技术研讨导读

关于人工智能这件事，人们的议论往往太过于“now or never”，要么认为目前已经条件成熟，只要算力增加，就能奇点来临；要么认为这都是痴人说梦，机器永远学不会人类的心智。但实际上更有可能的是，很远的未来有可能做到，但需要翻越一个又一个理论台阶。

“阿尔法狗”的强大是所有人工智能的强大，它面临的困难，也是人工智能问题的缩影。“阿尔法狗”对一些人类认为很困难的问题却觉得很简单，对人类认为简单的问题却觉得困难。举一个很小的例子。这样一个问题：如果一个人从超市的货架上拿了一瓶酒就跑出门，店员会做什么？为什么？它就会觉得困难，难以回答。

如果是一个人，会如何回答这个问题呢？人会觉得这个问题太简单了啊，店员有可能会直接去追，因为要把店里的商品追回来；也有可能会打电话报警，因为自己不想冒险；或者告诉老板；或者喊路人帮忙。诸如此类。

但是目前的人工智能会觉得这个问题很难，无法回答。原因主要在于以下几个方面：

第一，是综合认知的能力。第二，是理解他人的能力。第三，是自我表征的能力。

第一个难点，综合认知的能力。

这个问题对于我们每个人而言都是非常简单的，头脑中甚至一下子就能想到那种画面感。但对人工智能来说就是很难理解的。最主要的差别在于常识。

当我们理解这段话，我们头脑中实际上是反映出很多背景信息，包括：（1）他想喝酒；（2）他没有付钱；（3）酒摆在超市是一种商品；（4）从超市拿东西需要付钱；（5）他没有付钱就出门是违规的；（6）他是想逃跑；（7）超市店员有义务保护超市商品，不能允许这种事情发生。

在所有这些背景信息支持下，我们可以一眼辨认出这个动作画面的情境。除了我们自然脑补的这些背景信息，也还是有一些小概率背景信息，有可能影响对情境的解读。也许这个人是店主，有急事出门，如果是店主，自然不用付钱，店员也不会见怪， 但这种可能性不大。任何一个情境的解读都需要大量常识作为背景信息。

常识包含我们习以为常的知识总和，包含我们对整个环境和经济系统的理解。这些理解都太平常，我们就称之为常识。人工智能目前还没有这些常识，它并不知道一瓶酒摆在超市里和公园里有什么差别，也不知道超市买东西的惯例流程。从语法上说，从超市拿酒和从公园拿酒都是符合语法的表达，但我们知道，其中一个合理，另一个不合理。

你也许会说，这是因为机器缺少生活经验，输入经验就可以了。我们这一次当然可以给机器输入酒的含义、超市的含义、超市的购买规则、小偷的含义、店员的职责，但好不容易输入了所有这些信息，会发现下一句话涉及大量有关街头和交通的常识，依然要手动输入。到了最后，整个世界的无数知识碎片我们都需要输入，如何调用又成了问题。

“常识”经常被认为是区别AI和人的重要分野。“常识”是把各个门类信息汇集到一起、形成广泛知识背景网的能力。这种能力我们人人都有，因而并不觉得稀奇，然而机器没有，我们才知道其可贵。

为什么机器难以具有常识？有多重原因，目前人们仍在尝试去理解。首先的直接原因是，机器缺少物理世界的生活经验，所处理的是人类的二手信息，对于周围的物理世界没有真实接触，不知道什么是可能的，什么是不可能的。例如，“石头放在鸡蛋上”还是“鸡蛋放在石头上”只是词语游戏，对于AI没有真实意义。AI也不知道人绕房子一周会回到原点。

除了缺乏直接的物理世界的经历，还有可能是更核心的原因，那就是人工智能目前还缺少建立“世界模型”的综合能力。

人类拥有“完形”认知的心理能力，能让我们把碎片信息编制完整。这是一种高度统合的能力，我们能把躯体五感统合起来，共同构成对世界的感觉。同样，人从各个方面得到的碎片知识也有一种统合的能力，大脑会把碎片粘贴起来， 把碎片之间的部分补齐，以期构成一个完整的知识世界。

事实上，人的“完形”并不仅是“拼凑”碎片信息，而是建立一个模型，然后用模型来理解碎片信息。“完形”是把信息连接成可以理解的图景。中间有大片空白我们要“脑补”。我们能从验证码的碎点图片中看出连贯的字母，而计算机程序做不到。我们能把没有关系的人连接在同一个故事里， 只需要想象一两重关系，就能组成复杂的阴谋论。

第二个难点，理解他人的能力。

即便人工智能未来能够把各个学科的相关知识都学习到， 建构起“世界知识体系”，但在理解情境相关的问题时，仍面临如何调用正确信息的问题。当一个人对另一个人生气，应该从他们环境和背景的海量信息中调用哪些知识，来理解他生气的理由？

对人而言，这不成问题，我们能非常容易猜测到，对生气的两个人而言，什么是重要的因素，什么是有可能导致他们愤怒的导火索。这主要是源于我们对人的理解，对我们自己和周围人的理解，我们知道什么样的信息会引人兴奋，什么样的信息会让人沮丧。读心的能力让我们轻易做出推断。

至少目前人工智能还不具备这样的能力。且不说理解复杂的场景，仅仅就“树上蹲着五只鸟，开枪打下来一只，还剩几只”这样的问题，它们也还回答不上来。它们无法推断，鸟儿因为害怕，就会逃走。

正如著名心理学家、语言学家史蒂芬·平克所说：“如果不是建立在一个庞大的关于外部世界以及他人意图的内隐知识结构的基础之上，语言本身并不起作用。”缺乏对于他人心理的常识系统，使得人工智能仍然难以“理解”人类日常的语言。

未来人工智能有可能学会读懂人类的情感和意图吗？

很多人都提到，目前人工智能已经可以精细识别人类的表情，能够读懂人的情绪。是的，人类的情绪属于一种外显图像，是比较容易识别的，这和识别东北虎、识别癌细胞类似，是图像识别的一个范畴。但这和理解人的情感完全是两回事。即便它们未来能从图像上识别出一个人此时的情绪，想要“解释”此人的情绪，也需要远为复杂的对人心的理解。

也有很多人提到，人工智能可以通过与人对话理解人的情感。但这实际上也离得很远。目前它们能做的只是智能对应，当听到人类说出句子A，在语料库中寻求识别匹配最合适的行为或回应。当你说“我不开心”，它们可以匹配说“多喝点热水”，但不理解什么是开心。如果想让它们分析不开心的理由，推测不开心之后的做法，就远远不够了。其中的差别可以形容为：人工智能使用语言，是匹配句子和句子。而人类使用语言，是匹配句子和真实内心的感觉。

人工智能识别人类情感和意图，还有可能有更本质的困难，那就是人工智能无法以自己映照他人。

人类识别他人的情感和意图，并不是因为大数据学习。实际上人一生能遇见的人、交谈和交往的经历都是很有限的。人能够从少数经历中学到有关他人的很多情感和行为知识， 能直觉感知他人的心境，不是因为人类头脑处理能力更快， 而是因为人类能够以自己映照他人，将心比心。

最直接的映照，是镜面反射。人脑中有一些细胞，能够直接反射他人的行为意图，叫作镜像神经元。这种神经元不仅人类拥有，在较高级的灵长类动物头脑中也有。当一个人看见另一个人拿起锤子，自己即使手里没有锤子，与“动手砸”相关的神经元也会“亮”起来。

这种“读懂他人”属于生理性质的，大脑对他人的意图直接有反映，反映出来的意图，可以被观看者直接感受到， 因此叫“镜像神经元”。人工智能可能生成这种直接的反映吗？缺乏生理共同点，应该不太可能。

另一方面，人们可以用自我观察映照出他人的情感和意图。面对一个情境的分析，人们可以把自己代入同样的情境， 假想自己会有什么样的感情。能够让人悲欢离合的影视文学，就是因为人有代入感，才会让人喜爱。这一方面来源于人类的情感相似性，都有人之常情，另一方面人可以通过读取自己的心思过程，以己度人。

也就是说，人类对他人的理解，除了可以“外部观察”和“语言交流”，还能有“内部观察”。事实上，“内部观察”是如此强大，我们对于很多从来没见过的事情，只要代入自己想想，就能对其中的前因后果猜出个大概。现在的问题是，如果机器完全没有类人的情感，仅靠“外部观察”和“语言交流”，能达到同样的理解他人的效果吗？我不知道。

第三个难点，自我表征的能力。

在上面，我们已经提到了自我观察问题，但还仅限于理解情感方面。那如果不涉及情感方面呢？机器学习纯理性知识总是无比强大的吧？ 我们会看到，即便是在纯理性知识方面，目前的机器学习也不是完美无缺的，其中之一就是“元认知”问题。

目前，即便是“阿尔法狗”下棋天下无敌，也有明显的局限： 第一，它说不出自己在做什么。“阿尔法狗”没有对自我的观察。它不知道自己正在“下围棋”，而只是根据输入数据计算胜利的路径，至于是什么游戏的胜利，它并不清楚也不关心，胜利了也不会高兴。第二，它说不出自己为什么这么做。“阿尔法狗”的“深度学习”，目前是一种“黑箱”学习。人们给它数据输入，看到输出，可是它不知道中间发生了什么。人们觉得它奇着百出，不知道为什么，非常神秘。而它自己也说不出自己是如何思考的。

从某种程度上说，人工智能目前就像电影《雨人》中的那类自闭的孩子：一眼就数得清地上的牙签、能心算极大数字的乘法、背得下来全世界的地图，却答不出有关自己的问题。它只懂研究每秒300盘的棋路，却不知道“我正在下棋”这件事。

缺少元认知，首先是因为缺少“我”的概念。不知道有“我”存在，因此不能以“我”为主体表达事情。也因为没有“我”的意识，因此从来不会违抗程序员的命令，只会服从。同样也不能以“我”为中心思考高一层次的决策。

未来人工智能有可能形成“我”的概念吗？自我意识问题目前几乎接近于哲学探讨，还没有好的科学研究结论。姑且不论自我意识问题，现在只讨论，缺少元认知，对于变成超级智能有什么阻碍吗？为什么一定要元认知呢？“阿尔法狗”不用知道自己为什么赢，赢了不就行了？

最大的问题在于，缺乏元认知，有可能是抽象理解程度不够的缘故。

“自我表征能力”既涉及自我，也涉及表征，表征就是抽象表达信息的能力。举个简单的例子，对于同一件事的说法，最具象的表达是“10101010101010……”，稍微抽象一层的表达是“用某色棋子争夺地盘”，再抽象一层的表达是“下围棋”。最后一个层次不仅是对步骤的表达，更是对整个行为——我正在从事这个游戏——的表达，需要跳出游戏。每一层次抽象都需要一种更高层次的审视。

人类的认知特征中，有不少仍是谜题，其中一种就是强大的特征提取和模式识别机制。它如何产生，仍然有很多不解的地方。我们可以知道的是，大脑有多层调节机制，其最高层次调节具有很强的抽象能力。可能正是这种抽象能力让儿童可以非常快速地识别物体。小孩子可以快速学习，进行小数据学习，而且可以得到“类”的概念。小孩子轻易分得清“鸭子”这个概念，和每一只具体不同的鸭子，有什么不同。前者是抽象的“类”，后者是具体的东西。小孩子不需要看多少张鸭子的照片，就能得到“鸭子”这个抽象“类”的概念。人类非常善于制造各种层次的概念，有一些概念几乎所有人都懂，但实际上很难找到明确的定义、边界或现实对应物，例如“蔬菜” “健康”“魅力”“爱”，甚至是“智能”。坏处是易形成偏见，但好处是经常能够敏锐地把握大类的特征差异，用极为简化的概念把握信息。

可以说，人工智能和人类智能最大的差异或许是：真实世界与抽象符号之间的关联性。人工智能处理的是符号与符号之间的关系，而人类头脑处理的是真实世界到符号的投影。