·当代哲学问题探索·

人工智能可能具有想象力吗?

——以三木清哲学为出发点

徐 英 瑾
(复旦大学 哲学学院,上海 200433)

摘 要:最近以 ChatGPT 为代表的人工智能聊天软件正在成为舆论关注的焦点。此类人工智能产品是否能够在人—机对话中展现出真正的想象力呢?这就需要先对“何为想象力”这个问题进行哲学层面的界定。基于日本哲学家三木清的“构想力”理论,想象力的特征可以被归结为“具身性”“意志性”“实践性”“社会凝聚性”四个方面,其中“具身性”的地位尤其关键。但作为一个大型语言处理模型的 ChatGPT 本身就是非具身性的,这一点在根本上决定了它无法理解那些基于身体感受的语言表达式的真正含义,遑论对基于这种感受的其他的感受样态进行合理的想象。由于人工智能与碳基生命之间在物质构成上的本质性不同,ChatGPT 的这一缺陷也会在未来具有更多“具身性”特定的人工智能体中得到部分的保留。因此,从原则上看,人工智能永远不能拥有完整意义上的人类想象力。

关键词:ChatGPT;人工智能;深度学习;三木清;想象力;具身性
中图分类号:B0 文献标志码:A 文章编号:1002-462X(2023)04-0017-11

最近以 ChatGPT 为代表的人工智能聊天软件正在成为舆论关注的焦点。ChatGPT 的全称是“Chat Generative Pre-trained Transformer”(含义是“预训练的聊天生成转换器”),其本质是基于一种深度学习技术的人工智能聊天程序。ChatGPT 目前主要以文字方式与用户交互信息,能够完成诸如自动文本生成、自动问答、自动摘要等在内的多种任务,且据说已经具备了富有一定想象力的应答方式(比如,该软件能够根据不同用户的不同提问,从不同角度给予灵活的应答)。这就引发了下面这些具有哲学面相的问题:以 ChatGPT 为代表的当下人工智能软件已经具有了想象力吗?如果没有,为何没有?未来的人工智能又是否可能拥有这种能力?

要回答上述问题,我们就要对“何为想象力”这一问题进行哲学预研。“想象力”貌似是一个心理学话题,却在康德、胡塞尔、海德格尔等哲学大家那里都得到过严肃的讨论。而在 20 世纪的所有哲学家中,以“想象力问题”为自己最大学术立足点的哲学家,莫过于日本京都学派的成员之一三木清(1897—1945)。三木清按照他自己的话语习惯,将“想象力”写成“构想力”,[①] 并以此为主题完成了其哲学代表作《构想力的逻辑》。[②] 与德国古典哲

① 笔者更倾向于将三木清给出的日语表达式“構想力”翻译为“构想力”而不是“想象力”。这不仅仅是因为这样的翻译能够直接对应日语原文的汉字,也是因为日语中本是有“想象力”这个说法的,因此,如果我们将“構想力”翻译为“想象力”的话,就会造成某些混乱。另外,熟悉德语的三木清也有将“構想力”与德语表达式“Logik der Einbildungskraft”相互对应的意图,而德语中的“Einbildungskraft”的字面意思便是“构成图像的能力”。

② 1939 年,三木清出版了其哲学代表作《构想力的逻辑:第一部分》(『構想力の論理 第一』),而该书的第二卷直到 1948 年才以遗稿的方式出版。两卷在战后一般是合在一起出版的。参见三木清『三木清全集: 第 8 卷:構想力の論理』(岩波書店 1966 年版)。

基金项目:国家社会科学基金项目“对于通用人工智能与特定文化风土之间关系的哲学研究”(22BZX031);国家自然科学基金项目“探索研究 AI 伦理对科研环境的影响”(L2124040)

作者简介:徐英瑾,复旦大学哲学学院教授。

学、现象学对于“想象力”的先验主义讨论传统不同，三木清的“构想力”概念带有浓郁的历史唯物论与人类学、社会学色彩，因此，从三木清哲学出发，我们更容易对人类的“想象力”本质有一个更全面的理解。考虑到以 ChatGPT 为代表的当代人工智能技术是以“通用人工智能”为远期目标的，对于“人工智能是否（可能）具有想象力”这个问题的哲学思考，显然也应当以自身的触及面更广的三木清“构想力”理论为哲学参照系。

考虑到三木清哲学在国内的知名度还不高，下文的讨论将始自于对于其“构想力”概念的介绍，然后再迂回到对于人工智能的评论上。

一、三木“构想力”哲学概述

对康德哲学有所了解的读者或许知道，“想象力”是一个四两拨千斤的哲学问题。众所周知，康德知识论的基本观点便是预设知性与感性之间的区分：前者类似我们做月饼时候的模具，后者则类似被镶嵌到模具里的馅料，两者要互相配合才能构成完整的“月饼”（即知识）。不过，康德的麻烦也来了：知识的形式与质料既然是彼此异质的，又怎么可能毫无间隙地彼此结合在一起呢？所以，他就需要某些中介者来使得两者关系得以协调。换言之，这种像“月下红娘”一样的中介者必须既像知识范畴那样具有初步的形式，也要像感性材料那样可以在时—空形式中延展。康德最终找到的这位“月下红娘”便是“想象力”，也就是说，感性材料需要想象力的预加工，才能向知识的形式输送。

若读者觉得上述说法还是过于抽象的话，我就拿教学心理学中的案例来加以说明。一个老师如果要向一名幼童展现简单的加减法，那么，他就需要在幼儿质朴的心智与抽象的数学符号之间找到某些中介者。这一中介者便是类似积木这样的教具——积木彼此累加的图景既是感性的，又能为更加抽象的符号运算提供雏形。而在经过一段时间的算术训练后，积木叠加的图景就能内化为一种想象的对象——而我们之所以有时候会说“某些天文数字是难以想象的”，恰恰也便是因为日常生活的想象力练习一般不会触及这些大数。

不过，康德通过想象力为知性与感情搭桥的做法毕竟是“分而后合”，还是有“亡羊补牢”之嫌。更彻底的做法是将想象力视为知性与感性的共通根据，重构康德的知识论体系。这一努力在海德格尔的名著《康德与形而上学问题疑难》①中得到了初步展现。然而，海德格尔毕竟没有在不谈论康德的前提下独立地完成一部以想象力为主题的哲学著作。与之相较，熟悉德、法哲学脉络的三木清则试图将想象力，或者用他的术语来说即“构想力”的基础地位予以进一步凸显。他写道：

> 在构想力中，知性成分是与感性成分结合在一起的。根据里博（Théodule－Armand Ribot）②所言，构想力总是包含知性要素和感性要素，是两者的内在的统一。在构想力自身之中，包含着内在性且生成性的知性的要素，在这一点上，构想力是同感情有别的。因此，构想力的哲学，既非单纯的理性主义，也非单纯的非理性主义。构想力的逻辑，与其说是感情的逻辑，毋宁说是形象的逻辑。形象是动态发展之存在。构想力的逻辑并非静态的逻辑。之所以说“形象是动态的发展”，是因为它本就是通过综合感情与知性、主观与客观而生成的东西[1]46。

对于《构想力的逻辑》中这段话的解读，也可以结合三木清在 1939 年发表的一篇名为“历史的理性”的演讲来进行[2]249-269。三木清在这篇演讲中如此批评了那种黑格尔式的唯心主义历史哲学理论：黑格尔将个体视为世界理性展开自身时所使用的工具或者木偶，却全然忽略了个体的热情与欲望在历史发展中所扮演的角色。然而，在历史中真正起作用的毕竟是那些具身化的、能够实施行动的个体，因此，黑格尔的历史模型便是一个完全离地空转的车轮。但是，三木清也不想通过对于黑格尔的批判而直接滑向尼采式的唯意志主义，换言之，他不想由此破坏西方文化中两大要素之间的隐秘联系：其一乃是“逻格斯”（意思是

① 此书的中译本参见海德格尔《康德与形而上学疑难》，王庆节译，商务印书馆 2021 年版。

② 法国实证派心理学家（1839—1916）。——引者注。

“理论思维力”，亦与语言有关）；其二则是“帕索斯”（意思是“情绪感染力”）。由此，他的路径就是诉诸“构想力”这个概念。构想力一方面当然是带有情绪性的，但也并非与语言以及逻辑无关（譬如，作家的构想力就显然是在特定语言逻辑的约束下进行的）。此外，构想力的“造像能力”亦有能力使得被由此造出的“像”超越个体层面，而成为集体行动的黏合剂（譬如，在某些历史机缘的帮助下，一个人的梦想就能成为一个民族的梦想）——这就使得一种基于构想力的哲学可以自然地衍生出社会哲学与政治哲学的维度。这种基于构想力的哲学甚至还能具有科学哲学的维度，因为很多天才的科学理论恰恰都是在构想力的推动下产生的（如门捷列夫在梦中对于元素周期表的构想、凯库勒在梦中对于苯环的构想，等等）。

这里需要注意的是，在三木清的思想发展过程中，原本构成“逻格斯”与“帕索斯”之共通根源的概念并不是“构想力”，而是“基础经验”概念——此概念的出现，又明显是三木清受到作为京都学派的头号人物西田几多郎（1870—1945）的“纯粹经验”概念影响的后果。需要注意的是，受到马克思主义影响的三木清一开始就试图在一个更为接近历史唯物主义思维的向度上理解“基础经验”，而相关思想探索的时间则大约是在1928年。对于三木清与西田各自的“经验”观，独立学者斯庄巴克（Dennis Stromback）曾进行过非常仔细的比较[3]。结合他的研究成果后，笔者作出这样的判断：三木清的“基础经验”概念乃是对西田“纯粹经验”概念进行马克思主义化改造后的产物——随着时间的推移，为了进一步彰显自己的思想与西田之间的差异，他又放弃了“基础经验”这个提法，开始使用“构想力”这个更新的提法。因此，对于西田与三木清各自的“经验”观差异的了解，便能够为我们提供一扇窗户，以便理解三木清的“构想力”概念与马克思主义哲学之间的渊源。

现在先让我们来概述一下西田的“纯粹经验”是什么意思。① 非常简要地说，其是一种在禅宗式的冥想练习中达到“物我两忘”的特殊精神境界。而西田在成熟期间形成的“场所逻辑”，② 则是对于这种“纯粹经验”的逻辑特征的补充性描述。需要注意的是，与三木清的“构想力”概念一样，西田的“纯粹经验”亦含有统摄主—客对立，甚至是“逻格斯”与“帕索斯”之对立的意蕴（具体而言，其“逻格斯”的一面体现为“场所逻辑”，其“帕索斯”的一面则体现为与禅宗式的宗教审美体验之间的密切关联）。因此，与扮演同样功能的三木清式“基础经验”或“构想力”概念相比较，西田的这个概念便是一个恼人的理论竞争对手。

三木清与该理论对手竞争的方式却也并不复杂，而是直接亮出他心目中的“纯粹经验”的阶级属性，并由此发明了一个叫“无产者的基础经验”（無産者的基礎経験）的新表达：

> 使得无产者的基础经验的构造从根本上得到规定的，便是劳动（さて無産者の基礎經驗の構造を根源的に規定するものは勞働である）。无产者通过特定的交互方式，即感性的存在，同存在进行交涉。在这时候，如若劳动要维持其本质，那么那些劳动者所持有、并与之共同劳作的东西就不能是像“物在心中之映像”这般观念性的东西。在其存在之中，实践本质地必然地要求：实践对象乃是同进行实践者不同的独立的存在[4]。

换言之，三木清所说的“经验”并不来自禅宗高僧在茶室里的那种与现实隔绝的神秘宗教体验，而直接来自普通劳动者挥汗如雨的乡间与车间。因此，两个重要的因素就进入了三木清的“基本经验”概念的统摄范围：一是具身性，作为劳动者身体之一部分的双手与相关劳动工具的实际运作，才能使得主观的意识与外部物质世界得以达到统一；二是特定的社会建制，正因为现代工业条件下的劳动肯定是在复杂的社会分工模式下

---

① 这一概念出现在西田的代表作之一《善的研究》（何倩译，商务印书馆1965年版）之中。

② 这一概念出现在西田于1926年发表的论文《场所》中。此论文英译本收录于：Nishida Kitarō，*Place and Dialectic*，translated by John W. M. Krummel and Shigenori Nagatomo，Oxford University Press，2012.

进行的，所以，劳动者所产生的基础经验内容就不可能不打上具体历史条件的深刻烙印。而这两项要素都是西田的“纯粹经验”概念所相对匮乏的，却又分别对应三木清本人所心心念念的“帕索斯”与“逻格斯”——这又是因为身体的运作天然与情绪相关联，而任何分工形式又都天然具有一种内嵌式的分工逻辑。

在三木清将自己的核心哲学概念从“基础经验”转向“构想力”之后，两个新的因素也进入了他的理论视野：其一是基于神话特定民族国家的历史，其二则是模仿的力量。先来看第一个要素。为了理解“构想力”与特定民族历史之间的关联，请大家先思考一下美国学者本尼迪克特·安德森（Benedict Anderson）在其名著《想象的共同体——民族主义的起源与散布》[5]中所提出的观点。安德森指出，即使是一个很小的民族国家，如果没有集体想象力的参与，便无法在观念上成型。举个例子来说，尽管一个生活在北海道札幌的日本拉面师傅可能一辈子都没有见过住在冲绳的日本人中的大多数，但他依然会在观念上将一个住在冲绳县的与那国岛的不知名的日本渔民视为他的同胞。而前者之所以能够这么做，是因为他已经接受了“日本之为日本”的集体民族想象。从某种意义上说，安德森的这一思想已经被三木清所预报，而其具体的预报形式则是三木清对于神话问题的讨论。

很显然，神话所提供的想象，为民族国家的形成提供了重要的支点。具体到日本这个特殊的国家案例上，《日本书纪》对于“天御中主尊”“可美苇牙彦舅尊”等神之神谱的描述就对统一日本国民的民族意识起到了重要作用。而在一个更抽象的理论层面上，三木清楚地意识到神话的各种副产品，如相同的图腾崇拜、相同的神名、相同的神话叙事结构、相同的宗教音乐等，在凝聚人心方面所起到的作用。因为恰恰是这些作为中介物的观念副产品的存在，才能够将那些未必能够有直接物理接触的个体有机地联系起来。用三木清自己的话来说，“神话这种活动能够带来一种媒介性的力量，以便将一种不再被直接感受到的社会参与变得现实”[1]24。

再来看第二个因素：模仿。三木清既然是在个体之间彼此协同的语境中讨论构想力的，他就无法不提到模仿，因为构想力所构成的“像”的社会学传播是离不开一个个体对另外一个个体的行为的模仿的——而为了使得这种人传人的模仿不至于立即变得走样，特定的社会规则（特别是语言表达的规则）就应当被确定下来。因此，与神话所带来的“天马行空”色彩——同时也便是一种“帕索斯”色彩——相比，模仿活动便天然就更具备“逻格斯”的面相。关于模仿活动所应当具有的逻格斯形式，三木清特别提到了一些貌似很符合文化保守主义者胃口的字眼，比如“习俗”“习惯”“惯例”（日语“慣習”），等等[1]102-103。不过，他亦强调对于法则的尊重与创新之间的辩证关系，因为在他看来，任何技术创新本质上都是人类的个体欲望（帕索斯）在既定社会建制（逻格斯）的帮助下所实现的。或用他自己的话来说：

> 尽管自然法则总是在自然界中起效，但是自然本身却不能促成电灯或者电力汽车的发明。为了使得这些东西被发明出来，关于电力的法则就必须要先被发现；而这些法则要被发现，人类的欲望就首先要被导入。这些技术形式本身之被创制出来，其实就是作为客观法则与人类主观意志的综合体而出现的。任何历史事物的出现，概莫如此[2]72。

换言之，是特定民族中特定个体的求知欲，导致了相关科学规律的发现；而利用这些发现去改善生活、发财致富的欲望在一定社会范围内的普及（这种普及本身又会造成类似新教伦理之类的新惯例），则会进一步推动技术革命的发生。

有了上面的讨论做基础，现在我们就可以结合“通用人工智能”这个话题，对三木清“构想力”概念的以下四个面相进行概括。

（甲）涉身性。“构想力”天然涉及对图像的构造，而对于图像的感知又天然涉及与感知相关的那部分身体机能的正常运作。从这个角度看，一个通用人工智能机制要具有想象力，也就必须具有类似于人类身体的感知能力以及对于内部图像的构造能力。

（乙）意志性。“构想力”体现了人类愿意投

入某事的情绪。因此，一个通用人工智能机制要具有想象力，也就必须具有类似的情绪产生机制以及意志投射能力。

（丙）实践性。三木清心目中的构想力或想象力机制是在人类个体与物质世界的接触中产生的，而不是一种纯粹的主观精神活动。在这个过程中，构想力活动会随时根据主体从物质世界中获得的反馈改变其构想模式，因此将实践加以深入。由此，一个通用人工智能机制要具有想象力，也必须具备根据外部输入的变化灵活改变自己输出的能力，由此使得自己的内部图像与外部实在之间的间距被不断缩小。

（丁）社会凝聚性。在三木清的理论语境中，“构想力”通过众人对于共同构想物的分享而扮演了“社会凝结剂”的角色。因此，一个通用人工智能机制若要具有三木式的构想力，也需要通过类似的“图像分享机制”而强化某种意义上的社会团结性。

下面我就要论证，目前以 ChatGPT 为代表的人工智能软件是难以满足上述四项要求的。

二、ChatGPT 具有想象力吗

现在我们就转向对于 ChatGPT 技术的哲学本质的评估。与传统的深度学习技术一样，ChatGPT 系统在哲学上预设了经验论的知识形成模式，也就是说，根据该哲学预设，只要一个认知系统获得了关于外部环境的大量数据，该系统就能通过对于这些数据内部的统计学相似性而自动获取规律，并由此预测未来。说得更具体一点，深度学习技术通过对于人类的神经元网络结构的数学模拟，将“如何对既有数据进行统计学处理”这个问题转化为“如何对一个人工神经元网络进行训练”这个新形式。至于基于深度学习技术的 ChatGPT 技术的创新之处，则是引入了“预训练”这个新的数据处理阶段。所谓“预训练”，就是先将海量的语料“喂给”系统，却不告诉系统要完成什么任务，让系统自己琢磨不同语词的前后搭配关系。比如，系统在处理大量汉语语料后，若发现在出现短语“路遥知马力”之后，短语“日久见人心”出现的概率很高，系统就会在发现第一个短语后期待第二个短语出现。但需要注意的是，系统并不知道这个成语本身究竟是什么意思。而在经过复杂的“预训练”阶段后，程序员若再训练系统完成一些目的明确的任务（如完成一个求职报告），系统的后期训练成本据说就会变得比较小。但需要注意的是，使得这种较好的表现得以可能的“预训练”所需要的语料数量与算力消耗都是非常惊人的（目前 ChatGPT-3 的参数数量已经达到 1750 亿个，预训练需要的资料量是 45TB 文本，需要调用 2850000 个 CPU 以及 10000 个 GPU，训练总费用达到 1200 万美元[6]）。这也就使得类似 ChatGPT 之类的大型语言处理模型的开发门槛变得异常之高。

现在的问题是：这样的机制是否可能具有想象力？根据三木哲学的精神，答案应当是否定的。现在我们就根据前文给出的“构想力”标准来仔细比照 ChatGPT 在哪些方面无法“达标”。

前面已经提到，“构想力”或“想象力”需要对于某种图像具有感知能力。但目前的 ChatGPT 平台主要还是一个大型语言训练模型，因此是无法具有对于图形的感知力的。说得更形象一点，该平台的运作方式，非常类似于心灵哲学家杰克森（Frank Jackson）所提出的“黑白玛丽屋”的思想实验[7]：ChatGPT 就很像该思想实验所提到的那个叫“玛丽”的女孩，从出生起，她就始终被关在一间没有任何真实颜色样本的房间里学习抽象的颜色理论，并以一种与实践脱节的方式学会了对下述问题给出恰当的答案：波长最长的人类可见光是什么？橘色究竟是处在红色与黄色之间的颜色，还是处在绿色与黄色之间的颜色？你能设想有一种玻璃，既是乳白色的，又是透明的吗？靛蓝色与蓝色是一回事吗？

然而，即使玛丽能够正确地回答上述所有问题，她依然无法在心理活动中设想任何一种颜色图像，而这一点又在下面的事实中得到了相关的行为证据：假若将玛丽从上述房间中放出来，第一次进入外部世界的她显然会对生平第一次看到的颜色感到震惊。很显然，此刻的她无法将学到的关于颜色的语言知识与她看到的真实颜色相互联系，并在这种联系的基础上去成功地执行下述命

令:“当你看到红灯的时候,请立即刹车。”同样的道理,对 ChatGPT 来说,它当然抽象地知道“红灯亮,汽车停”的道理,但是它依然无法在实际的交通场景中自如地运用这一知识,因为它就像那位可怜的玛丽小姐一样,缺乏对于红灯的“图像知识”。因此,它就无法建立任何奠基于上述图像知识之上的对于未来行动的构想,比如假若一个国家将绿灯视为停车标志,交通状况又会变得如何呢?

对于我的上述诊断,当下主流人工智能技术的支持者会反驳说:只要训练数据足够大,大型语言处理模型就能够绕开对于图像构造能力的营建,而由此具备通用人工智能。具体而言,系统根本就不需要在内部视野中构造出“绿灯成为停车标志”这一图景,而只需要搜索系统是否获得了相关的语句训练资料,并由此给出合适的输出。支持这一思路背后的深层理由如下:既然语言表述本身也能包含对于未来的想象,那么,只要系统搜集的语料数据足够多,包含在这些语言数据中的人类的想象力产物也能随之被系统吸收,系统也能由此获得对于这些想象性场景的回应能力。也正是基于这种机制,目前的 ChatGPT 已经貌似具有对于部分构想性场景的细节描述能力,比如根据用户的描述,描写一下哥伦布假若穿越到今日的纽约会做些什么事情。

然而,根据笔者的实测,恰恰因为 ChatGPT 缺乏对于图像的实际构想能力,它对于可能性场景的细节描述能力是非常拙劣的。笔者曾要求 ChatGPT 构想下面一种场景,以便协助人类用户完成古装历史剧剧本的写作:假若古罗马政治家布鲁图斯刺杀凯撒的理由不仅是基于政治考量,而且还掺杂私情因素(比如他与凯撒爱上了同一个女人),那么,他与凯撒之间的对白究竟会是怎样的?面对这一“按要求写剧本”的任务,ChatGPT 的响应方式可谓非常“粗暴”:它让布鲁图斯直接对凯撒说出自己杀他就是因为争风吃醋,而完全不顾一个堂堂政治家在公开场合说出此话时所要面对的道德压力。看得更深一点,ChatGPT 之所以不理解这一点,是因为缺乏人类心理机制的它根本无法“看”到“羞耻感”所带来的心理现象,只能根据行为主义的粗鄙逻辑而将用户的要求在字面上展现出来。而“羞耻感”带来的微妙性却恰恰是“因为 x 而感到羞耻”这事本身却恰恰要求当事人在言语层面上回避对于“x”的提及,而不是将其表达出来,因为“羞耻”这词本身就包含有“不能被轻易公开”的语义。这一点乃是基于显白的语言描述的 ChatGPT 所完全不能理解的。

那么,我们是否能够用更多的语言训练材料让 ChatGPT 理解这一点?答案是否定的。因为深度学习机制只能在语言输入与语言输出之间建立起统计学联系,而人类想象力的活动方向却未必是可以被统计学规律所涵盖的。具体到“羞耻感”这一问题上,因为某人耻于谈及某事而“环顾左右而言他”的现象虽在生活中很常见,但作为掩护性事项的“他事”究竟为何,却无法通过大量语料的“轰炸”而让系统获知,只能通过对于当下对话逻辑的悉心体会才能得到合理的推断。在这种情况下,体会者真正需要做的事情,便是基于少量语料而对说话人的心理活动进行反向建模,而非对于大量类似语料进行反复的统计学处理。然而,基于少量语料作出符合人情(即常识心理学)的推断却不是 ChatGPT 所能做到的事情,因为 ChatGPT 本就只能处理语料而不懂人情,更不能施展基于人情的想象力。

更何况 ChatGPT 的运作所需要的前提——海量的语料并不总是充分的。目前 ChatGPT 所依赖的语料仅仅局限于 2021 年,因此,任何一项超越当下训练语料训练范围的任务都会让其表现变得捉襟见肘。譬如,假设世界上并不存在着任何一部关于“金星生活”的科幻小说作为训练语料,一个尝试写这个题材的人类科幻小说家难道能够仅依赖 ChatGPT 完成相关的写作任务吗?恐怕不能,因为在训练语料停止供应之处,就是 ChatGPT 停摆之所。这位人类科幻作家便只能通过他自己的努力学习关于金星的天文学知识,并在此基础上对下述问题进行构想:人类的身体将在什么样的设备帮助下,才能够在这样的一个恶劣的外星环境下维持生理机能呢——大气压强是地球的 92 倍,大气密度是地球的 100 倍,几乎没

有氧气，温度是 400～500 摄氏度，天上的“云”则都是硫酸构成的！而要回答这些历史上从未有人提出的问题，这位科幻小说家只能反复勾画草图，进行科学演算，甚至在书房内反复走动，自行“脑补”出各种可能的火星生活场景。所有这些思考，都预设了这位科幻作家是具有 ChatGPT 所不具备的“具身性”的——他能看、能走、能跳，并能设想一个与自己生理指标接近的宇航员在别的星球的重力环境下如何走与跳。换言之，写作者必须自己能够感知世界，才能想象自己出现在另外一个与地球不同的环境之中，亲身感知本身就能为想象力的展开提供某些替代性的素材，以弥补语料的不足（比如，一位科幻作家若无法想象宇航员在金星表面的重力感，他就可以通过“将自己在地球上的重力感减少 16.3%”这一方式来想象）。与之相较，缺乏感知力的 ChatGPT 在面对此类问题时乃是完全束手无策的。

面对上述反驳，主流人工智能技术支持者或许会退而求其次地回应说：纵然今天的 ChatGPT 缺乏具身性，但这未必意味着这一点是无法实现的。我们完全可以在 ChatGPT 系统上接驳很多传感器，这样，该系统便能具备感知力了。我们甚至可以将某个 ChatGPT 终端与可行走机器人结合起来，这样一来，ChatGPT 也就具有类似于人类身体的行动力了。

但这个理由依然是非常粗糙的，因为它预设了人类的认知机能也是按照“内部统计学机制加外部传感器”这样的机制建立的。而人类的认知机能显然要比这复杂得多，因为感觉本身显然是按照某种复杂的中介方式而与心智机器的核心部分互相联系的，对这一中介方式之玄奥的探索，恰恰也便构成了从康德到三木清的想象力（构想力）理论的聚焦点。按照当代认知语言学家的观点，感知与语言之间的中介形态便是“认知图式”（cognitive schema）。“图式”这个词的古希腊词源有“形状”的意思。其在康德哲学里的含义，则是指“想象力”机能产生的相对固定的时间样态，以便特定的纯粹知性形式（范畴）能够以此为中介，对特定的感性材料施加整合作用。而在认知语言学的语境中，“认知图式”指的则是特定的语言学模式的重复性特征的聚合，或者说得更专业一点，是“一系列语例中的共通性在得到强化后所获得的一些抽象的模板”[23]。比如，在英语中，“SOPHOMORE”这个概念就从属于由下列概念之矩阵构成的模板“TWO”“PERSON”，“KNOW”“YEAR”等[46]。① 此外，这些模板的内部结构往往得按照“意像式”（imagistic）的方式来加以把握。譬如，英语“ENTER”（进入）这个概念就可以被分析为数个意象图式的组合，包括“物体”（object）、“源点—路径—目标”（source-path-goal）与“容器—容纳物”（container-content）。三者结合的情况如图 1 所示[33]。

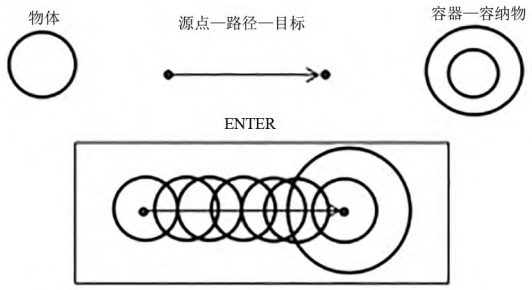

图 1 “ENTER”的认知图式

在认知语言学家看来，一个意图使用“ENTER”这个动词的语言使用者，只有先看到上述图像，才能将这个词真正用好。据此，如果他不能完成对于上述图形的构想，他就会丧失使用该词之词法的能力（即丧失“逻格斯”的力量），并无法掌握附加在这个词上的情感感受力（即丧失“帕索斯”的力量）。很显然，这是后世学界对于三木清哲学沿着语言学方向的一种具体化。考虑到对于认知图式自身的数码化重构具有的复杂性，这些问题是不可能通过在大型语言处理模型之外简单地接驳传感器而得到自动解决的。

上面的讨论显然已经足够说明 ChatGPT 为何不具有涉身性了（以及与此相关的图像构造能力），那么，ChatGPT 是否可能具有三木清版本的

① “SOPHOMORE”指的是大专院校二年级学生，字面意思是“知道得多一点”（这当然是相较一年级新生而言的）。

构想力所具有的意志性、实践性与集体凝聚性呢?

答案也是否定的。首先看“集体凝聚性”这一指标。从表面上看,ChatGPT 得到的语言训练资料是来自人类集体的,因此,ChatGPT 应当能够起到凝聚人类集体的作用。但看得更深一点,作为聊天程序的 ChatGPT 为了取悦特定的人类用户,往往会根据不同用户的提问倾向给出特定的应答,由此提高用户的“使用体验”。这样一来,它反而可能会在特定条件下强化人类用户的“信息茧房”效应,由此使得人与人之间的意见分歧变得更大。同时,该系统对于“喂料内容”的敏感性,也很容易使得其成为虚假信息的传播器,由此成为不同人群之间进行恶性斗争的工具。此外,笔者参与的机器实测也表明,这一系统的运作缺乏稳定性,往往会在不同时间对同一问题给出不同答案,很难使用户对其产生足够的信赖感。另外,也恰恰是因为 ChatGPT 自身的非涉身性,它亦无法通过对于基于人类共同运作身体模式的感性图式的展示来消弭人与人之间的分歧。

ChatGPT 亦不具有三木式“构想力”所要求的意志性与情绪性。人类基本情绪与意志的产生都与“维持生理机能”以及“散播基因”这两大目的紧密相关,从这个角度看,作为非生物体的 ChatGPT 并不具备使情绪与意志得以产生的上述生物学基础——它所能做的,无非就是搜集且处理与情绪以及意志表达相关的语料。很显然,ChatGPT 缺乏意志,乃是其缺乏具身性这一点的副产品。

ChatGPT 也不具有与外部实在相关的那种真正的实践性。这是因为:首先,它只能涉及语料,而不能涉及语料所谈论的客观事件(比如,它虽然能“阅读”菜谱,却无法做出哪怕一条真正能吃的糖醋鲤鱼);其次,它只能根据用户的语言反馈来修正自己的输出模式,而缺乏在悬置他人意见的前提下独立对世界进行探索的能力。

综上所述,ChatGPT 不能满足三木哲学关于“构想力”四个标准中的任何一项。因此,ChatGPT 也好,类似的主流深度学习技术也罢,都是缺乏想象力的。那么,未来的人工智能若不走 ChatGPT 的技术路线,是否能够具备想象力呢?

三、未来的人工智能否具有想象力

要回答上述问题,我们还是要按照三木哲学关于“构想力”的四项标准来看未来人工智能的发展。

先来看“具身性”标准。前文已指出,给既有的语言处理模型配上外部传感器并不难,难的是在两者之间构造出一个中介,这个中介既具有一定的图像性,又具有类似逻辑算子的可组合性,由此才能兼备三木清所说的“帕索斯”与“逻格斯”这两个因素。有人或许会认为目前的图像识别技术能够完成对于此类中介的构造。但需要指出的是,基于深度学习的主流图像识别技术所能做的,便是将输入图片上的任何像素都一股脑地“喂给”系统,以期“训练”系统能最终给出符合人类期待的图片标注。由于这种数据训练方式几乎是在不顾人眼识别物体的实际内在逻辑的前提下进行的,所以,此类系统的识别准确率貌似虽高,但一旦犯错,错误的类型却几乎是人类不太可能犯下的(比如,将某个角度的乌龟看成是冲锋枪,或是仅仅因为一条狗的出现背景是荒野就认为这条狗是狼[①])。换言之,此类系统在处理感觉材料时,缺乏对于物体分类的范畴性知识的先验牵导作用。

当然,从技术上看,将事物分类的范畴性知识输入计算机也并非不可能,但这里的难点是,要怎样将这些抽象的先验知识与感性对象加以联系。比如,系统即使通过了抽象的范畴分类而知道了电话机是一种通讯工具,也要知道电话机的外形是什么样的。然而,世界上可能出现事物的几何形状的数量又几乎是无穷无尽的,因此,如果采用“一器物对应一堆训练用图像数据”的方式来使得系统掌握事物名称与形象之间联系的话,费时费力不说,系统也将永远难以获得对于一种从未见过事物的外形的想象力。可以想到的一条解决路径,便是让系统掌握事物外形变化之道,以此做

① 相关讨论请参考:John Mark Bishop, “Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It”, *Frontiers in Psychology 11*, DOI: 10.3389/fpsyg.2020.513474.

到以不变应万变。在这个问题上，视觉心理学家彼得曼(Irvin Biederman)提出的“RBC理论”[9]是颇有参考价值的(RBC全称是“Recognition-by-Components”，意思是“通过部分来认知”)。根据其理论，人类视觉系统中预藏了大量的“几何离子”(geons①)，其数量小于或等于36个，其典型的代表如图2所示。

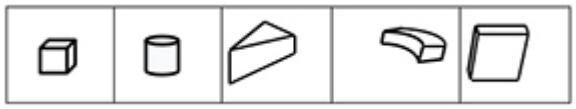

图2　五种典型几何离子的构成方式

几何离子数量虽然不会超过36个，但它们彼此之间两两结合的可能性就有74549种，若算上三三结合、四四结合的可能性，变换方式几乎是无穷无尽的。这种理论解释了人类视觉系统为何能够处理各种形状的新事物的外形——因为任何一种新事物的外形，几乎都能被分解为上述几何离子的构成方式。反过来说，几何离子组合方式的多样性又解释了人类为何能够对从未被看到的事物形状进行想象。因此，几何离子所构成的抽象性图像，便成为更低层次上的感觉材料与更高层面上的事物名称之间的中介物。

我们能否在人工智能语境中复制彼得曼的理论呢？没有任何原则上的障碍说明这样做是不可行的。实际上，彼得曼的理论就是从玛尔(David Marr)的“广义椎”理论[10]中发展出来的，而玛尔又被公认为计算视觉研究的奠基性人物，因此，RBC理论本就是在计算视觉理论的大传统中被提出的。而早在20世纪90年代，彼得曼本人就曾与同事合作，在当时的人工神经元网络技术基础上构造一个对于RBC理论的计算建模[11]。然而，除了少数后继者外，②这一技术路径的回应者寥寥。我个人认为主要原因有两方面：第一，主流用于图像识别的深度学习技术将识别对象的最终名称为目标，而不关注诸如“几何离子”之类中介物的构造。此粗暴思路虽然对人类视觉机制的内在机理视而不见，却因为切中了鼠目寸光的“市场需要”反而得到更好的发展。第二，RBC理论的建模无法脱离神经元网络深度学习技术带来的“无法灵活展现逻辑句法”的固疾，因此在该技术的拖累下，相关的建模成果反而无法体现出该理论在心理学层面上具有的优势：诸几何离子之间的离散性能够轻松解释“人类视觉几何语法”的创生性与灵活性。由此，此类建模成果也无法体现出对于主流视觉识别技术的明显优势。从这个角度看，对于RBC理论的计算建模，必须走一条与主流的深度学习路径不同的道路。笔者曾在“非公理化推演系统”(Non-Axiomatic Reasoning System)的基础上讨论过将RBC理论加以计算建模化处理的可能性[12]，但鉴于此话题过于技术化，在此不过多论述。

但即便我们能将RBC理论予以一种算法化实现，事情也还远远没完。人类的身体感受通道不仅只有视觉，还有听觉、触觉、味觉、嗅觉、动觉、痛觉、内感(如紧张感)等。要完全地模拟这些感受对人工智能研究来说是非常困难的，因为其中的某些感觉(如味觉与痛觉)是与人类机体的生物学运作方式明确相关，而人工智能体要在规避对于人类生物学机能全部复刻的前提下去实现对于这些感受通道的重建，可谓缘木求鱼。所以，我个人倾向于认为三木清关于构想力提出的“具身性标准”只能在人工智能的语境中得到一种局部实现。

上述结论直接影响我们对于余下诸标准与人工智能可能形态之间关系的判断。譬如，未来的人工智能是否可能具有人类意义上的那种牵涉到物理世界自身的实践能力？答案是：它只能在一定程度上具备此能力。比如，你的确可以指望未来的人工智能体能够像人类那样按照自主的规划去垒起一座小型建筑，甚至根据新观察到的情况(如土地的湿度对建材的影响)而自行想象对于原先计划的修正方案，但假若对于这些实践成果的评价标准涉及某种难以被人工智能复刻的人

① 这个词是由“geometrical ions”浓缩而成的。

② Dmitriy Maximov and Sekou A. K. Diane, Object Recognition by a Minimally Pre-Trained System in the Process of Environment Exploration, arXiv-CS-Artificial Intelligence.

类身体感受如味觉，那么人工智能对于人类实践能力的取代机制就会停止发挥作用。这也就是原则上不会有“机器人大厨”的理由——机器人固然可以根据固定的程序与固定的菜谱烹饪固定的菜肴，但由于其自身没有味觉，它们将始终无法自主地想象新美食的美味，并在这种想象的驱使下去开发新的菜品，最终再用自己的味觉验证这种新菜品是否可口。同样的道理，它们也不可能取代品酒师与炒茶师的工作。因此，它们所从事的实践事项的广度与深度都将远远不如人类的水准。

那么，未来的人工智能是否可能具有情绪与意志呢？从某种意义上说，这并非完全不可能，因为我们完全可以在算法层面上规定机器人也具有“自保”的需要，并且它也可能在这一需要的激励下产生一种“情绪”，也就是在生存受到威胁的情况下将认知资源集中于某个重大问题上的应激机制。反之，在运作资源相对宽裕的情况下，我们也允许系统产生“快乐、闲适”的情绪，以鼓励其从事一些与当下任务无关的新探索。但需要注意的是，人工智能所能产生的意志与情绪的种类相对有限。譬如，由于人工智能体没有进行繁殖的需要，所以笔者非常怀疑未来的人工智能是否有可能会产生基于“基因撒播”之需要的那些情绪，而理解这些情绪却又恰恰对处理那些与爱情有关的语料是非常重要的。这当然不是说自然选择进程本身不能被计算机所模拟（实际上，“遗传算法”早就是人工智能中的一个重要的流派），而是说人工智能本身并不必然地需要遗传的计算模拟机制以协助其“产生”下一代，而人类离开了遗传机制的作用，却一定会在生物学上灭绝。这两者之间的不对称性显然使得我们缺乏足够强的理由在人工智能的层面上去模拟人类的两性交往模式，而在此类模拟缺场的前提下，我们便又不得不面对这样一个后果：人工智能原则上不能理解人类基于两性交往模式而产生的各种情绪，以及基于这种情绪的各种文化。

那么，未来的人工智能体是否能够具有对于人类集体的凝聚功能呢？我们还当记得，在三木清哲学语境中，人类集体凝结剂的代表是神话，而神话的哲学本质又是对于同样的人生愿景的期望。不得不指出，如果缺乏对于生、老、病、死等共通的人生议题的关涉，我们便很难想象一种神话竟然能够起到凝聚一个民族的人心的作用——而不同民族的神话或宗教体系，其实也可以被视为是对于此类问题的不同解答方式。然而，作为非生物体的人工智能能够为解决人类此方面的焦虑提供终极答案吗？抑或本身没有严格意义上的生、老、病、死，而只具有“被启用”“变得陈旧”“出现故障”“被抛弃”等阶段的人工智能体，能够在多大程度上理解人类所关心的上述话题？浅层的理解或许是可能的，但考虑到人工智能体记忆完全可以通过数码复制方式而得到永生，它们对于人类基于死亡恐惧的种种意识形态构建恐怕很难产生深入的共情。这一点或许在根本上阻碍了人工智能体具备三木清赋予人类构想力的那种社会凝结作用。

总之，与人类的想象力与构想力相比，即使人类实现了通用人工智能目标，这样的人工智能体的想象力水平依然是不充分、不完全的。它们依然也只能在辅助的意义上参与人类的实践活动。而上述判断之所以能对几乎所有类型的人工智能体有效，又是基于一个无法被否定的概念性事实：人工智能体在原则上就不是生命体，因此，它不可能具备基于生命体才能具有的心智功能，特别是那些基于特定身体感官运作的想象力的特殊方面。

四、结语

假若早已于 1945 年就离世的三木清知道今日的 ChatGPT，他会说些什么呢？三木清是一位深受马克思主义影响的学者，因此，无产阶级的那种基于身体与物质世界之间直接接触的劳动模式才是他心目中最典型的实践方式。因此，在他看来，对于实践的谈论——由于其与物质世界之间的关系的间接性——就不可能取代真正的实践。类似的想法也为思想更接近马克思主义的另一位日本京都学派成员户坂润（1900—1945）所分享，他在《现代唯物论讲义》中亦指出，对于空间的先验条件的讨论不能取代实在的空间，就

像对于寿司的谈论不能取代对于寿司的真正食用一样。[①]因此，从他们的立场上看，ChatGPT 所能做的，无非是对于人类实践的谈论，却不能进入真正的实践领域，甚至不能进入真正的客观空间。毋宁说，ChatGPT 所能做的，便是将人类的言谈所已经制造出来的意识形态矩阵在规模上予以全面放大，由此进一步剥夺无产者自主反思社会现实的能动性与可支配时间。因此，从广义的西方马克思主义的光谱上看，ChatGPT 便是广义的“文化工业”的一部分，并因此不得不成为像阿多诺、本雅明、马尔库塞这样的文化批判者的批判对象。

不过，这并不意味着三木清与户坂润会抽象地反对一切人工智能体的研究。假若一种人工智能设备能够实际地增强人类改造自然的能力（正如机床、拖拉机、飞机所做的那样），作为马克思主义者的他们并没有理由反对这一点。但正如我们前面的分析所指出的，由于任何人工智能体的非生物性，它们都无法具有完整意义上的人类想象力（构想力），因此，它们就无法成为使得人类的庞大经济机器得以运作的“第一推动力”：需求（其本质就是主体对于获取更多资源之未来状况的一种构想力）。请注意，经济学需求本身就是生物性需求的抽象化，换言之，若没有衣、食、住、行与两性交往、后代繁殖方面的种种基本需求，以及衍生于这种基本需求的种种社会性需求（特别是对于社会地位的追求），整台经济机器都会停摆，遑论寄生于这台经济机器的人工智能产业。从这个角度看，恰恰是人类的想象力（构想力）自身的“帕索斯”成分使得人工智能得以被发明出来，而不会出现反过来的情况：人工智能自身的想象力会取代人类的“帕索斯”去推动人类经济机器的运作。因此，即使从长远来看，人工智能也不会发展到足以让我们忧心是否要赋予其法权意义上的“人格”的地步——因为法权意义上的人格恰恰是以完整的自由意志为前提的，而自由意志本身的存在又恰恰是以对于不同形式的未来的构想力为前提的，特别是与遗产有关的法律事务所要求的对于“死后状况”的那种构想。然而，本身未必会“死”的人工智能体又如何能充分地构想“死后状况”呢？由此，我们人类又该如何想象自己能与一些不会在真正的意义上去生、老、病、死的人工智能体分享平等的法权地位呢？既然这种想象本身就是非常牵强的，那么将人工智能始终定义为一种人类的生产工具（无论其有多高级），又有何不合理之处呢？

① 此书日语版收录于：『戸坂潤全集・第 3 巻』（勁草書房 1966 年版）。

参考文献：

[1] 『三木清全集：第 8 巻：構想力の論理』，东京：岩波書店 1966 年版。

[2] 『三木清全集：第 14 巻：評論 第 2』，东京：岩波書店 1967 年版。

[3] Dennis Stromback，“Miki Kiyoshi and the Overcoming of German and Japanese Philosophy”，*European Journal of Japanese Philosophy*，No.5，2020，pp.103-143.

[4] 『三木清全集：第 3 巻：唯物史観研究』，东京：岩波書店 1966 年版，第 45-46 页。

[5] 本尼迪克特・安德森《想象的共同体——民族主义的起源与散布》，吴睿人译，上海：上海人民出版社 2005 年版。

[6] Jie Zhou et al，*ChatGPT：Potential，Prospects，and Limitations.Frontiers of Information Technology & Electronic Engineering*（*2023*），https://doi.org/10.1631/FITEE.2300089.

[7] Frank Jackson，“Epiphenomenal Qualia”，*Philosophical Quarterly*，Vol.127，No.32，1982，pp.127-136.

[8] Ronald Langacker，*Cognitive Grammar：A Basic Introduction*，Oxford：Oxford University Press，2008，p.23.

[9] Irvin Biederman，“Recognition-by-Components：A Theory of Human Image Understanding”，*Psychological Review*，1987，pp.115-147.

[10] David Marr，*Vision：A Computational Investigation into the Human Representation and Processing of Visual Information*，San Francisco：W. H. Freeman and Company，1982.

[11] J. E. Hummel & I. Biederman，“Dynamic binding in a neural network for shape recognition”，*Psychological Review*，Vol.99，1992，pp.480-517.

[12] 徐英瑾《心智、语言和机器——维特根斯坦哲学与人工智能科学的对话》，北京：人民出版社 2022 年版，第 382-393 页。

责任编辑：朱　磊