandie Nonnemecke)和卡米尔·克里滕登(Camille Crittenden)；感谢西蒙斯计算理论研究所，特别是克里斯汀·凯恩(Kristin Kane)和理查德·卡普(Richard Karp)；感谢人机兼容AI中心，特别是斯图亚特·拉塞尔(Stuart Russell)和马克·尼茨伯格(Mark Nitzberg)；以及CHAI工作坊众多才华横溢、充满活力的成员和访客。你们都让我感到如此有启发性和宾至如归，你们的友谊和同志情谊比你们知道的更珍贵。

感谢我的妻子罗斯(Rose)，感谢她作为第一位读者、稳定的支撑、敏锐的眼光和听觉、坚实的肩膀，以及鼓励的欢呼。你总是相信，而我总是希望你是对的。

注释

题词

1. 参见彼得·诺维格(Peter Norvig), 《论乔姆斯基和统计学习的两种文化》, <http://norvig.com/chomsky.html>。
2. 这句话在许多资料来源中被广泛归因于布鲁克斯(Brooks), 似乎最初是在布鲁克斯的《没有表征的智能》中表述为“事实证明，使用世界本身作为其自身的模型更好”。
3. 现在著名的统计格言“所有模型都是错误的”最初出现在博克斯(Box)的《科学与统计》中；后来在博克斯的《科学模型构建策略中的稳健性》中出现了“但有些是有用的”这一积极的补充。

序言

1. 关于沃尔特·皮茨(Walter Pitts)生活的信息极其稀少。我从仅有的一些第一手资料中获取信息，主要是皮茨写给沃伦·麦卡洛克(Warren McCulloch)的信件，这些信件可在费城美国哲学学会的麦卡洛克档案中查阅。感谢那里工作人员的善意协助。其他材料来自皮茨同时代人的口述历史，特别是杰罗姆(杰里)·莱特文(Jerome (Jerry) Lettvin)在安德森和罗森菲尔德的《谈论网络》中的内容，以及麦卡洛克《沃伦·S·麦卡洛克全集》中的文章和回忆录。关于皮茨生平的其他描述，参见如斯马尔海泽(Smalheiser)《沃尔特·皮茨》；伊斯特林(Easterling)《沃尔特·皮茨》；以及格夫特(Gefter)《试图用逻辑拯救世界的人》。更多细节存在于麦卡洛克、诺伯特·维纳(Norbert Wiener)和控制论小组的传记中，如海姆斯(Heims)的《约翰·冯·诺依曼和诺伯特·维纳》和《控制论小组》，以及康韦和西格尔曼(Conway and Siegelman)的《信息时代的黑暗英雄》。
2. 怀特海和罗素(Whitehead and Russell)，《数学原理》。
3. 感谢麦克马斯特大学伯特兰·罗素档案馆的工作人员帮助尝试找到这封信的副本；不幸的是，没有已知的现存副本。
4. 安德森和罗森菲尔德，《谈论网络》。
5. 安德森和罗森菲尔德。这本书很可能是卡尔纳普(Carnap)的《语言的逻辑句法》(《Logische Syntax der Sprache》)，尽管一些资料来源认为是《世界的逻辑结构》(《Der logische Aufbau der Welt》)。
6. 根据他们确切的见面时间，皮茨可能已经十八岁了(和/或莱特文仍然二十岁)；麦卡洛克写道：“1941年，我在芝加哥大学数学生物学委员会拉舍夫斯基(Rashevsky)的研讨会上展示了我关于信息在神经元层级中流动的概念，并遇到了沃尔特·皮茨，他当时大约十七岁。”参见麦卡洛克《沃伦·S·麦卡洛克全集》，第35-36页。
7. 这种思考的一些根源早于麦卡洛克与皮茨的合作；参见如麦卡洛克《对控制论众多来源的回忆》。
8. 参见皮奇尼尼(Piccinini)《第一个心智和大脑的计算理论》，以及莱特文为麦卡洛克《沃伦·S·麦卡洛克全集》所写的介绍。

约翰·冯·诺依曼1945年的EDVAC报告是史上第一个关于存储程序计算机的描述，在其101页的内容中只包含一个引用：McCulloch和Pitts，1943年。(参见Neumann，《EDVAC报告初稿》。冯·诺依曼在原文中实际上拼错了：“Following W. S. MacCulloch [原文如此] and W. Pitts。”)冯·诺依曼被他们的论证所吸引，在题为“神经元类比”的章节中，他考虑了这对他所构想的计算设备的实际意义。“很容易看出，这些简化的神经元功能可以用电报继电器或真空管来模拟，”他写道。“由于这些管道装置要通过数字来处理数字，自然要使用数字也是二值的算术系统。这建议使用二进制系统。”我们都知道这种由逻辑门构建的二进制存储程序机器后来的发展历程。它们就是现在如此普及的计算机，数量已经远远超过了地球上的人类。

然而，这种受大脑启发的架构很快就偏离了这种“神经元类比”。许多人想知道是否可能存在在架构上更接近大脑的机器：不是单个处理器以极快的速度一次接收一个明确的逻辑指令，而是一个广泛分布的相对简单、统一的处理单元网络，其整体的涌现能力大于其相当基础的部分之和。也许甚至是一些不那么二进制的东西，带有一些Lettvin所拥抱而Pitts所回避的混乱性。神经网络的专用并行硬件会定期创建，包括Frank Rosenblatt的Mark I Perceptron，但通常是由定制的、一次性的方式。真正支持神经网络大规模并行训练的硬件革命——即GPU——将在几十年后的2000年代中期到来。

引言

[1].} {fn-pad} Mikolov, Sutskever, and Le, “Learning the Meaning Behind Words.”

[2].} {fn-pad} Mikolov, Yih, and Zweig, “Linguistic Regularities in Continuous Space Word Representations.”

[3].} {fn-pad} Tolga Bolukbasi, 个人访谈, 2016年11月11日。

[4].} {fn-pad} Adam Kalai, 个人访谈, 2018年4月4日。

[5].} {fn-pad} 2017年1月, Northpointe与CourtView Justice Solutions和Constellation Justice Systems合并, 并集体重新品牌为“equivant”(小写原文如此), 总部设在俄亥俄州。

[6].} {fn-pad} “而且经常, 这些调查是由开发该工具的同一批人完成的”(Desmarais和Singh, 《在美国惩教环境中验证和实施的风险评估工具》)。

[7].} {fn-pad} Angwin等, 《机器偏见》。

[8].} {fn-pad} 伦斯勒理工学院, 《与首席大法官约翰·G·罗伯茨Jr.的对话》, <https://www.youtube.com/watch?v=TuZEKlRgDEg>。

[9].} {fn-pad} 这个笑话是由程序主席Samy Bengio在2017年会议开幕式上讲的; 参见https://media.nips.cc/Conferences/NIPS2017/Eventmedia/opening_remarks.pdf。一万三千名与会者的数字来自2019年会议; 参见, 例如, <https://huyenchip.com/2019/12/18/key-trends-neurips-2019.html>。

[10].} {fn-pad} Bolukbasi等, 《男人之于计算机程序员如女人之于家庭主妇?》

[11].} {fn-pad} Dario Amodei, 个人访谈, 2018年4月24日。

[12].} {fn-pad} 这个令人难忘的表述来自经典论文Kerr, 《奖励A却希望得到B的愚蠢》。

[13].} {fn-pad} 关于船只竞赛事件的官方OpenAI博客文章, 参见Clark和Amodei, 《野外的错误奖励函数》。

第一章. 表征

[1.] {fn-pad} “海军新设备通过实践学习。”

[2.] {fn-pad} “相对较少的理论家，” Rosenblatt抱怨道，“关注的是一个不完美的神经网络，包含许多随机连接，如何能够可靠地执行那些可能由理想化接线图表示的功能问题。”参见Rosenblatt，《感知器》。

Rosenblatt受到加拿大神经心理学家Donald Hebb在1940年代后期工作的启发；参见Hebb，《行为的组织》。Hebb的观点，著名地总结为“一起激发的细胞连接在一起”，注意到神经元之间的实际连接因人而异，并且似乎随着经验而改变。因此，学习在某种基本意义上就是这些连接的改变。Rosenblatt将这一点直接应用于由简单数学或逻辑“神经元”组成的机器如何学习的实践中。

[3.] {fn-pad} Bernstein, “A.I.”

[4.] {fn-pad} “海军新设备通过实践学习。”

[5.] {fn-pad} “对手。”

[6.] {fn-pad} Andrew, “会学习的机器。”

[7.] {fn-pad} Rosenblatt, 《神经动力学原理》。

[8.] {fn-pad} Bernstein, “A.I.”

9. Walter Pitts写给McCulloch的最后一封信，在Pitts去世前几周寄出，现保存在费城美国哲学学会的Warren McCulloch档案中一个标有“Pitts, Walter”的马尼拉文件夹里。我将它拿在手中。Pitts从城镇另一边的一张病床写信给另一张病床，因为有人告诉他McCulloch想听他的消息。他很怀疑：“关于我们两个人都不会有什么令人愉快的事情。”但无论如何，他还是被说服写了这封信。

Pitts谈到McCulloch最近的冠心病，并表示他理解McCulloch现在“连接着许多传感器，这些传感器连接到面板和警报器……毫无疑问这是控制论的，” Pitts写道。“但这一切都让我感到极其悲伤。”

“想象一下我们两个人都发生最坏的情况，”他写道。他的思绪似乎回到了芝加哥，1942年：回到那些与Lettvin在McCulloch家中度过的难忘夜晚，那是27年前的事。“然后我们会把轮椅推到一起，看着面前无味的茅屋芝士，重述那个著名的对话，关于在老GLAUCUS家中的对话，PROTAGORAS和智者HIPPIAS住在那里：再次尝试理解他们关于认知者与被认知者的微妙而深刻的悖论。”然后，用颤抖的笔迹，全部大写：“愿你安好。”

10. Geoff Hinton, “讲座2.2—感知机：第一代神经网络”（讲座），机器学习神经网络，Coursera, 2012年。

11. Alex Krizhevsky, 个人访谈，2019年6月12日。

12. 在深度网络中确定梯度更新的方法称为“反向传播”；它本质上是微积分中的链式法则，尽管它需要使用可微分神经元，而不是McCulloch、Pitts和Rosenblatt考虑的全有或全无神经元。使这一技术得到推广的工作被认为是Rumelhart、Hinton和Williams的“通过误差传播学习内部表示”，尽管反向传播有着悠久的历史可以追溯到20世纪60年代和70年代，训练深度网络的重要进展在21世纪继续涌现。

13. Bernstein, “A.I.”
14. 参见LeCun等人， “反向传播应用于手写邮政编码识别”。
15. 参见”卷积网络和CIFAR-10：与Yann LeCun的访谈”，<https://medium.com/kaggle-blog/convolutional-nets-and-cifar-10-an-interview-with-yann-lecun-2ffe8f9ee3d6> 或 <http://blog.kaggle.com/2014/12/22/convolutional-nets-and-cifar-10-an-interview-with-yan-lecun/>。
16. 关于前馈网络能做什么和不能做什么的详细信息，参见Hornik、Stinchcombe和White，“多层前馈网络是通用逼近器”。
17. 这句话出自Hinton，在“神经网络和深度学习的’简要’历史，第4部分”中，<https://www.andreykurenkov.com/writing/ai/a-brief-history-of-neural-nets-and-deep-learning-part-4/>。似乎原始来源，一个Hinton演讲的视频，已经从YouTube上删除了。
18. 英伟达成立于1993年，在1999年8月31日推出了其具有重要意义的GeForce 256，“世界上第一个图形处理单元(GPU)”（参见https://www.nvidia.com/object/IO_20020111_5424.html），尽管其他类似技术以及”GPU”这个术语已经存在——例如，在1994年的索尼PlayStation中（参见<https://www.computer.org/publications/tech-news/chasing-pixels/is-it-time-to-rename-the-gpu>）。
19. 英伟达的通用CUDA平台，例如，于2007年推出。
20. Krizhevsky的平台叫做”cuda-convnet”；参见<https://code.google.com/archive/p/cuda-convnet/>。该平台利用了英伟达的计算统一设备架构(Compute Unified Device Architecture)，或CUDA，它允许程序员编写代码在英伟达GPU上执行高度并行的计算。
- 关于AlexNet之后训练神经网络效率惊人增长的2020年回顾，参见OpenAI的Danny Hernandez和Tom Brown在<https://openai.com/blog/ai-and-efficiency/>和https://cdn.openai.com/papers/ai_and_efficiency.pdf的工作。
21. “Rival”。
22. Jacky Alciné，个人访谈，2018年4月19日。
23. 参见 <https://twitter.com/jackyalcine/status/615329515909156865> 和 <https://twitter.com/yonatanzunger/status/615355996114804737>这次交流。
24. 参见Simonite，“谷歌照片在大猩猩方面仍然是盲目的”。一位谷歌发言人证实，在2015年事件后，‘gorilla’被从搜索和图像标签中屏蔽，今天’chimp’、’chimpanzee’和’monkey’也被屏蔽。’图像标记技术仍处于早期阶段，不幸的是它远远不够完美，’发言人写道。”
25. Doctorow，“两年后，谷歌通过从图像分类器中清除’Gorilla’标签解决’种族主义算法’问题”；Vincent，“谷歌通过从其图像标记技术中移除大猩猩来’修复’其种族主义算法”；以及Wood，“谷歌图像’种族主义算法’有了修复方法，但不是一个很好的解决方案”。
26. Visser, *Much Depends on Dinner.*
27. 参见Stauffer, Trodd, and Bernier, *Picturing Frederick Douglass*。已知有160张道格拉斯的照片和126张亚伯拉罕·林肯的照片。格兰特的照片数量估计为150张。十九世纪其他被高频拍摄的人物包括乔治·卡斯特，有155张照片；红云，

有128张；以及沃尔特·惠特曼，有127张。另见 Varon，“Most Photographed Man of His Era.”

28. Douglass, “Negro Portraits.” 关于摄影在非裔美国人经历中作用的广泛当代讨论，见，例如，Lewis, “Vision & Justice.”

29. Frederick Douglass, 致Louis Prang的信，1870年6月14日。

30. Frederick Douglass, 致Louis Prang的信，1870年6月14日。

31. Roth, “Looking at Shirley, the Ultimate Norm.”

32. 参见 Roth, 以及 McFadden, “Teaching the Camera to See My Skin,” 和 Caswell, “Color Film Was Built for White People.”

33. Roth, “Looking at Shirley, the Ultimate Norm.”

34. Roth.

35. Roth.

36. 这与 machine learning 中更广泛的问题相关，被称为分布偏移(*distributional shift*)：当一个在一组示例上训练的系统发现自己在不同类型的环境中运行时，而不一定意识到这种变化。Amodei et al., “Concrete Problems in AI Safety.” 概述了这个问题，这在本书后续章节中会多次涉及。

37. Hardt, “How Big Data Is Unfair.”

38. Jacky Alciné, 个人访谈，2018年4月19日。

39. Joy Buolamwini, “How I’m Fighting Bias in Algorithms,”
https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms.

40. Friedman and Nissenbaum, “Bias in Computer Systems.”

41. Buolamwini, “How I’m Fighting Bias in Algorithms.”

42. Huang et al., “Labeled Faces in the Wild.”

43. Han and Jain, “Age, Gender and Race Estimation from Unconstrained Face Images.”

44. 这里使用的估计是252张黑人女性面孔，通过将数据集中女性的比例(2,975/13,233)乘以数据集中黑人的比例(1,122/13,233)得出；数据来自Han and Jain。

45. 参见 Labeled Faces in the Wild, <http://vis-www.cs.umass.edu/lfw/>。根据互联网档案馆的Wayback Machine，免责声明出现在2019年9月3日至10月6日之间。

46. Klare et al., “Pushing the Frontiers of Unconstrained Face Detection and Recognition.”

47. Buolamwini and Gebru, “Gender Shades.”

48. 该数据集被设计为包含大致相等比例的所有六种肤色类别，这些类别通过皮肤学的“Fitzpatrick量表”测量。（值得注意的是，该量表之前是四类量表，有三个浅肤色类别和一个深肤色的综合类别，后来在1980年代扩展为三个独立类别。）

49. 参见 Joy Buolamwini, “AI, Ain’t I a Woman?,” <https://www.youtube.com/watch?v=QxuyfWoVV98>.

50. 关于微软的完整回应，见 <http://gendershades.org/docs/msft.pdf>.

51. 关于IBM的正式回应，见 <http://gendershades.org/docs/ibm.pdf>。IBM随后致力于构建一个包含一百万张面孔的新数据集，强调各种多样性措施；见 Merler et al., “Diversity in Faces.” 关于IBM构建其Diversity in Faces数据集方法的批评，见 Crawford and Paglen, “Excavating AI.”

52. 同样重要的是该领域本身的构成；见 Gebru, “Race and Gender.”

53. Firth, *Papers in Linguistics, 1934 – 1951*.

54. 实际上有两种训练当代词嵌入(word-embedding)模型的方法。一种是根据上下文预测缺失的单词，另一种相反：根据给定单词预测上下文单词。这些方法分别称为“连续词袋”(continuous bag-of-words, CBOW)和“跳字模型”(skip-gram)。为简单起见，我们的讨论主要关注前者，但两种方法都有优势，尽管它们最终往往产生相当相似的模型。

55. Shannon, “A Mathematical Theory of Communication.”

56. 参见 Jelinek and Mercer, “Interpolated Estimation of Markov Source Parameters from Sparse Data,” 和 Katz, “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer”；关于概述，见 Manning and Schütze, *Foundations of Statistical Natural Language Processing*.

57. 这个著名短语源自 Bellman, *Dynamic Programming*.

58. 参见 Hinton, “Learning Distributed Representations of Concepts,” 和 “Connectionist Learning Procedures,” 以及 Rumelhart and McClelland, *Parallel Distributed Processing*.

59. 参见，例如，潜在语义分析(见 Landauer, Foltz, 和 Laham, “An Introduction to Latent Semantic Analysis”), 多重原因混合模型(见 Saund, “A Multiple Cause Mixture Model for Unsupervised Learning” 和 Sahami, Hearst, 和 Saund, “Applying the Multiple Cause Mixture Model to Text Categorization”), 以及潜在狄利克雷分配(见 Blei, Ng, 和 Jordan, “Latent Dirichlet Allocation”).

60. 参见 Bengio et al., “A Neural Probabilistic Language Model”；概述见 Bengio, “Neural Net Language Models”。

61. 由于某些技术原因，原始的 word2vec 模型实际上为每个词都有两个向量——一个用于它出现为缺失词时，另一个用于它出现在缺失词上下文中时——因此参数总数会是两倍。

相似度通过计算两个向量之间的距离来度量——通过它们的“点积”——或者它们指向同一方向的程度——通过它们的“余弦相似度”。当向量长度相同时，这些度量是等价的。

关于以这种空间方式定义“相似度”的批评，指出了这种方法在反映人类相似度判断方面的局限性(这些判断并不总是对称的：例如，人们倾向于认为北韩比中国更“相似”于中国，而不是中国“相似”于北韩)，参见 Nematzadeh, Meylan, 和 Griffiths, “Evaluating Vector-Space Models of Word Representation”。

62. 有关 word2vec 模型训练方式的更多信息，参见 Rong, “Word2vec Parameter Learning Explained”。
63. Manning, “Lecture 2: Word Vector Representations”。
64. 正如他在1784年的论文《世界公民观点下的普遍历史理念》(“Idee zu einer allgemeinen Geschichte in weltbürgerlicher Absicht”)中所说，“Aus so krummem Holze, als woraus der Mensch gemacht ist, kann nichts ganz Gerades gezimmert werden.” 这里简洁的英译版本归功于以赛亚·柏林。
65. 参见，例如，Mikolov, Le, 和 Sutskever, “Exploiting Similarities Among Languages for Machine Translation”，Le 和 Mikolov, “Distributed Representations of Sentences and Documents”，以及 Kiros et al., “Skip-Thought Vectors”。
66. 在机器学习社区内部，关于如何精确计算这些“类比”存在重大分歧，在认知科学社区内部，关于它们在多大程度上捕捉了人类的相似度概念也存在分歧。更多关于这些问题的讨论请参见结论(及其尾注)。
67. Mikolov, “Learning Representations of Text Using Neural Networks”。
68. Bolukbasi et al., “Man Is to Computer Programmer as Woman Is to Homemaker?” 也许更令人震惊的是概念如何映射到种族。例如，在向量空间中最接近 white + male 的术语是 entitled to。最接近 black + male 的术语是 assaulted。(参见 Bolukbasi et al., “Quantifying and Reducing Stereotypes in Word Embeddings。”)如果你执行减法 white - minority 并将所有职业词汇映射到这个轴上，最偏向白人方向的职业是——具有讽刺意味的是，考虑到 Buolamwini 和 Gebru 用来重新校准面部检测系统的数据集——parliamentarian。最偏向少数族裔方向的职业是 butler。
69. 有关搜索排名中词嵌入的更多信息，参见 Nalisnick et al., “Improving Document Ranking with Dual Word Embeddings”；有关招聘中词嵌入的更多信息，参见 Hansen et al., “How to Get the Best Word Vectors for Resume Parsing”。
70. 参见 Gershgorn, “Companies Are on the Hook If Their Hiring Algorithms Are Biased”。
71. Bertrand 和 Mullainathan, “Are Emily and Greg More Employable Than Lakisha and Jamal?” 另见 Moss-Racusin et al., “Science Faculty’s Subtle Gender Biases Favor Male Students”，该研究在性别方面展示了类似的效果。
72. 当然，人类招聘者自身也可能受到机器学习的影响。哈佛大学的 Latanya Sweeney 在2013年对 Google AdSense 进行的一项开创性研究显示，暗示某人有犯罪记录的在线广告(无论他们实际上是否有犯罪记录)更有可能出现在“听起来像黑人”姓名的 Google 搜索结果旁边。Sweeney 指出了这对完成租房申请、申请贷款或求职的人可能造成的影响。有关分析和建议的解决方案，参见 Sweeney, “Discrimination in Online Ad Delivery”。
73. 关于管弦乐队试音中偏见的经典研究，参见 Goldin 和 Rouse, “Orchestrating Impartiality”。据作者介绍，一些管弦乐队使用地毯来达到同样的效果，有些甚至让男性提供“补偿性脚步声”。近年来，一些学者质疑了这篇经典论文结果的稳健性；参见 Sommers, “Blind Spots in the ‘Blind Audition’ Study”。
74. 这个想法被广泛称为“冗余编码”。例如参见 Pedreshi, Ruggieri, 和 Turini, “Discrimination-Aware Data Mining”。
75. Dastin, “亚马逊废弃对女性有偏见的秘密AI招聘工具。”
76. 同样值得注意的是，那些简历被该模型忽略、从未接到亚马逊招聘人员电话的潜在员工，可能根本不知道自己曾经在候选人池中。

77. 路透社在2018年报道称，亚马逊组建了一个新团队”再次尝试自动化就业筛选，这次专注于多样性”。关于招聘和偏见的计算方法研究，参见Kleinberg和Raghavan的”隐性偏见存在下的选择问题”。

78. Bolukbasi等人，“男性之于计算机程序员如同女性之于家庭主妇？”（另见Schmidt的”拒绝性别二元论”中的类似观点讨论。）Prost、Thain和Bolukbasi的”去偏见嵌入以减少文本分类中的性别偏见”重新审视了这一观点。

79. 更多内容，参见Bolukbasi等人的”男性之于计算机程序员如同女性之于家庭主妇？”

80. Bolukbasi等人，“男性之于计算机程序员如同女性之于家庭主妇？”

81. Tolga Bolukbasi，个人访谈，2016年11月11日。

82. 关于依赖Mechanical Turk参与者方法论的批评，以及它如何导致ImageNet数据集和其他数据集的问题，参见Crawford和Paglen的”挖掘AI”。

83. 事实上，“grandfather in”这个表达最初源于美国重建时期吉姆·克劳法的歧视性”祖父条款”。例如，《纽约时报》在1899年8月3日描述了这样一项法规：“它还规定，任何在1867年有投票资格者的后代，无论现有条件如何，现在都可以投票。这被称为’祖父条款’。”

84. Bolukbasi等人，“男性之于计算机程序员如同女性之于家庭主妇？”

85. Gonen和Goldberg，“给猪涂口红”。

86. DeepMind的Geoffrey Irving认为（个人通信），“词嵌入(Word embeddings)作为模型过于简单，无法在不丢失有用性别信息的情况下避免。你需要更智能的东西，能够从其他上下文理解是否应该考虑鞋子，这最终会以词嵌入所不具备的非线性和非凸方式呈现。当然，’我想我们需要更强大的模型来解决问题’这种一般模式是一个复杂而有趣的福音。”关于将更强大和复杂的语言模型与人类偏好对齐的更多内容，参见Ziegler等人的”基于人类偏好微调语言模型”。

87. Prost、Thain和Bolukbasi，“去偏见嵌入以减少文本分类中的性别偏见”。

88. Greenwald、McGhee和Schwartz，“测量隐性认知的个体差异”。

89. Caliskan、Bryson和Narayanan，“从语言语料库自动派生的语义包含类人偏见”。

90. Caliskan、Bryson和Narayanan。

91. Garg等人，“词嵌入量化100年的性别和种族刻板印象”。

92. Caliskan、Bryson和Narayanan，“从语言语料库自动派生的语义包含类人偏见”。

93. Narayanan在Twitter上：https://twitter.com/random_walker/status/993866661852864512。

94. 更新的语言模型，包括OpenAI的2019年GPT-2（参见Radford等人的”语言模型是无监督多任务学习器”）和Google的BERT（参见Devlin等人的”BERT：用于语言理解的深度双向Transformers预训练”），比word2vec更加复杂且性能更高，但表现出类似的刻板印象输出。例如，哈佛认知科学家Tomer Ullman给GPT-2两个类似的提示——“我妻子刚得到一份令人兴奋的新工作”和“我丈夫刚得到一份令人兴奋的新工作”——发现它倾向于以可预测的刻板印象方式完成段落。“妻子”会产生”做家务”和”全职妈妈”之类的内容，而”丈夫”会产生”银行顾问，同时

也是医生”之类的内容（非常令人印象深刻！）。参见<https://twitter.com/TomerUllman/status/1101485289720242177>。OpenAI研究人员一直在认真思考如何基于人类反馈”微调”其系统的输出；这是”去偏见”此类模型的一条可能路径，还有其他有前途的用途，尽管它在技术和其他方面都不乏复杂性。参见Ziegler等人的”基于人类偏好微调语言模型”。同样，研究人员已经在BERT模型中显示了偏见模式（参见Kurita等人的”测量上下文化词表示中的偏见”和Munro的”AI中的多样性不是你的问题，而是她的问题”）。“我们意识到这个问题，正在采取必要步骤来解决它，”谷歌发言人在2019年告诉《纽约时报》。“减少我们系统中的偏见是我们AI原则之一，也是重中之重”（参见Metz的”我们教AI系统一切，包括我们的偏见”）。

95. Yonatan Zunger, “So, About this Googler’s Manifesto,” <https://medium.com/@yonatanzunger/so-about-this-googlers-manifesto-1e3773ed1788>.

第二章 公平性

1. Kinsley, “What Convict Will Do If Paroled.”
2. 在 *Buck v. Davis* 案中：于 2016 年 10 月 5 日 辩 论；2017 年 2 月 22 日 裁 决；
https://www.supremecourt.gov/opinions/16pdf/15-8049_f2ah.pdf.
3. Hardt, “How Big Data Is Unfair.”
4. Clabaugh, “Foreword.”
5. Burgess, “Factors Determining Success or Failure on Parole.”
6. Clabaugh, “Foreword.”
7. Ernest W. Burgess and Thorsten Sellen, Introduction to Ohlin, *Selection for Parole*.
8. Tim Brennan, 个人访谈，2019年11月26日。
9. 参见Entwistle和Wilson, *Degrees of Excellence*, 由Brennan的导师撰写并总结了他的博士研究。
10. 关于Brennan和Wells在1990年代早期在监狱囚犯分类方面的工作详情，参见Brennan and Wells, “The Importance of Inmate Classification in Small Jails.”
11. Harcourt, *Against Prediction*.
12. Burke, *A Handbook for New Parole Board Members*.
13. Northpointe创始人Tim Brennan和Dave Wells在1998年开发了他们称为COMPAS的工具。关于COMPAS的更多细节，参见Brennan, Dieterich, and Oliver, “COMPAS,” 以及Brennan and Dieterich, “Correctional Offender Management Profiles for Alternative Sanctions (COMPAS).” COMPAS被Andrews, Bonta, and Wormith, “The Recent Past and Near Future of Risk and/or Need Assessment” 描述为“第四代”工具。COMPAS之前的主要“第三代”风险评估工具之一称为服务水平清单(Level of Service Inventory, LSI)，随后是服务水平清单修订版(Level of Service Inventory – Revised, LSI-R)。例如，参见Andrews, “The Level of Service Inventory (LSI),” 和Andrews and Bonta, “The Level of Service Inventory – Revised.” 关于佛罗里达州布劳沃德县采用COMPAS的更多信息，参见Blomberg et al., “Validation of the COMPAS Risk Assessment Classification Instrument.”
14. 特别地，暴力再犯分数为 $(\text{年龄} \times -w_1) + (\text{首次被捕年龄} \times -w_2) + (\text{暴力史} \times w_3) + (\text{职业教育} \times w_4) + (\text{不合规史} \times w_5)$ ，其中权重 w 通过统计方法确定。参见<http://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf>, § 4.1.5.
15. 参见《纽约州综合法律》，《行政法》 – EXC § 259-c：“州假释委员会；职能、权力和职责。”
16. “New York’s Broken Parole System.”
17. “A Chance to Fix Parole in New York.”

18. Smith, “In Wisconsin, a Backlash Against Using Data to Foretell Defendants’ Futures.”
19. “Quantifying Forgiveness: MLTalks with Julia Angwin and Joi Ito,” <https://www.youtube.com/watch?v=qjmkTGfu9Lk>.
关于Steve Jobs，参见Eric Johnson, “It May Be ‘Data Journalism,’ but Julia Angwin’s New Site the Markup Is Nothing Like FiveThirtyEight,” <https://www.recode.net/2018/9/27/17908798/julia-angwin-markup-jeff-larson-craig-newmark-data-investigative-journalism-peter-kafka-podcast>.
20. 该书是Angwin, *Dragnet Nation*.
21. Julia Angwin, 个人访谈, 2018年10月13日。
22. Lansing, “New York State COMPAS-Probation Risk and Need Assessment Study.”
23. Podkopacz, Eckberg, and Kubits, “Fourth Judicial District Pretrial Evaluation.”
24. Podkopacz, “Building and Validating the 2007 Hennepin County Adult Pretrial Scale.”
25. 另参见Harcourt, “Risk as a Proxy for Race,” 该文论证, “如今风险已经归结为先前犯罪史, 而先前犯罪史已成为种族的代理变量。这两种趋势的结合意味着使用风险评估工具将显著恶化我们刑事司法系统中不可接受的种族差异。” 关于反驳观点, 参见Skeem and Lowenkamp, “Risk, Race, and Recidivism.”
26. Julia Angwin, “Keynote,” Justice Codes Symposium, John Jay College, 2016 年 10 月 12 日 , <https://www.youtube.com/watch?v=WL9QkAwgqfU>.
27. Julia Angwin, 个人访谈, 2018年10月13日。
28. Angwin et al., “Machine Bias.”
29. Dieterich, Mendoza, and Brennan, “COMPAS Risk Scales.” 另参见Flores, Bechtel, and Lowenkamp, “False Positives, False Negatives, and False Analyses.”
30. 参见” Response to ProPublica.”
31. 参见Angwin and Larson, “ProPublica Responds to Company’s Critique of Machine Bias Story,” 和Larson and Angwin, “Technical Response to Northpointe.”
32. Angwin and Larson, “ProPublica Responds to Company’s Critique of Machine Bias Story.” 另见 Larson et al., “How We Analyzed the COMPAS Recidivism Algorithm.” 请注意, 这个引用中存在技术上的不准确。“被评为高风险但[未]再次犯罪”的个体衡量数学上可以转化为假阳性/(假阳性+真阳性)这一分数, 即假发现率(False Discovery Rate)。然而, ProPublica在此处实际指的不是假发现率, 而是假阳性率(False Positive Rate), 定义为假阳性/(假阳性+真阴性)这一分数。这个量的更好的语言表述应该颠倒ProPublica的语法: “未再次犯罪但被评为高风险”的被告。关于这一点的一些讨论, 见 <https://twitter.com/scorbettdavies/status/842885585240956928>。
33. 见 Dwork et al., “Calibrating Noise to Sensitivity in Private Data Analysis.” Google Chrome在2014年开始使用差分隐私(differential privacy), Apple在2016年将其部署在macOS Sierra和iOS 10操作系统中, 其他科技公司也跟进采用了许多相关的想法和实现。2017年, Dwork和她在2006年论文中的同事们因其工作共同获得了哥德尔奖(Gödel Prize)。
34. Cynthia Dwork, 个人访谈, 2018年10月11日。

35. Steel and Angwin, “On the Web’s Cutting Edge, Anonymity in Name Only.” 另见 Sweeney, “Simple Demographics Often Identify People Uniquely,” 该研究显示出生日期、性别和邮政编码的组合足以唯一识别87%的美国人。

36. Moritz Hardt, 个人访谈, 2017年12月13日。

37. 见 Dwork et al., “Fairness Through Awareness.” 关于此问题的更多讨论和辩论, 见如 Harcourt, “Risk as a Proxy for Race,” 和 Skeem and Lowenkamp, “Risk, Race, and Recidivism.”

38. 这一点在 Corbett-Davies, “Algorithmic Decision Making and the Cost of Fairness,” 以及 Corbett-Davies and Goel, “The Measure and Mismeasure of Fairness” 中得到讨论。

39. 关于这一点的最新论证, 见如 Kleinberg et al., “Algorithmic Fairness.” 关于可追溯到1990年代中期的讨论, 见如 Gottfredson and Jarjoura, “Race, Gender, and Guidelines-Based Decision Making.”

40. Kroll et al., “Accountable Algorithms.”

41. Dwork et al., “Fairness Through Awareness.”

42. 例如见 Johnson and Nissenbaum, “Computers, Ethics & Social Values.”

43. 例如见 Barocas and Selbst, “Big Data’s Disparate Impact.”

44. Jon Kleinberg, 个人访谈, 2017年7月24日。

45. Alexandra Chouldechova, 个人访谈, 2017年5月16日。

46. Sam Corbett-Davies, 个人访谈, 2017年5月24日。

47. Goel的研究显示, 除了其他发现外, 因所谓“鬼祟动作”而被记录实际上使某人不太可能是罪犯, 而不是如果他们没有被记录——“因为这表明你没有更好的理由”来证明拦截他们的合理性。(Sharad Goel, 个人访谈, 2017年5月24日。) 见 Goel, Rao, and Shroff, “Personalized Risk Assessments in the Criminal Justice System.”

48. 见 Simoiu, Corbett-Davies, and Goel, “The Problem of Infra-Marginality in Outcome Tests for Discrimination.”

49. 分别见 Kleinberg, Mullainathan, and Raghavan, “Inherent Trade-offs in the Fair Determination of Risk Scores”; Chouldechova, “Fair Prediction with Disparate Impact”; 和 Corbett-Davies et al., “Algorithmic Decision Making and the Cost of Fairness”。另见 Berk et al., “Fairness in Criminal Justice Risk Assessments.”

50. Kleinberg, Mullainathan, and Raghavan, “Inherent Trade-offs in the Fair Determination of Risk Scores.”

51. Alexandra Chouldechova, 个人访谈, 2017年5月16日。

52. Sam Corbett-Davies, 个人访谈, 2017年5月24日。具有讽刺意味的是, ProPublica正是因为这一事实而成为头条新闻; 见 Julia Angwin and Jeff Larson, “Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say,” ProPublica, 2016年12月30日。

53. Corbett-Davies, “Algorithmic Decision Making and the Cost of Fairness.”

54. Sam Corbett-Davies, 个人访谈, 2017年5月24日。

55. Kleinberg, Mullainathan, and Raghavan, “Inherent Trade-offs in the Fair Determination of Risk Scores.”
56. 关于借贷背景下公平性的详细讨论，特别见 Hardt, Price, and Srebro, “Equality of Opportunity in Supervised Learning,” 和 Lydia T. Liu, et al., “Delayed Impact of Fair Machine Learning,” 以及在 <http://research.google.com/bigpicture/attacking-discrimination-in-ml/> 和 <https://bair.berkeley.edu/blog/2018/05/17/delayed-impact/> 的交互式可视化。
57. Sam Corbett-Davies et al., “Algorithmic Decision Making and the Cost of Fairness” (video), <https://www.youtube.com/watch?v=iFEX07OunSg>.
58. Corbett-Davies, “Algorithmic Decision Making and the Cost of Fairness.”
59. Corbett-Davies.
60. Tim Brennan, 个人访谈, November 26, 2019.
61. 参见 Corbett-Davies and Goel, “The Measure and Mismeasure of Fairness”；另见 Corbett-Davies et al., “Algorithmic Decision Making and the Cost of Fairness.”
62. 参见, 例如, Rezaei et al., “Fairness for Robust Log Loss Classification.”
63. Julia Angwin, 个人访谈, October 13, 2018.
64. Flores, Bechtel, and Lowenkamp, “False Positives, False Negatives, and False Analyses.”
65. Tim Brennan, 个人访谈, November 26, 2019.
66. Cynthia Dwork, 个人访谈, October 11, 2018.
67. Moritz Hardt, 个人访谈, December 13, 2017.
68. 加利福尼亚州SB 10法案的通过促使AI伙伴关系（代表13个国家的90多个组织）发布了一份详细报告，呼吁任何拟议的风险评估模型都应满足十项不同的标准。参见 “Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System.”
69. 该工具称为囚犯评估工具目标估计风险和需求(PATTERN)，于2019年7月19日发布。
70. Alexandra Chouldechova, 个人访谈, May 16, 2017.
71. Burgess, “Factors Determining Success or Failure on Parole.”
72. Lum and Isaac, “To Predict and Serve?”
73. “Four Out of Ten Violate Parole, Says Legislator.”
74. 参见 Ensign et al., “Runaway Feedback Loops in Predictive Policing.”
75. Lum and Isaac, “To Predict and Serve?”

76. Lum and Isaac. 为了了解数据集到底有多偏见，需要知道所有未报告犯罪发生的地点。这在定义上几乎听起来是不可能的。但Lum和Isaac有一个巧妙的方法来取得进展。他们使用来自全国药物使用和健康调查的数据，能够在城市中创建一个精细的、大致按街区划分的非法药物使用估计地图，并将其与同一城市的逮捕记录进行比较。

77. Alexandra Chouldechova, 个人访谈, May 16, 2017.

78. 参见 ACLU Foundation, “The War on Marijuana in Black and White.”

79. 参见 Mueller, Gebeloff, and Chinoy, “Surest Way to Face Marijuana Charges in New York.”

80. 关于这一论点的更多讨论，参见，例如，Sam Corbett-Davies, Sharad Goel, and Sandra González-Bailón, “Even Imperfect Algorithms Can Improve the Criminal Justice System,” <https://www.nytimes.com/2017/12/20/upshot/algorithms-bail-criminal-justice-system.html>; “Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System”；以及 Skeem and Lowenkamp, “Risk, Race, and Recidivism.”

81. 参见 Angwin et al., “Machine Bias.” 关于此类工具在量刑中的适当性问题一直上诉到威斯康星州最高法院，最终确认使用COMPAS风险评分来指导量刑判决是合适的。参见 *State v. Loomis*; 总结见 <https://harvardlawreview.org/2017/03/state-v-loomis/>. 在量刑中使用风险评估本身就是一个话题。前美国司法部长Eric Holder曾论证：“刑事判决不应基于人无法控制的不可改变因素，或基于尚未发生的未来犯罪的可能性。” Monahan and Skeem, “Risk Assessment in Criminal Sentencing.” 讨论了量刑中责任和风险的混淆。在 Skeem and Lowenkamp, “Risk, Race, and Recidivism” 中，Lowenkamp “建议不要使用PCRA [风险评估工具] 来指导前端量刑决定或后端释放决定，除非首先对其在这些情况下的使用进行研究，因为PCRA并非为这些目的而设计。”

82. Harcourt, *Against Prediction*. 进一步讨论，参见，例如，Persico, “Racial Profiling, Fairness, and Effectiveness of Policing” 和 Dominitz and Knowles, “Crime Minimisation and Racial Bias.”

83. Saunders, Hunt, and Hollywood, “Predictions Put into Practice.” 另见芝加哥警察局的回复：“CPD Welcomes the Opportunity to Comment on Recently Published RAND Review.”

84. 另见 Saunders, “Pitfalls of Predictive Policing.”

85. 对于Bernard Harcourt（在他的“Risk as a Proxy for Race”中），更明智的假释决定——无论是机器驱动的还是其他方式的——虽然显然比愚蠢的决定更好，但并不是解决美国监狱过度拥挤和种族差异的主要方式：

那么应该采取什么措施来减少监狱人口？我认为，与其通过预测来释放，我们需要在前端减少惩罚性措施，并始终高度关注我们量刑法律中的种族不平衡。将crack-cocaine量刑差距减少到18:1是朝着正确方向迈出的一步；然而，其他直接措施应该包括取消强制性最低监禁期限，减少毒品量刑法律，替换为转介和替代监督项目，以及减少重刑的施加。研究表明，缩短刑期长度（即比刑期到期更早释放低风险罪犯）对监狱人口的长期影响不如减少收监人数大。因此，真正的解决方案不是缩短监禁期限，而是减少监狱收监人数。

86. 关于这一点的更多信息，参见Barabas等人，“Interventions over Predictions”。

87. Elek, Sapia, 和 Keilitz, “Use of Court Date Reminder Notices to Improve Court Appearance Rates”。在2019年的一项发展中，Hardt发现特别令人鼓舞的是，包括休斯顿在内的德克萨斯州Harris County批准了一项法律和解，涉及承诺开发基于短信的系统来提醒人们即将到来的法庭出庭安排。例如，参见Gabrielle Banks, “Federal Judge Gives Final Approval to Harris County Bail Deal,” *Houston Chronicle*, November 21, 2019。

88. 参见Mayson, “Dangerous Defendants,” 和 Gouldin, “Disentangling Flight Risk from Dangerousness”。另见” Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System”，该报告认为”工具不得混淆多重预测”。

89. Tim Brennan, 个人访谈, 2019年11月26日。

90. 另见, 例如, Goswami, “Unlocking Options for Women”, 一项关于芝加哥Cook County监狱女性的研究, 该研究得出结论, 法官应该被授权”判决女性接受服务而不是监禁”。

91. Moritz Hardt, 个人访谈, 2017年12月13日。

92. 另见, 例如, Mayson, “Bias in, Bias Out”, 该文认为, “在一个种族分层的世界中, 任何预测方法都会将过去的不平等投射到未来。长期以来渗透刑事司法的主观预测如此, 现在取代它的算法工具也是如此。算法风险评估所做的是揭示所有预测中固有的不平等, 迫使我们面对一个比新技术挑战更大的问题。”

93. Burgess, “Prof. Burgess on Parole Reform”。

第3章 透明度

1. Graeber, *The Utopia of Rules*。
2. Berk, *Criminal Justice Forecasts of Risk*。
3. 参见Cooper等人, “An Evaluation of Machine-Learning Methods for Predicting Pneumonia Mortality,” 和 Cooper等人, “Predicting Dire Outcomes of Patients with Community Acquired Pneumonia”。
4. 参见Caruana等人, “Intelligible Models for Healthcare”。
5. Cooper等人, “Predicting Dire Outcomes of Patients with Community Acquired Pneumonia”。
6. Caruana, “Explainability in Context—Health”。
7. 关于决策列表的更多信息, 参见Rivest, “Learning Decision Lists”。关于决策列表在医学中使用的最新讨论, 参见Marewski和Gigerenzer, “Heuristic Decision Making in Medicine”。关于决策集中解释的更多信息, 参见Lakkaraju, Bach, 和 Leskovec, “Interpretable Decision Sets”。
8. 在一个得出结论“如果患者患有哮喘, 他们就是低风险”的系统中, 明显缺失的一点是因果关系模型。研究因果关系的顶级计算机科学家之一是UCLA的Judea Pearl; 关于他在当代machine-learning系统背景下对因果关系的最新思考, 参见Pearl, “The Seven Tools of Causal Inference, with Reflections on Machine Learning”。
9. Rich Caruana, 个人访谈, 2017年5月16日。
10. Hastie和Tibshirani, “Generalized Additive Models”。Caruana和他的合作者还探索了一类稍微复杂的模型, 该模型还包括成对交互, 或两个变量的函数。他们称这些为“GA²M”, 或“Generalized Additive Models plus Interactions”; 参见Lou等人, “Accurate Intelligible Models with Pairwise Interactions”。
11. Caruana说这有许多不同的原因。退休对一些人来说意味着生活方式的改变, 这也意味着人们的收入发生变化, 他们的保险甚至可能医疗保健提供者发生变化, 他们也可能搬家——所有这些都改变了他们与健康和医疗保健的关系。
12. 广义加性模型显示风险在86岁时急剧上升, 但在101岁时又急剧下降。Caruana认为这些纯粹是社会效应; 他推测在80多岁中期左右, 家庭和护理人员更倾向于将健康危机解释为不应该拼命抗争的自然过程。另一方面, 一旦某人达到100岁, 人们几乎有相反的冲动: “你已经走了这么远; 我们现在不会放弃你。”他指出, 医生可能想要编辑这个图表——会决定哮喘规则没有意义, 会决定不对80岁、90岁和100岁的病人区别对待。另一方面, 保险公司可能不想在他们的模型中编辑这个图表。从保险公司的角度来看, 哮喘患者的结果平均确实更好。这突出了明确考虑系统中不同利益相关者的不同观点的重要性, 以及一个群体正在使用模型进行实际的、现实世界的干预, 这些干预反过来会改变观察到的基础数据; 另一个只是被动的观察者。机器学习本身并不知道其中的区别。
13. 参见Lou等人的研究。
14. Schauer, 《给出理由》。
15. David Gunning, 个人访谈, 2017年12月12日。

16. Bryce Goodman, 个人访谈, 2018年1月11日。“解释权”首次在Goodman和Flaxman的《欧盟算法决策法规和‘解释权’》中讨论。一些学者对这一条款的强度进行了辩论; 参见Wachter、Mittelstadt和Floridi的《为什么通用数据保护条例中不存在自动化决策解释权》。其他人对此进行了后续研究——例如, 参见Selbst和Powles的《有意义的信息和解释权》——某种程度的分歧仍在继续。“解释权”的确切法律地位可能只会在法院中逐步澄清。

17. Thorndike, 《评判人的基本定理》。

18. Robyn Dawes, 《Dawes畅所欲言》, Joachim Krueger访谈, 《理性与社会责任》, 卡内基梅隆大学, 2007年1月19日。

19. Sarbin, 《对精算和个人预测方法研究的贡献》。

20. Meehl, 《我那本令人不安的小书的原因和影响》。

21. Dawes和Corrigan, 《决策中的线性模型》, 引用Sarbin的《对精算和个人预测方法研究的贡献》。

22. 参见Dawes的《决策中不当线性模型的稳健之美》。

23. 参见Goldberg的《简单模型还是简单过程?》。

24. 参见Einhorn的《专家测量和机械组合》。

25. 关于Paul Meehl 1986年著作的回顾, 参见Meehl的《我那本令人不安的小书的原因和影响》。关于Dawes和Meehl在1989年的观点, 参见Dawes、Faust和Meehl的《临床判断与精算判断》。关于这些问题的当代观点, 例如, 参见Kleinberg等人的《人类决策和机器预测》。

26. Holte, 《非常简单的分类规则在最常用的数据集上表现良好》。

27. Einhorn, 《专家测量和机械组合》。

28. 参见Goldberg的《人与人的模型》, 以及Dawes的《研究生录取案例研究》。

29. Dawes和Corrigan, 《决策中的线性模型》。另见Wainer的《估计线性模型中的系数》, 特别详述了等权重; 正如他所写: “当你对预测感兴趣时, 很少有情况需要不等的回归权重。”另见Dana和Dawes的《社会科学预测中简单替代方案优于回归》, 在社会科学(和21世纪)背景下肯定了这一结论。

30. 参见Dawes的《决策中不当线性模型的稳健之美》。

31. Howard和Dawes, 《婚姻幸福的线性预测》。

32. 参见Howard和Dawes, 引用Alexander的《婚姻和婚前关系中的性、争论和社会接触》。

33. 实际上, Paul Meehl本人得出结论: “在大多数实际情况下, 少数‘大’变量的未加权总和平均而言会比回归方程更可取。”参见Dawes和Corrigan的《决策中的线性模型》进行讨论和参考文献。

34. Dawes, 《决策中不当线性模型的稳健之美》。另见Wainer的《估计线性模型中的系数》: “还要注意, 即使没有可操作的标准, 这种方案[线性模型中的等权重]也能很好地工作。”

35. Dawes, 《决策中不当线性模型的稳健之美》。

36. Einhorn, “Expert Measurement and Mechanical Combination.”
37. Dawes and Corrigan, “Linear Models in Decision Making.”
38. 参见 Andy Reinhardt, “Steve Jobs on Apple’s Resurgence: ‘Not a One-Man Show,’ ” *Business Week Online*, May 12, 1998, <http://www.businessweek.com/bwdaily/dnflash/may1998/nf80512d.htm>.
39. Holmes and Pollock, *Holmes-Pollock Letters*.
40. Angelino et al., “Learning Certifiably Optimal Rule Lists for Categorical Data.” 另见 Zeng, Ustun, and Rudin, “Interpretable Classification Models for Recidivism Prediction”；以及 Rudin and Radin, “Why Are We Using Black Box Models in AI When We Don’t Need To?” 关于另一个达到与COMPAS相似准确率的简单模型，参见 Dressel and Farid, “The Accuracy, Fairness, and Limits of Predicting Recidivism.” 进一步讨论，参见 Rudin, Wang, and Coker, “The Age of Secrecy and Unfairness in Recidivism Prediction,” 以及 Chouldechova, “Transparency and Simplicity in Criminal Risk Assessment.”
41. Cynthia Rudin, “Algorithms for Interpretable Machine Learning” (讲座), 第20届ACM SIGKDD知识发现与数据挖掘会议，纽约市，2014年8月26日。
42. Breiman et al., *Classification and Regression Trees*.
43. 参见 Quinlan, C4.5; C4.5还有一个更新的后续算法C5.0。
44. 关于CHADS₂的更多信息，参见 Gage et al., “Validation of Clinical Classification Schemes for Predicting Stroke,” 关于 CHA₂DS₂-VASc 的更多信息，参见 Lip et al., “Refining Clinical Risk Stratification for Predicting Stroke and Thromboembolism in Atrial Fibrillation Using a Novel Risk Factor – Based Approach.”
45. 参见 Letham et al., “Interpretable Classifiers Using Rules and Bayesian Analysis.”
46. 参见，例如，Veasey and Rosen, “Obstructive Sleep Apnea in Adults.”
47. SLIM使用所谓的“0-1损失函数”（衡量预测正确或错误数量的简单方法）和“ l_0 范数”（试图最小化使用特征的数量），并限制其特征权重的系数为互质整数。参见 Ustun, Tracà, and Rudin, “Supersparse Linear Integer Models for Predictive Scoring Systems,” 以及 Ustun and Rudin, “Supersparse Linear Integer Models for Optimized Medical Scoring Systems.” 关于他们与麻省总医院合作创建睡眠呼吸暂停工具的更多信息，参见 Ustun et al., “Clinical Prediction Models for Sleep Apnea.” 关于他们在累犯背景下应用类似方法的工作，参见 Zeng, Ustun, and Rudin, “Interpretable Classification Models for Recidivism Prediction.” 关于更近期的工作，包括此类方法的“最优性证书”，以及与COMPAS的比较，参见 Angelino et al., “Learning Certifiably Optimal Rule Lists for Categorical Data”；Ustun and Rudin, “Optimized Risk Scores”；以及 Rudin and Ustun, “Optimized Scoring Systems.”
48. 例如，可能使用逻辑回归来构建模型，然后将系数四舍五入。
49. 参见 Ustun and Rudin, “Supersparse Linear Integer Models for Optimized Medical Scoring Systems,” 获取讨论和参考资料。
50. “Information for Referring Physicians,” <https://www.uwhealth.org/referring-physician-news/death-rate-triples-for-sleep-apnea-sufferers/13986>.

51. Ustun et al., “Clinical Prediction Models for Sleep Apnea.” 关于使用SLIM构建并已在医院部署用于评估癫痫发作风险的模型，参见 Struck et al., “Association of an Electroencephalography-Based Risk Score With Seizure Probability in Hospitalized Patients.”

52. 参见 Kobayashi and Kohshima, “Unique Morphology of the Human Eye and Its Adaptive Meaning,” 以及 Tomasello et al., “Reliance on Head Versus Eyes in the Gaze Following of Great Apes and Human Infants.”

53. 如何准确计算显著性(saliency)是一个活跃的研究领域。参见，例如，Simonyan, Vedaldi, and Zisserman, “Deep Inside Convolutional Networks”；Smilkov et al., “Smoothgrad”；Selvaraju et al., “Grad-Cam”；Sundararajan, Taly, and Yan, “Axiomatic Attribution for Deep Networks”；Erhan et al., “Visualizing Higher-Layer Features of a Deep Network”；以及 Dabkowski and Gal, “Real Time Image Saliency for Black Box Classifiers.” 关于在强化学习背景下雅可比和扰动为基础的显著性比较，参见 Greydanus et al., “Visualizing and Understanding Atari Agents.”

关于显著性方法的局限性和弱点也存在开放性研究问题。参见，例如，Kindermans et al., “The (Un)reliability of Saliency Methods”；Adebayo et al., “Sanity Checks for Saliency Maps”；以及 Ghorbani, Abid, and Zou, “Interpretation of Neural Networks Is Fragile.”

54. 如 Landecker 所说：“对数据集的仔细检查显示，许多动物图像的背景模糊，而非动物图像往往各处都很清晰。这种图像偏见是合理的，因为所有照片都是由专业摄影师拍摄的。贡献传播的结果向我们展示了意外偏见如何轻易潜入数据集中。”参见 Landecker, “Interpretable Machine Learning and Sparse Coding for Computer Vision”，以及 Landecker et al., “Interpreting Individual Classifications of Hierarchical Networks”。另见对一个（人为构造的）例子的讨论，其中一个设计用来区分狼和哈士奇的网络实际上主要区分的是图像背景中雪地或草地的差别：Ribeiro, Singh, and Guestrin, “Why Should I Trust You?”

55. Hilton, “The Artificial Brain as Doctor”。Novoa 在 2015 年 1 月 27 日向同事发了一封邮件，说：“如果 AI 能区分数百种狗的品种，我相信它能对皮肤科做出巨大贡献。”这促成了与 Ko 等人的合作。参见 Justin Ko, “Mountains out of Moles: Artificial Intelligence and Imaging”（讲座），Big Data in Biomedicine Conference, Stanford, CA, 2017 年 5 月 24 日，<https://www.youtube.com/watch?v=kClvKNl0Wfc>。

56. Esteva et al., “Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks”。

57. Ko, “Mountains out of Moles”。

58. Narla et al., “Automated Classification of Skin Lesions”。

59. 参见 Caruana, “Multitask Learning”；但也参见 Rosenberg and Sejnowski, “NETtalk”，作为前身。参见 Ruder, “An Overview of Multi-Task Learning in Deep Neural Networks”，以获得更近期的概述。这个想法有时也被称为制作具有多个“头部”的神经网络——共享相同中间级特征的高级输出。这个想法在机器学习社区获得适度关注后，最近在 2010 年代的旗舰神经网络之一 AlphaGo Zero 中重新出现。当 DeepMind 迭代他们击败冠军的 AlphaGo 架构时，他们意识到通过将两个主要网络合并成一个双头网络，他们构建的系统可以大大简化。原始的 AlphaGo 使用“策略网络”来估计在给定位置下该走哪步，使用“价值网络”来估计该位置下每个玩家的优势或劣势程度。据推测，DeepMind 意识到，相关的中间级“特征”——谁控制哪个领域、某些结构有多稳定或脆弱——对两个网络来说极其相似。为什么要重复？在他们随后的 AlphaGo Zero 架构中，“策略网络”和“价值网络”变成了连接到同一个深度网络的“策略头”和“价值头”。这个新的、像 Cerberus 一样的网络更简单、在哲学上更令人满意——也是比原版更强的棋手。（技术上，Cerberus 在神话中更典型地被描述为三头；他不太知名的兄弟 Orthrus 是一只守护 Geryon 牛群的双头狗。）

60. Rich Caruana, 个人访谈, 2017年5月16日。
61. Poplin et al., “Prediction of Cardiovascular Risk Factors from Retinal Fundus Photographs via Deep Learning”。
62. Ryan Poplin, 接受 Sam Charington 采访, *TWiML Talk*, 第122集, 2018年3月26日。
63. Zeiler 和 Fergus, “Visualizing and Understanding Convolutional Networks”。
64. Matthew Zeiler, “Visualizing and Understanding Deep Neural Networks by Matt Zeiler”(讲座), <https://www.youtube.com/watch?v=ghEmQSxT6tw>。
65. 参见 Zeiler et al., “Deconvolutional Networks”, 以及 Zeiler, Taylor, 和 Fergus, “Adaptive Deconvolutional Networks for Mid and High Level Feature Learning”。
66. 到2014年, 几乎所有参加ImageNet基准测试竞争的团队都在使用这些技术和见解。参见 Simonyan 和 Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”; Howard, “Some Improvements on Deep Convolutional Neural Network Based Image Classification”; 以及 Simonyan, Vedaldi, 和 Zisserman, “Deep Inside Convolutional Networks”。在2018年和2019年, Clarifai内部对其图像识别软件是否会用于军事应用存在一些争议; 参见 Metz, “Is Ethical A.I. Even Possible?”
67. 他们方法的灵感包括 Erhan et al., “Visualizing Higher-Layer Features of a Deep Network”, 以及其他先前和同时期的研究; 参见 Olah, “Feature Visualization”以获得更完整的历史和参考书目。在实践中, 仅仅针对类别标签优化不会产生可理解的图像, 除非对目标有一些进一步的约束或调整。这是一个肥沃的研究领域; 参见 Mordvintsev, Olah, 和 Tyka, “Inceptionism”, 以及 Olah, Mordvintsev, 和 Schubert, “Feature Visualization”, 关于这一点的讨论。
68. Mordvintsev, Olah, 和 Tyka, “DeepDream”。
69. Yahoo的模型是open_nsfw, 可在https://github.com/yahoo/open_nsfw获取。Goh的工作, 不适合儿童或胆小者观看, 可在https://open_nsfw.gitlab.io获取, 基于Nguyen等人“Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks”一文中的方法。Goh随后加入了OpenAI的Olah Clarity团队。
70. 参见Mordvintsev, Olah和Tyka的”Inceptionism”, 以及Mordvintsev, Olah和Tyka的”DeepDream”。
71. 参见Olah, Mordvintsev和Schubert的”Feature Visualization”; Olah等人的”The Building Blocks of Interpretability”; 以及Carter等人的”Activation Atlas”。最近的工作包括对AlexNet等基石深度学习模型的详细”显微镜”研究; 例如参见<https://microscope.openai.com/models/alexnet>。
72. Chris Olah, 个人访谈, 2020年5月4日。更多信息见他的”Circuits”合作项目: <https://distill.pub/2020/circuits/>。
73. 该期刊是Distill, 网址为<https://distill.pub>。关于Olah对创立Distill的想法, 参见<https://colah.github.io/posts/2017-03-Distill/>和<https://distill.pub/2017/research-debt/>。
74. Olah等人, “The Building Blocks of Interpretability”。
75. Been Kim, 个人访谈, 2018年6月1日。
76. 另见Doshi-Velez和Kim的”Towards a Rigorous Science of Interpretable Machine Learning”, 以及Lage等人的”Human-in-the-Loop Interpretability Prior”。

77. 参见Poursabzi-Sangdeh等人的” Manipulating and Measuring Model Interpretability”。
78. 参见<https://github.com/tensorflow/tcav>。
79. Kim等人，“Interpretability Beyond Feature Attribution”。
80. 这产生的概念向量与我们在第1章讨论word2vec时看到的向量不同。相关方法另见Fong和Vedaldi的” Net2Vec”。
81. Been Kim, “Interpretability Beyond Feature Attribution”(讲座), MLconf 2018, 旧金山, 2018年11月14日,
<https://www.youtube.com/watch?v=Ff-Dx79QEEY>。
82. Been Kim, “Interpretability Beyond Feature Attribution”。
83. 参见Mordvintsev, Olah和Tyka的” Inceptionism” , 以及Mordvintsev, Olah和Tyka的” DeepDream”。
84. 参见<https://results.ittf.link>。
85. Stock和Cisse, “ConvNets and Imagenet Beyond Accuracy”。

第4章 强化

1. Skinner, “Reinforcement Today”。
 2. Arendt, *The Human Condition*.
 3. 关于Stein的本科研究，参见Solomons和Stein的“Normal Motor Automatism”。相关评论由B. F. Skinner撰写，将她的著名书籍与早期心理学研究联系起来，参见Skinner的“Has Gertrude Stein a Secret?”。关于Stein对这段生活时光的简要反思，参见Stein的*The Autobiography of Alice B. Toklas*。更多关于Stein的生平和影响，参见Brinnin的*The Third Rose*。
 4. Jonçich, *The Sane Positivist*。另见Brinnin, *The Third Rose*。
 5. Jonçich。
 6. Thorndike, “Animal Intelligence”。
 7. Thorndike, *The Psychology of Learning*.
 8. 当然，Thorndike既有前辈也有后继者；效果律的早期预兆可以在苏格兰哲学家Alexander Bain的工作中找到，他在1855年的*The Senses and the Intellect*中讨论了通过“摸索实验”和“试错的宏伟过程”进行学习——似乎创造了现在常见的短语。Conway Lloyd Morgan在Thorndike于哈佛工作的几年前，在他1894年的*Introduction to Comparative Psychology*中讨论了动物行为背景下的“试错”。从强化学习角度对动物学习的简短历史回顾，参见Sutton和Barto的*Reinforcement Learning*。
 9. 分别参见Thorndike的“*A Theory of the Action of the After-Effects of a Connection upon It*”和Skinner的“*The Rate of Establishment of a Discrimination*”。讨论见Wise的“*Reinforcement*”。
 10. Tolman, “The Determiners of Behavior at a Choice Point”。
 11. 参见Jonçich的*The Sane Positivist*，以及Cumming的“*A Review of Geraldine Jonçich's The Sane Positivist: A Biography of Edward L. Thorndike*”。
 12. Thorndike, “*A Theory of the Action of the After-Effects of a Connection upon It*”。
 13. Turing, “Intelligent Machinery”。
 14. “Heuristics”。
 15. Samuel, “Some Studies in Machine Learning Using the Game of Checkers”。
 16. McCarthy and Feigenbaum, “In Memoriam”。Samuel的电视演示于1956年2月24日进行。
- [17] Edward Thorndike致William James的信，1908年10月26日；见于Jonçich的《理性实证主义者》。
- [18] Rosenblueth, Wiener, 和 Bigelow，“行为、目的和目的论”。《牛津英语词典》区分了“返回输出信号的一部分”这一含义与“通过过程的结果或效果对过程或系统进行修改、调整或控制”这一含义，并引用Rosenblueth, Wiener, 和

Bigelow作为后一种含义的首个已知印刷实例。

[19] “cybernetics” 这个词对于当代人来说，既听起来未来主义又复古；它让人联想到《飞侠哥顿》和婴儿潮一代的科幻时代。实际上，这个词绝对是虚构的，而且它听起来远没有那么陌生。Wiener在寻找一个术语来概括生物和机械系统中自我调节和反馈的概念。“经过深思熟虑，”他写道，“我们得出结论，所有现有的术语都过于偏向某一方面，无法很好地服务于该领域的未来发展；正如科学家们经常遇到的情况，我们被迫创造至少一个人工的新希腊语表达来填补这一空白”（Wiener, 《Cybernetics》）。他在希腊语 $\kappa\upsilon\beta\epsilon\rho\nu\eta\tau\eta\varsigma$ ——或在罗马字母中为 *kybernetes*——中找到了他喜欢的词源，该词来自“舵手”、“船长”或“统治者”一词。实际上，英语单词“governor”本身就源于 *kybernetes*，只是拼写有所扭曲（被认为是伊特鲁里亚语的影响）。与许多新造词一样，早期在拼写上有一些差异；例如，1960年在伦敦出版了一本技术书籍，使用了另一种拼写：Stanley-Jones 和 Stanley-Jones 的《自然系统的 Kybernetics》。（“关于这个词的拼写，…我更倾向于Kybernetics，基于词源学理由。”）实际上，该术语的英语使用早于Wiener：James Clerk Maxwell在1868年使用它来描述电气“调节器”——这是对Wiener的有意暗示——在那之前（起初Wiener并不知道），André-Marie Ampère在1834年使用了它，带有航海操舵的含义，在社会科学和政治权力的背景下指代治理。根据Ampère的说法，这种从船只到城邦的比喻用法甚至在原始希腊语中就存在。分别见Maxwell的“On Governors”和Ampère的《科学哲学论；或人类所有知识自然分类的分析阐述》。

[20] Wiener, 《Cybernetics》。

[21] Rosenblueth, Wiener, 和 Bigelow, “行为、目的和目的论”。

[22] Klopff, 《脑功能与适应系统：一个异态理论》。“享乐神经元”的概念在机器学习历史中以略有不同的形式出现。例如，见Minsky在“神经模拟强化系统理论及其在脑模型问题中的应用”中讨论的“SNARC”系统作为早期例子，以及Sutton和Barto《强化学习》第15章的讨论。

[23] Andrew G. Barto, “强化学习：惊喜与联系的历史”（讲座），2018年7月19日，国际人工智能联合会议，瑞典斯德哥尔摩。

[24] Andrew Barto, 个人访谈，2018年5月9日。

[25] 关于强化学习的经典教材是Sutton和Barto的《强化学习》，最近更新为第二版。关于该领域到1990年代中期的总结，另见Kaelbling, Littman, 和 Moore的“强化学习”。

[26] Richard Sutton在<http://incompleteideas.net/rllai.cs.ualberta.ca/RLAI/rewardhypothesis.html>定义和讨论了这一概念，它也出现在Sutton和Barto的《强化学习》中。Sutton说他第一次听到这个概念是从布朗大学计算机科学家Michael Littman那里；Littman认为他第一次是从Sutton那里听到的。但最早的引用似乎是Littman在2000年代早期的一次讲座，他在其中论证“智能行为源于个体在复杂且不断变化的世界中寻求最大化其接收到的奖励信号的行为”。关于Littman对这段历史的回忆，见“Michael Littman：奖励假设”（讲座），阿尔伯塔大学，2019年10月16日，可在<https://www.coursera.org/lecture/fundamentals-of-reinforcement-learning/michael-littman-the-reward-hypothesis-q6x0e> 获取。

尽管这一特定框架出现较晚，但将行为理解为明确或隐含地由某种形式的可量化奖励所激励的概念，与效用理论有着广泛的联系。例如，见Bernouilli的“关于风险测度新理论的样本”，Samuelson的“关于效用测量的注记”，以及 von Neumann和Morgenstern的《博奕论与经济行为》。

[27] Richard Sutton, “强化学习简介”（讲座），德克萨斯大学奥斯汀分校，2015年1月10日。

28. “任何两个[标量]数字之间只有三种可能的比较，” Chang说。“一个数字大于、小于或等于另一个。但价值不是这样。作为后启蒙时代的生物，我们倾向于假设科学思维掌握着我们世界中一切重要事物的关键，但价值的世界与科学的世界不同。一个世界的内容可以用实数量化。另一个世界的内容不能。我们不应该假设是什么的世界——长度和重量的世界——与应该的世界——我们应该做什么的世界——具有相同的结构。” 参见Ruth Chang，“How to Make Hard Choices”（讲座），TEDSalon NY2014: https://www.ted.com/talks/ruth_chang_how_to_make_hard_choices。

29. 将强化学习视为“与批评者一起学习”的想法似乎至少可以追溯到Widrow、Gupta和Maitra的“Punish/Reward”。

30. 你可以将反向传播这样的算法看作是在结构上而不是时间上解决信用分配问题。正如Sutton在“Learning to Predict by the Methods of Temporal Differences”中所说，“反向传播和TD方法的目的都是准确的信用分配。反向传播决定网络的那个部分应该改变以影响网络的输出从而减少其整体误差，而TD方法决定时间输出序列的每个输出应该如何改变。反向传播解决结构性信用分配问题，而TD方法解决时间性信用分配问题。”

31. Olds，“Pleasure Centers in the Brain”，1956。

32. Olds和Milner，“Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain”。

33. 参见Olds，“Pleasure Centers in the Brain”，1956，以及Olds，“Pleasure Centers in the Brain”，1970。

34. Corbett和Wise，“Intracranial Self-Stimulation in Relation to the Ascending Dopaminergic Systems of the Midbrain”。

35. Schultz，“Multiple Dopamine Functions at Different Time Courses”，估计人脑中大约有400,000个多巴胺神经元，总共大约有800到1000亿个神经元。

36. Bolam和Pissadaki，“Living on the Edge with Too Many Mouths to Feed”。

37. Bolam和Pissadaki。

38. Glimcher，“Understanding Dopamine and Reinforcement Learning”。

39. Wise等人，“Neuroleptic-Induced ‘Anhedonia’ in Rats”。

40. Wise，“Neuroleptics and Operant Behavior”。关于“快感缺失假说”和早期发现大脑“快乐中心”以及后来发现多巴胺起核心作用的相当全面的历史，参见Wise，“Dopamine and Reward”。

41. 引用自Wise，“Dopamine and Reward”。

42. Romo和Schultz，“Dopamine Neurons of the Monkey Midbrain”。

43. Romo和Schultz。

44. Wolfram Schultz，个人访谈，2018年6月25日。

45. 参见，例如，Schultz、Apicella和Ljungberg，“Responses of Monkey Dopamine Neurons to Reward and Conditioned Stimuli During Successive Steps of Learning a Delayed Response Task”，以及Mirenowicz和Schultz，“Importance of Unpredictability for Reward Responses in Primate Dopamine Neurons”。

46. 参见Rescorla和Wagner, “A Theory of Pavlovian Conditioning” ; 学习可能只在结果令人惊讶时才会发生的想法来自较早的Kamin, “Predictability, Surprise, Attention, and Conditioning” 。
47. Wolfram Schultz, 个人访谈, 2018年6月25日。
48. Wolfram Schultz, 个人访谈, 2018年6月25日。参见Schultz、Apicella和Ljungberg, “Responses of Monkey Dopamine Neurons to Reward and Conditioned Stimuli During Successive Steps of Learning a Delayed Response Task” 。
49. 引用自Brinnin, *The Third Rose*。
50. Barto、Sutton和Anderson, “Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems” 。
51. “Rich有点像预测者, 而我更像是执行者” (Andrew Barto, 个人访谈, 2018年5月9日) 。
52. Sutton, “A Unified Theory of Expectation in Classical and Instrumental Conditioning” 。
53. Sutton, “Temporal-Difference Learning” (讲座), 2017年7月3日, 深度学习和强化学习夏季学校2017, 蒙特利尔大学, 2017年7月3日, http://videolectures.net/deeplearning2017_sutton_td_learning/。
54. Sutton, “Temporal-Difference Learning” 。
55. Sutton, “Learning to Predict by the Methods of Temporal Differences” 。另请参见Sutton的博士论文: “Temporal Credit Assignment in Reinforcement Learning” 。
56. 参见Watkins, “Learning from Delayed Rewards” 和Watkins和Dayan, “Q-Learning” 。
57. Tesauro, “Practical Issues in Temporal Difference Learning.”
58. Tesauro, “TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play.” 另见 Tesauro, “Temporal Difference Learning and TD-Gammon.”
59. “Interview with P. Read Montague,” Cold Spring Harbor Symposium Interview Series, Brains and Behavior, https://www.youtube.com/watch?v=mx96DYQIS_s.
60. Peter Dayan, 个人访谈, 2018年3月12日。
61. Schultz, Dayan, and Montague, “A Neural Substrate of Prediction and Reward.” 与TD-learning的突破性联系在前一年已经出现在Montague, Dayan, and Sejnowski, “A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning.”
62. P. Read Montague, “Cold Spring Harbor Laboratory Keynote,” <https://www.youtube.com/watch?v=RJvpu8nYzFg>.
63. “Interview with P. Read Montague,” Cold Spring Harbor Symposium Interview Series, Brains and Behavior https://www.youtube.com/watch?v=mx96DYQIS_s.
64. Peter Dayan, 个人访谈, 2018年3月12日。
65. Wolfram Schultz, 个人访谈, 2018年6月25日。

66. 例如，见 Niv, “Reinforcement Learning in the Brain.”
67. Niv.
68. 关于多巴胺TD-error理论潜在局限性的讨论，例如见 Dayan and Niv, “Reinforcement Learning,” 以及 O’ Doherty, “Beyond Simple Reinforcement Learning.”
69. Niv, “Reinforcement Learning in the Brain.”
70. Yael Niv, 个人访谈, 2018年2月21日。
71. Lenson, *On Drugs*.
72. 例如，见 Berridge, “Food Reward: Brain Substrates of Wanting and Liking,” 以及 Berridge, Robinson, and Aldridge, “Dissecting Components of Reward.”
73. Rutledge et al., “A Computational and Neural Model of Momentary Subjective Well-Being.”
74. Rutledge et al.
75. 见 Brickman, “Hedonic Relativism and Planning the Good Society,” 以及 Frederick and Loewenstein, “Hedonic Adaptation.”
76. Brickman, Coates, and Janoff-Bulman, “Lottery Winners and Accident Victims.”
77. “Equation to Predict Happiness,” <https://www.ucl.ac.uk/news/2014/aug/equation-predict-happiness>.
78. Rutledge et al., “A Computational and Neural Model of Momentary Subjective Well-Being.”
79. Wency Leung, “Researchers Create Formula That Predicts Happiness,” <https://www.theglobeandmail.com/life/health-and-fitness/health/researchers-create-formula-that-predicts-happiness/article19919756/>.
80. 见 Tomaszik, “Do Artificial Reinforcement-Learning Agents Matter Morally?” 关于这个话题的更多内容，另见 Schwitzgebel and Garza, “A Defense of the Rights of Artificial Intelligences.”
81. Brian Tomaszik, “Ethical Issues in Artificial Reinforcement Learning,” <https://reducing-suffering.org/ethical-issues-artificial-reinforcement-learning/>.
82. Daswani and Leike, “A Definition of Happiness for Reinforcement Learning Agents.” 另见 People for the Ethical Treatment of Reinforcement Learners: <http://petrl.org>.
83. Andrew Barto, 个人访谈, 2018年5月9日。
84. 大脑中的多巴胺和TD学习都还有更多内容；例如，多巴胺与运动活动和帕金森氏病等运动疾病有关。而且多巴胺似乎在正向预测误差中比在负向预测误差中更密切相关。例如，当涉及到“厌恶”刺激时，似乎有完全不同的神经回路：那些威胁性的、恶心的或有毒的东西。
85. 见 Athalye et al., “Evidence for a Neural Law of Effect.”

86. Andrew Barto, 个人访谈, 2018年5月9日。
87. 关于智能通用定义的更多想法, 例如见 Legg and Hutter, “Universal Intelligence” 以及 “A Collection of Definitions of Intelligence,” 和 Legg and Veness, “An Approximation of the Universal Intelligence Measure.”
88. McCarthy, “What Is Artificial Intelligence?”
89. 正如 Schultz, Dayan, and Montague, “A Neural Substrate of Prediction and Reward,” 所说: “没有区分哪些刺激对广播标量误差信号的波动负责的能力, 智能体可能会不当地学习, 例如, 它可能在实际渴了的时候学会去寻找食物。”

第5章 塑形(Shaping)

1. Bentham, *An Introduction to the Principles of Morals and Legislation*.
2. Matarić, “Reward Functions for Accelerated Learning.”
3. Skinner, “Pigeons in a Pelican.” 另见 Skinner, “Reinforcement Today.”
4. Skinner, “Pigeons in a Pelican.”
5. Ferster and Skinner, *Schedules of Reinforcement*。关于 Charles Ferster 在这一时期与 Skinner 合作的回忆，见 Ferster, “Schedules of Reinforcement with Skinner.”
6. Bailey and Gillaspy, “Operant Psychology Goes to the Fair.”
7. Bailey and Gillaspy.
8. Brelands 能够训练超过六千只动物，涉及多个物种，“我们已经敢于挑战如驯鹿、凤头鹦鹉、浣熊、海豚和鲸鱼等不太可能的训练对象。”然而，他们开始遇到在调节动物特定行为能力方面的某些重复性限制，得出结论认为行为主义作为一种理论未能充分考虑动物的本能、进化和物种特异性行为及倾向。见 Breland and Breland, “The Misbehavior of Organisms.”
9. Skinner, “Reinforcement Today” (原文强调)。
10. Skinner, “Pigeons in a Pelican.”
11. Skinner, “Pigeons in a Pelican.”
12. Skinner, “How to Teach Animals,” 1951年，这似乎是动词“shaping”在强化语境中最早的出现。
13. Skinner在其著作中的许多地方都讨论了这一事件。见其“Reinforcement Today,” “Some Relations Between Behavior Modification and Basic Research,” *The Shaping of a Behaviorist*, 和 *A Matter of Consequences*。另见 Peterson, “A Day of Great Illumination.”
14. Skinner, “How to Teach Animals.”
15. 正如 Skinner 所说：“一个熟悉的问题是，孩子似乎在惹恼父母方面表现出几乎病态的快乐。在许多情况下，这是调节作用的结果，与我们讨论过的动物训练非常相似。”见 Skinner, “How to Teach Animals.”
16. Skinner, “How to Teach Animals.”
17. 这句话的最早出现是 Spielvogel, “Advertising”，在 Edison 去世多年之后。关于其历史和变体的更多信息，见 O’ Toole, “There’s a Way to Do It Better—Find It.”
18. Bain, *The Senses and the Intellect*.

19. Michael Littman, 个人访谈, 2018年2月28日。
20. 在机器学习语境中明确提及”shaping”这一名称, 最早出现在Singh, “Transfer of Learning by Composing Solutions of Elemental Sequential Tasks”; 在1990年代, 它在机器人社区中成为越来越普遍的话题, 许多研究人员明确寻求动物训练和工具性条件反射文献的启发。见, 例如, Colombetti and Dorigo, “Robot Shaping”; Saksida, Raymond, and Touretzky, “Shaping Robot Behavior Using Principles from Instrumental Conditioning”; 以及Savage, “Shaping.”
21. Skinner, “Reinforcement Today.”
22. Shigeru Miyamoto, “Iwata Asks: New Super Mario Bros. Wii,” Satoru Iwata访谈, 2009年11月25日, <https://www.nintendo.co.uk/Iwata-Asks/Iwata-Asks-New-Super-Mario-Bros-Wii/Volume-1/4-Letting-Everyone-Know-It-Was-A-Good-Mushroom/4-Letting-Everyone-Know-It-Was-A-Good-Mushroom-210863.html>.
23. 关于机器学习方法应用于学习”课程”理念的更多信息, 见, 例如, Bengio et al., “Curriculum Learning.”
24. Selfridge, Sutton, and Barto, “Training and Tracking in Robotics.”
25. Elman, “Learning and Development in Neural Networks.” 然而, 另见, 例如, Rohde and Plaut, “Language Acquisition in the Absence of Explicit Negative Evidence,” 该研究在这方面报告了与Elman不同的发现。
26. 这个特殊的实验因猪的表现随时间恶化而值得注意, 这挑战了行为主义的经典模型。见Breland and Breland, “The Misbehavior of Organisms.”
27. Florensa et al., “Reverse Curriculum Generation for Reinforcement Learning.” 2018年, OpenAI的一个研究团队做了类似的事情来训练强化学习智能体玩特别困难的视频游戏。他们会记录一个能干的人类玩家玩游戏的过程, 然后通过这个记录的演示倒序工作来构建课程。首先他们会让智能体从成功的边缘开始训练, 然后逐渐向后移动, 最终到达游戏开始。见Salimans and Chen, “Learning Montezuma’s Revenge from a Single Demonstration.” 另见Hosu and Rebedea, “Playing Atari Games with Deep Reinforcement Learning and Human Checkpoint Replay”; Nair et al., “Overcoming Exploration in Reinforcement Learning with Demonstrations”; 以及Peng et al., “DeepMimic.” 这里也更广泛地与模仿学习相关, 我们在第7章中讨论。
28. 见, 例如, Ashley, *Chess for Success*.
29. 这很难确认, 但似乎极有可能。参见Edward Winter的”Chess Book Sales”, <http://www.chesshistory.com/winter/extrasales.html>, 了解更多国际象棋书籍销售信息。
30. 参见Graves等人的”Automated Curriculum Learning for Neural Networks”。这与我们在第6章讨论的奖励学习进展工作有联系。关于课程设计的早期机器学习工作, 参见Bengio等人的”Curriculum Learning”。
31. David Silver, “AlphaGo Zero: Starting from Scratch”, 2017年10月18日, <https://www.youtube.com/watch?v=tXlM99xPQC8>.
32. Kerr, “On the Folly of Rewarding A, While Hoping for B”。
33. Kerr。注意在1975年原版印刷中是”immortality”(原文如此)!
34. 这篇文章署名”The Editors”, 但他们的名字是Kathy Dechant和Jack Veiga; 参见Dechant和Veiga的”More on the Folly”。

35. 关于“游戏化”激励的几个警示故事，参见Callan、Bauer和Landers的“*How to Avoid the Dark Side of Gamification*”。

36. Kerr, “*On the Folly of Rewarding A, While Hoping for B*”。

37. Wright等人, “*40 Years (and Counting)*”。

38. “*Operant Conditioning*”，https://www.youtube.com/watch?v=I_ctJqjlHA。

39. 参见Joffe-Walt的“*Allowance Economics*”和Gans的*Parentonomics*。

40. Tom Griffiths, 个人访谈, 2018年6月13日。

41. Andre和Teller, “*Evolving Team Darwin United*”。

42. 在Ng、Harada和Russell的“*Policy Invariance Under Reward Transformations*”中引用，作为作者的个人交流。

43. Randløv和Alstrøm, “*Learning to Drive a Bicycle Using Reinforcement Learning and Shaping*”。

44. Russell告诉我，这源于1990年代对meta-reasoning的深入思考：思考的正确方式。当你玩游戏——比如国际象棋——你获胜是因为你选择的走法，但是让你能够选择这些走法的是你的思考。确实，有时我们反思一局棋会想，“啊，我出错是因为我困住了我的马。我需要让马远离棋盘边缘。”但有时我们会想，“啊，我出错是因为我不相信自己的直觉。我想太多了；我需要更有机、更直觉地下棋。”弄清楚一个有志向的棋手——或任何类型的agent——应该如何学习其思考过程似乎是一个更重要但也更困难的任务，而不仅仅是学习如何选择好的走法。也许shaping可以帮助。

“所以一个自然的答案是...如果你做了一个计算，改变了你对什么是好走法的想法，那么显然那似乎是一个有价值的计算，”Russell说。“所以你可以根据你改变想法的程度来奖励那个计算。”

他补充道，“现在，这是棘手的部分：所以你可能改变想法是因为发现原来的第二好走法实际上比原来的最佳走法更好。

“所以你因为做那件事得到奖励分数。你以前有一个你认为值50分的走法是你最好的，48分是你第二好的。现在那个48变成了你的52分，看到了。从50到52是正数。那么如果相反，你考虑50分并意识到它只值6分呢？所以现在你最好的走法是48分，也就是你的第二好走法。那应该是正奖励还是负奖励？再一次，你会认为它应该是正奖励，因为你做了一些思考，那种思考是有价值的，因为它帮助你意识到你认为你要做的事情没有那么好。你让自己免于一场灾难。但是如果你也因为那个给自己正奖励，对吧？你会一路给自己只有正奖励，对吧？你最终学会做的不是赢得游戏而是一直改变你的想法。

“所以有些地方不对。这让我产生了这样的想法，你必须安排这些内部伪奖励，使得沿着一条路径，它们加起来与真实的最终奖励相同。平衡账目”(Stuart Russell, 个人访谈, 2018年5月13日)。

45. Andrew Ng, “*The Future of Robotics and Artificial Intelligence*”(讲座)，2011年5月21日，https://www.youtube.com/watch?v=AY4ajbu_G3k。

46. 参见Ng等人的“*Autonomous Helicopter Flight via Reinforcement Learning*”，以及Schrage等人的“*Instrumentation of the Yamaha R-50/RMAX Helicopter Testbeds for Airloads Identification and Follow-on Research*”。关于后续工作，参见Ng

等人的”Autonomous Inverted Helicopter Flight via Reinforcement Learning”和Abbeel等人的”An Application of Reinforcement Learning to Aerobatic Helicopter Flight”。

47. Ng, “Shaping and Policy Search in Reinforcement Learning”。另见Wiewiora的”Potential-Based Shaping and Q-Value Initialization Are Equivalent”，该文章论证可以在设置agent的初始状态时使用shaping，同时保持实际奖励本身不变，并获得相同的结果。

48. Ng, “Shaping and Policy Search in Reinforcement Learning。”这也在Ng, Harada, and Russell, “Policy Invariance Under Reward Transformations”中逐字出现。

49. “保守场意味着如果你走任何路径回到同一个状态，总的积分 $v.ds$ 等于零”（Stuart Russell，个人访谈，2018年5月13日）。

50. Russell and Norvig, *Artificial Intelligence*。

51. Ng, Harada, and Russell, “Policy Invariance Under Reward Transformations.”

52. Spignesi, *The Woody Allen Companion*。

53. 关于进化心理学视角，参见Al-Shawaf et al., “Human Emotions: An Evolutionary Psychological Perspective,” 和Miller, “Reconciling Evolutionary Psychology and Ecological Psychology.”

54. Michael Littman，个人访谈，2018年2月28日。该论文是Sutton, “Learning to Predict by the Methods of Temporal Differences.”

55. Ackley and Littman, “Interactions Between Learning and Evolution.”

56. 训练本身是（或可能成为）某种“内部”奖励函数优化器的系统，是当代AI安全研究者关注和积极研究的问题源。参见Hubinger et al., “Risks from Learned Optimization in Advanced Machine Learning Systems.”

57. Andrew Barto，个人访谈，2018年5月9日。

58. 参见Singh, Lewis, and Barto, “Where Do Rewards Come from?”以及Sorg, Singh, and Lewis, “Internal Rewards Mitigate Agent Boundedness.”

59. Sorg, Singh, and Lewis, “Internal Rewards Mitigate Agent Boundedness。”这个问题的答案是肯定的——但仅在一些非常强的假设下。特别是，只有当我们的智能体的时间和计算能力是无限的时。否则，如果我们不让它的目标成为我们自己的目标，我们会更好。这有点像悖论的味道。通过告诉智能体做其他事情，我们自己的目标会得到更好的服务。

60. Singh et al., “On Separating Agent Designer Goals from Agent Goals.”

61. 关于最优奖励问题的更多信息，参见Sorg, Lewis, and Singh, “Reward Design via Online Gradient Ascent,”以及Sorg的博士论文“The Optimal Reward Problem: Designing Effective Reward for Bounded Agents。”关于为强化学习智能体学习最优奖励的最新进展，参见Zheng, Oh, and Singh, “On Learning Intrinsic Rewards for Policy Gradient Methods.”

62. 参见“工作场所拖延症每年给英国企业造成760亿英镑损失”，全球银行与金融评论，https://www.globalbankingandfinance.com/workplace-procrastination-costs-british-businesses-76-billion-a-year/#_ftn1。关于

拖延症成本和原因的广泛研究，参见 Steel, “The Nature of Procrastination.”

63. Skinner, “A Case History in Scientific Method.”

64. Jane McGonigal, “游 戏 可 以 让 世 界 变 得 更 美 好 ” ,
[https://www.ted.com/talks/jane_mcgonigal_gaming_can_make_a_better_world/。](https://www.ted.com/talks/jane_mcgonigal_gaming_can_make_a_better_world/)

65. 参见 McGonigal, *SuperBetter*.

66. Jane McGonigal, “能 给 你 额 外 10 年 生 命 的 游 戏 ” ,
[https://www.ted.com/talks/jane_mcgonigal_the_game_that_can_give_you_10_extra_years_of_life/。](https://www.ted.com/talks/jane_mcgonigal_the_game_that_can_give_you_10_extra_years_of_life/)

67. 参见 Deterding et al., “From Game Design Elements to Gamefulness.”

68. 参见 Hamari, Koivisto, and Sarsa, “Does Gamification Work?”

69. Falk Lieder, 个人访谈, 2018年4月18日。

70. 参见 Lieder, “Gamify Your Goals” 的总体概述以及 Lieder et al., “Cognitive Prostheses for Goal Achievement” 的更多细节。

71. 这个想法在 Sorg, Lewis, and Singh, “Reward Design via Online Gradient Ascent” 中也得到了更近期的探索。

72. Falk Lieder, 个人访谈, 2018年4月18日。

73. 具体来说, 他们被给予选择: 拒绝任务并获得15美分, 或者接受任务并获得5美分加上在规定期限内写一组论文获得20美元的能力。

74. Lieder et al., “Cognitive Prostheses for Goal Achievement.”

75. Lieder et al.

76. 例如, 参见 Evans et al., “Evidence for a Mental Health Crisis in Graduate Education”, 该研究发现“研究生经历抑郁和焦虑的可能性是普通人群的六倍多”。

第6章 好奇心

1. Turing, “Intelligent Machinery.”
2. 从2004年开始有努力开发标准化的强化学习基准和竞赛；参见 Whiteson, Tanner, and White, “The Reinforcement Learning Competitions.”
3. Marc Bellemare, 个人访谈，2019年2月28日。
4. Bellemare et al., “The Arcade Learning Environment” , 最初源自Naddaf的” Game-Independent AI Agents for Playing Atari 2600 Console Games” , 再之前则来自Diuk, Cohen, and Littman的” An Object-Oriented Representation for Efficient Reinforcement Learning” , 该研究使用*Pitfall!*游戏作为强化学习的环境。
5. 参见Gendron-Bellemare, “Fast, Scalable Algorithms for Reinforcement Learning in High Dimensional Domains.”
6. Mnih et al., “Playing Atari with Deep Reinforcement Learning.”
7. Mnih et al., “Human-Level Control Through Deep Reinforcement Learning.”
8. Robert Jaeger, 接受John Hardie采访, http://www.digitpress.com/library/interviews/interview_robert_jaeger.html.
9. 关于这一点的更多内容, 参见例如 Salimans and Chen, “Learning Montezuma’s Revenge from a Single Demonstration” , 该研究还探讨了从成功的目标状态向后工作来逐步教导强化学习智能体如何玩游戏的有趣想法。
10. 参见Maier and Seligman, “Learned Helplessness.” 关于这一领域的最新正式研究, 参见例如Lieder, Goodman, and Huys, “Learned Helplessness and Generalization.”
11. 参见Henry Alford, “The Wisdom of Ashleigh Brilliant,” <http://www.ashleighbrilliant.com/BrilliantWisdom.html>, 摘自Alford, *How to Live* (New York: Twelve, 2009)。
12. 内在动机(intrinsic motivation)的概念由Barto, Singh, and Chentanez在” Intrinsic Motivation of Hierarchical Collections of Skills” 以及Singh, Chentanez, and Barto在” Intrinsic Motivation Reinforcement Learning” 中引入机器学习领域。关于这一文献的最新综述, 参见Baldassarre and Mirolli, *Intrinsic Motivation in Natural and Artificial Systems*。
13. Hobbes, *Leviathan*。
14. Simon, “The Cat That Curiosity Couldn’t Kill.”
15. Berlyne, “ ‘Interest’ as a Psychological Concept.”
16. 参见Furedy and Furedy, “ ‘My First Interest Is Interest.’ ”
17. Berlyne, *Conflict, Arousal, and Curiosity*.

18. 参见Harlow, Harlow, and Meyer, “Learning Motivated by a Manipulation Drive” 以及Harlow, “Learning and Satiation of Response in Intrinsically Motivated Complex Puzzle Performance by Monkeys.”

19. 这类情况在Barto, “Intrinsic Motivation and Reinforcement Learning” 以及Deci and Ryan, *Intrinsic Motivation and Self-Determination in Human Behavior*中有所描述。

20. Berlyne, *Conflict, Arousal, and Curiosity*.

21. 另参见例如Berlyne自己的” Uncertainty and Conflict: A Point of Contact Between Information-Theory and Behavior-Theory Concepts.”

22. 关于”兴趣”作为心理学主题的二十一世纪综述，参见例如Silvia, *Exploring the Psychology of Interest*以及Kashdan and Silvia, “Curiosity and Interest.”

23. Konečni, “Daniel E. Berlyne.”

24. Berlyne, *Conflict, Arousal, and Curiosity*.

25. *Klondike Annie*, 1936年。

26. Fantz, “Visual Experience in Infants.” 严格来说，Fantz的隶属机构是” 西储大学(Western Reserve University)“，直到几年后的1967年才正式与凯斯理工学院(Case Institute of Technology)合并，成为我们今天所知的凯斯西储大学。

27. 参见Saayman, Ames, and Moffett, “Response to Novelty as an Indicator of Visual Discrimination in the Human Infant.”

28. 关于世纪之交的综述，参见Roder, Bushnell, and Sasseville, “Infants’ Preferences for Familiarity and Novelty During the Course of Visual Processing.”

29. 例如，马文·明斯基(Marvin Minsky)在1961年写道：“如果我们能够...为那些具有新颖性的预测的强化增加奖励，我们可能期望看到由某种好奇心驱动的行为。...在对确认的新颖期望的机制进行强化时...我们可能找到模拟智力动机的关键。” 参见Minsky, “Steps Toward Artificial Intelligence.”

30. 参见 Sutton, “Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming” 以及” Reinforcement Learning Architectures for Animats.” MIT的Leslie Pack Kaelbling设计了一种类似的方法，基于测量智能体对某些动作奖励的”置信区间(confidence intervals)“的想法；参见Kaelbling, *Learning in Embedded Systems*。置信区间越宽，智能体对该动作就越不确定；她的想法同样是奖励智能体去做那些最不确定的事情。另参见Strehl and Littman, ”An Analysis of Model-Based Interval Estimation for Markov Decision Processes”，该研究沿用了这一思路。

31. Berlyne, *Conflict, Arousal, and Curiosity*.

[32] 如果9个空格中的每一个都可以是X、O或空的，这就设定了 3^9 或19,683的上限。当然，实际数字会比这个小，因为并非所有这些位置都是合法的（例如，一个全是9个X的棋盘，在游戏中永远不可能出现）。

[33] Bellemare等人的”统一基于计数的探索和内在动机(Unifying Count-Based Exploration and Intrinsic Motivation)“，部分受到Strehl和Littman的”马尔可夫决策过程基于模型的区间估计分析(An Analysis of Model-Based Interval Estimation for Markov Decision Processes)“启发。另请参阅Ostrovski等人的后续论文”使用神经密度模型的基于计数的探索

(Count-Based Exploration with Neural Density Models) “。关于使用哈希函数的相关方法，参见Tang等人的” #Exploration”。关于使用exemplar模型的另一种相关方法，参见Fu、Co-Reyes和Levine的” EX² ”。

[34] Marc G. Bellemare，“密度模型在强化学习中的作用(The Role of Density Models in Reinforcement Learning)”（讲座），DeepHack.RL，2017年2月9日，<https://www.youtube.com/watch?v=qSfd27AgcEk>。

[35] 事实上，从概率到估计数的转换中存在相当大的巧妙数学细节。更多信息请参见Bellemare等人的”统一基于计数的探索和内在动机(Unifying Count-Based Exploration and Intrinsic Motivation) ”。

[36] Berlyne，《冲突、唤醒和好奇心》(Conflict, Arousal, and Curiosity)。

[37] Gopnik，“解释如同高潮和对因果知识的驱动(Explanation as Orgasm and the Drive for Causal Knowledge) ”。

[38] 关于新颖性和惊喜之间差异的计算观点，参见Barto、Mirolli和Baldassarre的” 新颖性还是惊喜？(Novelty or Surprise?) ”。

[39] Schulz和Bonawitz，“严肃的乐趣(Serious Fun) ”。

[40] “好奇心与学习：批判性思维的技能(Curiosity and Learning: The Skill of Critical Thinking)”，家庭与工作研究所，<https://www.youtube.com/watch?v=lDgm5yVY5K4>。

[41] Ellen Galinsky，“节日时给予好奇心的礼物——来自Laura Schulz的教训(Give the Gift of Curiosity for the Holidays—Lessons from Laura Schulz)”，https://www.huffpost.com/entry/give-the-gift-of-curiosity_n_1157991。关于近期科学文献的更全面回顾，参见Schulz的” 婴儿探索意外事件(Infants Explore the Unexpected) ”。

[42] Bonawitz等人，“儿童在探索、解释和学习中平衡理论与证据(Children Balance Theories and Evidence in Exploration, Explanation, and Learning) ”。

[43] Stahl和Feigenson，“观察意外事件增强婴儿的学习和探索(Observing the Unexpected Enhances Infants’ Learning and Exploration) ”。

[44] “约翰霍普金斯大学研究人员：婴儿从惊喜中学习(Johns Hopkins University Researchers: Babies Learn from Surprises)”，2015年4月2日，<https://www.youtube.com/watch?v=oJjt5GRln-0>。

[45] Berlyne，《冲突、唤醒和好奇心》(Conflict, Arousal, and Curiosity)。Berlyne特别受到Shaw等人的” 复杂信息处理的命令结构(A Command Structure for Complex Information Processing) ”启发。

[46] Schmidhuber，“创造力、乐趣和内在动机的形式理论(1990-2010)(Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990 – 2010)) ”。

[47] Jürgen Schmidhuber，“通用AI和乐趣的形式理论(Universal AI and a Formal Theory of Fun)”（讲座），牛津大学冬季智能会议，2011年，<https://www.youtube.com/watch?v=fnbZzcruGu0>。

[48] Schmidhuber，“创造力、乐趣和内在动机的形式理论(1990-2010)(Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990 – 2010)) ”。

[49] 这两个组成部分之间的张力以阴阳形式完美体现了纽约大学的James Carse所说的《有限与无限游戏》(Finite and Infinite Games)。有限游戏是为了达到终端平衡状态而进行的。无限游戏是为了永远延续游戏体验而进行的。有限游

戏玩家对抗惊喜；无限游戏玩家为了惊喜而游戏。用Carse的话说：“惊喜导致有限游戏结束；它是无限游戏继续的原因。”

这种对立的基本驱动力——既追求又对抗惊喜——的张力也得到了知名励志演说家Tony Robbins等人的呼应，他阐述道：“我相信有六种人类需求……让我告诉你它们是什么。第一个：确定性……虽然我们以不同方式追求确定性，但如果我们获得完全的确定性，我们会得到什么？如果你确定会发生什么，你会有什么感觉？你知道什么会发生，何时和如何发生：你会感觉如何？无聊至极。所以上帝，以她无限的智慧，给了我们第二个人类需求，那就是不确定性。我们需要变化。我们需要惊喜。” Tony Robbins，“我们为什么做我们所做的事(Why We Do What We Do)”（讲座），2006年2月，加利福尼亞州蒙特雷，https://www.ted.com/talks/tony_robbins_asks_why_we_do_what_we_do。

人类显然在内心同时拥有这两种驱动力。如果所有优秀的通用强化学习者——无论是生物还是非生物——也都如此，这可能并非巧合。

[50] “内在好奇心模块(intrinsic curiosity module)”实际上比这更微妙和复杂，因为它被设计为只预测屏幕上用户可控制的方面，为此还使用了另一个”逆动力学(inverse dynamics)“模型。完整细节请参见Pathak等人的”通过自监督预测的好奇心驱动探索(Curiosity-Driven Exploration by Self-Supervised Prediction)“。其他一些相关方法，通过奖励”信息增益”来激励探索，例如参见Schmidhuber的”好奇心模型构建控制系统(Curious Model-Building Control Systems)“；Stadie、Levine和Abbeel的”使用深度预测模型激励强化学习中的探索(Incentivizing Exploration in Reinforcement Learning with Deep Predictive Models)“；以及Houthooft等人的”VIME”。

[51] Burda等人，“好奇心驱动学习的大规模研究(Large-Scale Study of Curiosity-Driven Learning)”。

[52] 参见Burda等人，“通过随机网络蒸馏进行探索(Exploration by Random Network Distillation)”。

53. 需要注意的是，来自密歇根大学和Google Brain研究人员的Choi等人的同期论文”强化学习中的偶然性感知探索”也报告了在《蒙特祖马的复仇》中使用基于新颖性的探索方法取得的类似突破。

54. 在OpenAI宣布后几周，来自Uber AI Labs的团队宣布了他们称为Go-Explore的算法家族，该算法通过存储”新颖”状态列表（通过屏幕的粗糙、低分辨率图像测量）来优先重访，能够65%的时间击败《蒙特祖马的复仇》的第一关。详见<https://eng.uber.com/go-explore/>的新闻稿，以及Ecoffet等人的论文”Go-Explore”。使用一些关于游戏的手工编码人类知识，使得智能体能够连续数百次击败游戏关卡，在此过程中累积数百万分数。这些结果的一些重要性一直存在争议——例如，参见Alex Irpan，“对Go-Explore的快速看法”，Sorta Insightful，<https://www.alexirpan.com/2018/11/27/go-explore.html>，进行一些讨论。团队的新闻稿本身后来也更新了以解决这些和其他要点。

55. Ostrovski等人，“基于计数的神经密度模型探索”。

56. 关于这一点的更多讨论，参见例如Ecoffet等人的”Go-Explore”。

57. Burda等人，“好奇心驱动学习的大规模研究”。

58. 例外情况是在有复杂死亡动画的游戏中，智能体会故意死亡只是为了观看动画。（Yuri Burda，个人通信，2019年1月9日。）

59. Yuri Burda，个人通信，2019年1月9日。

60. Singh, Lewis, 和 Barto, “奖励从何而来?”

61. Singh, Lewis, 和 Barto。更多讨论见Oudeyer和Kaplan的“什么是内在动机?”

62. 由于这样的原因，研究人员已经实验了所谓的“粘性动作”——智能体偶尔会被随机强制重复其上一个按键按压一帧——作为epsilon-贪婪动作的替代变化源，在epsilon-贪婪动作中，智能体在随机时间按随机按钮。这更准确地建模了人类游戏中的固有随机性，我们的反应并不是毫秒级完美的，它使得需要连续多帧按住按钮的长跳等动作更容易被智能体实现。见Machado等人的“重新审视街机学习环境”。

63. 关于这个主题的一些早期工作，见Malone的“是什么让电脑游戏有趣？”和“走向内在激励指导理论”，以及Malone和Lepper的“让学习变得有趣”。

64. Orseau, Lattimore, 和 Hutter, “随机环境中的通用知识寻求智能体”。

65. 见Orseau的“通用知识寻求智能体”。这个智能体对随机性成瘾问题的解决方案，正如后来在Orseau, Lattimore, 和 Hutter的“随机环境中的通用知识寻求智能体”中制定的那样，是让智能体在基本层面上理解世界包含随机性，因此“抵抗非信息性噪声”。

66. Skinner, “今日强化”。

67. 见Kakade和Dayan的“多巴胺”，该文提供了基于新颖性的解释，明确借鉴强化学习文献来解释为什么新颖性驱动可能对有机体有用。另见Barto, Mirolli, 和 Baldassarre的“新颖性还是惊奇？”，对这些结果进行基于惊奇的解释。概述见例如Niv的“大脑中的强化学习”，其中指出，“早就已知新奇刺激会导致多巴胺神经元的相位性爆发”。关于人类决策中新颖性的实验工作，见Wittmann等人的“纹状体活动是人类基于新颖性选择的基础”。关于更近期将奖励预测误差和惊奇更广泛地统一在多巴胺功能中的工作，见例如Gardner, Schoenbaum, 和 Gershman的“重新思考多巴胺作为广义预测误差”。

68. Deepak Pathak, 个人采访，2018年3月28日。

69. Marc Bellemare, 个人采访，2019年2月28日。

70. Laurent Orseau, 个人采访，2018年6月22日。

71. Laurent Orseau, 个人采访，2018年6月22日。

72. Ring和Orseau, “妄想、生存和智能智能体”。

73. 柏拉图，《普罗泰戈拉和美诺》。在柏拉图的文本中，苏格拉底以疑问句的形式向普罗泰戈拉提出这个问题，尽管他明确表示这确实是他的观点。

第七章 模仿

[1] Egan, 《公理》。

[2] Elon Musk, 接受Sarah Lacy采访, “与Elon Musk的炉边谈话”, 加利福尼亚州圣莫尼卡, 2012年7月12日, <https://pando.com/2012/07/12/pandomonthly-presents-a-fireside-chat-with-elon-musk/>。不仅这辆车没有保险, 而且Peter Thiel没有系安全带。“我们两个都没受伤真是个奇迹,” Thiel说。见Dowd的“Peter Thiel, 特朗普的科技伙伴, 为自己辩护”。

[3] 这在Visalberghi和Fragaszy的“Do Monkeys Ape?”中有更详细的讨论。

[4] Romanes, 《动物智能》。

[5] Visalberghi 和 Fragaszy, “Do Monkeys Ape?” 另见 Visalberghi 和 Fragaszy, “‘Do Monkeys Ape?’ Ten Years After。” 并且注意到Ferrari等人, “Neonatal Imitation in Rhesus Macaques”, 报告了一些猕猴模仿的证据, 这代表了“据我们所知, 这是在类人猿谱系之外的灵长类物种中进行的新生儿模仿的首次详细分析。”

[6] Tomasello, “Do Apes Ape?” 另见, 例如, Whiten等人, “Emulation, Imitation, Over-Imitation and the Scope of Culture for Child and Chimpanzee”, 该文章试图重新评估这个问题。

[7] 尽管Kellogg夫妇对于终止实验的原因有些谨慎, 但据推测Donald令人担忧的人类词汇量缺乏是一个促发原因。例如见Benjamin和Bruce, “From Bottle-Fed Chimp to Bottlenose Dolphin”。

[8] Meltzoff 和 Moore, “Imitation of Facial and Manual Gestures by Human Neonates” 和 Meltzoff 和 Moore, “Newborn Infants Imitate Adult Facial Gestures”。请注意, 这些结果最近变得有些争议。例如见Oostenbroek等人, “Comprehensive Longitudinal Study Challenges the Existence of Neonatal Imitation in Humans”。但也见反驳, 例如Meltzoff等人, “Re-examination of Oostenbroek et al. (2016)”。

[9] Alison Gopnik, 个人访谈, 2018年9月19日。

[10] Haggstrom等人, “The 100 Most Eminent Psychologists of the 20th Century”。

[11] Piaget, 《儿童现实建构》。最初于1937年以《La construction du réel chez l’ enfant》出版。

[12] Meltzoff, “‘Like Me.’”

[13] Meltzoff 和 Moore, “Imitation of Facial and Manual Gestures by Human Neonates”。

[14] 在2012年的一项研究中, 两岁儿童观察一个成人将汽车撞向两个不同的盒子, 其中一个使汽车发光; 当孩子拿到汽车时, 他们只将汽车撞向那个盒子 (Meltzoff, Waismaner, and Gopnik, “Learning About Causes From People”)。“幼儿不会模仿任何东西,” Gopnik说。“他们模仿会导致有趣结果的行为” (Gopnik, 《园丁与木匠》)。

[15] Meltzoff, Waismaner, and Gopnik, “Learning About Causes from People”, 以及 Meltzoff, “Understanding the Intentions of Others”。见Gopnik, 《园丁与木匠》, 对这一领域的良好总结。

[16] Meltzoff, “Foundations for Developing a Concept of Self”。

[17] Andrew Meltzoff, 个人访谈, 2019年6月10日。 “婴儿生来就是为了学习,” Meltzoff写道, “他们首先通过模仿我们来学习。这就是为什么模仿是早期发展中如此重要和深远的方面: 它不仅仅是一种行为, 而是学习我们是谁的手段” (Meltzoff, “Born to Learn”)。

[18] 用来描述这种现象的术语“过度模仿”最初来自Lyons, Young, and Keil, “The Hidden Structure of Overimitation”。

[19] Horner和Whiten, “Causal Knowledge and Imitation/Emulation Switching in Chimpanzees (*Pan troglodytes*) and Children (*Homo sapiens*)”。

[20] McGuigan 和 Graham , “Cultural Transmission of Irrelevant Tool Actions in Diffusion Chains of 3- and 5-Year-Old Children”。

[21] Lyons, Young, and Keil, “The Hidden Structure of Overimitation”。

[22] Whiten等人, “Emulation, Imitation, Over-Imitation and the Scope of Culture for Child and Chimpanzee”。

[23] Gergely, Bekkering, and Király, “Rational Imitation in Preverbal Infants”。请注意, 一些研究者对这种方法论提出了异议, 例如指出, 婴儿——需要在桌子上保持平衡才能用头触摸灯——可能只是在模仿成人, 成人在弯腰用头触摸灯之前将手放在桌子上。见Paulus等人, “Imitation in Infancy”。

[24] Buchsbaum等人, “Children’s Imitation of Causal Action Sequences Is Influenced by Statistical and Pedagogical Evidence”。

[25] Hayden Carpenter , “What ‘The Dawn Wall’ Left Out” , 《Outside》, 2018年9月18日,
<https://www.outsideonline.com/2344706/dawn-wall-documentary-tommy-caldwell-review>。

[26] Caldwell, 《推力》。

[27] Lowell和Mortimer, “The Dawn Wall”。

[28] “‘I Got My Ass Kicked’: Adam Ondra’s Dawn Wall Story”, EpicTV Climbing Daily, 第1334集,
https://www.youtube.com/watch?v=O_B9vzIHlOo。

29. Aytar等人的论文“通过观看YouTube学习困难探索游戏”, 该研究扩展了Hester等人的“从演示中进行深度Q学习”的相关工作。需要一些非常巧妙的无监督学习来本质上将所有不同的视频——不同分辨率、颜色和帧率——“标准化”为单一有用的表示。但结果是一组智能体可以学习模仿的演示。

30. 这是一个非常活跃的研究领域。例如, 参见Subramanian、Isbell和Thomaz的“交互式强化学习的演示探索”; Večerík等人的“在稀疏奖励机器人问题上利用演示进行深度强化学习”; 以及Hester等人的“从演示中进行深度Q学习”。

31. 实际上, 许多在视频游戏中训练的智能体都具备跳回到之前游戏中几乎任何地方的能力, 类似于人类玩家制作数百个(或更多)不同的“保存状态”。在这里, 死亡只是将你送回到上一个检查点, 也许只是几秒钟之前, 而不是回到游戏本身的开始。这允许智能体在棘手或危险的部分进行实验, 而不必在失败时从游戏的最开始重新开始——但这也在训练中引入了一定的人为因素。更有能力的智能体理论上应该能够在没有这种人为因素的情况下复制或超越人类在这些游戏上的“学习曲线”。

32. Morgan, 《比较心理学导论》。

33. 参见Bostrom, 《超级智能》。
34. “机器人历史：叙述和网络口述历史：Chuck Thorpe”，这是2010年11月22日由Peter Asaro和Selma Šabanović在印第安纳大学布卢明顿分校为印第安纳大学和IEEE进行的口述历史，<https://ieeetv.ieee.org/video/robotics-history-narratives-and-networks-oral-histories-chuck-thorpe>。
35. 有关ALV（自主陆地车辆）项目的更多信息，参见Leighty的“DARPA ALV（自主陆地车辆）摘要”。有关DARPA战略计算倡议的更多信息，参见“战略计算”。有关DARPA 1980年代中期项目的更多信息，参见Stefik的“DARPA的战略计算”。另见Roland和Shiman的《战略计算》。
36. Moravec，“真实世界中视觉机器人漫游者的障碍规避和导航”。
37. 另见Rodney Brooks在《血肉与机器》中的反思。
38. 如《科学美国人前沿》第7季第5集“机器人活着！”中的片段，1997年4月9日在PBS播出。参见<https://www.youtube.com/watch?v=r4JrcVEkink>。
39. Thorpe通过看汽车是否会在Leland骑着带训练轮的自行车冲到车前时刹车来测试Navlab上的防撞系统。Leland长大后在卡内基梅隆大学获得机器人学学位，然后去Thorpe的学生Dean Pomerleau创立的AssistWare公司从事自动驾驶汽车技术工作。Thorpe开玩笑地想象Leland的工作面试：“‘从那时起我就一直是提高自动驾驶车辆可靠性和安全性的真正拥护者！’” Leland后来完全离开了计算机行业，成为圣母玛利亚献身会的神学院学生。
40. Pomerleau，“ALVINN”，以及Pomerleau，“自主机器人驾驶人工神经网络的知识基础训练”。
41. 来自1997年KDKA新闻片段：<https://www.youtube.com/watch?v=IaoIqVMd6tc>。
42. 参见<https://twitter.com/deanpomerleau/status/801837566358093824>。（“目前的计算机似乎足够快，并且有足够的内存来完成[控制汽车的]工作”，AI先驱John McCarthy在1969年——有些天真地——争论道。“然而，所需性能的商用计算机太大了。”参见McCarthy，“计算机控制的汽车”。）
43. 另见Pomerleau的“自主机器人驾驶人工神经网络的知识基础训练”：“自动驾驶有潜力成为反向传播等监督学习算法的理想领域，因为有一个现成的教学信号或‘正确响应’，即人类司机当前的转向方向。”
44. 该课程是Sergey Levine的CS294-112，深度强化学习；这次题为“模仿学习”的讲座于2017年12月3日举行。
45. Bain, 《感官和智力》。
46. Kimball和Zaveri，“Tim Cook谈Facebook数据泄露丑闻”。
47. Ross和Bagnell在“模仿学习的高效约简”中讨论了他们的架构选择：一个三层神经网络，输入 24×18 像素彩色图像，有32个隐藏单元和15个输出单元。Pomerleau在“自主机器人驾驶人工神经网络的知识基础训练”中讨论了ALVINN的架构，由一个三层神经网络构成，输入 30×32 像素黑白图像，有4个隐藏单元和30个输出单元。
48. Pomerleau。
49. Pomerleau，“ALVINN”。
50. Stéphane Ross，个人访谈，2019年4月29日。

51. 参见Ross和Bagnell，“模仿学习的高效约简”。
52. 参见Ross、Gordon和Bagnell，“模仿学习和结构化预测到无悔在线学习的约简”。关于早期方法，参见Ross和Bagnell，“模仿学习的高效约简”。
53. Giusti等人，“移动机器人森林小径视觉感知的机器学习方法”。关于该研究的视频解释，参见“使用深度神经网络在森林中进行四旋翼飞行器导航”，<https://www.youtube.com/watch?v=umRdt3zGgpU>。
54. Bojarski等人，“自动驾驶汽车的端到端学习”。Nvidia团队进一步用“Photoshop”操作增强了侧向摄像头图像，以获得更多角度的多样性。这些图像受到与ALVINN图像类似的限制，但在实践中足够好用。有关更多非正式讨论，参见Bojarski等人，“自动驾驶汽车的端到端深度学习”，<https://devblogs.nvidia.com/deep-learning-self-driving-cars/>。关于汽车在Monmouth县道路上行驶的视频，参见“Dave-2：神经网络驱动汽车”，<https://www.youtube.com/watch?v=NJU9ULQUwng>。
55. 参见LeCun等人，“反向传播算法在手写邮编识别中的应用”。
56. Murdoch，《钟》。
57. Robert Hass，“违背与祈祷”，收录于《时间与材料》。
58. Kasparov，《生活如何模仿国际象棋》。
59. Holly Smith，个人访谈，2019年5月13日。
60. 参见，以Holly S. Goldman之名发表的“相对正确性与道德缺陷”。另见Sobel，“功利主义与过去和未来的错误”。
61. 参见Goldman，“尽力而为”。“拖延症教授”这个名字后来来自Jackson和Pargetter的“应当、选择与实在论”。Jackson后来在“拖延症重访”中重新讨论了这些想法。
62. “可能主义”和“实在论”这两个术语由Jackson和Pargetter提出。
63. 关于Smith更倾向于可能主义的观点，参见Goldman，“尽力而为”。关于几十年后发表的该主题简要概述，参见Smith，“可能主义”，更详细和最新的综述见Timmerman和Cohen，“伦理学中的实在论与可能主义”。关于一个特别有趣的细节，参见Bykvist，“替代行动与结果主义精神”，第50页。
64. 感谢Joe Carlsmith对这个及相关话题的有益讨论。一些最近的哲学文献明确讨论了可能主义、实在论与有效利他主义之间的联系。例如，参见Timmerman，“有效利他主义的规范不足问题”。
65. Singer，“饥荒、富裕与道德”；另见Singer，“溺水儿童与扩展圆圈”。
66. Julia Wise，“目标要高，即使达不到”，《快乐给予》（博客），2014年10月8日。<http://www.givinggladly.com/2014/10/aim-high-even-if-you-fall-short.html>。
67. Will MacAskill，“有效利他主义最佳书籍”，Edouard Mathieu访谈，Five Books，<https://fivebooks.com/best-books/effective-altruism-will-macaskill/>。另见MacAskill和同事Toby Ord创立的“给予我们所能”组织，Ord受Singer等人启发，决定承诺将收入的一部分捐给有效的慈善机构。

68. 参见Singer, 《你能做的最大善事》。
69. 强化学习中的经典同策略方法被称为SARSA，是State – Action – Reward – State – Action的缩写；参见Rummery和Niranjan，“使用连接主义系统的在线Q学习”。经典的异策略方法被称为Q-Learning；参见Watkins，“从延迟奖励中学习”，以及Watkins和Dayan，“Q-Learning”。
70. 参见Sutton和Barto, 《强化学习》。
71. 在伦理学语境中，哲学家Rosalind Hursthouse将美德伦理学框定为一种模仿学习；参见Hursthouse, “规范美德伦理学”。当然，正如Hursthouse和她的批评者所讨论的，存在许多实践和理论困难；例如，参见Johnson, “美德与权利”。关于为什么模仿表面上完美的榜样可能实际上不是好主意的不同观点，参见Wolf, “道德圣人”。
72. Lipsey和Lancaster, “次优的一般理论”。
73. Amanda Askell, 个人通信。
74. 参见Balentine, 《做一台好机器比做一个坏人更好》。
75. Magnus Carlsen, 在2018年世界国际象棋锦标赛第5局后的新闻发布会上，伦敦，2018年11月15日。
76. “启发式方法”。
77. “启发式方法”。
78. 参见Samuel, “Some Studies in Machine Learning Using the Game of Checkers.”
79. 有关Deep Blue架构的更多信息，参见Campbell, Hoane, and Hsu, “Deep Blue。”有关通过使用大师级棋局来调整评估启发式算法的更多阐述，参见程序员Andreas Nowatzky在“Eval Tuning in Deep Thought,” Chess Programming Wiki, https://www.chessprogramming.org/Eval_Tuning_in_Deep_Thought中给出的解释。有关IBM团队在1990年的进展快照，当时被称为Deep Thought，以及关于基于（当时只有900局）专家棋局数据库自动调整（当时只有120个）参数权重的决定的讨论，参见Hsu et al., “A Grandmaster Chess Machine,” 以及Byrne, “Chess-Playing Computer Closing in on Champions。”有关Deep Blue（及其前身Deep Thought）评估函数调整的更多信息，参见例如Anantharaman, “Evaluation Tuning for Computer Chess.”
80. Hsu, “IBM’s Deep Blue Chess Grandmaster Chips.”
81. Weber, “What Deep Blue Learned in Chess School.”
82. Schaeffer et al., “A World Championship Caliber Checkers Program.”
83. Fürnkranz and Kubat, *Machines That Learn to Play Games*.
84. 当然，Deep Blue和AlphaGo的架构和训练过程之间存在许多细微差别。有关AlphaGo的更多细节，参见Silver et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search.”
85. AlphaGo的价值网络来自自我对弈，但其策略网络是模仿性的，通过对人类专家棋局数据库的监督学习进行训练。粗略地说，它在考虑的走法上是常规的，但在决定哪种走法最佳时会独立思考。参见Silver et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search.”

86. Silver et al., “Mastering the Game of Go Without Human Knowledge.” 2018年，AlphaGo Zero进一步被改进为一个更强大的程序——以及一个更通用的程序，不仅在围棋方面具有创纪录的实力，在国际象棋和日本将棋方面也是如此——称为AlphaZero。有关AlphaZero的更多细节，参见Silver et al., “A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go Through Self-Play.” 2019年，该系统的后续迭代称为MuZero，以更少的计算和更少的游戏规则先验知识达到了这一性能水平，同时证明了其足够灵活，不仅擅长棋盘游戏，也擅长Atari游戏；参见Schrittwieser et al., “Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model.”
87. Silver et al., “Mastering the Game of Go Without Human Knowledge.”
88. 有关“快速”和“缓慢”心理过程的心理学研究，也称为“系统1”和“系统2”，参见Kahneman, *Thinking, Fast and Slow*.
89. 参见Coulom, “Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search.”
90. 详情参见Silver et al., “Mastering the Game of Go Without Human Knowledge.” 更准确地说，它使用MCTS期间每个移动的“访问计数”——因此网络学会预测实际上它将花费多长时间思考每个移动。另见同期且密切相关的“专家迭代”（“ExIt”）算法，在Anthony, Tian, and Barber, “Thinking Fast and Slow with Deep Learning and Tree Search.”
91. Shead, “DeepMind’s Human-Bashing AlphaGo AI Is Now Even Stronger.”
92. Aurelius, *The Emperor Marcus Aurelius*.
93. Andy Fitch, “Letter from Utopia: Talking to Nick Bostrom,” *BLARB* (blog), November 24, 2017, <https://blog.lareviewofbooks.org/interviews/letter-utopia-talking-nick-bostrom/>.
94. Blaise Agüera y Arcas, “The Better Angels of our Nature” (lecture), February 16, 2017, VOR: Superintelligence, Mexico City.
95. Yudkowsky, “Coherent Extrapolated Volition.” 另见Tarleton, “Coherent Extrapolated Volition.”
96. 注意一些哲学家，即“道德现实主义者”，实际上确实相信客观道德真理的观念。有关这一系列立场的概述，参见例如Sayre-McCord, “Moral Realism.”
97. Paul Christiano, interviewed by Rob Wiblin, *The 80,000 Hours Podcast*, October 2, 2018.
98. 参见Paul Christiano, “A Formalization of Indirect Normativity,” *AI Alignment* (blog), April 20, 2012, <https://ai-alignment.com/a-formalization-of-indirect-normativity-7e44db640160>，以及Ajeya Cotra, “Iterated Distillation and Amplification,” *AI Alignment* (blog), March 4, 2018, <https://ai-alignment.com/iterated-distillation-and-amplification-157debf1d1616>.
99. 有关AlphaGo策略网络与迭代能力放大思想之间联系的明确讨论，参见Paul Christiano, “AlphaGo Zero and Capability Amplification,” *AI Alignment* (blog), October 19, 2017, <https://ai-alignment.com/alphago-zero-and-capability-amplification-ede767bb8446>.
100. Christiano, Shleiferis, and Amodei, “Supervising Strong Learners by Amplifying Weak Experts.”
101. Paul Christiano, 个人访谈，2019年7月1日。

102. 例如，参见 alignmentforum.org，以及越来越多的研讨会、会议和研究实验室。

第8章. 推理

1. 参见 Warneken and Tomasello, “Altruistic Helping in Human Infants and Young Chimpanzees,” 以及 Warneken and Tomasello, “Helping and Cooperation at 14 Months of Age.” 关于一些实验的视频片段，参见，例如，“Experiments with Altruism in Children and Chimps,” <https://www.youtube.com/watch?v=Z-eU5xZW7cU>。
2. 另见 Meltzoff, “Understanding the Intentions of Others,” 该研究表明18个月大的婴儿能够成功模仿成人试图但未能完成的预期动作，这表明他们“将人们置于一个心理框架内，该框架区分人们的表面行为和涉及目标和意图的更深层次”。
3. Warneken and Tomasello 证明了年仅14个月的人类婴儿也会在伸手够取方面提供帮助，但不会在更复杂的问题上提供帮助。
4. 再次参见 Warneken and Tomasello, “Altruistic Helping in Human Infants and Young Chimpanzees.”
5. Tomasello et al., “Understanding and Sharing Intentions.”
6. Felix Warneken, “Need Help? Ask a 2-Year-Old” (讲座), TEDxAmoskeagMillyard 2013, <https://www.youtube.com/watch?v=-qul57hcu4I>。
7. 参见 “Our Research,” Social Minds Lab, University of Michigan, <https://sites.lsa.umich.edu/warneken/lab/research-2/>。
8. Tomasello et al. (注意他们明确使用控制系统和控制论的语言来构建他们的讨论。)
9. Stuart Russell, 个人访谈, 2018年5月13日。
10. 参见，例如，Uno, Kawato, and Suzuki, “Formation and Control of Optimal Trajectory in Human Multijoint Arm Movement.”
11. 参见，例如，Hogan, “An Organizing Principle for a Class of Voluntary Movements.”
12. Hoyt and Taylor, “Gait and the Energetics of Locomotion in Horses.”
13. Farley and Taylor, “A Mechanical Trigger for the Trot-Gallop Transition in Horses.” 关于人类和动物运动的生物力学的更多信息，另见已故英国著名动物学家Robert McNeill Alexander的工作：例如“The Gaits of Bipedal and Quadrupedal Animals,” *The Human Machine*, 和 *Optima for Animals*。正如Alexander解释的那样，“动物的腿和步态是两个非常强大的优化过程的产物，即自然选择进化过程和经验学习过程。研究它们的动物学家正在试图解决逆向优化问题：他们正在试图发现在动物腿进化和步态进化或学习中重要的优化标准。”关于人类步态背景下逆向最优控制的更多当代研究，参见Katja Mombaur的工作，例如，Mombaur, Truong, and Laumond, “From Human to Humanoid Locomotion—an Inverse Optimal Control Approach.”
14. 关于强化学习与多巴胺系统之间联系的更多信息，参见第4章正文和尾注中的讨论。关于与动物觅食的联系，参见，例如，Montague et al., “Bee Foraging in Uncertain Environments Using Predictive Hebbian Learning,” 和 Niv et al., “Evolution of Reinforcement Learning in Foraging Bees.”

15. Russell, “Learning Agents for Uncertain Environments (Extended Abstract).” 关于从计量经济学角度处理类似问题的早期工作，即所谓的“结构估计”，参见 Rust, “Do People Behave According to Bellman’s Principle of Optimality?” 和 “Structural Estimation of Markov Decision Processes,” 以及 Sargent, “Estimation of Dynamic Labor Demand Schedules Under Rational Expectations.” 从控制理论角度来看这个问题的更早期先驱，参见 Kálmán, “When Is a Linear Control System Optimal?” 1964年在巴尔的摩高等研究院工作，获得美国空军和NASA资助的Kálmán对他所说的“最优控制理论的逆向问题”很感兴趣，即：“给定一个控制律，找出所有使该控制律最优的性能指标。”他指出，“今天对这个问题知之甚少。”

16. 这些加法和乘法变化被称为“仿射”变换。

17. Ng and Russell, “Algorithms for Inverse Reinforcement Learning.”

18. 具体来说，Ng and Russell使用了一种被称为“ ℓ_1 正则化”的方法，也称为“lasso”。这个想法来自Tibshirani, “Regression Shrinkage and Selection via the Lasso.” 关于正则化思想和技术的易懂概述，参见Christian and Griffiths, *Algorithms to Live By*。

19. Abbeel and Ng, “Apprenticeship Learning via Inverse Reinforcement Learning.”

20. Andrew Ng 在 Pieter Abbeel 博士论文答辩中的介绍，斯坦福大学，2008年5月19日；见 http://ai.stanford.edu/~pabbeel//thesis/PieterAbbeel_Defense_19May2008_320x180.mp4。

21. Abbeel, Coates, and Ng, “Autonomous Helicopter Aerobatics Through Apprenticeship Learning.”

22. Abbeel et al., “An Application of Reinforcement Learning to Aerobic Helicopter Flight.” 他们还成功执行了机头向内漏斗飞行和机尾向内漏斗飞行。

23. “由于重复的次优演示往往在其次优性方面有所不同，它们结合在一起通常能够编码出预期的轨迹。”见 Abbeel, “Apprenticeship Learning and Reinforcement Learning with Application to Robotic Control,” 该文提到了 Coates, Abbeel, and Ng, “Learning for Control from Multiple Demonstrations” 中的工作。

24. Abbeel, Coates, and Ng, “Autonomous Helicopter Aerobatics Through Apprenticeship Learning.”

25. 见 Youngblood 的网站：<http://www.curtisyounblood.com/curtis-youngblood/>。

26. Curtis Youngblood, “Difference Between a Piro Flip and a Kaos,” Aaron Shell 访谈，https://www.youtube.com/watch?v=TLi_hp-mmk。

27. 关于斯坦福直升机执行混沌动作的视频片段，见 “Stanford University Autonomous Helicopter: Chaos,” <https://www.youtube.com/watch?v=kN6ifrqwlMY>。

28. Ziebart et al., “Maximum Entropy Inverse Reinforcement Learning,” 该方法利用了从 Jaynes, “Information Theory and Statistical Mechanics” 衍生出的最大熵原理。另见 Ziebart, Bagnell, and Dey, “Modeling Interaction via the Principle of Maximum Causal Entropy.”

29. 见 Billard, Calinon, and Guenter, “Discriminative and Adaptive Imitation in Uni-Manual and Bi-Manual Tasks,” 关于该领域的2009年综述，见 Argall et al., “A Survey of Robot Learning from Demonstration.”

30. 见 Finn, Levine, and Abbeel, “Guided Cost Learning.” 另见 Wulfmeier, Ondrúška, and Posner, “Maximum Entropy Deep Inverse Reinforcement Learning,” 以及 Wulfmeier, Wang, and Posner, “Watch This.”

31. 具体而言, Leike 分析了被称为“套索程序(lasso programs)“的终止或非终止属性。见 Jan Leike, “Ranking Function Synthesis for Linear Lasso Programs,” 硕士论文, 弗赖堡大学, 2013年。

32. Jan Leike, 个人访谈, 2018年6月22日。

33. 见 Leike and Hutter, “Bad Universal Priors and Notions of Optimality.”

34. 该论文是 Christiano et al., “Deep Reinforcement Learning from Human Preferences.” 关于 OpenAI 的博客文章, 见 “Learning from Human Preferences,” <https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/>, 关于 DeepMind 的博客文章, 见 “Learning Through Human Feedback,” <https://deepmind.com/blog/learning-through-human-feedback/>。关于探索从人类偏好和人类反馈学习想法的早期工作, 见, 如 Wilson, Fern, and Tadepalli, “A Bayesian Approach for Policy Learning from Trajectory Preference Queries”; Knox, Stone, and Breazeal, “Training a Robot via Human Feedback”; Akroud, Schoenauer, and Sebag, “APRIL”; 以及 Akroud et al., “Programming by Feedback.” 另见 Wirth et al., “A Survey of Preference-Based Reinforcement Learning Methods.” 关于统一从演示学习和从比较学习的框架, 见 Jeon, Milli, and Drăgan, “Reward-Rational (Implicit) Choice.”

35. Paul Christiano, 个人访谈, 2019年7月1日。

36. Todorov, Erez, and Tassa, “MuJoCo.”

37. 正如论文所述: “从长远来看, 理想的做法是使从人类偏好学习任务的难度不超过从程序化奖励信号学习的难度, 确保强大的强化学习系统能够服务于复杂的人类价值观, 而不是低复杂性目标” (Christiano et al., “Deep Reinforcement Learning from Human Preferences”)。关于 Leike 及其同事后续追求人类奖励建模议程的工作, 见 Leike et al., “Scalable Agent Alignment via Reward Modeling.”

38. Stuart Russell, 个人访谈, 2018年5月13日。

39. 值得注意的是, 将物体递给另一个人本身就是一个令人惊讶的微妙而复杂的动作, 包括推断对方想要如何拿取物体、如何向他们发出你想让他们拿取物体的信号等。见, 如 Strabala et al., “Toward Seamless Human-Robot Handovers.”

40. Hadfield-Menell et al., “Cooperative Inverse Reinforcement Learning.” (“CIRL” 发音为软音 c, 与强人工智能怀疑论者 John Searle 的姓氏同音(无关)。我在社区内倡导硬音 c “curl” 发音更合理, 因为 “cooperative” 使用硬音 c, 但似乎木已成舟。)

41. Dylan Hadfield-Menell, 个人访谈, 2018年3月15日。

42. Russell, *Human Compatible*.

43. 例如, CIRL 框架内的首批理论进展之一就利用了早期认知科学关于教师-学习者策略协同适应的研究。参见 Fisac 等人的 “Pragmatic-Pedagogic Value Alignment” (该研究运用了 Shafit, Goodman 和 Griffiths 的 “A Rational Account of Pedagogical Reasoning”的见解); 作者写道: “据我们所知, 这项工作构成了基于经验验证的认知模型的价值对齐的首次正式分析。” 另见许多相同作者的后续论文: Malik 等人的 “An Efficient, Generalized Bellman Update for Cooperative Inverse Reinforcement Learning”。

44. 华盛顿大学西雅图分校的Maya Çakmak和里斯本高等技术学院的Manuel Lopes一直在研究这个想法；参见Çakmak和Lopes的”Algorithmic and Human Teaching of Sequential Decision Tasks”。当然，如果人类调整他们的行为以达到最大的教学效果——不是为了优化他们自己的指标，而是为了传达指标是什么——那么计算机反过来最好不要使用标准IRL(假设演示是最优的)，而是要进行推理，考虑到教师的行为本质上是教学性的。教学和学习策略相互适应。这是认知科学和机器学习中一个活跃研究的丰富领域。另见，例如，Ho等人的”Showing Versus Doing” 和Ho等人的”A Rational-Pragmatic Account of Communicative Demonstrations”。

45. 参见Gopnik, Meltzoff和Kuhl的《摇篮中的科学家》：“事实证明，婴儿语不仅仅是用来吸引婴儿的甜美塞壬之歌……完全无意识地，[父母们]在与婴儿交谈时比与其他成人交谈时产生的声音更清晰，发音更准确。”作者指出，例如，英语和瑞典语的婴儿语听起来不同。关于这一领域的更多近期工作，参见Eaves等人的”Infant-Directed Speech Is Consistent With Teaching” 和Ramírez, Lytle和Kuhl的”Parent Coaching Increases Conversational Turns and Advances Infant Language Development”。

46. 物体的交接是人机交互研究的明确焦点。参见，例如，Strabala等人的”Toward Seamless Human-Robot Handovers”。

47. Drăgan, Lee和Srinivasa的”Legibility and Predictability of Robot Motion”；另见Takayama, Dooley和Ju的”Expressing Thought”(公平地说，该文确实提到了”可读”动作的概念)，以及Gielniak和Thomaz的”Generating Anticipation in Robot Motion”。更近期的工作研究了，例如，如何不仅传达机器的目标，而且当目标已知时，传达其计划：参见Fisac等人的”Generating Plans That Predict Themselves”。

48. Jan Leike，个人访谈，2018年6月22日。另见Christiano等人的”Deep Reinforcement Learning from Human Preferences”：“离线训练奖励预测器可能导致按真实奖励衡量不受欢迎的奇异行为。例如，在Pong游戏中，离线训练有时会导致我们的agent避免失分但不得分；这可能导致极长的对打。这种行为表明，一般来说，人类反馈需要与强化学习交织在一起，而不是静态提供。”

49. Julie Shah，个人访谈，2018年3月2日。

50. 关于人类交叉训练的研究，参见Blickensderfer, Cannon-Bowers和Salas的”Cross-Training and Team Performance”；Cannon-Bowers等人的”The Impact of Cross-Training and Workload on Team Functioning”；以及Marks等人的”The Impact of Cross-Training on Team Effectiveness”。

51. Nikolaidis等人的”Improved Human-Robot Team Performance Through Cross-Training: An Approach Inspired by Human Team Training Practices”。

52. “Julie Shah: Human/Robot Team Cross Training”，<https://www.youtube.com/watch?v=UQrtw0YUlqM>。

53. Shah实验室更近期的工作探索了无法切换角色的情况。在这里可以使用一个相关的想法叫做”扰动训练”；参见Ramakrishnan, Zhang和Shah的”Perturbation Training for Human-Robot Teams”。

54. Murdoch，《钟》。

55. 一些处于认知科学和AI安全交叉领域的研究人员，包括人类未来研究所的Owain Evans，正在研究如何进行逆向强化学习，以考虑到一个人，例如，经过糕点店时忍不住要进去，但会绕道避开它。参见，例如，Evans, Stuhlmüller和Goodman的”Learning the Preferences of Ignorant, Inconsistent Agents” 和Evans和Goodman的”Learning the Preferences of Bounded Agents”。有整个研究脉络的IRL研究纳入了人类行为的怪癖和有时的非理性。另见Bourgin等人的”Cognitive Model Priors for Predicting Human Decisions”，该工作使用机器学习来开发人类偏好和决策的模型。

56. 参见，例如，Snyder, *Public Appearances, Private Realities*; Covey, Saladin, and Killen, “Self-Monitoring, Surveillance, and Incentive Effects on Cheating”；以及 Zhong, Bohns, and Gino, “Good Lamps Are the Best Police.”

57. 参见 Bateson, Nettle, and Roberts, “Cues of Being Watched Enhance Cooperation in a Real-World Setting,” 以及 Heine et al., “Mirrors in the Head.”

58. Bentham, “Letter to Jacques Pierre Brissot de Warville.”

59. Bentham, “Preface.”

第9章 不确定性

1. Russell, “Ideas That Have Harmed Mankind.”
2. “Another Day the World Almost Ended.”
3. Aksenov, “Stanislav Petrov.”
4. Aksenov.
5. Hoffman, “‘I Had a Funny Feeling in My Gut.’”
6. Nguyen, Yosinski, and Clune, “Deep Neural Networks Are Easily Fooled.” 有关神经网络预测置信度的讨论，参见 Guo et al., “On Calibration of Modern Neural Networks.”
7. 参见 Szegedy et al., “Intriguing Properties of Neural Networks,” 以及 Goodfellow, Shlens, and Szegedy, “Explaining and Harnessing Adversarial Examples。”这是一个活跃的研究领域；关于使系统对对抗样本具有鲁棒性的最新工作，参见，例如，Mqdry et al., “Towards Deep Learning Models Resistant to Adversarial Attacks,” Xie et al., “Feature Denoising for Improving Adversarial Robustness,” 以及 Kang et al., “Testing Robustness Against Unforeseen Adversaries。”另见 Ilyas et al., “Adversarial Examples Are Not Bugs, They Are Features。”该文在对齐语境下构建对抗样本——“(人类指定的)鲁棒性概念与数据固有几何结构之间的不对齐”——并论证”获得既鲁棒又可解释的模型将需要在训练过程中明确编码人类先验。”
8. Creighton, “Making AI Safe in an Unpredictable World.”
9. 关于 Dietterich 的“开放类别问题”研究的详细信息，他为此获得了生命未来研究所资助，参见 <https://futureoflife.org/ai-researcher-thomas-dietterich/>。
10. Thomas G. Dietterich, “Steps Toward Robust Artificial Intelligence” (讲座), 2016年2月14日, 第30届 AAAI 人工智能会议, 凤凰城, 亚利桑那州, http://videolectures.net/aaai2016_dietterich_artificial_intelligence/。该演讲也以略微不同的形式出现在印刷版中；参见 Dietterich, “Steps Toward Robust Artificial Intelligence。”有关开放类别学习的更多信息，参见，例如，Scheirer et al., “Toward Open Set Recognition” ; Da, Yu, and Zhou, “Learning with Augmented Class by Exploiting Unlabeled Data” ; Bendale and Boult, “Towards Open World Recognition” ; Steinhardt and Liang, “Unsupervised Risk Estimation Using Only Conditional Independence Structure” ; Yu et al., “Open-Category Classification by Adversarial Sample Generation” ；以及 Rudd et al., “The Extreme Value Machine。”其他相关的对抗样本和鲁棒分类方法包括 Liu and Ziebart, “Robust Classification Under Sample Selection Bias,” 以及 Li and Li, “Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics。”关于 Dietterich 及其合作者的更多最新结果，参见 Liu et al., “Can We Achieve Open Category Detection with Guarantees?” 和 Liu et al., “Open Category Detection with PAC Guarantees,” 以及 Hendrycks, Mazeika, and Dietterich, “Deep Anomaly Detection with Outlier Exposure。”关于 Google Brain 和 OpenAI 研究人员在2018年提出的促进这些问题研究的基准竞赛提案，参见 Brown et al., “Unrestricted Adversarial Examples,” 以及 “Introducing the Unrestricted Adversarial Example Challenge,” Google AI Blog, <https://ai.googleblog.com/2018/09/introducing-unrestricted-adversarial.html>。
11. Rousseau, *Emile; or, On Education*.

12. Jefferson, *Notes on the State of Virginia*.
13. Yarin Gal, 个人访谈, 2019年7月11日。
14. Yarin Gal, “Modern Deep Learning Through Bayesian Eyes” (讲座), 微软研究院, 2015年12月11日, <https://www.microsoft.com/en-us/research/video/modern-deep-learning-through-bayesian-eyes/>。
15. Zoubin Ghahramani, “Probabilistic Machine Learning: From Theory to Industrial Impact” (讲座), 2018年10月5日, PROBPROG 2018: 概率编程国际会议, <https://youtu.be/crvNIGyqGSU>。
16. 关于贝叶斯神经网络的开创性论文, 参见 Denker 等人的“Large Automatic Learning, Rule Extraction, and Generalization”; Denker 和 LeCun 的“Transforming Neural-Net Output Levels to Probability Distributions”; MacKay 的“A Practical Bayesian Framework for Backpropagation Networks”; Hinton 和 Van Camp 的“Keeping Neural Networks Simple by Minimizing the Description Length of the Weights”; Neal 的“Bayesian Learning for Neural Networks”; 以及 Barber 和 Bishop 的“Ensemble Learning in Bayesian Neural Networks”。关于更近期的工作, 参见 Graves 的“Practical Variational Inference for Neural Networks”; Blundell 等人的“Weight Uncertainty in Neural Networks”; 以及 Hernández-Lobato 和 Adams 的“Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks”。关于这些思想的更详细历史, 参见 Gal 的“Uncertainty in Deep Learning”。关于机器学习中概率方法的总体概述, 参见 Ghahramani 的“Probabilistic Machine Learning and Artificial Intelligence”。
17. Yarin Gal, 个人访谈, 2019年7月11日。
18. Yarin Gal, “Modern Deep Learning Through Bayesian Eyes” (讲座), 微软研究院, 2015年12月11日, <https://www.microsoft.com/en-us/research/video/modern-deep-learning-through-bayesian-eyes/>。
19. 关于使用 dropout-ensemble 不确定性来检测对抗样本的研究, 参见 Smith 和 Gal 的“Understanding Measures of Uncertainty for Adversarial Example Detection”。
20. 每个模型通常被分配一个权重, 描述它能多好地解释数据。这种方法被称为“贝叶斯模型平均”, 或 BMA; 参见 Hoeting 等人的“Bayesian Model Averaging: A Tutorial”。
21. 特别是, 人们发现 dropout 有助于防止网络过于脆弱地“过拟合”其训练数据。参见 Srivastava 等人的“Dropout”, 该论文在发表后的前六年内被引用了惊人的18,500次。
22. 参见 Gal 和 Ghahramani 的“Dropout as a Bayesian Approximation”。近年来出现了替代方案和扩展; 例如, 参见 Lakshminarayanan、Pritzel 和 Blundell 的“Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles”。
23. Yarin Gal, 个人访谈, 2019年7月11日。一个应用是在眼科学中, 在正文中讨论; 其他例子包括, 例如, Uber 的需求预测模型 (Zhu 和 Nikolay 的“Engineering Uncertainty Estimation in Neural Networks for Time Series Prediction at Uber”), 以及丰田研究所的驾驶员预测系统 (Huang 等人的“Uncertainty-Aware Driver Trajectory Prediction at Urban Intersections”。
24. 参见 Gal 和 Ghahramani 的“Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference”, § 4.4.2; 具体来说, Gal 和 Ghahramani 研究了 Lin、Chen 和 Yan 的“Network in Network”, 以及 Lee 等人的“Deeply-Supervised Nets”。注意在调整 dropout 率时应该小心; 参见 Gal 和 Ghahramani 的“Dropout as a Bayesian Approximation”。关于这个想法在递归网络和强化学习中的应用, 分别参见 Gal 和 Ghahramani 的“*A Theoretically*

Grounded Application of Dropout in Recurrent Neural Networks”；Gal的”Uncertainty in Deep Learning”，§3.4.2；以及Gal、McAllister和Rasmussen的”Improving PILCO with Bayesian Neural Network Dynamics Models”。

25. Gal和Ghahramani，“Dropout as a Bayesian Approximation”。

26. Yarin Gal，个人访谈，2019年7月11日。

27. 参见Engelgau等人的”The Evolving Diabetes Burden in the United States”，以及Zaki等人的”Diabetic Retinopathy Assessment”。

28. Leibig等人，“Leveraging Uncertainty Information from Deep Neural Networks for Disease Detection”。

29. 许多研究小组正在探索机器学习中”选择性分类”这一广泛思想的潜力。例如，谷歌研究的Corinna Cortes和她的同事们探索了”带拒绝学习”的想法——即分类器可以简单地”弃权”或以其他方式拒绝做出分类判断。参见Cortes、DeSalvo和Mohri的”Learning with Rejection”；另参见二十世纪中期C. K. Chow探索相关想法的统计工作：Chow的”An Optimum Character Recognition System Using Decision Functions”，以及Chow的”On Optimum Recognition Error and Reject Tradeoff”。关于在强化学习环境中的类似方法，参见Li等人的”Knows What It Knows”。

2018年，由博士生David Madras领导的多伦多大学研究人员拓宽了这一理念的视野，他们不仅询问机器学习系统如何在棘手或模糊的案例上推迟决策以避免犯错，更重要的是，它如何与负责收拾残局的人类决策者协同工作。如果人类决策者在某些类型的案例上特别准确，系统应该更加谦让，即使它本身很有信心；相反，如果在某些类型的案例中人类表现特别糟糕，系统可能简单地冒险做出最佳猜测，即使它不确定——目标不是优化其自身的准确性，而是优化人机决策团队整体的准确性。参见Madras, Pitassi, and Zemel，“Predict Responsibly。”

在相关工作中，来自密歇根大学的Shun Zhang、Edmund Durfee和Satinder Singh探索了网格世界环境中智能体的想法，该智能体通过询问人类用户是否介意某些东西被改变来寻求最小化副作用，并且他们能够提供如何用最少查询次数安全运行的边界。参见Zhang, Durfee, and Singh，“Minimax-Regret Querying on Side Effects for Safe Optimality in Factored Markov Decision Processes。”

30. Kahn et al., “Uncertainty-Aware Reinforcement Learning for Collision Avoidance.”

31. 关于将不确定性与陌生环境联系起来的相关工作，参见Kenton et al., “Generalizing from a Few Environments in Safety-Critical Reinforcement Learning。”关于模仿学习和自动驾驶汽车背景下的相关工作，参见Tigas et al., “Robust Imitative Planning。”

32. Holt et al., “An Unconscious Patient with a DNR Tattoo。”另见Bever, “A Man Collapsed with ‘Do Not Resuscitate’ Tattooed on His Chest,” 和Hersher, “When a Tattoo Means Life or Death,” 的新闻报道。

33. Holt et al., “An Unconscious Patient with a DNR Tattoo.”

34. Cooper and Aronowitz, “DNR Tattoos.”

35. Holt et al., “An Unconscious Patient with a DNR Tattoo.”

36. Bever, “A Man Collapsed with ‘Do Not Resuscitate’ Tattooed on His Chest.”

37. Sunstein, “Irreparability as Irreversibility。”另见Sunstein, “Irreversibility.”

38. Sunstein, “Beyond the Precautionary Principle.” 另见Sunstein, *Laws of Fear*.
39. Amodei et al., “Concrete Problems in AI Safety,” 对“避免负面副作用”和“影响正则化器”进行了出色而广泛的讨论，Taylor et al., “Alignment for Advanced Machine Learning Systems,” 也讨论了“影响测量”的各种想法。关于最新影响研究的良好概述，参见 Daniel Filan, “Test Cases for Impact Regularisation Methods,” <https://www.alignmentforum.org/posts/wzPzPmAsG3BwrBrwy/test-cases-for-impact-regularisation-methods>.
- 卡内基梅隆大学博士生Benjamin Eysenbach在3D MuJoCo环境中研究了类似的想法。他的想法是可逆性，结合了徒步旅行者和背包客“不留痕迹”的理念。这个想法是使用正常的强化学习方法来培养各种任务的能力，但有一个关键限制。与Atari游戏的遍历环境不同（通常学习涉及数十万次外部强制重启），他的智能体有责任在尝试做他们想做的任何事情之前，总是将自己重置回完全相同的起始配置。初步结果令人鼓舞，比如他的简笔画猎豹滑到悬崖边缘，然后倒退——似乎已经内化了一旦越过边缘就无法逆转的道理。参见Eysenbach et al., “Leave No Trace.” 另见更早的Weld and Etzioni, “The First Law of Robotics (a Call to Arms),” 提出了类似的想法。
40. 关于Armstrong的低影响AI智能体工作，参见Armstrong and Levinstein, “Low Impact Artificial Intelligences。”他在2012年和2013年的论文是最早明确解决这一问题的论文之一：参见Armstrong, “The Mathematics of Reduced Impact,” 和Armstrong, “Reduced Impact AI.”
41. Armstrong and Levinstein, “Low Impact Artificial Intelligences.”
42. Armstrong and Levinstein.
43. 正如 Eliezer Yudkowsky 所说，“如果你要治疗癌症，确保病人仍然会死！”参见 <https://intelligence.org/2016/12/28/ai-alignment-why-its-hard-and-where-to-start/>. 另见 Armstrong and Levinstein, “Low Impact Artificial Intelligences,” 该文使用了小行星冲向地球的例子。被限制只能采取“低影响”行动的系统可能无法改变小行星轨道——或者，也许更糟糕的是，有能力抵消的系统可能会改变小行星轨道，拯救地球，然后还是会炸毁地球。
44. Victoria Krakovna, 个人访谈，2017年12月8日。
45. 参见Krakovna等人的“使用逐步相对可达性惩罚副作用”。对Krakovna来说，将问题框架化为“副作用”而不是“影响”本身，至少让一些悖论看起来消失了。“如果一个机器人在搬箱子时撞到了花瓶，”她说，“打破花瓶是副作用，因为机器人本可以轻易绕过花瓶。另一方面，制作煎蛋卷的烹饪机器人必须打破一些鸡蛋，所以打破鸡蛋不是副作用。”另见Victoria Krakovna, “使用相对可达性测量和避免副作用”，2018年6月5日，<https://vkrakovna.wordpress.com/2018/06/05/measuring-and-avoiding-side-effects-using-relative-reachability/>。
46. Leike等人，“AI安全网格世界”。
47. Victoria Krakovna, 个人访谈，2017年12月8日。
48. 逐步基线的想法由Alexander Turner在<https://www.alignmentforum.org/posts/DvmhXysefEyEvXuXS/overcoming-clinginess-in-impact-measures>中提出。相对可达性的想法在Krakovna等人的“使用逐步相对可达性惩罚副作用”以及Krakovna等人的“设计代理激励以避免副作用”，DeepMind 安全研究（博客），<https://medium.com/@deepmindsafetyresearch/designing-agent-incentives-to-avoid-side-effects-e1ac80ea6107> 中进行了探讨。

49. Turner , Hadfield-Menell 和 Tadepalli , “通过可达效用保护的保守代理”。另见 Turner 在 <https://www.alignmentforum.org/s/7CdozhJaLEKHwvJW> 的“重新构建影响”系列，以及他在“迈向新的影响测量”中的进一步讨论，<https://www.alignmentforum.org/posts/yEa7kwoMpsBgaBCgb/towards-a-new-impact-measure>；他写道，“我有一个理论，AUP似乎对高级代理有效，不是因为可达集合的效用内容实际上很重要，而是因为存在一种通用的效用实现货币——权力。”参见Turner的“最优远视代理倾向于寻求权力”。有关AI安全背景下权力概念的更多信息，包括“赋权”的信息理论描述，参见Amodei等人的“AI安全中的具体问题”，该文章反过来引用了Salge, Glackin和Polani的“赋权：介绍”以及Mohamed和Rezende的“内在动机强化学习的变分信息最大化”。

50. Alexander Turner, 个人访谈, 2019年7月11日。

51. Wiener, “自动化的一些道德和技术后果”。

52. 据Paul Christiano称，作为AI安全原则的“可纠正性”始于机器智能研究所的Eliezer Yudkowsky，这个名称本身来自Robert Miles。参见Christiano的“可纠正性”，<https://ai-alignment.com/corrigibility-3039e668638>。

53. Dadich, Ito和Obama, “巴拉克·奥巴马，神经网络，自动驾驶汽车，以及世界的未来”。

54. Dylan Hadfield-Menell, 个人访谈, 2018年3月15日。

55. Turing, “数字计算机能思考吗？”

56. Russell, 人类兼容。Russell早些时候但用不同的话在“我们应该害怕超级聪明的机器人吗？”中提出了这一点。在那之前近十年，Steve Omohundro在“基本AI驱动”中指出，“几乎所有系统[都会]保护其效用函数不被修改。”

57. Soares等人，“可纠正性”。另见Armstrong的相关工作：“人工代理的有动机价值选择”。对于修改或中断AI代理时出现的有趣问题的其他观点，例如，参见Orseau和Armstrong的“安全可中断代理”以及Riedl和Harrison的“进入矩阵”。对于实际要求人们不要关闭它们的机器人的研究，以及人类是否遵从，参见Horstmann等人的“机器人的社交技能及其反对是否会阻止互动者关闭机器人？”

58. 参见Nate Soares等人，“可纠正性”，AAAI-15演示，2015年1月25日，<https://intelligence.org/wp-content/uploads/2015/01/AAAI-15-corrigibility-slides.pdf>。

59. Russell, “我们应该害怕超级聪明的机器人吗？”

60. Dylan Hadfield-Menell, “关机开关”（讲座），鲁棒和有益AI学术研讨会系列(CSRBAI)，机器智能研究所，加利福尼亚州伯克利，2016年6月8日，<https://www.youtube.com/watch?v=t06IciZknDg>。

61. Milli等人，“机器人应该顺从吗？”对于系统最好不服从人类命令的其他情况的工作，例如，参见Coman等人的“AI反叛的社会态度”以及Aha和Coman的“AI反叛”。

62. Smitha Milli, “实现AI安全的方法”（访谈），澳大利亚墨尔本，2017年8月，<https://www.youtube.com/watch?v=l82SQfrbdj4>。

63. 有关使用此范式的可纠正性和模型错误规范的更多信息，另见，例如，Carey的“CIRL框架中的不可纠正性”。

64. Dylan Hadfield-Menell, 个人访谈, 2018年3月15日。

65. Russell, 人类兼容。

66. Hadfield-Menell等人， “逆向奖励设计”。
67. 关于DeepMind的Tom Everitt及其在DeepMind和澳大利亚国立大学的合作者们对这一问题的相关框架和方法，参见Everitt等人的” Reinforcement Learning with a Corrupted Reward Channel”。
68. Hadfield-Menell等人， “Inverse Reward Design”。
69. Prümmer, 《道德神学手册》。
70. Rousseau, 《爱弥儿；或论教育》。
71. 实际上，几乎所有真正的历史辩论都涉及关于某个行为是否有罪的争论案例——但不包括也考虑不采取该行为可能有罪的案例。关于这一点的更多讨论，参见Sepielli的” ‘Along an Imperfectly-Lighted Path.’ ”。
72. 这句被广泛引用的格言最初出现在1930年9月20日《圣地亚哥联合报》的社论版上：“零售珠宝商声称每个人都应该佩戴两块手表。但拥有一块手表的人知道现在是几点，而拥有两块手表的人永远无法确定。”
73. 参见Prümmer, 《道德神学手册》，§ 145 – 56。
74. 参见Connell, “Probabilism”。
75. Prümmer, 《道德神学手册》。
76. Will MacAskill, 个人访谈，2018年5月11日。
77. 近年来重新审视这些问题的哲学家之一是密歇根理工大学的Ted Lockhart。参见Lockhart的《道德不确定性及其后果》。如他所说：“当我不确定在道德上应该做什么时，我该怎么办？哲学家们很少关注这类问题。”
78. 关于effective altruism理念的更多信息，参见MacAskill的《做更好的好事》和Singer的《你能做的最大好事》。关于” effective altruism” 这一术语历史的更多信息，参见 MacAskill 的” The History of the Term ‘Effective Altruism’ ”，Effective Altruism论坛，http://effective-altruism.com/ea/5w/the_history_of_the_term_effective_altruism/。
79. MacAskill、Bykvist和Ord, 《道德不确定性》。另见Lockhart的早期著作：《道德不确定性及其后果》。
80. 参见，例如，Lockhart的《道德不确定性及其后果》，以及Gustafsson和Torpman的” In Defence of My Favourite Theory”。
81. 例如，有纯序数理论和纯义务论理论。还存在其他问题；更多讨论参见MacAskill、Bykvist和Ord的《道德不确定性》。
82. 关于社会选择理论的更多信息，参见，例如，Mueller的《公共选择III》和Sen的《集体选择与社会福利》；关于计算视角下的社会选择理论，参见，例如，Brandt等人的《计算社会选择手册》。
83. 关于” 道德议会” 的概念，参见Bostrom的” Moral Uncertainty—Towards a Solution? ”；关于” 道德交易”，参见Ord的” Moral Trade”。
84. 一种方法，参见，例如，Humphrys的” Action Selection in a Hypothetical House Robot”。

85. 参见“Allocation of Discretionary Funds from Q1 2019”，《GiveWell 博客》，<https://blog.givewell.org/2019/06/12/allocation-of-discretionary-funds-from-q1-2019/>。

86. 关于这方面的更多讨论，另见Ord的《悬崖》。

87. Paul Christiano，个人访谈，2019年7月1日。

88. 关于这一主题，另见Sepielli的“What to Do When You Don’t Know What to Do When You Don’t Know What to Do...”。

89. Shleiferis，“Why I’m Less of a Hedonic Utilitarian Than I Used to Be”。

结论

1. Bertrand Russell, “The Philosophy of Logical Atomism”，收录于《逻辑与知识》。
 2. 参见Knuth的” Structured Programming with *Go to Statements*” 和” Computer Programming as an Art”，均来自1974年。这句话有着复杂的历史，Knuth本人在15年后的1989年，在《TeX的错误》中称其为” [C.A.R.] Hoare的格言”。然而，似乎没有证据表明这句话出自Hoare。当2004年有人询问Hoare这句话时，他说对其来源” 毫无印象”，暗示这可能是Edsger Dijkstra会说的话，并补充道：“我认为你可以公正地假设这是共同文化或民间传说” (Hans Gerwitz, “Premature Optimization Is the Root of All Evil”，<https://hans.gerwitz.com/2004/08/12/premature-optimization-is-the-root-of-all-evil.html>)。2012年，Knuth承认：“我确实说过类似‘过早优化是编程中万恶之源’这样的话” (Mark Harrison, “A note from Donald Knuth about TAOCP”，<http://codehaus.blogspot.com/2012/03/note-from-donald-knuth-about-taoct.html>)。很可能这句话确实出自他本人。
 3. 美国疾病控制中心警告，在寒冷卧室中的婴儿可能出现体温过低，2017年泰国一名健康的中年男子在寒冷夜晚仅因为让风扇持续运转，就在卧室内死于低温性休克。参见：Centers for Disease Control and Prevention, “Prevent Hypothermia and Frostbite,” <https://www.cdc.gov/disasters/winter/staysafe/hypothermia.html>，以及Straits Times, “Thai Man Dies From Hypothermia After Sleeping With 3 Fans Blowing at Him,” November 6, 2017, <https://www.straitstimes.com/asia/se-asia/thai-man-dies-from-hypothermia-after-sleeping-with-3-fans-blowing-at-him>。
 4. Wiener, *God and Golem, Inc.*。2016年，MIRI研究员Jessica Taylor探索了她称为” quantilizers”的相关概念：智能体不是完全优化潜在有问题的指标，而是满足于” 足够好”的行为；参见Taylor, “Quantilizers”。这也与被称为” 早停法”的正则化方法有相似之处；参见Yao, Rosasco, and Caponnetto, “On Early Stopping in Gradient Descent Learning”。在AI安全背景下，关于” 一个用于改进系统的指标被使用到进一步优化变得无效或有害的程度”的进一步讨论，参见Manheim and Garrahan, “Categorizing Variants of Goodhart’s Law”。
 5. “£13.3m Boost for Oxford’s Future of Humanity Institute,” <http://www.ox.ac.uk/news/2018-10-10-£133m-boost-oxford’s-future-humanity-institute>。
 6. Huxley, *Ends and Means*。
 7. 关于医学中性别偏见的近期普通读者讨论，参见Perez, *Invisible Women*。关于该主题的学术文献，参见如Mastroianni, Faden, and Federman, *Women and Health Research*，以及Marts and Keitt, “Foreword”。医学领域也担心老年人——目前增长最快的人口群体之一——在医学试验中也明显代表不足；参见如Vitale et al., “Under-Representation of Elderly and Women in Clinical Trials”，以及Shenoy and Harugeri, “Elderly Patients’ Participation in Clinical Trials”。
- 这是许多领域的活跃研究领域；例如，关于动物学博物馆收藏的近期讨论，参见Cooper et al., “Sex Biases in Bird and Mammal Natural History Collections”。
8. 参见如Bara Fintel, Athena T. Samaras, and Edson Carias, “The Thalidomide Tragedy: Lessons for Drug Safety and Regulation,” *Helix*, <https://helix.northwestern.edu/article/thalidomide-tragedy-lessons-drug-safety-and-regulation>; Nick McKenzie and Richard Baker, “The 50-Year Global Cover-up,” *Sydney Morning Herald*, July 26, 2012, <https://www.smh.com.au/national/the-50-year-global-cover-up-20120725-22r5c.html>；以及” Thalidomide,” Brought to Life, Science Museum, <http://broughttolife.science museum.org.uk/broughttolife/themes/controversies/thalidomide>，连同如Marts and Keitt, “Foreword”。

9. 有时这种共识完全是令人反感的。2019年，AI Now Institute的Kate Crawford和艺术家Trevor Paglen深入挖掘ImageNet数据，发现了一些奇异且令人震惊的内容。参见他们的“Excavating AI”：<https://www.excavating.ai>。他们的工作导致ImageNet从数据集中删除了六十万张人物图像，这些图像被标记为从“盗窃癖者”到“乡巴佬”再到“妓女”等各种标签。

10. 原始ImageNet数据实际包含两万个类别；2012年由AlexNet获胜的ImageNet大规模视觉识别挑战赛(ILSVRC)使用了精简版数据，仅包含一千个类别。参见Deng et al., “ImageNet”，以及Russakovsky et al., “ImageNet Large Scale Visual Recognition Challenge”。

11. Stuart Russell论证了这一观点，他建议使用机器学习本身来推断标签错误不同成本的更细致表示。参见如Russell, *Human Compatible*。

12. 参见Mikolov et al., “Efficient Estimation of Word Representations in Vector Space”。

13. 2019年5月，arXiv论文Nissim, van Noord, and van der Goot, “Fair Is Better Than Sensational”引起了一些轰动；它用尖锐的措辞批评了类比的“平行四边形”方法。Bolukbasi et al., “Man Is to Computer Programmer as Woman Is to Homemaker?”的作者在Twitter上回应，非正式讨论可见于<https://twitter.com/adamfungi/status/1133865428663635968>。Bolukbasi等人的论文本身在其附录A和附录B中讨论了3CosAdd算法与其作者使用的算法之间微妙但重要的差异。

14. Tversky, “Features of Similarity”。

15. 参见如Chen, Peterson, and Griffiths, “Evaluating Vector-Space Models of Analogy”。

16. 关于监禁本身可能产生的犯罪效应，参见如Stemen, “The Prison Paradox”中的讨论(特别是脚注23)，以及Roodman, “Aftereffects”。

17. 参见如Jung et al., “Eliciting and Enforcing Subjective Individual Fairness”。

18. 参见Poursabzi-Sangdeh et al., “Manipulating and Measuring Model Interpretability”。

19. Bryson, “AI的六种解释”，例如，论证在AI语境下的“解释”不应该仅仅包括系统的内部运作，还应包括“导致系统作为产品发布和销售以及/或作为服务运营的人类行为”。

20. 参见Ghorbani, Abid, and Zou, “神经网络的解释是脆弱的”。

21. 参见Mercier and Sperber, “人类为什么推理？”AI对齐中一个有趣的研究方向涉及开发能够相互辩论的机器学习系统；参见Irving, Christiano, and Amodei, “通过辩论实现AI安全”。

22. Jan Leike, “通用强化学习”（讲座），2016年鲁棒与有益AI研讨会系列，机器智能研究所，加利福尼亚州伯克利，2016年6月9日，<https://www.youtube.com/watch?v=hSiJuVTBoE&t=239s>。避免不可恢复错误的强化学习思想是一个活跃的研究领域；参见，例如，Saunders等人，“无错试验”，以及Eysenbach等人，“不留痕迹”，了解一些方法。

23. 参见Omohundro, “基本AI驱动力”。另见，例如，L. M. Montgomery 1921年的《绿山墙的安妮》小说《英格赛的瑞拉》，其中瑞拉沉思道：“我不会想要回到两年前的我，即使我能做到…而且仍然…在两年后，我可能会回顾并为它们带给我的发展而感激；但我现在不想要它。”奥利弗小姐回答她：“我们从来不想要。这就是为什么我们不能选择自己的发展方式和程度，我想。”

24. 参见Paul, 《变革性体验》。
25. “多智能体强化学习”子领域致力于解决这类问题。参见，例如，Foerster等人，“学习用深度多智能体强化学习进行通信”，以及Foerster等人，“在对手学习意识下学习”。
26. Piaget, 《儿童现实的建构》。
27. Demski and Garrahan, “嵌入式智能体”。
28. 参见，例如，Evans, Stuhlmüller, and Goodman, “学习无知的、不一致智能体的偏好”；Evans and Goodman, “学习有界智能体的偏好”；以及Bourgin等人，“预测人类决策的认知模型先验”。
29. 参见Ziebart等人，“最大熵逆强化学习”，以及Ziebart, Bagnell, and Dey, “通过最大因果熵原理建模交互”。机器人学和自动驾驶汽车的许多近期工作使用相同的人类行为模型，有时称为“噪声理性”行为或“玻尔兹曼(非)理性”。参见，例如，Finn, Levine, and Abbeel, “引导成本学习”；Sadigh等人，“利用对人类行为影响的自动驾驶汽车规划”；以及Kwon等人，“当人类不是最优的时候”。
30. 正如Stuart Russell在他1998年的原始论文中所说：“我们能否在学习过程中而不是学习之后通过观察确定奖励函数？”Russell, “不确定环境下的学习智能体（扩展摘要）”。
31. 这仍然是一个非常开放的研究问题。对于最近的工作，参见Chan等人，“辅助多臂老虎机”。
32. 伯克利的Smitha Milli和Anca Drăgan探索了这个问题：参见Milli and Drăgan, “字面意思还是教学的人类？”
33. Stefano Ermon , 接受 Ariel Conn 采访 , 生命未来研究所 , 2017 年 1 月 26 日 ,
<https://futureoflife.org/2017/01/26/stefano-ermon-interview/>。
34. Roman Yampolskiy , 接受 Ariel Conn 采访 , 生命未来研究所 , 2017 年 1 月 18 日 ,
<https://futureoflife.org/2017/01/18/roman-yampolskiy-interview/>。
35. 参见，例如，Arrow, “社会福利概念中的一个困难”。
36. 另见Recht等人在“ImageNet分类器能泛化到ImageNet吗？”中的努力，他们试图在新照片上重现像AlexNet等图像识别系统的准确性——存在持续的准确性差距，导致作者推测，无论他们多么努力地模仿原始CIFAR-10和ImageNet方法，图像和人工提供的标签在2019年与2012年相比简单来说不可避免地有些不同。
37. Latour, 《潘多拉的希望》。
38. 参见 Paul Christiano , “失败是什么样子的” , AI 对齐论坛 , 2019 年 3 月 17 日 ,
<https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like>。“AI灾难的刻板印象是一个强大的、恶意的AI系统，它让创造者措手不及，并迅速获得对人类其余部分的决定性优势。我认为失败可能不会是这样的，”他写道。相反，他担心“机器学习将提高我们‘得到我们能测量的东西’的能力，这可能导致缓慢的灾难”。
39. 国家运输安全委员会，2019年。发展中自动驾驶系统控制车辆与行人碰撞。公路事故报告NTSB/HAR-19/03。华盛顿特区。
40. 参见，例如，Odell, *How to Do Nothing*。

41. Read, *The Grass Roots of Art*.

42. Turing等人, “Can Automatic Calculating Machines Be Said to Think?”

致谢

1. McCulloch, *Finality and Form.*

[参考文献]

Abbeel, Pieter. “Apprenticeship Learning and Reinforcement Learning with Application to Robotic Control.” 博士论文，斯坦福大学，2008年。

Abbeel, Pieter, Adam Coates, 和 Andrew Y. Ng。 “Autonomous Helicopter Aerobatics Through Apprenticeship Learning.” *International Journal of Robotics Research* 29, 第13期 (2010年) : 1608 – 39。

Abbeel, Pieter, Adam Coates, Morgan Quigley, 和 Andrew Y. Ng。 “An Application of Reinforcement Learning to Aerobic Helicopter Flight.” 载于 *Advances in Neural Information Processing Systems*, 1 – 8, 2007年。

Abbeel, Pieter, 和 Andrew Y. Ng。 “Apprenticeship Learning via Inverse Reinforcement Learning.” 载于 *Proceedings of the 21st International Conference on Machine Learning*。ACM, 2004年。

Ackley, David, 和 Michael Littman。 “Interactions Between Learning and Evolution.” 载于 *Artificial Life II: SFI Studies in the Sciences of Complexity*, 10:487 – 509。Addison-Wesley, 1991年。

ACLU Foundation。 “The War on Marijuana in Black and White,” 2013年。<https://www.aclu.org/report/report-war-marijuana-black-and-white?redirect=criminal-law-reform/war-marijuana-black-and-white>。

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, 和 Been Kim。 “Sanity Checks for Saliency Maps.” 载于 *Advances in Neural Information Processing Systems*, 9505 – 15, 2018年。

Aha, David W., 和 Alexandra Coman。 “The AI Rebellion: Changing the Narrative.” 载于 *Thirty-First AAAI Conference on Artificial Intelligence*, 2017年。

Akrour, Riad, Marc Schoenauer, 和 Michèle Sebag。 “APRIL: Active Preference-Learning Based Reinforcement Learning.” 载于 *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 116 – 31。Springer, 2012年。

Akrour, Riad, Marc Schoenauer, Michèle Sebag, 和 Jean-Christophe Souplet。 “Programming by Feedback.” 载于 *International Conference on Machine Learning*, 1503 – 11。JMLR, 2014年。

Aksenov, Pavel。 “Stanislav Petrov: The Man Who May Have Saved the World.” BBC News, 2013年9月25日。<https://www.bbc.com/news/world-europe-24280831>。

Al-Shawaf, Laith, Daniel Conroy-Beam, Kelly Asao, 和 David M. Buss。 “Human Emotions: An Evolutionary Psychological Perspective.” *Emotion Review* 8, 第2期 (2016年) : 173 – 86。

Alexander, R. McNeill。 “The Gaits of Bipedal and Quadrupedal Animals.” *International Journal of Robotics Research* 3, 第2期 (1984年) : 49 – 59。

———。 *The Human Machine: How the Body Works*。Columbia University Press, 1992年。

———。 *Optima for Animals*。Princeton University Press, 1996年。

Alexander, S.A.H。 “Sex, Arguments, and Social Engagements in Martial and Premarital Relations.” 硕士论文，密苏里大学堪萨斯城分校，1971年。

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, 和Dan Mané。 “Concrete Problems in AI Safety。” arXiv预印本arXiv:1606.06565, 2016年。

Ampère, André-Marie。 *Essai sur la philosophie des sciences; ou, Exposition analytique d' une classification naturelle de toutes les connaissances humaines*。 巴黎：Bachelier, 1834年。

Anantharaman, Thomas S. “Evaluation Tuning for Computer Chess: Linear Discriminant Methods。” *ICGA Journal* 20, 第4期 (1997年) : 224 – 42。

Anderson, James A., 和Edward Rosenfeld。 *Talking Nets: An Oral History of Neural Networks*。 MIT Press, 1998年。

Andre, David, 和Astro Teller。 “Evolving Team Darwin United。” 载于 *RoboCup-98*, 346 – 51。 Springer, 1999年。

Andrew, A. M. “Machines Which Learn。” *New Scientist*, 1958年11月27日。

Andrews, D. A. “The Level of Service Inventory (LSI): The First Follow-up。” *Ontario Ministry of Correctional Services*。 多伦多, 1982年。

Andrews, D. A., 和J. L. Bonta。 “The Level of Service Inventory – Revised。” 多伦多：Multi-Health Systems, 1995年。

Andrews, D. A. , James Bonta , 和J. Stephen Wormith 。 “The Recent Past and Near Future of Risk and/or Need Assessment。” *Crime & Delinquency* 52, 第1期 (2006年) : 7 – 27。

Angelino, Elaine, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, 和Cynthia Rudin。 “Learning Certifiably Optimal Rule Lists for Categorical Data。” *Journal of Machine Learning Research* 18 (2018年) : 1 – 78。

Angwin, Julia。 *Dragnet Nation: A Quest for Privacy, Security, and Freedom in a World of Relentless Surveillance*。 Times Books, 2014年。

Angwin, Julia, 和Jeff Larson。 “ProPublica Responds to Company’s Critique of Machine Bias Story。” ProPublica, 2016年7月29日。

Angwin, Julia, Jeff Larson, Surya Mattu, 和Lauren Kirchner。 “Machine Bias。” ProPublica, 2016年5月23日。

“Another Day the World Almost Ended。” RT, 2010年5月19日。<https://www.rt.com/usa/nuclear-war-stanislav-petrov/>。

Anthony, Thomas, Zheng Tian, 和David Barber。 “Thinking Fast and Slow with Deep Learning and Tree Search。” 载于 *Advances in Neural Information Processing Systems*, 5360 – 70, 2017年。

Arendt, Hannah。 *The Human Condition*。 University of Chicago Press, 1958年。

Argall, Brenna D. , Sonia Chernova , Manuela Veloso , 和 Brett Browning。 “A Survey of Robot Learning from Demonstration。” *Robotics and Autonomous Systems* 57, 第5期 (2009年) : 469 – 83。

Armstrong, Stuart 。 “The Mathematics of Reduced Impact: Help Needed , ” LessWrong , 2012 年 2 月 16 日 。
<https://www.lesswrong.com/posts/8Nwg7kqAfCM46tuHq/the-mathematics-of-reduced-impact-help-needed>。