

何为理性思考？



2024.03

参考答案·月刊

2024年3月

参考月刊

ANSWERS

写在前面

本次月刊的作者是美国决策理论家、人工智能研究专家，埃利泽·尤德科夫斯基，很多人将他视为我们这个时代最伟大的思想家之一。

他在美国有一个非常著名的理性主义社区 LessWrong.com，社区讨论主题集中于“如何理性思考”以及“如何在生活中作出最佳决策”，思考浓度非常之高，我第一次访问时，简直惊掉下巴，在简中世界从来没有见过。

本次策展内容是作者在社区中的过往博客合集《理性：从人工智能到僵尸》的第一册《地图和疆域》（Map and Territory）的序章部分，全集共有六册，可以算是关于现代理性思维的最佳作品之一。

作者用有趣易懂的方式向读者解释了高度复杂的哲学和科学概念。本月刊可以算是有史以来，字最少但是最难懂的一期。但是，如果你仔细读完并加以吸收，我相信你一定会迈入理性思考的新世界。

本月刊由参考答案策展翻译。

「参考答案」是一项为终身学习者打造的内容策展编辑服务，专注分享更具长期价值的内容，希望成为读者的优质信息捕手。

欢迎在小红书/即刻/微信搜索「参考答案 Answers」，了解更多内容。

本内容仅供学习交流，非商业使用，版权归原作者所有。

PART1：我所说的“理性”是什么意思？

Eliezer Yudkowsky

我的意思是：

1. 认识理性（Epistemic rationality）：系统地提高信念的准确性。
2. 工具理性（Instrumental rationality）：系统地实现你的价值观。

当你睁开眼睛并环顾四周，你会根据桌子的位置找到你的笔记本电脑，根据墙壁来确定书柜的位置。如果你的视力或大脑有问题，你可能

会误以为某个地方有书柜，但当你走过去拿书时，才意识到其实那里并没有书柜。

这就像是你持有一个错误的信念，并且拿着一张与实际领土不符的世界地图。而认识理性就是要构建一张精确的地图。信念与现实之间的这种一致性通常被称为“真理”（[truth](<https://www.readthesequences.com/The-Simple-Truth>))，我也乐于将其称为真理。

另一方面，工具理性关注的是如何实践，如何将未来引导至你所期望的方向。这是一种选择行动的艺术，以实现更符合你偏好的结果。我有时也把这称作“成功”（winning）。

因此，理性涉及形成准确信念并作出有成效的决定。

这里的追求“真理”（truth）并不意味着忽略不确定性或者只关注直接证据。

从本质上讲，观察你周围的房间并在心里构建地图，与相信地球有一个熔融的核心或者朱利叶斯·凯撒是秃顶的，其实并没有本质区别。

这些问题在空间和时间上都与你相距甚远，可能看起来比关于你家书柜的问题更为抽象和模糊。然而，无论是关于2015年地球核心的状态，还是关于公元前50年凯撒是否秃顶的情况，背后都有其事实依据。这些事实可能会对你产生实质性的影响，哪怕你永远没有机会亲眼见到凯撒或触摸到地球的核心。

而且，追求“成功”并不意味着必须以牺牲他人为代价。人生规划可以是以合作和自我奉献为

核心，而不仅仅是竞争。在这里，所说的“你的价值观”（Your values）涵盖了你所关心的一切事物，包括对他人的关心。这并不局限于自私或狭隘的价值观。

当人们声称某人是理性的时候，这通常只是一种更加强烈的说法，意在表达“我认为某人是正确的”或“我认为某人是好的”。那么，我们为何还需要一个额外的词汇“理性”来描述“正确”（true）和“好”（good）呢？

我们也可以类似的逻辑来质疑“真实”

（true）这个词。当我们在描述“雪是白色的”时候，没必要还特意加上“这是真的”，因为说“雪是白色的”已经隐含了这一点。真理这一概念的价值在于，它允许我们更准确地预测和解释实验或现实世界中的现象。如果没有“真实”

（true）或“精准”（accurate）这类概念的支撑，我们很难得出那些实用的、普遍适用的观点或总结。

“理性人总是做出能最大化期望收益的决定”这句话基于所谓的“工具理性”。然而，说“吃蔬菜是理性的”大概也可以用“吃蔬菜有好处”或“吃蔬菜对你有益”来替代。我们需要像“理性”这样的词，以便更好地了解哪些思考方式能更有效地帮我们找到真理或价值，以及在哪些方面我们还不够好。

有时，实验心理学家会发现人们的推理方式很奇怪。例如，有些人认为“比尔是爵士乐手”的可能性要低于“比尔既是会计又是爵士乐手”的可能性，这似乎说不通。因为如果比尔既是会计又能演奏爵士乐，那么他显然也符合是爵士

乐手的条件。关键在于，我们应该如何正确指出这种判断的错误呢？

实验心理学家会使用这两个黄金标准：概率论和决策理论。

概率论是我们进行理性思考的黄金规则。不论是确定书架的具体位置、计算地球核心的温度，还是估算凯撒大帝头上的头发数量，概率论都以公平且不偏不倚的方式处理这些问题。它们本质上都是同一个问题：如何处理证据和观察，以此来修正（或“更新”）我们的个人信念。同样地，决策理论是指导理性行动的法则，无论我们的目标是什么，可选择的行动方案如何，它的适用性都是一样的。

用“ $P(\text{某事件})$ ”来表示“某某事件发生的概率”，而 $P(A, B)$ 则用来表示“A事件和B事件同时发生的概率”。根据概率论的基本原理，我们有 $P(A) \geq P(A, B)$ 。因此，认为 $P(\text{比尔演奏爵士乐})$ 的概率小于 $P(\text{比尔演奏爵士乐, 比尔是会计师})$ 的概率，这种判断是错误的。

为了保持讨论的技术性，我们可以将这种概率判断称为非贝叶斯的。在数学上，那些符合明确定义原则的信念和行为，被称为“贝叶斯”（Bayesian）的。

（译者注：贝叶斯理论（Bayesian）是一种统计学方法，用于在给定证据的情况下估计一个事件的概率。该理论依赖于对先验知识和新证据的结合，以更新对某一假设概率的认识。）

现代理性的概念并非仅限于文字推理。就像我之前提到的，你睁开眼睛，环顾四周，然后在脑海中构建一个房间的模型，这个房间里靠墙放着一个书架。这种理性概念足够广泛，它包括了你的眼睛和大脑的视觉区域，作为映射世界的工具。同样，它也涵盖了你那无需言语的直觉。数学并不在乎我们是否用同一个英文词汇“理性”来指代斯波克或贝叶斯主义。数学模型关注的是实现目标或是描绘世界的有效方法，而不是这些方法是否符合我们对“理性”应有的预设概念和固有成见。

这样的解释并没有彻底澄清实际应用中“理性”所承载的意义，究其原因，主要有两点：

首先得说，对于现实世界中的大多数问题，想要完整应用贝叶斯理论在计算上几乎是不可能

的。没人能做到精确无误地完成所有的数学运算，这就像你不可能通过计算夸克的活动来准确预测股市的走势一样。

要想在现实世界中发现真理并创造价值，光靠遵循数学公式是远远不够的，这需要更深层次的思考和更高超的技巧。探寻真理和创造价值的过程，实际上是一门艺术，它要求我们认识到自身的局限，克服偏见，避免自我欺骗，保持情绪稳定，以便直面真相并采取必要行动，如此等等。

其次，有时连数学的应用本身也会遭到质疑。拿概率论和决策理论来说，这些数学工具旨在帮助我们理解世界和做出选择。但在某些场合下，人们会对这些工具是否真的适用产生疑问。

在概率论中，有一些问题涉及到观察者数量不确定的情况，这使得概率的计算变得格外复杂。而在决策理论中，问题则涉及到其他智能体能够预测你的决策，这也引发了对决策理论的质疑。在这种情况下，试图通过重新定义“理性”一词来解决问题是不起作用的。因为这样做最终只会引出另一个问题：为什么别人必须接受这一新的定义呢？

我之所以对概率论不感兴趣，并不是因为它被视为来自拉普拉斯传下的神圣真理。相反，我对贝叶斯式的信念更新——也就是采用奥卡姆剃刀原则——感到着迷，是因为我相信这种思考方式能帮我们更系统、更具有“准确性”，即能更好地反映出世界实际情况的地图。

然后，还有一些思考方式的问题，似乎既不属于概率论也不属于决策论的范畴——比如，一旦你掌握了真相，应该怎样看待这个真相。在这里，再次尝试以特定的方式定义“理性”，并不能提供答案，反而只是预设了一个答案。

我在这儿，不是来跟你争论什么词语的意思，哪怕这个词是“理性”。我们用字母串连起来的意义，形成词汇，其实质是为了促进人与人之间的沟通，目的在于把一个人脑海中的想法，传达给另一个人。通过控制某个词与某个概念之间的联系，你既不能改变现实，也无法证明某个想法的正确与否。

因此，如果你能理解我通常所指的“理性”及其下属概念“认识理性”和“工具理性”的含义，那么我们的沟通就算是成功了：我们已经围绕如

何定义“理性”进行了讨论，这正是我们需要探讨的全部内容。剩下来讨论不是去纠结于“ratio-na-li-ty”（理性）这几个音节应该承载什么具体含义，而是要讨论什么样的思考方式才算是好的。

如果你说：“对我来说相信X是（认识上的）理性的，但实际上事实是Y”，这样的话，你对“理性”这个词的使用可能与我所理解的含义不同。（比如，我认为“理性”应当在反思时保持一致——无论是“理性地”审视证据，还是“理性地”思考你如何处理这些证据，都不应该得出截然不同的结论。）

同样，如果你发现自己在说：“对我来说，从工具理性的角度，X是理性的选择，但我认为正确的做法是Y”，那么你几乎可以肯定，你对

“理性”或“正确”这两个词的理解与我不同。我使用“理性”一词时，旨在从规范性的角度挑选出理想的思维模式。

在这种情况下，或者在任何人们对词义存在分歧的情况下，你应该用更具体的语言来替代“理性”，比如说：“自私的做法是逃跑，但我希望你至少能尝试一下把孩子从铁轨上救下来”，或者说“按照因果决策理论的常规表述，面对纽康问题时，你应该选择两个盒子，但我更愿意选择100万美元。”这样做能更清楚地表达你的观点，避免了不必要的词义纷争。

（译者注：因果决策理论（Causal decision theory）是一种决策理论，它主张个体在做决策时应该基于其行为可能导致的后果来选择行

动。换句话说，它强调考虑因果关系，即一个行动如何会影响结果，来做出最有利的选择。

纽康问题（[Newcomb's Problem](<https://www.readthesequences.com/Newcombs-Problem-And-Regret-Of-Rationality>)) 是一个著名的思想实验，用于探讨预测、自由意志和因果关系在决策中的作用。问题的基本设定是这样的：有两个盒子，A和B。盒子A是透明的，里面有1000美元；盒子B是不透明的，里面可能有1百万美元，也可能什么都没有。一个超能力者已经预测了你会如何选择。如果预测你只选择盒子B，那么盒子B里就有1百万美元；如果预测你选择两个盒子，盒子B就为空。现在你需要做出选择：只选择盒子B，或者两个盒子都选择。）

实际上，我建议你重读这篇文章，把所有“理性”一词替换成“foozal”（译者注：这个“foozal”是作者编造的词），看看这样是否改变了文章想要传达的含义。如果真的有所变化，那么我的建议就是：不要追求理性，而应该追求foozality（foozal性）。

“理性”这个词在很多场合下都能准确表达我的想法，特别是那些不那么极端的情形，而“非理性”也是如此。因此，在这些情况下，我不会对使用这些词感到忌讳。

然而，我们应该小心不要过度使用这个词。光是大声念叨，并不能为我们带来任何好处。如果你过于频繁地谈论“道”，你可能永远也达不到它。

PART2：感受理性

Eliezer Yudkowsky

在现代社会里，人们普遍把“理性”看作是和感性背道而驰的东西。人们通常认为，无论是忧伤还是欢乐，只要是情感，就和逻辑相悖。然而，我翻遍了概率论的定理，也没发现哪条规则能证明理性的表现就是无动于衷或面无表情。

理性和感性真的是水火不容的吗？显然并非如此。我们的情感，实际上是根据我们对现实的理解而产生的。

比如，如果我相信我的已故弟弟奇迹般地复活了，我自然会感到无比的喜悦；但当我醒来，

意识到这只是一场梦，悲伤便接踵而至。我在梦中的幸福因残酷的现实而化为泡影。清醒时的悲伤，却是基于理性的；没有任何真理能够否定这份悲伤。正如P. C. Hodgell所言：“能被真相摧毁的，就应该被摧毁。”

理性探究的第一步，就是去追问“世界的真相是什么”（how-the-world-is），然后这种探询会像病毒一样扩散到所有基于这一真相的其他思考中。你对“世界的真相”的看法，可以关系到任何你认为在现实中存在的事物，无论其存在与否，只要它有能力影响其他事物。

举个例子，如果你坚信衣柜里藏着一个精灵，它的乐趣就是把你的鞋带搞乱，这其实也反映了你对“世界真相”的一种看法。你的鞋子是摸得着看得见的，如果真有什么东西能左右

你的鞋带，那么它也必定是实实在在的，它是构成我们所谈论的这个宇宙广阔因果关系网的一部分。

对于小精灵搞乱你鞋带的行为而感到愤怒，并不仅仅是因为对“世界真相”的看法。比如说，你可能是个修行深厚的佛教徒、接受过脑部手术的病人，或者天性就是个心态极好的人，在你发现鞋带被搞乱时，可能根本就不会有任何愤怒的情绪。

在这种情况下，你对世界的预期并没有发生变化——你仍然会期待着打开衣柜时，看到自己的鞋带被完整的系在一起。无论你是感到愤怒还是保持平和，这都不应该影响你的判断，因为衣柜里发生的事实并不会随着你的情绪而改

变；虽然，保持这样冷静的思维确实需要一番努力。

但愤怒这种情绪其实和你对“世界真相”的认识紧密相关，你之所以会愤怒，是因为你相信是那个精灵搞乱了你的鞋带。这种对理性的看法就像病毒一样传播开来，从你最初是否相信是小精灵搞乱了你鞋带的这个问题，扩散到因此而产生的愤怒情绪。

提高理性水平——也就是对“世界真相”的理解更加精确——可能会削弱我们的情感，也可能让情感更加强烈。有时候，人们会通过否认现实，逃避那些能引发强烈情感反应的观点，从而避免面对激烈的情绪波动。如果真是这样，那么当你开始学习如何更加理性地思考，训练

自己不去回避和否认事实时，你的感情会变得更加强烈。

在我年轻时，我总是纠结于是否应该强烈地感受——这是否被允许，是否得体？我认为这种困惑并不完全是因为我对理性有着错误的理解。即使是那些并不追求成为理性主义者的人，也有类似的困扰：他们在快乐时会怀疑自己是否真的被允许快乐；在悲伤时，他们也总是不确定是否应该逃避这种情感。

自从苏格拉底以来，至少是从那时起，显得有教养和世故的方式一直是不让人看见你对任何事情的强烈关心。在一个注重礼节的社会里，公开表达情感往往被看作是不得体的。你应该看看那些因为我对理性的深刻关注而投向我的异样目光。我认为，这并不是因为这个话题特

别罕见，而是因为他们不习惯看到一个成熟而理智的成年人对任何事情都展现出深切的关心。

然而，我现在明白了，强烈地感受并没有错。自从我接受了“能被真相摧毁的，就应该被摧毁。”的原则，我也领悟到“被真理滋养的，应当茁壮成长”。

当美好的事情发生时，我会充满喜悦，而不会再对我感到快乐是否合理感到迷惑。当感到悲伤时，我不会用虚假的安慰或是空洞的希望来逃避悲伤。

我思考人类的过去和未来，想到我们历史上数以十亿计的死亡，那些痛苦与恐惧，那些寻找答案的探索，以及那些从血泊中伸出的颤抖之

手。我幻想着有一天我们能将星辰变为城市，想象那时我们能成为什么样的存在，所有的黑暗与光明——我知道我永远也无法完全理解它，也无法用言语来完整表达。

尽管我有许多哲学思考，但我仍然觉得表达强烈情感很尴尬，你听到这些可能也会感到不自在。但我现在明白，情感其实是理性的一部分。

PART3：为什么要追求真理？并 且...

Eliezer Yudkowsky

在讨论“如何克服偏见”时，我们触及到一个根本性的议题：为何我们要追求真理（幸运的是，很少有人去质疑[真理的本质](<https://www.readthesequences.com/The-Simple-Truth>))。我们将思维定义为追求理性和激励自己，而激励我们前行的动力，正是判断事物对错的基石——即我们追求真理的初衷。

书中有这样一句话：“好奇心是第一美德。”好奇心或许不是追求真理的唯一理由，但它却拥有一种特别的、值得尊敬的纯净感。如果是好

奇心驱动你，那么你就会按照每个问题对你个人的吸引程度来判定其优先级。

那些挑战性更大、失败几率更高的问题，正因为它们更能激起你的兴趣，就显得更值得去探索和投入，哪怕它们比那些简单得多的问题要困难得多。

正如之前提到的，人们习惯将理性和情感看做是彼此的对立面。作为一种情感，好奇心经常会被一些人质疑，认为它不是理性的一部分。就我个人而言，我认为，如果一个情绪建立在错误的信念之上，或者更准确地说，建立在导致错误的认知行为之上，那么这种情绪就是“非理性”的：“假如一个熨斗靠近你的脸，你以为它是热的，但实际上它是冷的，那么你的恐惧就是不理性的。

反之，如果熨斗靠近你的脸，你以为它是冷的，但实际上它是热的，那么你的镇定就是不理性的。“因此，那些基于正确信念或合理的认知过程产生的情绪，可以被视为“理性情绪”。这种观点的优势在于，它允许我们把镇定看作一种情感状态，而不只是某种默认的的状态。

当人们探讨“情感”与“理性”的关系时，他们往往在考虑两种思维模式：系统1，即基于直觉做出的快速判断，以及系统2，即经过深度思考后做出的慢速判断。

值得注意的是，经过深度思考后做出的慢速判断并非总是对的，基于直觉上做出的快速判断也不总是错的。因此，弄清楚这种二元对立与“理性”的区别至关重要。这两种思维模式都能

促进我们追求真理，也都可能阻碍我们，关键在于我们如何应用。

追求真理除了出于单纯的好奇心之外，还有什么其他理由吗？例如，如果你有一个明确的目标——比方说造一架飞机——那么你就需要了解空气动力学的相关知识。

举个更生活化的例子，如果你想喝巧克力牛奶，你就需要知道附近的超市是否有售，这样你才能决定是走路去买还是去别的地方买。如果追求真相是为了这样的原因，你给予问题的优先级会体现出你对信息实用价值的预期——即，不同答案可能对你的决策产生多大的影响，这个决策对你而言有多么重要，以及你多么希望找到答案。这些考量会直接影响你原本默认的选择偏好。

仅仅因为真理的实用性而追求真理，可能看起来不够纯粹——我们难道不应该仅仅因为真理本身的价值而渴望它吗？然而，基于实用目的的探索是极其重要的，它为我们提供了一种外部验证的方式：如果你建造的飞机坠落，或者你发现商店里没有巧克力牛奶，这都直接指出了你的错误。这种经验给你提供了反馈，让你了解哪些思考方式是行得通的，哪些是行不通的。

纯粹的好奇心固然珍贵，但当初次的神秘感消散后，它可能就不会那么专注于验证答案了。好奇心这种情感，早在古希腊时代之前就已经存在。但推动人类稳步走向科学探索的，是我们发现某些思考方式能够揭示世界运行的真理。至于纯粹的好奇心，古老的神话和英雄故

事同样能够满足这种渴望，而过去的人们也对此表示满意。

除了出于好奇和实际应用，追寻真理的动力确实还有其他来源。我想到的第三个理由是道德上的考量：人们相信寻求真理本身就是一种高尚、重要且值得追求的行为。这种观念赋予了真理一种内在的价值，但它与出于好奇心的态度完全不同。好奇心让我们对未知充满兴趣，而道德信念则让我们认为深入探寻是一项义不容辞的责任。

在这种道德观念的指导下，我们可能会期望他人也应参与探究真相，如果有人故意避开真相，我们可能会对其持批评态度。因此，我将这种追求真理的态度定义为“道德动机”，它认为探究真理对社会具有深远的重要性，是每个

人都应承担的责任。这一动机不仅影响我们认为哪些真理最为重要——不仅是哪些最实用或最有趣，还包括在何种情况下我们寻求真理的责任感最强烈。

我之所以不同意将道德视为追求理性的动机，不是因为我否认道德本身的理念，而是因为这种做法可能带来一系列问题。

把理性视为一种默认习得的道德责任，容易让人们采取一些不恰当的思维方式，这就像在舞蹈中默认会踩到别人脚一样。以《星际迷航》中的斯波克先生（Mr. Spock）为例，他代表了一个理想化的理性形象。斯波克总是表现得非常冷静，即使在这种冷静完全不合时宜的情况下。他还会提出一些极不准确、未经校准的概率预测（比如斯波克会说：“舰长，如果我

们直接驾驶企业号飞进那个黑洞，我们的生存概率仅为2.234%。”

但实际上，大多数情况下企业号都安然无恙。谁会在这种情况下提供一个精确到小数点后三位的概率呢？这个预测概率和实际情况的差距不是一个数量级，而是两个数量级）。这种对理性的刻板印象可能让很多人对于完全接受理性持有保留意见。将理性变成一项道德义务，等同于给它带来了无限的自由度，就像任何一种部落习惯那样。

这种情况下，人们可能在做出错误决定后，不仅不会进行反思，反而会愤怒地坚称自己的行为是正确的。这样，把道德当作追求理性的动机，不仅没能促进真正的理性思考，反而可能妨碍了对错误的认识和改正。

为了“提高”我们的理性技巧，超越古老狩猎采集社会的思维模式，我们必须深入学习如何正确思考。我们给自己编写的新心理程序，最初是从系统2出发的，即负责深度思考的部分，并且只能逐渐地训练到系统1中，那是掌管我们快速反应的底层神经网络。为了不让像偏见这样的问题影响我们，我们会在系统2中培养出一种明确的规则或原则，即坚决不采用那些可能导致错误判断的思考方式。

如果我们想要追求真理，就必须采用一些特别的思维方式，这些方式就是我们所说的理性技巧。这意味着我们要学会如何跨过那些可能阻碍我们看清真相的障碍，比如克服我们的偏见和先入为主的观念...

PART4:什么是偏见?

Eliezer Yudkowsky

偏见，是我们在追求真理过程中经常遇到的障碍之一。它之所以成为障碍，正是因为我们的终极追求是真理。然而，并非所有障碍都能被称作偏见。

如果我们一上来就问“什么是偏见？”，实际上这种提问方式本身就颠倒了问题的顺序。就像俗语说的那样，“疯狂有四十种形式，但理智只有一种。”真理是一条狭长的小径，可命中的目标区域极小。

可能“她爱不爱我”听起来像个非此即彼的问题，但实际上，公式 $E = mc^2$ 在所有方程式

的海洋里不过是一粒微尘，就像一张彩票在无数张彩票中幸运中奖一样。失败并非意外；相反，真正的成功因其难以预见而显得尤为珍贵，这本身就是一个值得探寻的谜题。

我们在一开始的时候，并不会因为偏见被看作是一件有害或不合时宜（Just Not Done）的坏事，就自然而然地承担着要“消除偏见”的道德责任。如果有人因为受到周围社会的影响而采取自以为理性的行动，他们可能会落入一个误区：认为自己应该机械地执行一些任务，却从未真正理解这些行为背后的深层原因。这种情况，正如我小时候读到的在[《别闹了，费曼先生》](<https://book.douban.com/subject/1037602/?from=rec>)一书中所描述的那样，不仅效果不佳，还带有误导性，而且考虑的并不周全。

（译者注："Just Not Done" 更多地被用来表达某种行为或举止在特定社会、文化环境中被认为是不恰当或不可接受的，通常指违反了未成文的社会规范或礼仪。这个表达方式强调了某些事情虽然没有明确的法律或规则禁止，但在社会习惯或礼节上是不被认可的。）

相反，无论出于什么原因，我们都想了解真相，却发现各种障碍阻碍我们获取真相。比如，它们可能源自于计算能力不足或信息获取成本高昂。有趣的是，许多障碍因为相似的特征而被归为一类，就像是在通往真相的障碍中形成了一个特定的聚集区域，我们称之为“偏见”。

“偏见是什么？”如果我们尝试从过往的经验中找出一系列具体的例子，来形成一个紧凑的定

义或测试标准，可能是徒劳的。有时候，用一些具体的例子来说明，让听众自己去理解，可能比试图给出一个详尽的定义更有效。

比如，对于一个刚开始研究火的科学家来说，直接指向篝火说“那边橙色明亮的东西，就是火”，可能比使用复杂的定义如“火是物质燃烧时的化学反应，释放出热量和光”更为直观。

同样，我们不应该因为无法精确定义某个概念就选择忽视它。我可能记不住广义相对论的具体方程式，但这并不妨碍我理解如果我走下悬崖，我会坠落这一自然法则。对于偏见也是如此，即便我们不能简洁地界定它是什么，这并不意味着它对我们的影响就减少了。因此，我们可能会指出合取谬误（conjunction fallacies）、过度自信（overconfidence）、可用性启发式（availability heuristics）和代表性

启发式（representativeness heuristics）、基本率忽略（base rate neglect）等概念，并简单地讲：“就是这些东西。”

（译者注：

1. 合取谬误（Conjunction fallacy）是一种逻辑错误，发生在人们错误地认为两个事件同时发生的概率比其中任何一个单独发生的概率都要高的情况。这种谬误通常在描述更具体、更详细的情况时出现，因为详细的情况听起来更具说服力，即使它们发生的概率实际上是更低的。

2. 可用性启发式（Availability heuristic）是一种心理学概念，指的是人们能够根据从记忆中

易于调取的信息来评估事件发生的概率或频率的倾向

3. 代表性启发式（Representativeness heuristic）是一种认知心理学概念，指的是人们在评估一个对象属于某个类别或在做出概率判断时，依赖于对象与某个类别的相似度或典型性。

4. 基本率忽视（Base rate neglect）是一个心理学和决策理论中的概念，指的是在判断或决策时，人们倾向于忽略或轻视整体频率（基本率）信息，而更多地依赖于特定情况或证据。这种倾向导致在评估概率时出现偏差，因为基本率是了解任何事件发生可能性的重要信息。）

话说回来，我们好像给所有阻碍我们认识真理的障碍贴上了“偏见”的标签，这些障碍并不是因为获取信息的成本或计算能力有限造成的。其实，很多时候，这些障碍源自我们自己心理机制的特定构造。

有时候，我们的思维方式就是为了反对准确认知而设计的，比如在政治辩论中为了胜出而采取的策略。或者，我们因为想要融入群体，就不自觉地去相信别人所相信的，这也是一种偏离真相的选择压力。有些情况下，我们的大脑就像是运行着某种算法，这可以帮助我们处理一些问题，但同时也会引发错误。

比如说，可用性启发式本身并没有什么不对，但它有时会让我们做出一些错误判断。这些失误往往都是容易被发现并且能够简单描述的。

当我们的大脑做出一样的失误判断时，并且在经过仔细的实验检验或者深度思考（系统2）之后被指出来，我们就把这种失误称为“偏见”。这种认知失败，并不是因为我们的大脑做得不够，而是因为我们的思维方式——即我们的“认知机器”的结构造成的。

“偏见”区别于那些源自我们所持信念或道德责任的错误。当我们意识到某个观点或行为与自己的信仰或道德标准不符时，我们称之为“错误”，而不是“偏见”。这类错误一旦被我们察觉，通常更容易被纠正。值得注意的是，尽管这些错误的深层原因可能仍旧是某种偏见。

相比之下，“偏见”并非由个人大脑损伤或特定文化习俗所导致的错误。它们根植于所有人类共有的心理机制中。

柏拉图被认为有偏见，并不是因为他不了解广义相对论（General Relativity）——在他的时代，这样的知识根本无法获得，所以他的无知并不能归咎于他的心理机制。

然而，如果柏拉图坚定地认为哲学家会成为更好的统治者，仅仅因为他自己是哲学家，这种信念如果是基于一种广泛存在的、为了个人利益而形成的政治本能，而不是因为他被教导说每个人都有道德义务把自己分内的工作做到最好，或者是由于他童年时期的某些误解，那么这种信念就构成了偏见，不管是否有人指出过这一点给柏拉图。

纠正这种偏见可能需要付出高昂的代价，有时甚至可能无法纠正。但是，当我们深入剖析这种心理机制，识别出某些错误背后存在的可追

溯的因果关系；当这些问题看起来更多是由于我们思维模式的进化而非仅仅因为信息缺失或错误信息所导致时；我们便将其定义为偏见。

个人而言，我认为我们的目标应该是培养个人的理性技能，以及提高我们探索真理的能力。我们所面临的挑战，在于要积极地实现对真理的追求，而不仅仅是避免失败。

错误的可能性是广泛的，形式多样，而这些错误构成了一个难以完全描述的广阔空间。正如同“对于某个苹果而言是这样，对于另一个苹果可能就不是这样”所示，对一个具体事物的理解可能远比试图概括所有事物的理解来得深刻。成功的路径相对狭窄，因此我们能够更具体、更精确地讨论它。

尽管我不反对讨论定义——正如你所见，但我们必须记住，这不是我们的主要目标。我们之所以聚在这里，是因为人类对于真理有一种伟大的追求，这源于我们对知识的渴望和好奇心。因此，让我们共同努力，克服前进道路上的所有障碍，不论我们是否将其称作“偏见”。

PART5：可用性启发式

Eliezer Yudkowsky

“可用性启发式”是一种心理判断策略，人们依据能够轻易想到的信息实例来评估事件发生的频率或可能性。

1978年，利希滕斯坦（Lichtenstein）、斯洛维奇（Slovic）、菲施霍夫（Fischhoff）、莱曼（Layman）和库姆斯（Combs）联手进行了一项颇具影响力的研究，名为“致命事件的判断频率”，旨在探究人们在评估风险严重性或比较两种危险哪个更常发生时所犯的常见错误。

他们的研究揭示了一个有趣现象：大多数人对于事故和疾病造成死亡的频率有着严重的误判，而且往往认为他杀比自杀更为常见。然而，实际情况却大相径庭：疾病导致的死亡人数是事故死亡人数的十六倍之多，自杀的比率也是他杀的两倍。

导致这些扭曲信念的显著原因可能是，人们更倾向于讨论谋杀而不是自杀，这使得谋杀的信息更加易于被记住。同样，事故之所以给人深刻印象，很可能是因为它们的戏剧性质，这也让人们更容易回想起来。

1979年，库姆斯和斯洛维奇的后续研究揭示了一个有趣的现象：人们对概率的偏见与两家报纸报道的偏向性紧密相关，这表明新闻报道的内容与人们记忆中的偏差之间存在密切的联

系。然而，这项研究并没有明确解释是因为报纸更频繁报道谋杀案而让人们更容易记住，还是因为谋杀案本身更加引人注目从而被报纸更多报道。但无论是哪种情况，可用性启发式偏见都在其中发挥了作用。

在过去，人们获取信息的渠道大多是直接经验或是从身边人那里听来的，这时信息的选择性相对较小。但是在今天的互联网时代，信息在被我们获取到前会经过多层筛选，这个过程极大地增加了选择性报告的情况。与我们的祖先相比，我们现在生活在一个存在大量信息和实践频繁发生的世界里。但我们实际能够接触到的信息却相对有限。这种筛选效应的加剧，很可能会导致我们对信息的可用性偏见更加严重。

在日常生活中，我们遇到像比尔·盖茨这样的成功人士的机会几乎微乎其微。然而，受到媒体选择性报道的影响，我们很容易不自觉地把自己的成就与他的成功相比较，进而感受到一种由于对比产生的不满。虽然遇见比尔·盖茨的实际存在率极低，仅为0.00000000015，但我们接收到关于他的信息远比这个数字要多。另一方面，全球有19%的人每天生活费不足1美元，但我怀疑我们阅读的博客中有五分之一是由他们撰写的。

依赖于可用性的信息处理方式会导致一种[荒谬的偏见](https://www.greaterwrong.com/lw/j4/absurdity_heuristic_absurdity_bias/): 我们不会回忆起从未发生过的事件，因而认为这类事件的发生概率为零。

例如，当近期没有发生洪水时，人们就会拒绝购买洪水保险，即便该保险有很多补贴，价格也远低于其应有的价值。如昆鲁瑟等人

（Kunreuther et al.）所述，这种对洪水威胁反应不足可能是因为人们无法想象从未发生过的洪水，他们被自己的经验所困——最近发生的洪水似乎为损失的上限设了一个界限，导致他们忽略了那些需要重视以减少未来潜在损失的问题。

伯顿等人（Burton et al.）的研究报告指出，建造水坝和堤坝虽然能降低洪水发生的频率，但这反而会导致人们产生一种错误的安全感，从而减少了采取预防措施意愿。虽然水坝的建设减少了洪水发生的次数，但一旦洪水发生，其造成的损失却远比没有水坝时要大，导致的年平均损失反而增加了。这反映了一个现

象：智者会从小规模的危险中推断出大规模危险的可能性，但过去小规模灾害的经历却往往成为了人们感知风险上限的依据。

这意味着，一个社会如果在预防小灾害方面做得很好，就可能不会对重大风险采取相应的行动。一旦常规的小洪水被消除，人们就可能在洪泛平原上进行建设，误以为风险已经得到了控制。这种情况下，社会将常见的小灾害视为风险的最大界限，专注于防范这些小规模的常发灾害，而忽略了罕见但潜在的大规模灾害。

这表明，依靠记忆来指导我们对过去事件概率的判断，甚至预测未来，往往是不可靠的。

PART6：繁琐的细节

Eliezer Yudkowsky

这些只不过是给原先那些乏味又不够打动人心的说法，增添了艺术的真实感.....

所谓的合取谬误，就是人们往往会高估两个事件同时发生的可能性，即便根据逻辑定律，这种同时发生的概率 $P(A,B)$ 应该是低于单独某个事件发生的概率 $P(B)$ 的。拿1981年的一个实验来说，有68%的参与者更相信“里根会给单亲妈妈提供联邦资助的同时，还会削减对地方政府的联邦支持”，而不是简单地相信“里根会给单亲妈妈提供联邦资助”。

[经过一系列精心设计的实验，我们排除了其他可能的解释](https://www.greaterwrong.com/lw/jj/conjunction_controversy_or_how_they_nail_it_down/)，最终确认了一个关键观点：合取谬误之所以发生，是因为人们倾向于用“代表性”来判断事物，而非基于实际的概率。换句话说，当我们往故事里添加更多细节时，这个故事就显得更加合情合理，仿佛更加符合其发生的真实过程。

例如，当我们说里根既会支持单亲妈妈，又会削减对地方政府的支持时，这种陈述听起来更具有说服力。这就是说，一个观点的不足可以通过另一个观点的合理性来补充，二者在我们的心中达到了一种“平衡”。

因此，多添加一些细节可以让故事听起来更加真实，即使可能会让整个事件听起来发生的可能性更小。

如果真有那么一天，我们可能会目睹一些听上去头头是道却又离奇莫测的未来学说的诞生，或者是人们对一些未曾验证、却被吹得天花乱坠的观点轻易地信以为真。当这些直观的对比出现时，如果能发现并有意识地纠正那些逻辑上的破绽，的确能够解决一些问题。但这样的处理方式只是权宜之计，治标不治本。

在[1982年的一项研究](https://www.greaterwrong.com/lw/ji/conjunction_fallacy/)中，预测专家们在不知情的情况下，系统性地给出了“俄罗斯入侵波兰”后美苏中断外交关系的可能性高于“美苏中

断外交关系”这一单独时间的概率判断，尽管这两个情况是分别被不同实验组提出的。

当预测者们没有意识到他们正在评估合取谬误时，他们作为一个群体该如何采取措施避免这种谬误？如果他们不知道这项实验的目的是探讨合取谬误，他们又该如何提高自己在做出概率判断时的准确性？

仅仅针对特定情况提出解决方案，并不能根治整个问题。这种陷阱只是表面的症状，而非问题的本质。

如果预测者们没有办法直接进行比较，甚至不知道有人会检验他们是否犯下了合取谬误，他们能采取的一个策略就是格外留意“和”这个

词。他们需要对此保持高度警觉——不仅仅是小心，而是要完全避免它。

这意味着，即便他们不知道研究人员后续可能专门测试他们是否陷入合取谬误，他们也应该能够敏感地意识到，当两个事件被连结在一起提出时，这种预测的复杂度急剧上升，对于任何试图让他们支持这种复杂预测的请求，都应该表示惊讶。根据实验的具体情况，他们需要对这种预测的可能性进行大幅度的降低——至少要降低四倍。

这种思考方式同样适用于分析“美国和苏联中断外交关系”的潜在原因。这不仅仅是一个“无缘无故的断交”场景，而是要求预测者深挖背后可能的导火索。换言之，他们需要从简单的

“美苏无缘无故断交”推进到探索背后可能的详细原因。

当评估“里根政府将向未婚母亲提供联邦支持，同时削减对地方政府的联邦支持”这一情境时，参与者同样应对“和”这个词表现出惊讶。进一步来说，他们应该将两个假设的荒谬性进行累加，而非平均处理——这里的荒谬性被视作对数概率。

这意味着他们的思考过程应该是：“里根可能会也可能不会削减对地方政府的支持（荒谬性评为1位），但他支持未婚母亲的可能性极低（荒谬性评为4位）。所以，总体荒谬性评为5位。”或者是，“里根不可能支持未婚母亲，这个主张立即被否定。第二个提议只会让整体情况更加复杂。”

设想这样一个场景：我们面前有一个六面骰子，其中四面涂成了绿色，两面则是红色。如果有人要在这个骰子连续掷出二十次的结果中，选择一个序列进行投注，选项有：(1) rgrrrr，(2) grgrrrr，或者(3) grrrrrr。有趣的是，多达60%到70%的人会选择下注在序列2上，尽管从逻辑上讲，选择序列1对他们更有利，因为任何包含序列2的情况，序列1也一定会发生。

那么，人们如何能做出更好的选择呢？单纯地提醒他们注意某些特点似乎只是临时解决方案，并不能触及问题的核心。或许，通过精确计算概率似乎能够根治这个问题，但现实是我们不可能总是这么做。

许多参与者之所以做出错误的选择，是因为他们错误地认为：“哇，第二个序列中绿色和红色的比例更高，我应该选择它！”但是，如果他们能够从另一个角度思考，意识到“哇，第一个序列更短，因此出现的概率更高！”那么他们就能够获得启发式的胜利。

他们必须从奥卡姆剃刀原则中深刻体会到一种情感上的重压——认识到每增加一个细节，就像是给自己的肩膀再添一份负担，好似每一次增加都是对命运的一次赌注。

（译者注：奥卡姆剃刀（Occam's Razor）是一个解决问题的原则，它认为在所有可能的解释中，最简单的解释往往是正确的。）

记得有一次，我跟一个深陷于他那套充满幻想的未来学说中的人交流。这家伙对于在自己的理论里添油加醋，堆砌一些听起来很炫酷但复杂得让人摸不着头脑的细节特别有兴趣。

我尝试向他解释，那些听起来震撼人心、让人难以置信的理论为什么没能打动我。我跟他讲了什么是逻辑谬误，并特意拿出一个实验来说明我的观点：“中断关系±入侵波兰”。我对他说：“你想啊，宇宙出于某种原因自我复制的可能性，和通过黑洞实现复制相比，哪个的可能性更大？如果说高级文明能够创造黑洞，那也只不过是因为宇宙的自然进化让这一切变为可能。”听我这么一说，他似乎恍然大悟，只是轻轻地“哦”了一声。

（译者注：将“暂停国际关系”与“入侵波兰”这两个完全不相干的概念相加减——作者是在用夸张的手法，来讨论复杂理论或观点在现实中的不可行性或逻辑上的漏洞。）

直到那一刻，他才意识到，那些看似给故事增色的细节，其实变成了一种额外的负担。但实际上，这些细节应该是增强论证的证据，让整个叙述显得更加真实可信。比如，有人跟你提了一堆奇奇怪怪的想法，其中之一就是关于宇宙复制的。然后，他们还拿出了一些所谓的证据来支持“宇宙复制论”。但这些并不能成为[包装交易谬误](<https://en.wikipedia.org/wiki/%20Package-deal%20fallacy>)，尽管它们是在讲述一个故事的过程中被提出的。

（译者注：包装交易谬误（Package-deal fallacy）是一种逻辑谬误，它发生在将一组本质上不相关或者仅部分相关的事物捆绑在一起，当作是不可分割的整体来接受或拒绝时。这种谬误忽略了个别成分之间的区别和独立性，导致了错误的推理和结论。）

你需要仔细审视这些细节。对于每一个细节，你都必须独立评估，然后问自己：“我们怎样才能确定这个细节是准确的？”例如，某人描述了一个场景，人类因纳米技术而陷入战争，中国拒绝遵守国际控制协议，从而触发了一场军备竞赛……但等一下，你怎么能确定是中国呢？这是你口袋里的水晶球告诉你的，还是你喜欢扮演未来预言家的角色？所有这些细节，包括那些特定的细节，它们的来源是什么？

因为正如古语所言：

如果你有能力减轻自己的负担，那么你就应该去做。

每一根稻草都具备压垮你的能力。

PART7：规划谬误

Eliezer Yudkowsky

丹佛国际机场的开放延期了16个月，成本超出预算20亿美元（还有一种说法是31亿美元）。欧洲战斗机台风，一个由几个欧洲国家共同参与的防御项目，延迟交付了54个月，成本从预计的70亿美元飙升至190亿美元。而悉尼歌剧院，可能是历史上最著名的预算超支建筑之一，原计划1963年完工，预算700万美元，实际上直到1973年才竣工，最终成本达到了1.02亿美元。

这些看似孤立的灾难，是不是因为选择可用性
(selective [availability])(<https://>

www.readthesequences.com/Availability))

——我们只关注其中一部分而格外引人注目呢？它们确实可能是官僚主义或政府激励机制失败的体现。然而，这些情况同时也映射了一种在个人规划实验中被广泛复现的认知偏见。

（译者注：选择可用性（Selective Availability），通常在特定领域或上下文中使用。在涉及网络选择可用性、选择性可靠度或选择性效益的讨论中，该术语可能会被提及。）

布勒等人对学生进行了一项研究，询问他们预计何时能完成自己的学术项目。具体来说，研究者让学生估计他们认为完成个人项目的概率分别达到50%、75%和99%时所需的时间。你

能猜到有多少学生实际上达到或低于他们对于50%、75%和99%完成概率的时间预测吗？

- 当被问及在50%的概率水平下，只有13%的学生按时完成了他们的项目；

- 在75%的概率水平下，完成项目的学生比例为19%；

- 而在99%的概率水平下，只有45%的学生（不到一半！）完成了项目。

布勒等人（Buehler et al.）指出，99%的完成概率尤其值得注意。即便是在被要求进行高度保守的预测时，学生们对于时间估计的自信远远超出了他们的实际完成情况。学生在时间管理和任务预测上的信心远远超出了他们的实际成就。

计划谬误（planning fallacy）揭示了人们在进行计划时的一个有趣现象：他们往往认为自己能够制定出完美的计划，这本身就带有一丝幽默。

纽比·克拉克等人（Newby-Clark et al.）的研究为这种现象提供了进一步的证据。他们对受试者进行了两种不同的询问：

- 他们基于现实的“最佳猜测”场景，以及
- 关于他们所希望的“最佳情况”情景的预测

结果表明，这两种预测产生了难以区分的结果。

当人们基于进行现实的“最佳猜测”时，他们往往梦想着一切都能够井井有条地进行，不会有

任何“意外”的耽搁或“不可预料”的灾祸发生——这其实跟他们所期望的“理想情况”没什么两样。

然而，事实常常比我们所想象的“最糟糕情况”还要糟糕。

不同于大多数的认知误区，我们其实掌握了一些很好的方法来克服这种所谓的“计划谬误”。虽然这些方法可能无法解决丹佛国际机场那样级别的混乱局面，但它们对许多个人计划乃至一些小型组织的项目都颇为有效。这其中的关键就在于采取“外部视角”（outside view）而非“内部视角”（inside view）。

人们习惯于从手头工作的特殊性和独特性出发来做出预测，并围绕如何完成这项工作构建一番蓝图——这正是我们所说的“计划”。

当我们打算完成某项任务时，往往会细致规划：确定在哪里、何时、如何行动；评估需要投入的时间和资源；设想从起步到成功完成的每一步。这些都属于“内部视角”，它们没有将意外的延误和无法预见的困难考虑在内。正如之前提到的，即便是让人们设想“最坏的情形”，也往往不足以抵消他们的乐观倾向——他们没有想到足够多的潜在问题。

当我们从“外部视角”出发时，我们有意忽略了项目的特殊和独特之处，只是简单地探询类似项目在过去通常需要多长时间来完成。这种方式可能初听起来有些违反直觉——毕竟，从

“内部视角”看问题能提供更多的细节信息，让人自然而然地认为，一个基于尽可能多的数据做出的、量身定做的预测应当更为准确。然而，实际上，采用外部视角能够为我们提供一个更客观、更全面的规划预期。

但研究显示，当受试者的想象变得越发详细时，他们的乐观程度也随之上升（准确度却下降）。布勒等人进行的实验要求一组参与者具体描述他们的圣诞节购物计划——包括地点、时间和方式。结果显示，这组人平均预计会在圣诞节前一周左右完成购物。而另一组仅被问及他们预计何时能完成圣诞购物的参与者，他们的平均回答是在圣诞节前四天。实际上，两组参与者都平均在圣诞节前三天完成了购物。

同样，在一项跨文化研究中，布勒等人发现，日本学生希望能在截止日期前十天完成论文，但实际上他们大多在截止日期前一天才完成。当询问他们以往完成类似任务的时间时，他们的回答是：“截止日期的前一天。”这些发现凸显了外部视角相对于内部视角的重要性和影响力。

类似的发现还表明，那些对项目细节不那么了解但拥有相关经验的局外人，往往不如项目的实际规划者和执行者那么乐观，但他们给出的预测却更为准确。

因此，如果你手头的任务“大体上”与你之前遇到的项目相似，那么有一个相当可靠的方法可以帮你避免计划上的误判。你所需要做的，就是询问以往类似项目大约花费了多少时间，而

不去考虑你当前项目的“任何”特别之处。更好的做法是，向有经验的局外人询问类似项目需要多长时间。

这样，你可能会得到一个让人觉得有些吓人的答案，同时这个答案也会让你清楚地认识到，你之前为何会低估了特定任务所需的时间。这个答案往往是正确的，所以接受它，然后据此进行计划。

PART8：透明度错觉：为什么 没有人理解你？

Eliezer Yudkowsky

在[《事后偏见》](https://www.greaterwrong.com/lw/il/hindsight_bias/)中，人们总误以为事件的发展和结果本应该是可以预见的。一旦我们知晓了结果，就倾向于用这个结果去重新解读事情的发展过程。甚至在别人提醒我们时，我们也很难摆脱这种已有的解读，去理解那些并不知情的人的感受。更不用说与他们产生情感上的共鸣了。

与此紧密相连的还有“透明度错觉”。我们对自己所说的话了如指掌，自然而然就期待其他人也能理解。在我们自己的文字面前，由于我们对其真实意图了如指掌，似乎理解起来轻而易举。然而，对于那些只能依赖于文字本身，而无法获得更多背景信息的读者来说，往往难以对我们说的话展现出足够的理解和同情。

琼（June）向马克（Mark）推荐了一家餐馆。马克去那儿用餐后，可能会有两种截然不同的体验：一种是觉得食物一般，服务也不怎么样；另一种则是食物非常美味，服务也极为出色。不管体验如何，马克都在琼的答录机上留言说：“琼，我刚在你推荐的餐厅吃过晚饭，我得说，实在是太棒了，简直太棒了。”

在基萨尔（Keysar）的研究中，当向一部分受试者展示马克的用餐体验较差的情境时，59%

的人认为他的留言是在讽刺，相信琼能够感受到这种讽刺。而在另一部分被告知马克体验很好的受试者中，只有3%的人认为琼会觉得马克的留言含有讽刺意味。这项研究似乎还特意让受试者听了一段实际的语音信息。基萨尔进一步指出，如果受试者知道餐厅很糟糕，但马克试图掩饰自己的真实感受，他们倾向于认为琼不会从留言中读出讽刺意味。这揭示了人们对信息的解读极大受到他们所了解的背景信息和预期的影响。

他们认为，无论马克是在隐藏自己的糟糕经历，还是坦诚分享自己的好时光，琼都能敏锐地感知到其中的真实意图，甚至包括那些讽刺的成分。因此，大家都觉得马克传递的信息是透明无误的。就像他们假设的那样，琼总能准确捕捉到马克希望她了解的所有意图。

基萨尔和布莱（Bly）对一批参与者说，“鹅高高挂起”，现在已经没有人这么说了，这句古老的俗语的意思是前途光明；而另一批参与者则被告知，它意味着前途暗淡。之后，他们询问参与者，对这个习语一无所知的人更可能认为这句话表达的是哪种含义。每组人都认为，他们被告知的那个意义会被外人视为这个习语的“标准”含义。

（测试的其他习语还包括“叔叔来访”、“走钢丝”和“身处薰衣草之中”。英语啊，真是个充满乐趣的语言。）

（译者注：

1. The goose hangs high: 直译是鹅高高挂起，通常表示情况良好，预示着好事将要发生，或者形势乐观。

2. Come the uncle over someone: 被用来描述某人以权威或者高人一等的姿态对待他人，类似于“摆出长辈的架子”。

3. To go by the board : 这个习语来源于航海，原意是指船上的物品或者部件落入海中、被丢弃。引申为比喻意义时，它意味着放弃、丢弃或者失去。比如说，如果一个计划“走在木板上”，那么这个计划就被放弃了。

4. To lay out in lavender : 这个表达在过去用来指存放衣物时在其中放置薰衣草，以防虫蛀，保持衣物香气。比喻意义上，它可以用来

描述对某人或某事进行严厉的批评或处理，就像把它们“铺在薰衣草上”一样，可能含有让事情变得更好或更整洁的意味。)

基萨尔和亨利（Henly）的研究着眼于一个问题：当人们说话时，他们是倾向于低估还是高估听众理解他们所说的内容的程度，亦或是他们能够准确判断？

在这项研究中，参与者需要用一个模糊不清的句子来描述一幅图画，这幅图画清楚地展示了一个场景——一个男人在追赶一位骑自行车的女士。随后，研究者让这些参与者猜测，他们的话有多少被听众准确理解了。结果发现，有72%的参与者相信他们传达的信息被听众理解了，但实际上只有61%的情况下是被准确理解的。在听众没能理解的情况下，有46%的说话

者错误地认为他们被理解了；而在那些听众真正理解了的情况下，只有12%的说话者误以为他们的话未被理解。

在另一项研究中，只有56%的参与者认为，即使是偶然听到他们谈话的人也能理解他们的意思。这说明大多数人并不自动假设别人会理解他们的话，普遍认为不存在过度相信他人理解自己的倾向。

基萨尔和巴尔（Barr）提到的一个例子很能说明问题：在德国即将入侵波兰的前两天，英国首相张伯伦向希特勒发出了一封信，信中清楚地警告，如果德国侵略波兰，英国将会参战。但是，由于这封信用了非常礼貌的外交语言，希特勒却错误地将其解读为英国想要和解的迹象，结果导致了德国的军事行动。

这个例子告诉我们，当别人误解了我们认为非常明确的表达时，我们不应该急于责备他们。这可能是因为我们的表达实际上比我们自己意识到的更加含糊。

PART9：预期推理距离较短

Eliezer Yudkowsky

远古时期，智人的生活圈子主要由最多200人的狩猎采集者群体构成，这个圈子就是我们祖先所生存的环境，又称环境进化适应（EEA）。那时候，文字尚未诞生，人们通过口头和记忆的方式，把知识一代代传下去。

（译者注：环境进化适应（Evolutionary Adapted Environment），是一个进化心理学中的概念，用来描述我们的祖先适应并演化的环境。）

在这种环境下，知识和信息是没有秘密可言的，几乎所有的背景知识都被视为共有财富。

除了极个别必须严格保密的信息，大家分享的都是彼此知道的内容。

在我们祖先生存的环境里，人与人之间的联系紧密到几乎无法想象的程度，几乎从未超过一步的距离。

比如，当你发现了一个新的绿洲，你不需要向你的同伴解释什么是绿洲，为何喝水很重要，或者怎么走过去。绿洲的位置可能只有你知道，那是你的私人知识；但关于绿洲的基本概念和为何需要水，这些对大家来说都是显而易见的，属于共有的背景知识。

在那样的环境下，解释任何事物几乎都不需要从头开始讲起你的概念。最多的时候，你也就

只需引入一个新概念，而不是同时解释两个或更多。

在我们祖先生存的环境里，抽象的学科是不存在的。那时候，人们不会把大量的证据汇集起来，总结成优雅的理论，然后通过书籍来传播。因为那样的理论，其结论往往需要依靠与大家共有的知识背景，通过上百步的逻辑推理才能理解。换句话说，那个时候，人们获取和理解知识的方式，完全建立在直接的、具体的生活体验上，而非抽象的理论和书面文字。

在我们祖先生存的环境里，如果有人在缺乏明确证据的情况下断言某事，这个人往往会被认为是在撒谎或者根本就是个愚蠢的人。你肯定不会去想，“哦，或许这个人掌握了一些充足

的、我所不知的背景知识”，因为在那样的环境下，这种情况基本上是不可能发生的。

相反，如果你提出了一个显而易见的观点，而对方却看不到这一点，那么问题通常出在对方身上——要么他们实在是理解能力有限，要么他们故意在固执己见，试图惹恼你。

更进一步说，如果有人在没有任何明显支持的情况下说出某些话，并且还“期待”你去相信，当你表示怀疑时他们又表现得非常愤怒，那么这种人很可能就是在“发疯”。

结合[透明度错觉](<https://www.readthesequences.com/Illusion-Of-Transparency-Why-No-One-Understands-You>)和[自我锚定](<https://www.greaterwrong.com/lw/kf/selfanchoring/>)的概念，我们可以更深

入地理解为什么科学家在与非专业观众乃至其他学科的科学家人沟通时会遇到显著的障碍。通常，当沟通失败时，我们会发现解释者只后退了一步，而实际上需要后退两步甚至更多。或者，当听众需要更多步骤来理解某个概念时，他们却期待能够一步到位。在这样的情境下，双方都似乎误以为从大家已经认同的知识到掌握新知识之间的逻辑距离非常短。

例如，一位生物学家在向一位物理学家解释进化论的合理性时，可能会说这是“最简单的解释”。但是，并非每个人都深谙从牛顿到爱因斯坦的科学史，这段历史给“最简单的解释”这个表述赋予了深刻的含义——那是理论诞生之初的力量之言，铭刻于理论的基石之上。

对其他人而言，“这是最简单的解释！”这句话可能听起来不过是一个有趣但难以说服人的论点；它在理解办公室政治或修理汽车时，并不显得是一个强有力的工具。这位生物学家可能过于迷恋自己的观点，傲慢到无法接受其他听起来同样合理的解释。（如果这对我来说听起来合理，那么对我团队中任何理智的成员来说也应该是如此。）

从生物学家的视角看，他们可能会觉得一开始接受进化论可能会有些难度——但是，尽管生物学家已经说了它是最简单的解释后，还是会有人拒绝接受进化论。此时生物学家就会认为，那些非科学家简直就是不明事理，与他们讨论这个话题完全是浪费时间。

要提出一个有效的论点，首先需要基于听众“已经知道或接受”的知识，然后一步步地展开推理“路径”。如果你没有回溯到足够远的基础上去解释，那么你的解释无异于自言自语。

如果在任何时候，你提出的论点在没有明确的逻辑支持下就进行了断言，听众很可能会觉得你疯了。

当你对某个观点的重视程度明显超过听众当前认为合理的范围时，也容易造成沟通的困难。例如，你深信“最简单的解释”是支持进化论的强有力论据（这确实是事实），但对于那些从未深刻理解奥卡姆剃刀原则的人来说，这只不过是个颇有趣味的小观点而已。

同时，最好不要让观众感觉到你认为自己的理解与他们之间存在着巨大的推理鸿沟，或者你认为自己掌握了他们无法理解的特殊背景知识。当观众对进化心理学或任何专业论证一知半解时，由于认知偏差的存在，他们往往会低估了理解一个概念所需跨越的逻辑距离，这种情况下，他们更可能觉得你在居高临下，而非进行平等的交流。

如果你认为你可以用几句话就简要解释清楚“系统性低估推理距离”这个概念，那么我不得不向你传达一些可能让人感到失望的消息...

（译者注：系统性低估推理距离用于描述人们在推理过程中，由于各种原因（如认知偏见、信息不足或心理限制等），常常低估从已知信息推断出未知结论所需的逻辑步骤或难度。这

种系统性的低估可能导致推理的不准确或不完整，从而影响决策的质量和效果。)

PART10：看到自己缺陷的镜头

Eliezer Yudkowsky

阳光穿越万里空间，最后落在你的鞋带上，又反射回空中。当中有些光子钻进你的瞳孔，击中视网膜，激发起一阵阵神经冲动。这些冲动沿着神经路径，传送到大脑的视觉处理中心，在那里，收集到的光学信息经过处理，再构建成一个三维模型，这个模型告诉你：你的鞋带开了。这就是你相信鞋带真的解开了的原因。

这就揭示了“刻意理性”（deliberate rationality）的奥秘——整个过程其实并没有什么神秘，它完全可以被理解。你能够理解自己是如何看到鞋带的，能够思考是什么样的思

思维过程让你产生了与现实相符的认知，哪些思维过程又可能偏离了现实。

（译者注：刻意理性指的是一个人有意识地运用理性和逻辑来指导自己的思考和决策过程。这种方法强调通过批判性思维、逻辑分析和证据评估，有意识地提升和优化自己的思考质量，而不是仅仅依赖直觉或习惯性反应。）

老鼠虽然能看见东西，但它们无法理解“看见”的过程。而你，能理解视觉这一现象，因此你能做到老鼠做不到的事。你是不是觉得很神奇，所以不妨花点时间好好感受一下！

老鼠虽能看见世界，但它们不知道自己拥有能产生视觉的大脑结构，因此，面对视觉错觉时无法自我纠正。它们的世界充满了猫、洞穴、

奶酪和捕鼠器，但它们却无法意识到这一切都是大脑构建的现实。

简单来说，老鼠无法“拍摄”到自己视觉系统的“快照”。而我们人类则不同，我们能观察到一些诸如旋转的轮子之类的[奇异图像](<http://www.richrock.com/gifs/optical-illusion-wheels-circles-rotating.png>)，并认识到我们所看到的部分其实是我们视觉系统的反映。这意味着，你不必完全信任你的眼睛所看到的，但你得明白自己“有”眼睛，你需要在心理上区分地图与领土、感官与现实。

这种能力，虽看似微不足道，实际上在动物王国中极为罕见。简单说来，科学的核心理念就是通过反思性推理寻找一种更可靠的方式，让我们的思想内容更真实地反映外部世界。

这是老鼠所无法创造的。当我们思考“进行可复制的实验来证伪理论”的过程时，我们能理解它为何有效。科学不是站在现实生活和普通人理解之外的独立权威。科学不仅仅是实验室里的事情。科学本身，就是一个将大脑与现实相连的可理解过程。

当我们深入思考时，科学的意义便显现出来。但老鼠因为无法进行深度思考，所以它们没有科学。这一点，我们作为人类不应视为理所当然，它不仅仅是科学共同体的力量所在，更是我们每个人作为个体所拥有的潜在力量。

确实，理解我们思考的机制可能比理解蒸汽机要复杂一些，但这并不意味着两者是完全不同的任务。都是通过观察、实验和理论来理解和利用自然界的规律。

我曾在EFNet的#哲学频道提出一个问题：“你认为未来20年内会发生核战争吗？如果你认为不会，原因是什么？”一位回答者认为，在未来100年内都不会发生核战争，因为“目前所有能决定核战争的人都对此没有兴趣。”当我追问为何将时间延长至100年时，他回答说：“出于纯粹的希望。”这种对话展示了人类思维的复杂性和我们对未来的希望，同时也反映了通过对话和思考，我们如何构建和理解我们的世界观。

（译者注：EFNet 是一个早期的互联网中继聊天（IRC）网络，它在90年代初期到21世纪初期非常流行。IRC 是一种在线聊天的形式，允许用户通过各种频道（chatroom）进行实时文字交流。在这些频道中，用户可以围绕特定主题进行讨论和交流。）

当我们深入反思我们的思考过程时，我们可以理解为什么核战争的想法会让人感到不安，以及为什么某人的大脑会拒绝接受这种可能性。然而，如果你设想存在着十亿个宇宙——按照埃弗里特的多世界解释或泰格马克

（Tegmark）的宇宙分类——这种思考过程并不会系统地将乐观主义者与未发生核战争的那些宇宙版本联系起来。

虽然有人可能会认为，由于他们抱有希望，因此会更加努力工作，为全球经济注入活力，从而帮助避免国家之间的愤怒和绝望情绪，进一步防止了核战争的发生。但实际上，这种个人的希望与全球避免核战争的直接联系是微乎其微的。

在这里，我们需要引入贝叶斯定理来定量评估这种联系的可能性。即使乐观主义真的能以某种方式影响世界，其影响力也不可能大到足以将核战争的可能性降低20%，无论这种乐观态度如何影响个人的行为和信念。因为一个事件对正确判断的贡献非常有限，所以一个人的信念大体上还是会对他们对世界的认知产生混淆。

（译者注：

1. 埃弗雷特的多世界解释（Everett's Many-Worlds Interpretation，简称MWI）是量子力学的一种解释，由物理学家休·埃弗雷特在20世纪50年代提出。这一理论与传统的哥本哈根解释形成对比，提出了一个极为大胆的观点：量子测量过程中并不发生波函数的坍缩，相

反，宇宙在每次量子事件发生时都会分裂成多个版本，每个可能的测量结果都在这些分裂出的宇宙中实现。

2. 马克斯·泰格马克 (Max Tegmark) 提出的宇宙分类是对可能存在的不同类型宇宙的一种系统化描述。泰格马克将宇宙分为四个等级，每个等级基于不同的理论和假设，涵盖从我们当前可观测的宇宙到极端的多宇宙概念。

- Level 1: 相同的物理定律和常数的宇宙
- Level 2: 不同物理常数的宇宙
- Level 3: 多世界解释下的宇宙
- Level 4: 不同数学结构的宇宙

1. 贝叶斯定理 (Bayes's Theorem) 是概率论中的一个重要定理，用于描述在给定先验知识的情况下，如何通过新的证据来更新某个假设的概率，以托马斯·贝叶斯 (Thomas Bayes) 的名字命名。)

追求快乐的过程中，我们应该更多地向内探索，而非仅仅关注外部环境——这能够让我们更深入地了解自己，而非简单地作为对环境反应的证据。我并不是反对追求幸福，但这种追求应该[基于]([https://](https://www.readthesequences.com/Feeling-Rational)

www.readthesequences.com/Feeling-Rational)你对这个世界的真实理解，而不是通过扭曲你的认知来实现。

如果你能明白这一点——如果你能看到希望在很大程度上是如何影响你的基础思维的——如

果你能认识到你的思维实际上是一个有缺陷的“地图制作引擎”——那么你就能够进行反思性的修正。

我们的大脑，无论是老鼠的还是人类的，都像是观察现实的有缺陷的镜头。但人类的大脑与众不同之处在于，它能够认识到自身的缺陷——系统性的错误、偏见等，并对这些问题进行二阶校正。

这种能力，在实际应用中，使得这个“镜头”变得更加强大。虽然它依然不完美，但确实变得更加有力量。