天问

Tong Zhang

网上常有人发帖争论，大模型还是 world model。

两年前，我在知乎上也问过一个问题：大模型可以问出一个好的科学问题吗?

今天跟同学聊起人工好奇心。我们的实验大概都很 negative。当 world model 和 environment 之间存在很大的 gap 时，一个 Agent 是否应该主动 explore? 如果应该，它能否由一个具体的问题驱动?

什么是好的问题?

过去，一个好的问题可能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，他只是顺着一个回答继续问下去，直到回答的人发现，自己相信的几件事情无法同时成立。

在这种对话里，问题的作用并不是显得深刻，而是暴露某个习以为常的 assumption。

但科学还要再往前一步。

到了科学里，含混的提问还不够。我们需要把它落实为一次可以观察的干预：如果我改变这里，世界会发生什么?

假设我们对同一个现象有两个看起来都合理的解释。一个有用的问题应当帮助我们找到尽可能小的 intervention，使两个解释给出不同的预测，再由 observation 作出裁决。

因此，我现在更愿意用“能否区分不同的 world model”来判断一个问题，而不是看它覆盖了多大的未知。

从这个角度看，人工好奇心实验的 negative result 并不奇怪。

我们常用 novelty、prediction error 或 entropy 奖励 Agent。可是当模型和环境的 gap 很大时，surprise 到处都是。Agent 可以反复晃动摄像头，可以盯着随机噪声，可以钻进一个混乱而不可预测的角落。它看起来非常好奇，却没有获得任何可以迁移的知识。

问题在于，不可预测的东西未必值得理解。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。这就是所谓的 noisy TV problem：噪声永远新鲜，却不会让模型变得更好。

我们后来觉得，更有用的探索目标也许不是模型最不知道的地方，而是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：“我现在无法在什么解释之间作出选择? 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?”

我暂时用三个条件来判断这样的问题：它可以被行动回答，能够排除某些解释，而且答案会改变之后的行为。如果不能落实为行动，它还不是一个实验问题；如果不能排除解释，探索就只是收

集数据；如果不影响后续行为，答案对 Agent 也没有实际作用。

语言模型压缩了人类曾经提出过的问题。它知道类比、反事实、矛盾和隐喻，能够从一个领域借来另一领域的概念。它也许很擅长提出“像好问题的问题”，但这些问题可能无法被实验回答，甚至与当下的 environment 毫无关系。

World model 则提供了另一部分能力：如果采取某个动作，之后可能发生什么？哪些观察符合现有解释，哪些观察会使解释崩溃。

两年前，我问大模型能不能提出好的科学问题。今天再想，只会“问”大概是不够的。问题还要落实为一次有代价的行动，而观察结果也应该真的更新提问者的判断。

如果一个 Agent 能够不断提出并检验这样的问题，那么 world model 与 environment 之间的 gap 就能成为下一轮实验的线索，而不只是模型的缺陷。

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》