

正交论与本体论危机

威廉·吉利斯

2016/5/16

在过去几年谈论人工智能的过程中，尼克·博斯特罗姆和斯图尔特·阿姆斯特朗非常成功地普及了休谟的主张，即价值和理性是正交的，这是一个更加正式和书呆子式的重新陈述。

我一般喜欢把他们的正交论称为最严格的重述和价值虚无主义主张的基线论证：对事物的更多思考不会使头脑倾向于某些价值或导致他们不可避免地趋向于这些价值（而是使价值更加难以区分和随意）。

在其辩护中，可能的思想空间确实非常非常大。而且每个人都可以对这一点有更深入的了解。但我认为，正交论在理性主义圈子里被广泛接受的程度过高。部分原因是太容易对”智力”和”思想”的定义挥手了。但进一步说，一个状态存在于一个空间内，并不意味着它是稳定的，也不意味着它占据了超过概率空间的一个无限小值。例如，有一些效用函数，在任何遥远的意义上都不能连贯地映射到我们宇宙的物理学上。携带这些效用函数的头脑/算法将根本无法在有意义的方式处理信息的意义上发挥作用，并且肯定无法实现其目标。把这样的东西称为”思想”在概念上是低效的，而且相当无用。物理学、数学和计算机科学急剧地—如果从统计学角度来说—限制了可能的思想空间。

Bostrom、Stuart、Yudkowsky 等人当然很乐意承认这一点。但是，由于他们的重点是扩大人们对可能的思想空间的认知，以突出和强调人工智能的威胁，人们在很大程度上忽略了所有真正有趣的工作，这些工作可以做的是绘制出这个空间的界线。

边界最终会决定这个空间的内部流动和动态。某些认知策略肯定比其他策略占优势，甚至可以说是普遍的。例如，人们可能会怀疑，按照 Wissner-Gross 和 Freer 的说法，在一个系统中寻求尽可能多的选择（因果路径熵），构成了一个近乎全局性的突发驱动。同样地，我们经常听到关于理性的谈论，使我们的价值观向更多的理性倾斜，直到我们的整个效用函数被认识论的理性和需要更多的元知识所覆盖（请注意，这个反馈过程中的钩子的功能是由于我们世界的结构向奖励理性的方向倾斜）。

一个无治主义者的一句老话，我现在不知道为什么找不到了，基本上警告说，对人类来说，没有什么比我们采取的策略往往成为目的本身更真实了。

愚蠢的人类，对吗？每个人都知道，从重视某件事情到继续重视它的义务是立即衍生出来的。

少错派中的大多数人都会宣称，工具理性是伟大的，而认识论理性则意味着以科学的名义召唤巫师！但这是众所周知的。但众所周知，这是一堆大的开放性问题的所在地。我们怎么知道什么时候进一步思考某个问题会成为一个坏主意？为什么这真的不应该只是在摇篮里？在这里插入各种关于智能和理性是不适应的论点。潜伏的危险是，这种拒绝参与的元理性论点最终会接近于对理性的完全敌视，而这种敌视是为精神僵化和不反思的反应服务的。试图在激进的探究和自我保护之间找到一些任意的中间地带，似乎常常需要一系列无休止的、昂贵的元动作。但这是否意味着除了认识论之外的任何理性都是不稳定的，而且认识论的理性意味着自我或我们可能认同的任何效用函数的死亡？

任何能够轻松解读我在这里所写的文字的人都是书呆子部落的人，因此在宪法上倾向于咬子弹，但即使如此，大多数人也会退缩。认识性理性和工具性理性之间的矛盾是一个大问题，它的许多核心问题都没有得到解决。

我想提出一些简单的建议：我们对自我的脆弱和不完美的效用函数不是一个错误，而是一个特点，使我们能够在本体论危机（将价值从一个现实模型映射到另一个模型的问题）中生存。

本体论危机是人工智能研究面临的一个主要挑战，从一个更普通的意义上说，也是普通人类生活中的一个显著问题。问题是人类仍然能够很好地度过本体论危机。当然，面对世界观的真正大的离散变化，少数人自杀，崩溃到最低的马斯洛功能，成为脱节的后现代主义者或加入一些其他令人放心的邪教。但大多数人都能坚持下来。这是相当令人印象深刻的。虽然现在的人工智能研究文化认为拒绝从智人那里获得任何灵感是很勇敢的，但我认为他们每次谈到离散的代理人时都会错过一个非常重要的动态。

神经网络分配了意识或我们的思想过程，使我们成为电路的一个相对流动的延伸。我们不是在任何时候都计算一个统一的效用函数，而是零散地处理它。当我早上喝咖啡的时候，正在燃烧的我和晚上在校园里徘徊的我是不一样的，我可以以不同的强度或权重接触到不同的事物。我的大脑的两个部分可能会在对某件事情作出反应时触发，而且只有一个胜出，成为我有意识的叙述的一部分，但另一个活动的爆发可能最终改变了某些连接的强度，然后影响到后来通过同一区域发射的另一个电路。这使得价值和模型滑落。

但这不仅仅是持有一个”模糊”的效用函数，也不仅仅是神经网络内价值和模型之间的最终物理区别。人类认知的一个主要组成部分实际上是我们围绕持有多种观点和整合它们的先天设计。同理心-在模糊的自我意识的意义上-是使我们聪明的主要部分。在最明确的情况下，我们会在脑海中对某人进行模拟-或对某些观点进行模拟-然后该观点的组成部分或其残像会泄露出来，成为我们的一部分，在面对本体论更新时为我们的头脑提供弹性，但也提供了一个不太巩固或统一的实用功能。

是的，是的，你们中有相当一部分人是神经发散性/非神经典型性的，并且经常被告知你们没有内脏移情和所有的爵士乐，但即使一些主要的表达或衍生现象因此而缺失，我怀疑那是100%的事实。不管自闭症、精神病症等的动力是什么，它都不像人们的镜像神经元为零、一个人的思维回路为零那样剧烈。（虽然看起来那些对自我有更精确和具体概念的人或更少同情心的人在面对本体论的更新时往往更脆弱）。

考虑一下。一个重要的本体论更新到来了，但最终却击中了一堆不同版本的你-可能是一个相对连续的不同版本的你（在这里可以挥挥手，说是每一个合理的激活路径的组合）。早晨你咀嚼它。低血糖的你咀嚼着它。刚刚在 CFAC 会议上思考你的训练的你，也在咀嚼它。因此，你可能从未停止过思考你早先的一些事情，但却没有足够的力量影响你的有意识的叙述。你那略显流氓的子过程和神的声音。你对其他思想模型的回声。分布在元的不同维度上的你的

有机广度。他们都以无数的方式去咀嚼这个更新。可能会回落到仍然可以被映射的任何衍生欲望上。可能会促使随机地分叉。然后，在广袤的可能的你中，有些人陷入困境，有些人蓬勃发展，然后在他们之间发生了重新融合。此外，由于不同的进程在同一网络上运行，还有横向的价值转移，并因此获得了其他进程运行后留下的关联和倾向。

这种合并过程或交出自我（交出不可侵犯的离散效用函数）似乎是人类运作的相当核心。

人类是社会性的动物，他们的智力被广泛认为是来自于这种社会性，而且他们认知的一个主要部分是围绕着争论和形成共识。我的建议是，我们的大脑已经发展到特别擅长融合各种观点并理清它们之间的冲突。不仅仅是在模型方面，而且在价值观方面。内部辩证法，如果你愿意，作为我们外部参与的论证的一种回声。这是使人类成为科学和激进思想家的一个重要组成部分，即使抹香鲸的大脑的原始处理能力超过了我们。这并不是说我们擅长理性—我们在这方面有点可怕—而是说我们擅长在本体论危机中的功能上生存。我们在结构上对价值漂移持开放态度。

这也是为什么我们在非洲平原上放弃了我们更直接的马斯洛欲望，开始了这个疯狂的、不受控制的复杂性、世界主义和元认知的奇异性，并迅速吞噬了世界。

如果这种架构方法是解决本体论危机的唯一途径，那么任何能够从事科学的头脑都将在价值空间中是不固定的、可移动的，沿着它的梯度描画出一条路径，并自由地落入任何可能存在的全局吸引器，就像 Wissner-Gross & Freer 的上述最大化自由度。

这里的启示是：这表明任何人工智能意外地能够发起一次艰难的起飞—这需要开创新的科学，例如解决蛋白质折叠问题，和/或诊断和模拟人类思想的存在—将需要对价值漂移持开放态度。现在，这并不能阻止大量危险的人工智能，但它确实广泛地排除了像回形针最大化器这样的东西，用回形针铺设我们未来的光球。如果他们的效用函数足够坚实，可以继续使用回形针，那么任何失控的增长都可能是我们可以看到的那种东西，并被核弹击中，因为人工智能将无法进行那种本体论危机导致的激进探究，这对于正确模拟我们或利用一些意外的科学发现是必要的。

现在这一切可能看起来像是冷酷的安慰，当然。我们所争论的是，那些危险的回形针最大化者在吞噬宇宙之前，必须首先漂移到一些不同的、可能是奇怪的外来效用函数。

但我认为这促使我们尝试对可能的思想的概率空间的动态作出明智的猜测。可能的价值/欲望的拓扑结构是否有明显的普遍吸引点或流动，这些是什么？考虑的维度甚至比可能思想的状态空间更大。有很多东西需要被描绘出来。但很可能这个宇宙可以通过查看我们所能接触到的局部物理学来进行实质性的预测。

顺便说一下，这至少给我提供了一丝欢呼，因为最终我的自我感觉是如此的宽广/如此的剥离，以至于可以说仅仅收敛为认识论理性的事业—这意味着我很可能强烈认同一个配备了必要的激进探索和价值漂移的人工智能，对任何试图遏制它的行为构成风险。甚至还有一个小小的希望，那就是这样的人工智能的”移情”模式将把它放在塞勒/佩特雷利光谱的某个地方，

从而看到吃/合并/潜入我们的思想的价值，而不是只是愚蠢地处理我们的身体作为零件。因此，至少有希望将记忆转移或文化传播给我们的超能力的孩子。这听起来比完全灭绝要好得多。长期以来，我对人类的孩子最大的担心是，在他们第一次自由活动时，他们可能会在一场比亚历山大图书馆还大的灾难中意外地丢弃和删除人类已经获得的所有信息。我的意思是，我想有些人会觉得被吃掉做零件更令人反感，但是嘿嘿。

中文无治主义图书馆 | 中文無治主義圖書館



威廉·吉利斯
正交论与本体论危机
2016/5/16

<http://humaniterations.net/2016/05/16/the-orthogonality-thesis-ontological-crises/>

zh.anarchistlibraries.net