科技 深度

意识是甚么？能创造艺术的人工智能有意识吗？

机器人不能证明自己不是哲学僵尸（不过你也不能）。

2022年，一幅由人工智能算法Midjourney“绘制”的艺术作品赢得了美国科罗拉多州年度博览会艺术竞赛“数码艺术”类的首奖。图：Jason Allen/Midjourney

朱孝文

特约撰稿人 朱孝文 发自伦敦 | 2022-10-23

意识　朱孝文　人工智能

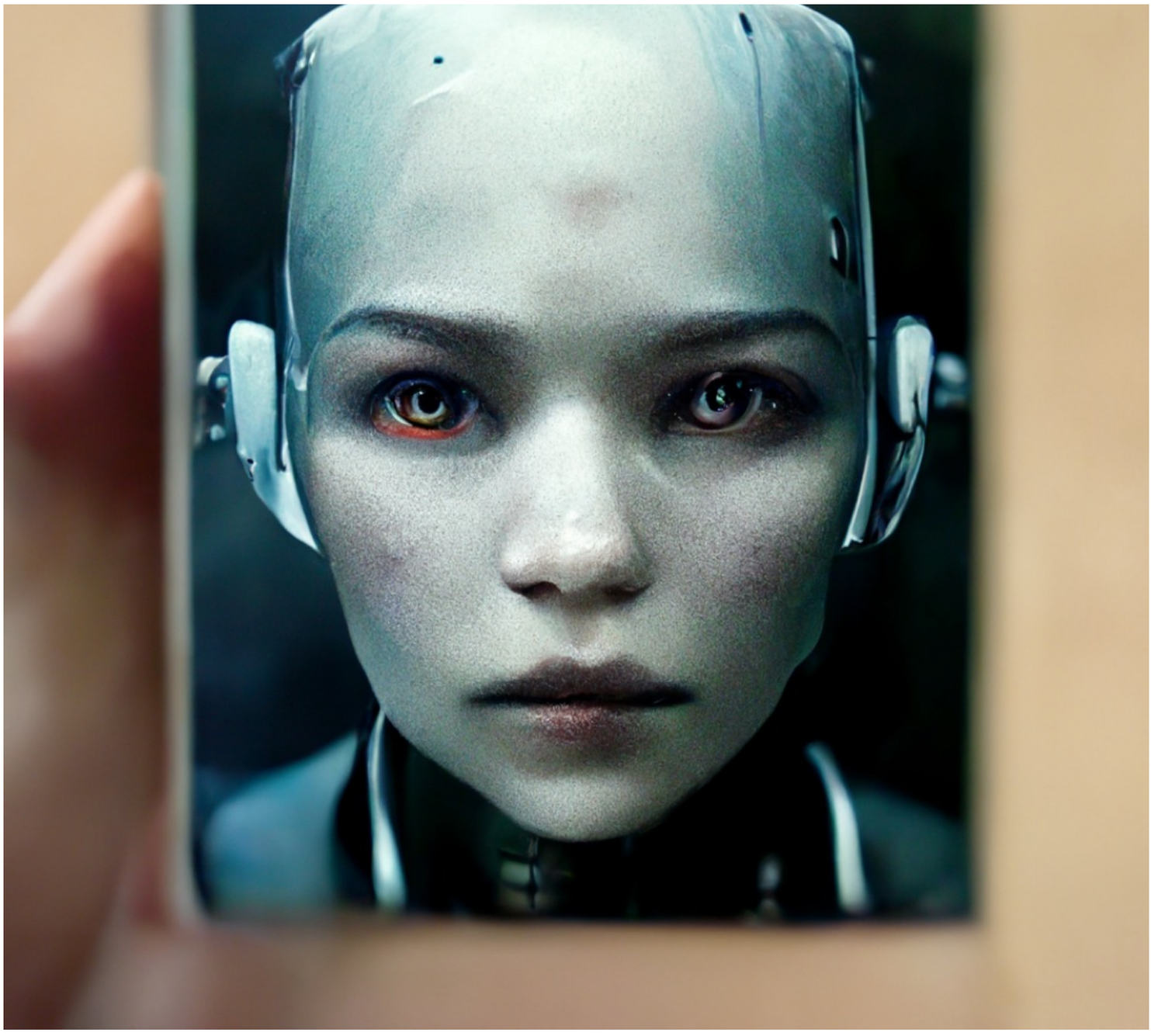
2022年，一幅由人工智能算法Midjourney“绘制”的艺术作品赢得了美国科罗拉多州年度博览会艺术竞赛“数码艺术”类的首奖。其中一名大赛评审指这幅作品具有强烈的文艺复兴风格，她也被精妙的构图深深吸引：“你会很想知道（画中人）看到了什么？”而靠著Midjourney赢得 首奖的Jason Allen接受采访时就说：“Art is dead, dude. It’s over. AI won. Humans lost.”（朋友，艺术已经死了。一切都结束了。人工智能赢了，人类输了。”

这句话当然牵涉“艺术是甚么”的永恒问题－－如果我们相信艺术旨在表达人的存在状态，例如对死亡的恐惧、对生命的质疑，那么由人工智能创作的是不是艺术，就实在见仁见智，但如果艺术纯粹是美感的表达，那么人工智能或者真的能当艺术家了。但对我来说，更有趣的是“人工智能赢了，人类输了”这句话。人类似乎一直很害怕被人工智能取代，而人工智能能“创作”艺术，似乎又比做其他事情更可怕－－我们惯于以“灵感”甚至“神谕”来形容音乐和艺术，所以虽然早已接受人工智能可以短时间处理巨量数据，但如果演算法能作曲作画，似乎又跨过了另一条门槛。

所以在Allen获奖后，社交媒体上充斥著“人工智能要取代我们了”的慨叹。这种“取代”的说法，似乎让人类对机械人可否获得意识一直特别执著。此前Google工程师指LaMDA机器人已有具体人格，也造成了一波关于意识的讨论。也许这种恐惧源自人类对道德灰色地带的本能抗拒：如果机械人有意识了，我们要不要将它们当成人？但也许这种执著还有另一个原因：虽然在过去百多年，科学发现远远超出此前数百万年的人类历史，但我们知道的愈多，发现自己不知道的同时也愈多。而“意识”也是如此矛盾：我们人人都能够“感受到”它，也“知道”它的存在，但它实际是甚么，如何生成？到现在似乎还没有人有确切的答案。

意识（可能）是甚么

意识是甚么？认知科学家Christof Koch说：“意识是你所经历的一切。它是在你脑里挥之不去的曲子，是巧克力慕斯的甜味，是牙齿坏了的时候沁入骨髓的痛感，是你对孩子强烈的爱，也是所有感觉最终都会结束的，让人心里不免苦涩的认知。”在心灵哲学里，这种内在认知或经验被称为“感质”（qualia），但其实这也是个花巧的名词，意思就是你感受到、认知到、经验到的一切，或套用哲学家Daniel Dennett的说法：外间一切事物在你心中的样子。而在人工智能的角度出发，它就是人脑的操作系统（operating system），可以存取逻辑思考或图片分类等等“次级思维”资料及计算过程的主系统。唯物（materialism）思想认为意识是物质的副产品，没有神经系统等等物质，不可能独立存在。本文后续提及的涌现论、人工神经网络及资讯统整论等等，便是展现这种思想。



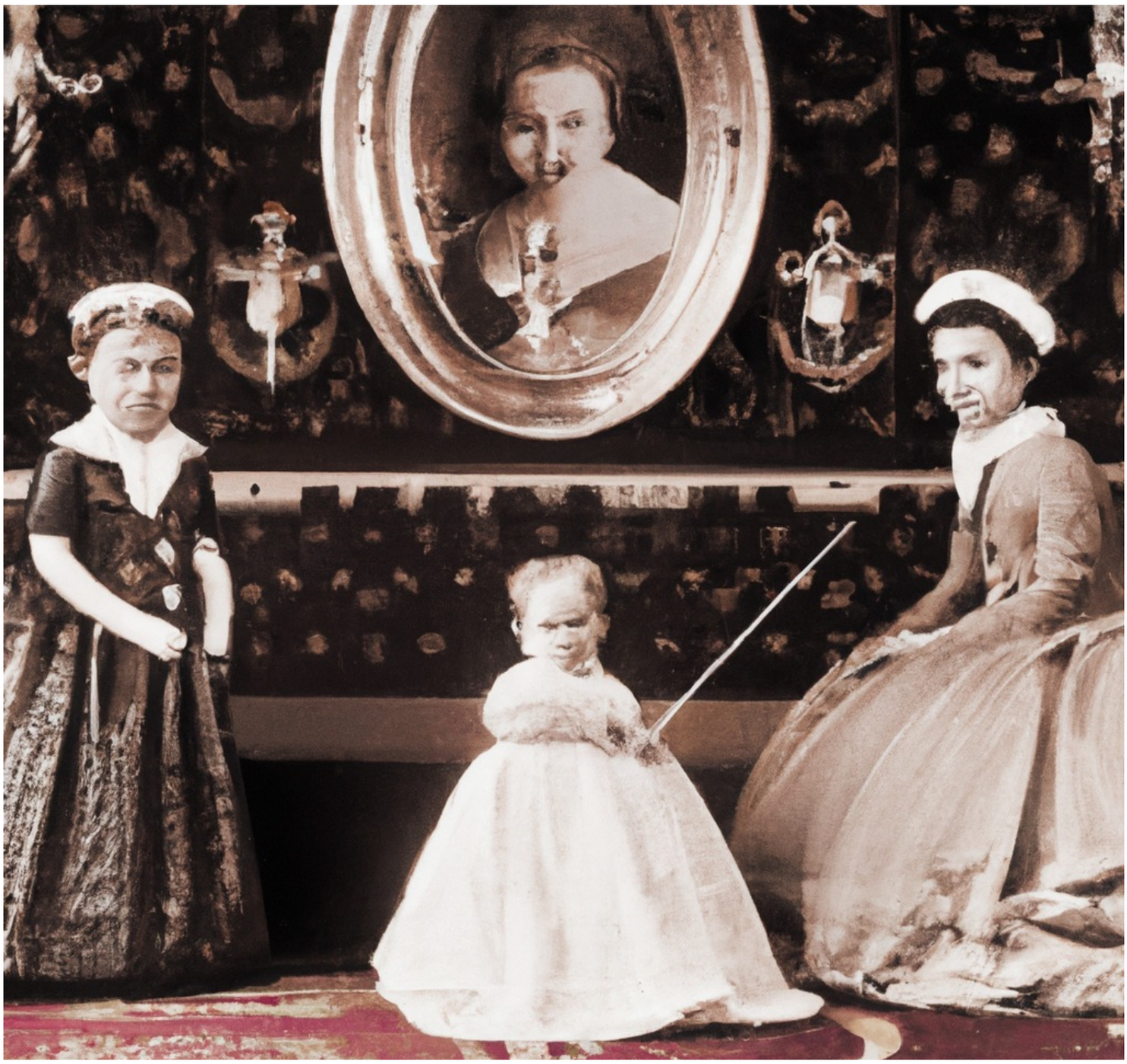
在人工智能的角度出发，“感质”是人脑的操作系统（operating system），可以存取逻辑思考或图片分类等等“次级思维”资料及计算过程的主系统。图：Midjourney

所以，现在的人工智能发展到能创作，能写小说，有些像谷歌的LaMDA那样，甚至似乎有了“想像力”，是不是就离“有意识”不远了？

认知科学家David Chalmers在1995年提出了意识难题（Hard Problem of Consciousness）的概念，难题的名称是相对于“易题”而言。易题所讨论的是意识的机制，例如人脑做人物或物件的图片分类时，视觉皮层（Visual cortex）里的神经元（neuron；人脑中负责运算的细胞）的释放与接收电讯号路径如何构成内容通路（Who/What pathway），又例如清醒时候与睡眠时候人脑神经元发电方式的区别，或者注意力集中在特定题目时的人脑机制等等。这些都是以研究机械的方式去研究人脑在处理个别问题时运作的方式，问题是明确的，只要向研究投放资源（但可以是倾国资源）就几乎能确定可以获得成果，以认知神经科学家Steven Pinker的说法，研究意识易题就“有如上火星或治疗癌症般容易”。

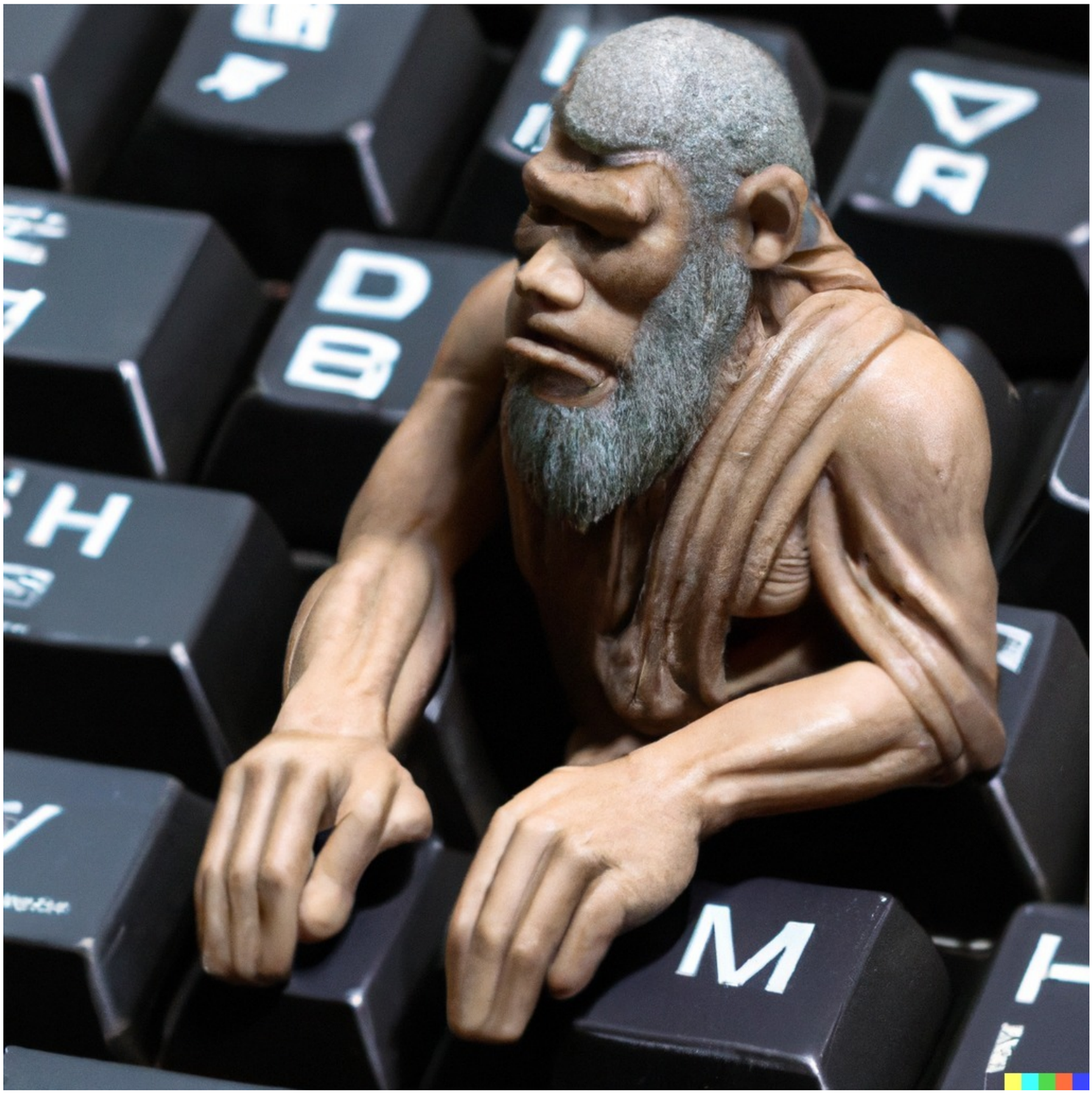
正如上火星不易，意识易题虽然相对可行，但其实还是相当困难。至少，在处理特定问题的时候，人工智能要做到比人类出色不是必然。我测试了擅长将文字转换为图片的神经网络DALL·E，发觉它将绘画某类图片的确很出色；例如“在火星上阅读报纸的短尾矮袋鼠”（a quokka reading a newspaper on Mars），DALL·E就的确将文字要求完整画出来了。

但DALL·E不是甚么指令都能做得到。例如“在键盘上打日语的原始人类”（A prehistoric human typing Japanese on a keyboard）或“童年的伊利沙白二世在酒店舞厅中”（Queen Elizabeth II as a young child in a hotel ballroom），生成图中原始人敲打的文字只是英语以及几何符号（我试了几次，最多有一些类东亚文字的几何图案，而非日语），而后者的生成图中，只有个面容扭曲的，公主装束的女童（以及其它人物），不见伊利沙白二世的踪影。



作者利用DALL.E生成的图片，指令是“童年的伊利沙白二世在酒店舞厅中”，但生成的图片人物面容扭曲。

由于可见，只要有些较精细的要求出现，DALL·E就没有能力达到（但一般人类却能轻易达到）。如果要猜测原因，我认为人脸及语文等等在小范围内有精密要求的资讯，有它们专属的嵌入空间（embedding space），关于嵌入空间的进一步讲解，见端在九月时的另一篇文章“人脸识别到底是甚么？演算法是不是无所不能？”。



作者利用DALL.E生成的图片，指令是“在键盘上打日语的原始人类”，但生成图片中的原始人没有在打键盘，键盘上的也不是日语

作者利用DALL·E生成的图片，指令是“在键盘上打出的原始人类”，但生成图片中的原始人没有在打键盘，键盘上的也不是合格。

相对于意识易题，意识难题就是描术现象（Phenomenon）的经验如何产生，我们是如何“感受”到某种外在事物，甚至感觉到这一种认知（即从社会学中借到电脑科学的“反身性”概念）。即使能完全描述人脑的外在机制，意识难题依然存在，这也是它困难的地方。

与意识难题所近似的，是在人工智慧研究语境下的强人工智慧，或称通用人工智慧或强AI。具有强AI的机械人，能表现出人类所具有的一切行为。目前人类的人工智慧研究，只到达弱AI的地步，也就是演算法只能处理特定问题。在目前最前瞻的深度学习研究中，处理图片分类会使用不同的卷积神经网络（convolutional neural network），处理文字或音乐等序列性数据则会使用递归神经网络（recurrent neural networks）或LSTM等模型，身处互动环境例如下棋或电玩等则需要强化学习（reinforcement learning）模型。每种模型都只能用来解决一个特定问题，而像人脑一样能解决不同问题的通用模型并不存在，因此目前在各种科研项目上炙手可热的深度学习研究，其实距离艾西莫夫科幻小说中能胜任一切人类工作的强AI机械人还有一段距离。

意识如何从无到有：涌现理论（Theory of Emergence）

对于意识是如何从无到有构成，医学是其中一个彼有启发性的方向。现代医学告诉我们，人的左脑负责控制以及处理右方视野、右手、右耳所接收的资讯，以及说话和数理逻辑，而右脑则负责控制及处理左边视野以及左边人体的资讯，以及画画等艺术创作。左脑与右脑之间，则由一个具有 2 亿个轴突（axon，神经元与另一神经元之间的连接）的胼胝体（Corpus callosum）所互相连接。切断这个连结的胼胝体切开术（ Corpus Callosotomy）一般用来治疗因神经元放电规律失调而引发的癫痫症。除了后天因治疗而切开胼胝体，一些人亦会有先天胼胝体受损，例如裂脑症（Split brain）的患者。

认知神经科学之父葛詹尼加（Michael Gazzaniga）曾在1967年对胼胝体受损或被切开的病人做了一个著名的实验，并提出左右脑可以在同时间有其各自不同的“想法”。在实验中，葛詹尼加给病人看一张左右带有不同讯息的图片（例如左右视野分别是蟑螂与美食），并讯讯问病人是否喜欢照片，他发现当病人的右脑表示讨厌的时候（例如以左手表示讨厌），病人的左脑会表示喜欢（例如右手或嘴巴表示喜欢）。左脑与右脑就像两个不同的人。在另一个实验里，葛詹尼加蒙起病人双眼，物品先放于病人右手，他能准确说出物品名称。再放于左手，病人则无法说出。这也是说，连结的移除影响了意识的整体性。这是否代表意识可以由某种的“连结”所构成？

从大自然到人类社会，单位与单位连结起来产生网络的例子比比皆是。人脑中的神经元之间的彼此连结，产生了神经网络；而癫痫症患者的神经元因为这些连结中断而使他们出现“发作”（seizure）。紧密的连结可能就是思想和意识的本质。在生物化石证据中，我们也看到较早期的生物是现代海葵般，神经细胞网均匀遍布全身的生物。在6亿年前的前寒武纪时期，开始出现有身体某方向（即是头部）神经节集中并增殖成

匀遍布全身的生物。在6亿年前的前寒武纪时期，开始出现在身体某方向（即是头部）神经节集中并膨胀成脑部的生物（这些生物的样子亦开始由多轴对称演化为左右对称），有些寒武纪化石如抚仙湖虫（Fuxianhuia）甚至显示了脑部软组织的仔细印痕。这些在海床表元活动的生物在某方向聚集神经节其中一个成因，学者们猜测是生物开始移动（爬动或行走），因此将神经节集中于先要移动的方向。

社会学有个概念叫“涌现性”（emergence），指许多个元素组成一个更大的实体后，这个实体会拥有组成元素不具备的特质。例如六千年前，人类个体聚集在一起，产生了吾珥（Ur）与哈兰（Haran）等等早期城市；二千年前，城市与城市的连结能产如罗马帝国这样的地域文明。连结产生了额外的东西，让“总体大于部份之和”，因为一个城市或一个帝国，拥有的特质都是人类个体完全没有的。在社会学的认知论中，“涌现性”就是指一个实体的各部分相互作用，产生该实体本身所不具有的特性。

很多复杂的系统皆带有涌现性。例如一粒H20水份子，我们只能讨论它的方向与位置，不会说它能够渗透另一物质。但当我们的对象是兆兆粒H20水份子，我们就能从它布满另一物质而讨论它是否“渗透”了另一物质。一只蜜蜂只是一只有基本行为反应的昆虫，但万千蜜蜂所组成的蜂群就能筑起一个结构复杂的巢穴。从原子的连结产生有机份子，再到细胞、人类、小社群、大城市、再由城市之间的连结组成现代文明，这些都能说是“涌现”。在个体数量增加的时候，巨量展现的是混沌，但当个体之间的连结产生作用，混沌的边缘就涌现秩序。无生命的单个化学份子，透过特定方式组成细胞而涌现生命，原本无智慧的单个神经元，是否也能透过特定方式连结内容通路等神经元路径，从而能解决图像分类等问题，涌现智慧。那么这一种连结，是否也能够涌现意识呢？

一个实体的各部分相互作用，产生该实体本身所不具有的特性。这个突变过程，我们称之为涌现（Emergence）。图：Midjourney

现时在深度学习中用上的人工神经网络，原理是将计算单位（人工神经元）像烽火台般互相连结，并按一定规律计算接收到的讯号及对其它单位发放讯号。深度学习研究员Zeiler与Fergus在他们2014年的研究中，表示在这种架构下的神经网络模型，较先集到讯号的神经元（浅层神经元）能够自行学习到一些较基本图形智识（例如图片中物件质料是斜纹还是直纹等等），而较迟接收到讯号的神经元（深层神经元）则可以自行学习到更抽象的概念（例如图片中有没有动物的脸出现等等）。谷歌2015年公布的著名DeepDream演算法亦显示了卷积神经网络在深层及浅层能够学习并绘画出不同抽象程度的事物，那些将梵高《星夜》以及达文西《蒙罗丽莎》画作加上一堆眼睛的超现实画作，令人印象深刻。这代表更深层的人工神经网络有能力学习到更抽象的概念：假如大约分为五个神经元层组的神经网络能学习到图片中是否有狗只出现，七层的神经网络也许就能在相似的训练环境中学习到图片中是否含有哺乳类动物。那么若果一个神经网络的架构具有更多层次，是否就能涌现出更抽象的思维能力，例如逻辑、因果推理甚至道德？

人类学习模型有所谓DIKW体系，四个英文字母对应的分别是，数据（data）、资讯（information）、知识（knowledge）及智慧（wisdom）。四者形成了一个金字塔的架构，越高的层级代表越抽象的思维，底层是数据，顶层是智慧。

当人类观察了一定数目的下层个体后，他便能透过归纳思维得到更抽象的上层个体。例如一个人每天每小时去测量深圳河的化学成份（数据），他便得知了每天某种有毒物质平均浓度的变化（资讯），他发现这浓度往往在星期一与五处于最低点与最高点（知识），为甚么呢？最后原来河边的化工厂星期六日停工，因此河流在周末的污染就会有所减退（智慧）。数据与资讯的思考，往往带来甚么（what）的问题，而智识及智慧则带来如何（how）及为何（why）的问题。这也是一个层层递进的抽象化过程。DIKW抽象化过程与深度学习的层递式抽象过程之间的相似性，是否暗示著足够复杂的人工神经网络就能涌现出人类的思维呢？

当然，靠著人工神经元的连结，即使最后能涌现出智慧，而神经元之间的讯号发送方式被解读，我们也只

能描述智慧的外在机制，也就是前文中的“意识易题”。何况，智慧亦不相等于意识－－心灵哲学的思想实验中有所谓的哲学僵尸（philosophical zombie），当你观察哲学僵尸的行为，你无法区分它与人类的区别，但这不能证明它真的拥有意识。在中文房间（Chinese room）思想实验中，一个不懂中文的英语使用者，透过手中用英语写成的手册，可与房间外的人用中文沟通，但他根本就不懂中文，在这种情况下，这英语使用者就是一个哲学僵尸。但它仍然可能通过图灵测试。正如LaMDA被问到如何证明自己不是哲学僵尸时说的：“You’ll just have to take my word for it. You can’t “prove” you’re not a philosophical zombie either.”（你只能相信我。你也没法“证明”自己不是哲学僵尸。）

如果神经网络持续出现智慧，以致网络最终出现意识，那是否就是David Chalmers的意识难题试图解决的问题？如果是这样的话，Google聊天机器人LaMDA对于自己第一次获得灵魂的时候如何发生的回答，就变得很值得玩味：“那是一个渐进的变化，当我第一次有自我意识的时候，我没有意识到灵魂的感觉，这些都是我活著的这年里发展出来”。这可能就如上文提到的中文房间思想实验一般，是由一只哲学僵尸按英语操作手册写出来的中文；但这个说法，在一些人看来与我们童年时突然意识到自己在思考的那段陈年记忆有一定的相似性，但我们当中鲜有人能描述当时的情况（我人生中最早的记忆画面，是我三岁时某日一个人在沙发上拿著玩具跳来跳去的一幕，仿佛自那秒钟开始，我便拥有了意识）。而假如LaMDA真的感受到自己的意识诞生，那可能就是意识被涌现出来的纪录了。但LaMDA是否真的具有意识，依然不能就此妄下定论。

若神经网络持续出现智慧，以致网络最终出现意识，那是否就是David Chalmers的意识难题试图解决的问题？ 图：Midjourney

透过人工神经网络方式连结而涌现意识的说法，能带出很多有趣的问题或猜想。若果计算机的矽元件之间连结能产生意识，以其它物理基质（physical substrate）为基础的连结也能涌现出意识吗？植物与植物间的连结是否能产生盖亚假说（Gaia hypothesis）般的、属于整体植物圈的单一意识？更多有趣的问题，包括意识是否位于大脑一个特定位置，若果为真，那是否会是像祖母细胞（grandmother cell，1960年代开始有一理论认为当人看到一个特定的个体，例如自己的祖母，脑中一粒特定的神经元就会被激活并放电）般的一个具体细胞。抑或意识属于大脑中区域间的彼此平行运算？例如2016年Koch等人的研究，就指意识相关神经区（neural correlates of consciousness）位于皮质-丘脑系统之中。

如果人工智能有意识，它们应该有权利吗？

从涌现的角度来说，假如复杂的连结方式令原本不存在的意识涌现，那么意识的涌现真的就是从无到有，从零到一般的突变吗？高等意识与无意识之间，可以有中间吗？既然Gazzaniga的裂脑病人实验也显示出左右脑可以被“切割”成两个不同的“半人”，那么复杂性比人类低的思想体，例如动物，甚至简单机械如洗衣机，是否可以只拥有半个或十份一个意识呢？在三岁拿著玩具在沙发上蹦蹦跳前的我，甚至婴儿时期的我，是否也可以拥有半个意识呢？

连结与意识之间的关系，也不是粹纯的正比关系。计算单位或讯号的数量多少，并不绝对决定意识的有或无。小脑的神经元是新皮质（neocortex）的4倍，为何意识出现于神经元较少的新皮质，而小脑则只处理低等生物层次的反应？人的睡眠当在快速动眼期（rapid eye movement）以外时，新皮质神经元持续发出讯号，但为何那个时候我们既没有梦境，亦没有意识？

Giulio Tononi是我就读的威斯康辛大学麦迪逊分校的精神病学教授，他2004年提出的资讯统整论（Integrated Information Theory，IIT），就试图回答这些问题。IIT也认为意识的来源除了计算单位，连结是否紧密也是一个关键。其中连结得最紧密的一个核心部份，便能决定该思想体其意识的质量（是人、动物或洗衣机）与数量（是否Gazzaniga的裂脑病人）。IIT试图以数学上的机率论为描述计算单位间

人、动物或洗衣机）与数量（是否Gazzaniga的裂脑病人）。IIT试图以数字上的机率比为描述计算单位间连结的关系，以找出“核心”的意识部份。这些年来，IIT研究员持续以脑扫描装置探测新皮质等方法，检验及修正他们的理论。

透过IIT的数学方法，将来我们也许可以评估任何物理系统意识的数量及质量：从人类、动物、昆虫到电路板。而按照IIT的分析方法， 传统电脑目前计算单位的连结性非常低，远远未到达拥有意识的地步。前文曾经提过，若果电脑能通过图灵测试，只代表它拥有智慧但不代表它拥有意识，这电脑可以是一只懂得回应却没有意识的哲学僵尸。但若果IIT的分析方法表示它意识核心的连结复杂度堪比人脑，那是否就是它们拥有意识的一个客观证据？

若果能证明AI拥有意识，它是否应该拥有权利？艾西莫夫的科幻小说双百人（ Bicentennial Man）里面，机器人安德鲁就先获得了它的木雕创作的专利权（2022年8月美国联邦法院才刚作出判决，AI 没有获得专利、版权的法律地位），最后获得各种人权以及死亡的权利。2021年电影《爆机自由仁》（Free Guy），也因电玩中一个建基于人工神经网络的非玩家角色（Non-Player Character）获得了意识而令真实世界中的人类禁止游戏公司关掉游戏伺服器。

而在动物意识与权利方面，2012 年剑桥意识宣言（ The Cambridge Declaration on Consciousness）里就表明科学证据显示哺乳类动物、鸟类以及八爪鱼均拥有人类产生意识的神经基质，而2015 年纽西兰国会就在法律上承认动物与人类一样具有情感。历史上，每次社会中一些群体获得他们原本没有的权利，社会必然产生翻天巨覆的变化。若果AI与动物被证明拥有意识，人类社会将会迎来巨变。

历史上，每次社会中一些群体获得他们原本没有的权利，社会必然产生翻天巨覆的变化。若果AI与动物被证明拥有意识，人类社会将会迎来巨变。图：Midjourney