人工智能原理与技术第 1 次作业

2154312 郑博远

阅读图灵关于 AI 的原始论文(Turing，1950)。在该论文中，他讨论了一些对于他提出的事业以及他的智能测试的潜在的异议。哪些异议现在仍有分量？图灵的反驳是否合理？你能想到在他撰写该论文以后的发展引起的新异议吗？在该论文中，他预测到 2000 年以前，计算机将有 30%的机会通过 5 分钟的图灵测试，测试由不熟练的询问者进行。你认为当今计算机能有多少可能性？再过 50 年呢？

图灵在 1950 年关于人工智能的论文(Computing Machinery and Intelligence)提出了经典的图灵测试，即让机器参与“模仿游戏”假装人类，人类审讯者能否正确辨别出机器与人类，用以代替“机器能否思考”这一问题。在对“模仿游戏”与离散状态机进行介绍之后，图灵在文章的第 6 部分花费许多笔墨进行了对当时关于人工智能异议的驳斥与说明。下面我将分别对图灵文章中的 9 种异议与图灵的反驳进行概括，并分析下列异议在当今的分量：

1. 来自神学的异议 (The Theological Objection)

这种异议认为思考是上帝赋予人类灵魂的独有功能，因此动物或者机器都不能思考。图灵的驳斥如下：他一方面认为神学的论据漏洞百出，这既体现在伽利略时期经文的种种错误上，也可以用不同宗教间观点不同来驳斥；另一方面，他认为即使从神学的角度出发，上帝有能力也没有理由不给机器或动物赋予灵魂。我认为图灵对这部分的驳斥是充分且令人信服的。

在科学技术日益发达、人们受教育程度普遍提升的今天，宗教信仰与神学色彩已经日渐消退，因此这部分异议在当今社会分量已然不大。

2. “鸵鸟”式的异议 (The "Heads in the Sand" Objection)

这种异议并非质疑人工智能的发展可能性，而是认为机器思维的后果太可怕了——若机器能够思考，则人类将失去高人一等的优越感。我十分认同图灵所持有的观点，即这种论点不足以反驳，而应受到安慰。我认为随着人工智能技术的发展，社会上“鸵鸟”式异议的分量可能不减反增。“鸵鸟”式害怕的原因将不仅局限于失去思考的优越性；随着 AI 的发展，重复性、机械性的劳动力将很可

能率先被人工智能所取代。因此，当下社交媒体上人们关于人工智能取代人类岗位的担忧并不鲜见，且往往多于对享受其便利的期许。不过我认为，公众对人工智能的担忧并不会对人工智能发展的迅猛趋势有过多影响。

3. 来自数学的异议 (The Mathematical Objection)

这种异议指出，以哥德尔定理（即任何可以进行一定数量初等算术的一致形式系统是不完备的）为首的一些数理逻辑证明都表明离散状态机能力的局限。在阅读图灵的论文时，我对哥德尔不完备定理与人工智能能力局限性的关联不甚清楚，因此我又查阅了卢卡斯-彭罗斯关于哥德尔定理的佐证。卢卡斯认为，可以将特定符号与机器的特定状态相关联，由此将推理规则与机器的状态转换相关联。从而可以认为机器证明为真的主张，将“对应于可以在相应的形式系统中证明的定理”。若为这样的形式系统构造哥德尔句子，由于哥德尔句子无法在系统中证明，因此机器将无法将此句子作为算术真理生成。

图灵对此的反驳是，尽管机器的能力有限是既定的事实，但人类智慧不受限制在此异议中仅作陈述而未被证明。若机器与人类都有相同的局限性，则这样的局限性无法作为在“模仿游戏”中的判断依据。另一方面，图灵指出人们过于重视机器出错而沾沾自喜，却忽视了人们也通常会出差错。

我认为来自数学的异议在当今仍然是一个悬而未决的争议。卢卡斯等学者通过对哥德尔不完备定理的延申企图证明人类思维不属于图灵机（人类思维若是图灵机，永远能够找到其系统中的哥德尔句子使该系统无法判断真实性；人类思维可以判断，因而不是图灵机），从而得到人类思维比计算机的优越性。但卢卡斯也受到了许多不同方面的质疑，如有人认为人脑遵循的算法比较复杂以至于当下无法找到该系统中的哥德尔句子，也有人认为人类思维或许并不遵循一致性……关于该方面的许多争论我可能无法完全理解消化，但我认为这样的异议在当今与可见的未来仍都将占有重要的分量，需要进一步的研究证明以佐证。

4. 来自意识的论点 (The Argument from Consciousness)

这种异议认为机器并不存在意识，不能感受到思想与情绪。即便人工智能能够创造如协奏曲等作品，但其的创作过程仅是拼凑涂鸦而成，无法感受到创造的喜悦等情绪。我认为时至今日，来自意识的论点在大众讨论中占很大的分量。许

多人认为人工智能的创作是僵硬死板的拼凑，而没有创作者个人情感的注入；但我认为随着 AI 的不断发展，未来的人工智能能够颠覆人们对于机器意识情感的认知。图灵关于来自意识的论点的驳斥我认为尤为精彩：他指出，按照这样的方式，判断人或机器是否具有思维的方式唯有成为那个人或机器，这便会陷入唯我论喋喋不休的争执中，颇有些“子非鱼，安知鱼之乐”的意味。图灵也没有停留在“唯我论”简单的回避，而是给出了自己的判断。他认为只要机器能够回答阐释创作的意图等具体详实的问题，就可以认为机器具备创作的意识。

5. 来自各种能力限制的观点 (Arguments from Various Disabilities)

这种异议认为机器能够完成许多不同的功能，但在某某方面存在缺陷无法达到（如幽默感、坠入爱河，甚至享受草莓和奶油等等）。图灵对这种观点的反驳在于，人们对大量现有机器的认识导致了对人工智能的刻板印象。我相信这一点在图灵的年代应该尤为明显，即人们所认识到的机器往往都从事单调简单的小任务，因此当时的人们从大量生活经验中进行归纳总结，得到“机器绝对不可能做到某某事情”的简单偏见，图灵也指出这样的科学归纳或许不适用于日常生活中。我认为图灵的反驳是合理且有效的，即他认为只要机器有足够充足的容量便能完成各式各样的不同任务。随着计算机技术的迅猛发展，我认为这种反驳观点的分量正逐渐减小。实际上，当下计算机的容量早已远远超越图灵的预计，绘画、写作、剪辑等等层出不穷的技术也正在被 AI 所掌握。随着年轻一代在成长中形成新的固有印象，人们对人工智能的发展前景将会持更开放包容的态度。

6. Lovelace 夫人的异议 (Lady Lovelace's Objection)

这种异议认为计算机只能够按照人们所制定的程式与指令，而没有任何创作的意图。我个人认为图灵关于这部分异议的反驳不太能够让我信服。他首先指出，“太阳底下无新事”，人们所谓的“独创性工作”也是基于教育发展的结果。其次他指出，如果创新指的是让人吃惊，那么他也经常为机器超乎他想像所惊讶。我认为这样的说法主要围绕“创新”的定义进行驳斥，意义不太大，并没有像之前的回答那样落到实处。其实我认为，图灵后文中开创性的提出让机器学习的方式便是对这一异议的最好回应，此处其实也有略带一笔，应该是出于文章篇幅的安排没有在此详细说明。我认为当下对于人工智能独创性的异议应该越来越少，

因为身边越来越多的 AI 模型已然能够展现出不俗的创造能力：如根据风格与关键词作画等等。尽管现阶段仍然被诸如作品是素材拼贴的言论诟病，但我认为人类在作为初学者时也需经历模仿到创造的阶段，加以时日 AI 应该能够在创作领域达到人们所满意的“创新”程度。

7. 来自神经系统连续性的论点(Argument from Continuity in the Nervous System)

这种异议指出，神经系统并非离散状态机，因而离散的状态机无法模拟神经系统的行为。图灵对该异议的驳斥是，以微分分析机这样的连续机器为例，尽管离散状态机并非连续机器，但其可以通过使其输出值受到某种概率分布的扰动来预测连续设备的输出。实际上，包括人脑在内的系统的观测精度都是有限的，因此只要在足够的计算资源下进行模拟便能得到较为满意的结果。我认为当今这样的反驳分量应该在逐渐减小，因为许多神经网络的模型已经能够较好的完成计算视觉等等领域的识别任务，用离散状态机模拟神经系统已经取得了不错的成绩。

8. 来自行为非形式化的论点 (The Argument from Informality of Behaviour)

这种异议认为人类的行为逻辑并不是遵从某种特定的规则，无法制定出一个人在某种情况下应该做什么的行为准则，因此人不可能是机器。我认为图灵关于这个论点的反驳也是充分且精彩的。图灵将这样的反对意见总结为：“如果每个人都有一套行动规则来调控其生活，那么他与机器就相差无几了。但不存在这样的规则，因此人不能成为机器”，并敏锐的指出了这样论述中不周延的中项。随后图灵紧接着提出，尽管他承认人类的行为不存在某种行为准则（rules of conduct），但人类行为可能服从于某种行为规律（laws of behaivour）。图灵举出了一个简单的例子，即他在曼彻斯特的计算机上设置了一个仅使用 1000 个存储单位的小程序，其中提供十六位数的机器在两秒钟内回复另一个数字。图灵认为，即便是如此简单的规律，也是难以从表象中预测的；因此，人类行为是否服从于某种行为规律仍需要大量的科学观察，而不能妄下定论。

9. 来自超感官认知的论证 (The Argument from Extrasensory Perception)

这一部分的论证主要围绕如何在图灵测试中排除超感官认知的影响从而完善理论的严谨性，与人工智能可行性关系不大，因此我不做赘述。

当今的新异议：

当下的讨论与异议集中于如何对待将来社会中的人工智能，使其更好地为人所用。这部分异议主要关注于人工智能的伦理问题与法律监管、人工智能数据集所涉及到的隐私安全和数据滥用问题、人工智能所造成的失业等问题。

也有人对图灵测试是否能较好的测试出“机器能否思考”这一问题提出质疑。其中较为著名的是“中文房间”，即假想在一个密闭的房间中，有一个完全不懂任何中文的美国人。他凭借所有的中文字符集与如何处理这些中文字符的规则书，按照规则书将问题的正确回答拼凑出来后传递出去。按照图灵测试，这个美国人将被认为成能够理解中文的智能，但实际上他对中文一无所知。实际上我认为图灵测试对于审问者的提问水平、通过测试的概率等定义都较为模糊，或许需要更为确切的标准或更好的测试方式来判定机器是否具有思维。

对人工智能通过图灵测试的预计：

早在 2014 年，英国雷丁大学便宣布聊天机器人尤金·古斯特曼首次通过了“图灵测试”，成功让人类相信它是一个 13 岁的男孩。在 5 分钟不熟练的询问者和机器的对话中，其中 33%的回答让裁判认为与他们对话的是人而非机器。但我认为，仅凭这样的测试便简单得出人工智能已然能够通过图灵测试是过于草率的。实际上，该程序冒充的是一个来自乌克兰、英语非母语的 13 岁小孩，这便大大降低了图灵测试的难度。我认为，包括时下风靡的 ChatGPT 在内的许多人工智能通过图灵测试的概率仍然不容乐观，精心设计的问题仍然很可能让它们露出马脚；另一方面，计算机似乎也能通过提前预设的心理学上的语言引导欺骗人类审问者，但这似乎并不意味着机器思考水平能力的提升。此外，即使人工智能发展能够到达如假包换的水平，审问者随机选择的概率也将维持在 50%上下，有些人对计算机 90%以上给通过图灵测试的愿景似乎无法达到。

尽管人工智能发展的水平不及图灵在上世纪中所乐观估计的那样，但近十年来仍然取得了突破性的进展，也改变了许多人对机器能否思考这一问题的认知。因此我乐观地相信，50 年后，将有许多计算机能够达到图灵所期许的“智能化”的水平，人工智能将就现有水平有颠覆性的进步。