人机协同下可信内容合成

邓晃煌，贺宝仪，吴 飞
（浙江大学计算机学院，浙江杭州 310058）

摘 要：ChatGPT在工程上创新性地整合了大数据、大模型和大算力，按照“共生则关联”的原理，挖掘出自然语言中单词和单词共生概率知识，辅以人类反馈信息，以机器智能实现了统计关联下的语言快速合成。ChatGPT推动人工智能由识人辨物和预测决策等技术赋能向内容合成这一新领域跃升，即人工智能内容合成（Artificial intelligence generated content，AIGC）。AIGC会塑造内容生产的新范式，成为智能数字交往的有力手段，并悄然引发一种文明范式的转型。ChatGPT进行内容合成的计算机理及其所引发的伦理道德潜在风险，将使人机协同下可信内容合成更具有价值。

关键词：内容合成；自注意力机制；潜在风险；人机协同

中图分类号：G206 **文献标识码**：A **文章编号**：2096-8418（2023）04-0024-08

一、人造智能：清晰描述后计算模拟

1955年8月，时任达特茅斯学院数学系助理教授、1971年度图灵奖获得者麦卡锡（John McCarthy），时任哈佛大学数学系和神经学系青年研究员、1969年度图灵奖获得者明斯基（Marvin Lee Minsky），时任贝尔实验室数学家、信息论之父香农（Claude Shannon），时任IBM信息研究主管、IBM第一代通用计算机701主设计师罗切斯特（Nathaniel Rochester）四位学者向美国洛克菲勒基金会递交了一份题为“关于举办达特茅斯人工智能夏季研讨会的提议（a proposal for the Dartmouth summer research project on artificial intelligence）”的建议书，希望洛克菲勒基金会资助拟于1956年夏天在达特茅斯学院举办的人工智能研讨会，研究“让机器能像人那样认知、思考和学习，即用计算机模拟人的智能”的研究。[1]

在这份建议书中，“人工智能（artificial intelligence）”这一术语被首次使用。该建议书对能够实现“人工智能或人造智能”的原因进行了如下描述：学习的每个方面或智能的大多数特性原则上都可以被精确描述，从而用机器来模拟。

大多数学科都有必须遵守的最基本的命题或假设，这些命题或假设不能被省略和被违反，即学科发展的第一性原理。比如，牛顿经典力学中“引力和惯性”以及达尔文进化论中“物竞天择、适者生存”都是需要遵守的第一性原理。在人工智能研究中，对智能行为过程的精确描述或许可以作为类似于第一性原理需要遵守的原则，也就是说以机器为载体来展示人类智能或生物智能，需要对智能行为发生过程予以清晰描述，从而通过程序设计语言被机器按序执行。

这份建议书同时详细列举了7个在达特茅斯会议中需要重点讨论的问题，它们分别如下：自动计算机、计算机编程、神经网络、计算的复杂度、智能算法的自我学习与提高、智能算法抽象能力、智能算法随机性与创造力。

人工智能登上人类历史舞台后，研究者们围绕智能行为的模拟进行了诸多研究，形成了如下多种人工智能方法[2][3][4]：以符号主义为核心的逻辑推理、以问题求解为核心的探寻搜索、以数据驱动为

作者简介：邓晃煌，男，讲师；贺宝仪，女，博士研究生；吴飞，男，教授，博士生导师。

核心的机器学习、以行为主义为核心的强化学习、以博弈对抗为核心的群体智能（两人及以上）。符号主义人工智能将概念（如命题等）符号化，从若干判断（前提）出发得到新判断（结论）；问题求解的探寻搜索依据已有信息来寻找满足约束条件的待求解问题的答案；数据驱动的机器学习方法则是从数据出发，从数据中发现数据所承载语义（如概念）的内在模式，利用学习得到的内在模式完成识别和分类等任务；行为主义为核心的强化学习根据环境所提供的奖罚反馈来学习所处状态可施加的最佳行动，在“探索（未知空间）—利用（已有经验）（exploration vs. exploitation）”之间寻找平衡，完成某个序列化任务，具备自我学习能力；博弈对抗则推动机器学习从“数据拟合”优化解的求取向“均衡解”的求取迈进。

1965 年，诺贝尔物理学奖获得者费曼（Richard Feynman）曾经说过：“不可造者，未能知也（What I cannot create，I do not understand）。”透过今天计算机外在之形了解其计算之内禀，更能感叹计算之伟大、计算之局限！

二、内容计算合成机理

近年来，人工智能领域出现了若干科技创新现象级产品，如耳熟能详的 AlphaGo、AlphaFold 和 ChatGPT，这些现象级产品表现出较强的内容合成能力（即“无中生有”）。AlphaGo 根据当前落子局势，从已有落子的学习中合成一个策略，以更好应对当前落子。AlphaFold 从蛋白质的基因序列和其三维空间结构的配对数据中进行学习后，按照给定的基因序列输入，合成一个刻画生命功能的蛋白质三维结构。ChatGPT 这一复杂的神经网络大模型，按照“共生则关联”挖掘所得单词之间共生概率，实现统计意义下的语言合成。

ChatGPT 的成功并非一蹴而就，而是源自于以深度学习为代表的人工智能技术的长期积累。其核心在于谷歌公司于 2017 年将自注意力（self-attention）机制引入所构造的 Transformer 神经网络结构，这一结构可更高效挖掘句子或篇章中单词与其上下文单词之间因共生概率而形成的关联关系。[5]

为了训练 Transformer，OpenAI 主要采取了三种方法。首先，采取一种被称为“完形填空”的训练方法。给定任意一个自然语言句子，从中“移除”一个单词，然后让模型根据剩下单词所形成的上下文来预测最合适的“填空词”，以便完成填空任务。这一“完形填空”过程在人工智能领域被称为“自监督学习（self-supervised learning）”。自监督学习在人工智能中具有重要作用。图灵奖获得者杨乐昆（Yann LeCun）曾表示，自监督学习让人工智能推理更像人类，因为人类和动物是通过自监督模式获得新知识，具备学会了学习（learning to learn）的能力。

其次，为了让 ChatGPT 完成聊天问答任务，在训练得到 Transformer 的基础上，OpenAI 研究者提出了一种“提示学习”（prompt）方法来让 Transformer 具备“聊力”。在提示学习中，研究者设计“提示样例”教人工智能模型学习更流畅合成语言。提示样例可形象理解为“知识模板”，让 ChatGPT 从中掌握各种“闲聊”固有套路。比较有意思的是，目前出现了一种编写“提示案例”的工程师工作岗位（prompt engineer），被一些媒体称为人工智能的私语者（AI whisperer），OpenAI 创始人之一山姆·阿尔特曼（Sam Altman）说：“这一职业需要强技能的水平。”

最后，为了进一步提高模型合成语言性能，ChatGPT 还引入了人类反馈中强化学习（Reinforcement Learning from Human Feedback，RLHF）的技术，将人类反馈作为一种监督信息输入给模型，对模型参数微调，提高语言模型回答的真实性和流畅性。

还要说明的是，微软公司将其收购的开源及私有软件项目托管平台 GitHub 中数十亿行源代码开放给 OpenAI，用来训练 ChatGPT 的逻辑，使得 ChatGPT 从程序代码中学习到思维链（chain of thought），因此 ChatGPT 所合成的语言中鲜见前后矛盾的语句。

ChatGPT是大数据、大模型和大算力的工程性创新整合，体现了“数据是燃料、模型是引擎、算力是加速器”的深度学习特色：

* 大数据：ChatGPT的训练使用了45TB的数据、近1万亿个单词（大概是1351万本牛津词典所包含的单词数量），同时包括数十亿行源代码。

* 大模型：ChatGPT的前身GPT-3模型参数高达1750亿。如果将这些模型参数全部打印在A4纸张上，一张一张叠放后，其高度将超过上海中心大厦632米的高度。

* 大算力：训练ChatGPT所耗费的算力大概是3640 PetaFLOP/s-days，即用每秒能够运算1000万亿次的算力对模型进行训练，需要3640天完成。

在热烈讨论ChatGPT时，不由让人想起另外一个人工智能现象级产品AlphaGo在2016年以4∶1击败李世石的场景。虽然AlphaGo在迎战李世石之前，已经几乎“阅览完毕”人类选手所有围棋比赛的棋局，且每天通过自我对弈来“华山论剑”（AlphaGo每天可完成200万次的自我对弈），但李世石在其战胜AlphaGo的唯一对局中落下了人类选手几乎不可能落子的一招，这一落子是AlphaGo之前从未见过的，使AlphaGo无法从容应对而落败。

可见，“数据有多大，智能就有多强”是计算独大模式下人工智能算法不可避免的局限性。一旦数据无法覆盖某些场景，则人工智能算法就会在这些场景中失效。这是因为算法无法理解数据背后所承载的机理、缺乏一种“灵气”。

三、计算优势：试错与暴力

博弈对抗是检验人工智能能力大小的标杆。1997年5月，国际象棋冠军卡斯帕罗夫和IBM公司的“深蓝（Deep Blue）”计算机程序展开了一轮令全球瞩目的人机大战。结果，深蓝计算机发挥出色，以2胜3平1负的总比分战胜了卡斯帕罗夫，成为首个在标准比赛时限内击败国际象棋世界冠军的电脑系统，这是人工智能领域一个里程碑事件。

在象棋比赛中，深蓝需要判断某一时刻棋局落子会对整个棋局胜负带来怎样的影响。即基于已知规则，深蓝要从当前棋局出发，尽可能向前搜索更多可能的未来棋局，以便掌握更多信息来对当前落子的优劣进行判断。由于可选棋局众多，尽管深蓝平均每秒能够对1亿个棋局进行判断评估，还是无法在规定时间内计算得到当前棋局对胜负的潜在影响。1950年，香农（Claud Shannon）发表了一篇有关国际象棋编程的论文。在这篇论文中，香农估算国际象棋比赛中落子选择从第一次移动时的20种会增加到第二次移动时的400种，在第六次移动时可能的落子选择达到1.19亿种。香农通过估算，认为国际象棋的落子总数为10的120次方种，远远超出了10的82次方这一宇宙原子总数。这就是国际象棋中“组合爆炸”难题。

为了从海量可能落子中选择一种合适落子，以克服组合爆炸挑战，深蓝采用了由1971年图灵奖获得者约翰·麦卡锡（John McCarthy）发明的“阿尔法-贝塔（Alpha-Beta）”剪枝搜索算法。简单来说，该算法主动“剪掉”对胜败不产生任何影响的棋局，从而减少搜索空间以提高搜索效率，解决了组合爆炸难题。可以看到，搜索是很重要的一种人工智能答案求解方法。“你见，或者不见，我就在那里，不悲不喜”，解决某个问题的答案就在那里，需要运用搜索之术来获得。

实际上，计算机在其诞生之初，就展示了其突破组合爆炸问题的优势。二战期间，德国研制出了密码机，能将明文自动转换为密码（密文），再通过无线电或电话线路传送出去，这就是有名的恩尼格玛机（enigma machine）。每一次U型潜艇深海的奔袭及每一次残酷的进攻都由恩尼格玛机加密传输和解密阅读。这些通信信号即使被截获，也就是一堆毫无意义的乱码。恩尼格玛密码机是一种电子系统和机械系统的组合，看上去像一台很复杂的打字机，由按键、转子、插接板和导线连接。恩尼格玛机将

“每加密一个字母就更换一次密码表并且永不重复”的信息加密追求发挥到了极致。据估计，恩尼格码的密钥总数是 1590 万万亿种。更为困难的是，德军是每天更换 1 次密钥，也就是说要在 24 小时内验证 1590 万万亿种密钥的可能性，这几乎是难以想象的困难。

英国科学家、计算机理论模型之父图灵（Alan Twring）在二战期间深信“机器创造出来的密码怪兽，只有用机器才能战胜”，提出密码破译的过程可用机械方法自动处理。1940 年初，在波兰数学家和密码学家马里安·雷耶夫斯基（Marian Adam Rejewski）通过暴力搜索方法机械式验证恩尼格玛机所有转子组合思路的基础上，图灵研究通过概率学原理来推断不同密钥组合的可能性，优先运算出可能性最高的组合，研制成功了被称为“炸弹（Bombe）”或“图灵炸弹（Turing Bombe）”的计算机，极大加快了二战结束的进程。

作为另外一种人工智能现象级产品 AlphaGo，其在战胜李世石之前已自我对弈了千万次比赛。通过试错（trial & error）与暴力（brute force）穷尽遍览围棋落子布局，对模型参数不断进行调整。据估计，李世石一生仅能参加数千场比赛。人类智能以“经一蹶者长一智”模式，从有限尝试中学习经验，人类智能和机器智能在智能优化方面存在巨大差距。

与上述人工智能模型类似，ChatGPT 以完形填空预测、提示样例调教和人类相关反馈对齐等机制来训练调优参数，是计算机在优化人工智能模型时试错与暴力的一种体现。

四、内容计算合成的潜在风险

产生式算法推动人工智能由识人辨物和预测决策等向内容合成跃升，使得人工智能能够被普通用户使用。但由于这一技术架构在数据驱动为模式下的深度学习基础上，存在着解释性差、失真而失信、难以由果溯因等不足，从而带来潜在的应用风险。笔者将分别介绍这些风险产生的技术原因以及人工智能的伦理风险。

（一）模型解释性差

在物理学过去的发展历史中，还原论（reductionism）的观点一直是物理学工作者进行研究的最基本的指导原则，其将一切复杂系统中出现的各种现象都归结为最基本组成单元以及单元之间相互作用的基本规律，或者说将复杂现象还原为简单构成，然后再从简单重建复杂。譬如，气体、液体和固体都被分解为分子或原子，原子又被分解为原子核和电子，原子核被分解为质子和中子，质子和中子又被分解为夸克。但是，以还原论为基础来研究和讨论复杂系统的合作现象时，却遇到了前所未有的挑战。1977 年诺贝尔物理学奖获得者安德森在美国《科学》杂志发表了“More is different”文章中深刻讨论了物理世界中所具有的涌现这一现象[6]：还原论假说从来都不意味着建构论（constructionist）假说，将所有事物还原为简单的基本定律的能力并不意味着从那些基本定律出发重建整个宇宙的能力。事实上，粒子物理学家告诉我们越多关于基本定律的本质，这些基本定律和其他学科的问题的关联就越少，和社会问题的关联也越少。在面对尺度（scale）和复杂性（complexity）的双重孪生难题时，以还原论为基础的建构论的假定完全崩溃了。

与之对应的是，我们对深度学习研究中所提出的 MCP 神经元、感知机模型、深度神经网络、自注意力机制等，均可通过相应数学模型来刻画。但一旦将这些基本单元或基本机制组合成庞大复杂的 ChatGPT，其体现了难以解释的涌现能力（emergence），即：模型参数较少时（如十亿、百亿等），ChatGPT 无法表现其应有能力；而一旦模型参数增加到千亿级别时，模型能力瞬间涌现和提升。机器学习模型具有涌现能力意味着重要的科学意义，因为如果涌现能力是永无尽头的，那么只要模型足够大，类人人工智能的出现就是必然。但是，研究表明，不存在明确表征可证明哪类任务最具涌现，并且 ChatGPT 在逻辑推理和因果推断任务中的涌现能力最低。微软公司因此把其称为现象级产品（phenome-

nological)。[7]

(二) 失真而失信

ChatGPT以“共生则关联”为标准对模型进行训练，会通过单词和单词之间的概率关联产生非事实的合成结果（也有学者将其称为人工智能幻觉，hallucination）。例如，ChatGPT一本正经地回答“林黛玉倒拔垂杨柳”这样啼笑皆非的问题、捏造法学家性骚扰丑闻文章、发布澳洲某位官员因涉及澳大利亚储备银行子公司受贿案而入狱的虚假文章等，这些非事实结果产生的原因在于ChatGPT内容合成机理是“机械式共生匹配”。

1913年，法国科学家埃米尔·波雷尔（Emile Borel）发表了《静态力学与不可逆性》（*La mécanique statique et l'irréversibilité*）论文，[8] 阐释了“无穷”概念：让我们想象一下，假设猴子经过训练学会了随意按下打字机的按钮，现在让猴子们在一位文盲领班监督下工作。如果无限多猴子在无限多打字机上随机乱敲，并持续无限久时间，那么在某个时候，将会有只猴子打出莎士比亚全部著作。这被称为“无限猴子定理”。“无限猴子定理”试图说明“随机+无限=一切皆可能”。一旦在单词之间建立了概率关联，很显然从“无穷”角度而言，ChatGPT可以合成任意的内容。但是，这一通过随机概率将语言序列拼接在一起的合成模式也被若干学者称为随机鹦鹉（stochastic parrot）。

据估计，全球高质量文本数据的总存量在5万亿token左右（一个token可理解为单词或符号等基本单元），这包括了世界上所有的书籍、科学论文、新闻文章、百科、公开代码以及网络上经过筛选的达标数据（如网页、博客和社交媒体）。按照目前ChatGPT消耗数据的速度，人工智能算法可能在一个数量级内，耗尽世界上所有有用的语言训练数据供应，而不得不使用ChatGPT合成的语言来训练ChatGPT。这将是一个多么令人魔幻的时刻！

可见，“数据有多大、智能就有多强”是计算独大模式下人工智能算法不可避免的局限性。一旦数据无法覆盖某些场景，则人工智能算法就会在这些场景中失效。这是因为算法无法理解数据背后所承载的机理，“数据独大、机理式微”这一计算模式缺乏“智慧之灵气”。

(三) 由果溯因弱

哲学上把现象与现象之间那种“引起和被引起”的关系，叫做因果关系，其中引起某种现象产生的现象叫做原因，被某种现象引起的现象叫做结果，如“力，形之所以奋也”。因果分析和推断是一种重要的获取知识的手段，是人类智能的关键组成。回答诸如“吸烟是否导致癌症”和“某个广告发布是否导致了某个商品销量上涨”等问题时，往往需要因果推理的能力。

1973年，美国科学院院士、加州大学伯克利分校统计系教授彼得·毕克（Peter Bickel）在美国《科学》上杂志发表了一篇有趣论文来讨论“伯克利分校录取新生时性别歧视”的困惑。[9] 在文章中，彼得教授对申请1973年秋季入学的12763名学生进行了统计，发现有8442个男生申请者和4321个女生申请者。在统计了当年度所录取学生中男生和女生人数后，发现当年度男生录取率为44%，远高于女生录取率35%。因此，从这个数据可以得出“伯克利分校当年在男生和女生录取中性别歧视昭然若揭”这一结论，也就是女生更难被录取。

然而，如果单独统计每个院系的录取情况，就会发现，对于每个院系而言，男生录取率和女生录取率相差无几，甚至对伯克利六个最大院系分别统计男生和女生录取率，竟然有四个院系女生录取率大于男生录取率。

也就是说，将所有新生按照院系分组后，统计得到男生和女生的录取率，与不按照院系分组统计男生和女生的录取率结果正好相反。这就是著名的辛普森悖论（Simpson's paradox）。辛普森悖论反映了总体数据集上成立的某种关系却在分组数据集合中“反其道而行之”这一怪异现象。彼得教授认为，在伯克利男生和女生录取率这个案例中，产生悖论原因在于女生更愿意申请那些竞争压力更大的院系

（比如英语系），而男生更愿意申请那些相对容易进的院系（比如工程学系）。在分析伯克利分校录取率时，不应该只看到男生和女生这个性别因素，还应该知晓“专业选择”这一因素会对新生录取产生作用。

辛普森悖论的重要性在于告诉如下道理：很多时候我们看到的数据并非反映现象全貌的数据，如果忽略产生数据的“潜在变量”，可能会改变已有结论，而我们常常却一无所知。比如伯克利招生录取中专业选择就是一个潜在变量。从观测结果中寻找引发结果的原因，“知其然且知其所以然”，由果溯因，就是因果推理。

“横看成岭侧成峰”，通过机械匹配来挖掘单词之间概率关联的语言大模型，显然会因为忽略了单词所组成句子和篇章的内部隐性规律（如主题等），将会带来错误之虞。

（四）伦理风险

“伦理”一词来源于希腊语的“道德”（character，性格）一词和拉丁语的“风俗（customs）”一词。这两个单词结合在一起，刻画了个体之间如何互动，用来描述人和人之间言行的道德与准则。“伦”表示人际各种关系，“理”则说明人际关系是有条有理。有原则有标准的伦理学就是对道德、道德问题及道德判断所作的哲学思考。伦理学是哲学的一个课题，是对道德、道德问题和道德判断所做的哲学思考。

伦理一词在中国最早见于《礼记·乐记》中“乐者，通伦理者也”。在中国语境中，对“伦”的诠释主要有三重含义：其一，“伦”者从“人”从“仑”，许慎《说文解字》训“伦”为“辈也”，讲“伦”起码要两人以上，“伦”即指人与人之间的辈分次第关系，单个人无所谓“伦”。其二，“伦”通“乐”，如《礼记·乐记》曰“乐者，通伦理者也”，强调音乐与伦理、美与善的相通性，或者说“乐”是通伦理的最佳方式，伦理以愉悦与和谐为要。其三，“伦”同“类”，如郑玄注“伦理”曰：“伦，类也，理之分也”，强调“伦”的本质是一种“类”的“分”。可见，“伦”在中国文明中就是一种关系，一种规则，一种秩序。

科学技术作为人类理性实践的结晶，对人类社会发展产生越来越深刻的影响，其产生和发展始终伴随着伦理观念、社会文化的演变。近代科学技术是伴随着理性精神、人文精神的兴起和传播而崛起的，通过发现和应用新知识为人类谋幸福是近代科学兴起的原动力之一。科技以前所未有的程度渗透进人类社会，甚至对政治、文化等产生深刻影响。当科学技术的探索与应用符合伦理规范，被引导至向善、负责任的方向，会更好地促进社会发展和人类福祉提升。倘若科学技术探索和应用打破了伦理底线，则可能给社会造成巨大危害。

传统的科技发展往往采取一种所谓的“技术先行或占先行动径路（proactionary approach）”模式，以发展技术为优先原则，体现出一种强大的工具理性，即“通过缜密的逻辑思维和精细的科学计算来实现效率或效用的最大化”。这种对技术效用单一维度的追求导致了科技异化现象，技术发展逐渐时而偏离“善”的方向，进而引发了一系列伦理风险。为确保科技发展的正当性与合理方向，在科学的社会建构思潮影响下，科技伦理应运而生。

科技伦理是科技活动需要遵循的价值理念和行为规范。人类已进入科技和信息时代，相较于传统工业时期以安全性为表征的技术风险，关涉人类福祉、公正等核心价值的伦理风险正成为当代科技发展引发的主要消极后果。

随着物联网、移动终端、互联网、传感器网、车联网、穿戴设备等的流行，计算与感知已经遍布世界，与人类密切相伴。网络不但遍布世界，更史无前例地连接着个体和群体，开始快速反映与聚集他们的意见、需求、创意、知识和能力。世界已从“物理世界—人类社会”二元空间结构（Physics world-Human Society）演变为“信息空间—物理世界—人类社会”三元空间结构 CPH（Cyber Space- Physics

world- Human Society)。在 CPH 三元空间中，带来的伦理学讨论不再只是人与人之间的关系，也不是人与自然界既定事实之间的关系，而是人类与自己所发明一种产品在社会中所构成的关联。因此，对于科技本身，需要既考虑其技术属性、又考虑其社会属性，形成人机共融所形成社会形态应遵守道德准则和法律法规。

一般而言，社会意识在很多时候落后于社会存在，如历史上有名的“红旗法案（Red flag traffic laws）”。19 世纪，汽车刚被发明出来时，被大众认为是一种怪物，不少人忧心忡忡地认为汽车会给人类社会带来重重危险。为此，1865 年，英国议会专门通过了一部被称为“红旗法案”的《机动车法案》。这个法案规定：每一辆在道路上行驶的机动车，必须由 3 个人驾驶，其中一人必须在车前面 50 米以外做引导，同时这个人还要用红旗不断摇动为机动车开道（提示人们危险将近），并且汽车速度每小时不能超过 6.4 公里。红旗法案的结果是把汽车当马车用，使得汽车工业发展几乎处于停滞状态，在人类汽车发展史上留下了令人深思的一页。

这种对技术进步所带来的负面感觉被称为“技术恐惧”（technophobia），好比 19 世纪末 20 世纪初，在拥有电力设施的家庭里工作的女佣十分担心电力设施会如同蒸汽锅炉或煤气厂一样在房屋内突然爆炸。

2023 年 3 月底，美国生命未来研究所（Future of Life Institute）公布一封公开信，呼吁所有 AI 实验室立即暂停训练比 GPT-4 更强大的 AI 系统至少 6 个月。这一公开信提出了如下四个令人深刻思考的问题：是否应该让机器用宣传和谎言充斥我们的信息渠道？是否应该自动化所有工作？是否应该发展最终可能超过并取代我们的非人类思维？是否应该冒险失去对文明的控制？

在这些问题中，机器通过机械匹配所合成的非事实内容已经出现，另外三个问题离我们还很久远。从长远来看，应该肯定这封公开信所作出的思考，但不应对这封信断章取义。

如何看待一项新技术的发展，这是做技术预测时必须要有的一项认知准备。遗憾的是，我们人类总是习惯于线性思维（这符合人类自然的认知模式：节省能量与快速计算），但这种认知配置很容易出现认知偏差。其中最常见的认知偏差是对于技术近期与远期影响的判断出现不对称性，这个关系被美国科学家罗伊·阿玛拉（Roy Amara）发现，并形成了所谓的阿玛拉法则（Amara's Law）。所谓的阿玛拉法则是指：短期内我们倾向于高估技术的影响，长期内我们低估技术的影响。

五、人机协同创新

人类进入信息时代以来，先后通过键盘（文字输入）和鼠标（图形界面）与信息交互，ChatGPT 这一模型基座的出现使得人们可用自然语言与信息世界交互。人工智能这一能力第一次在聊天、虚拟助理、语言翻译和内容生成等方面助力每一个普通用户，成为像水和电一样的通用资源。

随着 ChatGPT 与不同领域 App 结合，整合不同领域的实时数据，如旅行软件、购物软件、支付软件、在线订餐平台等，这样 ChatGPT 就演变为 xGPT。未来人们不必再自己去下载和使用各种 App 和网站，互联网的统一入口将由不同 xGPT 开启。每个人都可通过自然语言连接所需互联网信息，极大改变了生活、生产和工作的生态。

美国数学家、哲学家诺伯特·维纳（Norbert Wiener）在 1950 年出版了一本极具洞察力和先见之明的著作《人有人的用处：控制论与社会》，目的就是希望人类在技术世界的环绕中更有尊严、更有人性，而不是相反。未来将是人和人工智能共同进化的时代，人和人造物之间将如影随形、协作共进、相得益彰。机器是人类创造出来的，人类的作用就是在人和机器共处的社会中，不断用知识强化优化机器。我们需要用进化的观点去看待这个过程，最大限度发展种种可能性，而不是陷入“人机相斗”和“人机相害”的臆想中。

不论怎样，人类始终是人工智能高度、广度和深度的总开关和决定者。因此，我们一方面要警惕将人工智能等同于人类大脑的不切实际之举和“人工智能奴役人类”等杞人忧天之思，另一方面要善于利用人工智能这一帮手，在人机协同中进行更好的内容创造，创造更加美好的未来。

当前人工智能中知识引导方法长于推理（但其难以拓展）、数据驱动模型善于预测识别（但其过程难以理解）、策略学习手段能对未知空间进行探索（但其依赖于搜索策略），大模型提供了一种群智交互机制，可望建立人在回路的机器学习框架，最终充分协调数据驱动下归纳知识指导中演绎及行为探索内顿悟等不同学习手段和方法。

以 ChatGPT 为代表的复杂神经网络大模型这类新型人造物的出现，将给人类社会诸多生产、生活模式带来一次大变革。但这也为另外更多的奇妙“多样性”打开了一扇窗户，因为“人有人的作用”[10]。

参考文献：

[1] McCarthy, J., Minsky, M. L., Rochester, N. & Shannon, C. E. A proposal for the dartmouth summer research project on artificial intelligence, Retrieved March 12, 2023, from https://doi.org/10.1609/aimag.v27i4.1904.

[2] Jordan, M. & Mitchell, T. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349 (6245): 255-260.

[3] Yann, L., Yoshua, B. & Geoffrey, H. (2015). Deep learning. *Nature*, 521 (7553): 436 - 444.

[4] Pan, Y. (2016). Heading toward artificial intelligence 2.0. *Engineering*, 2 (4): 409- 413.

[5] Ashish, V., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan, N. G., Łukasz, K. & Illia, P. (2017). Attention is all you need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc, 6000-6010.

[6] Anderson, P. W. (1973). More is different. *Science*, 177 (4047): 393-396.

[7] Sébastien, B., Varun, C., Ronen, E., Johannes, G., Eric, H., Ece, K., Peter, L., Yin, T. L., Yuanzhi, L., Scott, L., Harsha, N., Hamid, P., Marco, T. R. & Yi, Z. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. Retrieved March 12, 2023, from https://doi.org/10.48550/arXiv.2303.12712.

[8] Émile, B. (1913). La mécanique statique et l'irréversibilité. *Journal de Physique Théorique et Appliquée*, 3 (1): 189-196.

[9] Peter, J., Bickel, E. A. H. & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187 (4175): 398-404.

[10] 吴飞. 走进人工智能 [M]. 北京：高等教育出版社，2022：167-175.

[责任编辑：高辛凡]