

什么是一个好的问题

Tong Zhang

网上常有人发帖争论，未来究竟属于大模型，还是属于 world model。两年前，我也在知乎上问过一个问题：大模型可以提出一个好的科学问题吗？

最近和同学聊起人工好奇心。我们的实验结果大多是 negative 的：当 world model 与 environment 之间存在很大的 gap 时，Agent 确实会去 explore，但这种探索经常没有意义。它被新奇、误差和不确定性吸引，四处行动，收集了大量 observation，却没有因此更理解世界。于是问题不再只是 Agent 会不会探索，而是它能不能被一个好的问题驱动。

什么是一个好的问题？过去，一个好的问题可能首先是苏格拉底式的提问。苏格拉底并不急着告诉别人什么是正义、勇敢和善，而是顺着对方的回答继续追问，直到那个人发现，自己相信的几件事情无法同时成立。一个问题之所以好，不是因为它听起来深刻，而是因为它迫使我们放弃一个习以为常的 assumption，或者承认原来的解释并不完整。

但科学还要再往前一步。世界不会替我们解释自己，它只会在 observation 和 intervention 中留下差异。一个科学问题最终必须能够变成某种操作：如果我改变这里，之后会发生什么？如果我们对同一个现象有两个看起来都合理的解释，真正好的问题就应该找到一个尽可能小的 intervention，使两个解释给出不同的 prediction，再让 environment 作出裁决。

因此，好的问题不是最大的未知，也不是最令人惊讶的现象。它是能够最大程度地区分不同 world model 的最小实验。一个巨大的未知往往无从下手，一个很小的实验却可能让一整套解释失效。科学并不总是从更大的问题开始，它经常从一个足够准确的小问题开始。

我们常用 novelty、prediction error 或 entropy 奖励 Agent，但当模型与环境之间的 gap 很大时，surprise 到处都是。Agent 可以不停晃动摄像头，可以盯着随机噪声，也可以钻进一个混乱而不可预测的角落。它看起来非常好奇，prediction error 也一直很高，但它没有获得任何能够迁移的知识。

一个纯粹追逐 prediction error 的 Agent，喜欢的可能不是科学，而是电视机里的雪花。这就是 noisy TV problem：噪声永远新鲜，却永远不会使模型变得更好。真正的好奇心不应该寻找模型最不知道的地方，而应该寻找一次行动之后，最可能改变模型信念的地方。

它问的不是：“还有什么东西是我没见过的？”而是：“我现在无法在哪几个解释之间作出选择？什么是区分它们的最小动作？得到答案以后，我会改变什么决定？”这三个问题缺一不可。一个问题必须能够被行动回答，必须能够排除一部分解释，答案也必须改变之后的行为。缺少第一点，它还没有进入科学；缺少第二点，探索只是收集数据；缺少第三点，答案便无关紧要。

当 world model 与 environment 之间存在 gap 时，我们通常把它看作模型的缺陷，想办法用更多数据把它填平。但 gap 也可以是问题出现的地方。Agent 先发现自己的解释在哪里发生冲突，再

设计一个最小动作让环境作出回答。只要它能够不断完成这个过程，gap 就不再只是错误，而会成为科学的起点。

很久以前，屈原也曾把问题直接抛向天空。《天问》几乎通篇都在提问。面对天地与历史，他没有急着编造一个圆满的答案，而是把自己与世界之间的 gap，一条一条写了下来：

遂古之初，谁传道之？
上下未形，何由考之？
冥昭瞢暗，谁能极之？
冯翼惟像，何以识之？
明明暗暗，惟时何为？
阴阳三合，何本何化？
圜则九重，孰营度之？
惟兹何功，孰初作之？
斡维焉系？天极焉加？
八柱何当？东南何亏？
九天之际，安放安属？
隅隈多有，谁知其数？

屈原《天问》