图灵测试与人工智能

黄铂钧

微软亚洲研究院

bojhuang@microsoft.com

2014 年 6 月

最近一段时间一条“机器首次通过图灵测试”的消息[1]引起热议。多家媒体报导说一个聊天机器人程序伪装成一个名为 Eugene Goostman 的 13 岁乌克兰小男孩在图灵测试中成功骗过了 30 位人类参与者中的 10 位。热捧这一事件的人认为这是一项划时代的壮举，标志着人工智能的重大突破。而质疑声中有的怀疑这个事件的有效性，认为在主办方不愿透露更多细节的情况下不能确定聊天机器人是否真的通过了图灵测试；而有的则怀疑图灵测试本身，认为图灵测试相对于人工智能其实意义不大。本文认为，图灵测试是通用人工智能的一个有效的充分条件（且其自身并未试图定义智能的概念）。但正是出于它的重要性和难度，我们才应该对图灵测试有关的结果格外审慎。

什么是图灵测试？

在一篇 1950 年发表的著名论文《Computing Machinery and Intelligence》中，数学家阿兰·图灵详细讨论了“机器能否拥有智能？”的问题。有趣的是，作为计算机科学与人工智能领域共同的先驱，图灵成功定义了什么是机器，但却不能定义什么是智能。正因如此，图灵设计了一个后人称为图灵测试的实验。**图灵测试的核心想法是要求计算机在没有直接物理接触的情况下接受人类的询问，并尽可能把自己伪装成人类。**如果“足够多”的询问者在“足够长”的时间里无法以“足够高”的正确率辨别被询问者是机器还是人类，我们就认为这个计算机通过了图灵测试。图灵把他设计的测试看作人工智能的一个充分条件，主张认为通过图灵测试的计算机应该被看作是拥有智能的。

具体就操作层面来说，图灵在他的论文原文中是这样定义图灵测试的[2]：

“我们称下面这个问题为“模仿游戏”。游戏参与者包括一个男人，一个女人，以及一个任意性别的询问者。询问者与另两个人待在不同的房间里，并通过打字的方式与他们交流，以确保询问者不能通过声音和笔迹区分二者。两位被询问者分别用 X 和 Y 表示，询问者事先只知道 X 和 Y 中有且仅有一位女性，而询问的目标是正确分辨 X 和 Y 中哪一位是女性。另一方面，两位被询问者 X 和 Y 的目标都是试图让询问者认为自己是女性。也就是说，男性被询问者需要把自己伪装成女

[1] http://www.bbc.com/news/technology-27762088

[2] 为清楚起见，这段摘录并非逐字翻译，且语句顺序也稍有调整，具体可参考原文第一节。

性，而女性被询问者需要努力自证。现在我们问：如果我们把“模仿游戏”中的男性被询问者换成计算机，结果会怎样？相比人类男性，计算机能否使询问者更容易产生误判？”

这里有几个细节值得注意，它们在很大程度上决定了图灵测试的有效性。（1）首先，图灵测试中询问者与被询问者之间进行的并不是普通的日常聊天，询问者的问题是以身份辨别为目的。这种情况下询问者通常不会花费时间寒暄和拉家常，而是会开门见山地说“为了证明你的身份，请配合我回答下面问题…”。事实上，目前网络上聊天机器人有时能够以假乱真，往往是采用了在用户在不知情的情况下尽量把谈话引到没有鉴别力的话题上的策略（例如“谈谈你自己吧”）。（2）其次，图灵测试中人类被询问者的参与是必不可少的，她的存在是为了防止计算机采取“消极自证”的策略，例如拒绝正面回答问题，或者答非所问闪烁其词，就像一个真正的不合作的人所做的一样。在这种情况下，另一个积极自证的人类被询问者可以保证询问者总是有足够的信息做出判断。类似的情况也适用于当计算机试图模仿正在牙牙学语的幼童或头脑不清的病人等“特殊人类”时。（3）另外，图灵测试的原则是要求询问的交互方式本身不能泄露被询问者的物理特征。在图灵所处的年代这几乎只能全部通过基于文本的自然语言来完成，因此图灵限定测试双方基于打字进行交流。但在多媒体技术发达的今天，视频、音频、图片等等“虚拟内容”都可以通过计算机以非物理接触的形式呈现（这当然是 60 年前的图灵不能预知的！）。因此，允许询问者在图灵测试中使用多媒体内容作为辅助材料进行提问（例如“请告诉我这个视频的笑点在哪儿”）似乎是对原始图灵测试定义的一个自然合理的补充[3]。（4）最后，今天一般意义上理解的图灵测试不再严格区分人类参与者的性别。通常我们允许人类被询问者是任意性别，而询问者的目标也随之变成辨别哪一位被询问者是人类。

除此之外，完成一次具体的图灵测试还要注意很多操作细节，例如多少人参与测试算“足够多”，多长的讯问时间算“足够长”，多高的辨别正确率算“足够高”，如何挑选人类询问者和被询问者才能代表“人类”的辨别和自证能力，等等。由于图灵测试的巨大影响力，几十年来一直有人尝试挑战它，不时就会传出“某某计算机程序成功通过图灵测试”的消息。我想，正是对于意义深远的实验，我们才理应格外审慎。只有在仔细检查上面所列和其他一些重要细节之后，我们才能对其结果的有效性做出正确判断。类似几年前“超光速实验”那样的闹剧应该尽量避免。

图灵测试与人工智能是什么关系？

如果有一天机器真的通过了图灵测试，这到底意味着什么？这个问题涉及到图灵测试与人工智能的关系。的确，几乎所有有关人工智能的书籍都会谈到图灵测试，但一个经常被误解的地方是，**图灵测试是作为一个人工智能的充分条件被提出的，它本身并没有，也从未试图定义智能的范畴**。这一点图灵在他的论文里写的很清楚：

“机器能否拥有智能，为了回答这个问题我们应该首先定义‘机器’和‘智能’。一种可能性是根据大多数普通人的日常理解去定义这两个概念，但这样做是危险的。……在这里我并不打算定义这两个概念，而是转而考虑另一个问题，它与原问题密切相关，同时可以被更清楚无疑地表

[3] 参见 Total Turing Test 及相关工作。

达。……（图灵测试的描述）……可能有人会说这项测试对机器而言过于严格——毕竟人类也无法反过来成功伪装成机器，这只需检查算术的速度和正确度即可辨别。难道被认为拥有智能的机器就不能表现出和人类不同的行为么？这是一个很有力的反对意见，但至少不管怎样，假如我们有能力制造出一个可以成功通过测试的机器的话，也就无需为这个反对意见烦恼了。”

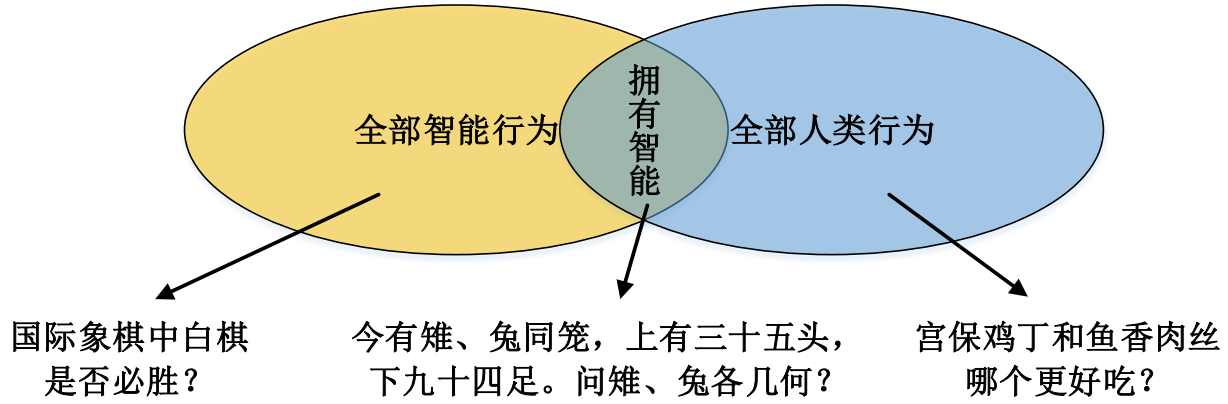

图 1：智能行为与人类行为的关系

借助集合的概念我们可以更容易地理解图灵测试与人工智能的关系。如图 1 所示，“所有智能行为”对应的集合和“所有人类行为”对应的集合既有交集又互有不同。在全部智能行为中有一些是人类靠自身无法做到的（比如计算出国际象棋中白棋是否必胜），但无论如何人类都被认为是有智能的，因此，在各方面都能达到“人类水平”— 也就是完成两个集合的交集部分—就应该被认作是“拥有智能”的。[4]另一方面，人类行为并不总是和智能相关。图灵测试要求机器全面模拟“所有人类行为”，其中既包括了两个集合的交集，也包括了人类的“非智能”行为，因此通过图灵测试是 “拥有智能”的一个有效的充分条件。

图灵本人对机器能够通过他的测试相当乐观，他大胆预测*“到 2000 年左右时，一台拥有 1GB 内存或类似规模的计算机可以在接受普通人 5 分钟的询问之后，使他们的判断正确率不超过 70%”*。然而直到 2014 年的今天，仍然没有任何机器被公认为已经通过图灵测试。有趣的是，这一失败事实反而还带来了一个我们再熟悉不过的应用 - 图形验证码。（每一次输入验证码都是一次图灵测试！）

图灵测试问题的进展缓慢与目前人工智能学界对图灵测试这个“充分条件”的研究热情不高有关。[5]这一部分上由于主流人工智能研究与图灵测试所追求的目标之间存在差异，同时也因为图灵

[4] 但反之未必，不一定非要达到人类水平才能被认作是智能的。

[5]一般认为人工智能学科正式成型于 1956 年的一次著名的研讨会前后，也就是说图灵测试实际上提出于人工智能领域诞生之前。正如 Stuart Russell 和 Peter Norvig 在一本人工智能的经典教科书中所写，在随后的 60 年间，整体而言“人工智能研究者们在图灵测试方面只投入了很少的精力”。

测试本身难度巨大。下面我们通过人工智能研究的三个重要特征来进一步讨论图灵测试与人工智能之间的异同，以及为什么图灵测试不大可能在短时间内解决。

图 2：图形验证码

- 主流人工智能研究关注智能体的外部行为，而不是产生该行为的内部过程

在这方面图灵测试的思想和人工智能学界是完全一致的。只关注外部行为是一个典型的功能主义/行为主义风格的做法，事实上这也是一个人工智能经常被外界所指摘的地方。严格的“主观思考”定义要求智能体具有自我意识。但一方面，从严格的科学方法讲，我们甚至并不真的确定是否有客观证据证实“意识”的存在。更重要的是，人们发现智能行为和主观思考完全可以被看作是两个独立的问题来考虑，二者并不必要纠缠在一起。具体来说，可以从数学上证明任何一台数字计算机的行为都可以用查表的方式机械地模拟。假设我们真的制造了一台具有“意识”的机器 A，我们总可以制造另一台机器 B 以查表的方式来机械地模拟 A 的内部运行，问题是 B 是否具有意识？如果每一台“拥有”意识的机器都能被一台 B 这样的“机械查表式”的机器所模拟，那么我们就无法通过外部行为来断定一个机器在内部上是真的在“思考”还是只是在模拟“思考”的过程，[6] 因此“是否拥有意识”从行为主义的角度也就成了相对独立的“另外一个问题”。同时，“拥有意识的机器总可以被没有意识的机器模拟”也说明“拥有意识”并不能给机器带来额外的“行为能力”，这进一步降低了“拥有意识”在行为主义者眼中的重要性。

基于外部行为与主观思考之间的独立性，主流人工智能研究和图灵测试把实现外部行为作为唯一目标，这样的观点被称为弱人工智能观点。我们知道每个学科的研究都基于一个“基本假设”展开。比如支撑物理研究的基本假设是“万物运转都受一套普适的、永恒的规律所约束”，而物理研究的目的“只是”找出这套规律是什么。类似的，“弱人工智能假设”(weak AI hypothesis) 认为经过良好设计的计算机**可以**表现出不低于人类智能水平的外部智能行为。可以说主流人工智能研究是以弱人工智能假设为出发点，研究如何实现这样一个计算机。

[6]一个有趣的不同是，人类研究“动物意识”（包括人类自身）的方法恰恰是通过观察动物在特定环境下的外在行为。这背后隐含的假设是我们相信没有意识的动物并不会“有意识地”装出一副有意识的样子（当然！），而这一假设对机器（或者机器的制造者）而言却并不一定成立。

图 3：机械查表式”的机器 – 西尔勒的“中文屋子”实验

- 主流人工智能研究关注如何模拟人类的纯粹智能活动，而不是全部脑力活动

就像前面提到的，人类的脑力活动 (mental process) 不仅包括智能，同时具有情感、审美能力、性格缺陷、社会文化习惯等等一系列“非智力特征”。因为图灵测试的模仿对象是普通人，事实上它对这些非智力特征的要求甚至可能还高过对纯粹智力的要求——作为一个普通人，他/她完全有可能对国际象棋一窍不通，但却不大可能从照片分辨不出美女/帅哥来。

当然，“非智力特征”的引入本身并不妨碍图灵测试成为一个有效的充分条件，但除非我们假设所有这些“非智力特征”都是拥有智能之后的必然产物，否则不得不承认图灵测试确实在机器智能这个核心问题之外加入了过多充满挑战却又显得不那么相关的因素。就像《人工智能》这本经典教科书里写到的，*“航空领域试图制造性能良好的飞机，而不是使飞机飞得如此像鸽子以至于可以骗过其他鸽子。”*人工智能研究确实应该更多关注与智力活动相关的抽象功能和一般原则。

- 人工智能的最终目标是能够综合适应“人类所在环境”的单一智能体，而不是专门解决特定数学问题的算法

在这一点上图灵测试与人工智能研究的最终目标也是一致的，只不过现有的人工智能水平离这一目标还相去甚远。事实上“综合模拟人类的智力活动”正是人工智能区别于其他计算机科学分支的地方。我们通过比较人工智能软件与传统软件来说明这一点。首先从最广义的角度看，传统软件也应属于人工智能的范畴：实际上很多早期的计算机科学家，比如图灵，就是以人工智能为动力展开对计算机科学的研究。所谓“计算”本来就是诸多人类智能活动中的一种。一个从未接触过计算机的人也许很难说清“从一个数列中找出所有素数”和“从一张照片中找出一只狗”哪个更有资格代表“智能”（前者属于传统软件范畴，后者属于传统人工智能范畴）。但另一方面，传统软件并不代表人工智能的全部内涵。粗略讲，我们可以认为传统软件对应了这样一类“计算

问题”，它们的共同特点是，问题本身是用一个算法（或非构造性的数学描述）来描述的，而对它们的研究主要关注在如何找到更好的算法。[7]而我们称之为“人工智能问题”的问题可以理解为另一类“计算问题”，它们的共同特点是无法用算法或从数学上对问题进行精确定义，这些问题的“正确答案”从本质上取决于我们人在面对这类问题时如何反应。对于人工智能问题，我们可以基于数学模型或计算模型来设计算法，但问题的本质并不是数学的。

通用人工智能（Artificial General Intelligence）基于弱人工智能假设，以全面模拟人类的所有智力行为为目标。注意到图灵测试作为一个充分条件，是不可能在通用人工智能真正实现之前得到解决的。另一方面，可以说现有每一个 AI 分支的成功都是通过图灵测试的必要条件，而它们中的大部分还没有达到“人类水平”。因为我们不可能穷尽所有人类智能行为，必须依赖**有限个**具有通用性的模型和算法来实现通用智能。目前人们仍然只能基于一些简单初等的模型来设计学习、推理、和规划算法。这些 AI 分支的研究都默认基于针对自己领域问题的弱人工智能假设，而支撑这些子领域研究的动力往往是其巨大的社会实用价值。它们固然已经在很多具体应用领域成绩斐然，但看起来离图灵测试所要求的水平仍然相差甚远。

2235 2237 2239 2241 2243 2245 2247 2249 2251 2253

图 4：一排包含素数的数列和一张包含狗的照片

（本文部分摘录自发表于《NEWTON 科学世界》2014 年第 3 期的文章“什么是人工智能？”。文中图片部分引自互联网。）

[7]需要注明是，对传统软件的研发同样也并不是计算机科学的全部内涵，就像“计算机”的概念远远不只是“电子硬件”。计算机科学的根本问题是“什么是计算”。而人工智能，作为计算机科学的重要分支，可以认为主要研究“智能是不是计算”的问题。