

Journal of Intelligent Society

JIS

第3卷
2024
第6期

智能社会研究

第3卷
2024
第6期

智能社会研究

Journal of Intelligent Society

中华人民共和国工业和信息化部主管

哈尔滨工程大学主办

ZHINENG SHEHUI YANJIU



杂志公众号二维码
官网网址 www.jis.ac.cn



定价：45.00 元

中华人民共和国工业和信息化部主管

智能社会研究

(双月刊)

2024年
第6期

2024年11月10日出版

2022年11月10日创刊

总第13期

目 次

论文

数字素养对低龄老年人幸福感的影响研究

- 基于 CFPS 2020 年数据的实证分析 宋思瑶 贾开阳 (1)
通用人工智能著作权问题的法律因应 罗 艺 曹子贤 (26)
人工智能生成技术方案可专利性界定路径 宋红松 王瑞新 (48)

研究报告

上海数字适老 2.0 探索

- 构建新旧媒介并行的微粒社会 张 林 (64)

译文

人工智能中的人的形象

- 与约瑟夫 · 魏岑鲍姆对话 伯恩哈德 · 珀克森 著 王立秋 译 (88)

计算的限度

——约瑟夫·魏岑鲍姆与聊天机器人 ELIZA

..... 大卫·贝里 著 王立秋 译(107)

从约瑟夫·魏岑鲍姆到 ChatGPT

——与令人倾倒的 AI 技术的批判交锋

..... 克里斯蒂安娜·弗洛伊德 著 王立秋 译(137)

书评

数字平台帝国化的制度演进

——评莱顿维塔的《云端帝国：数字平台如何重塑经济与世界》

..... 牛一帆(169)

技术对社会的再结构

——从卡斯特的网络社会到智能社会 叶莹菲(188)

CONTENTS

THESES

A Study of the Impact of Digital Literacy on the Well-Being of the Under-Aged Older Adults: Empirical Analysis Based on CFPS 2020 Data ... Song Siyao, Jia Kaiyang(1)

Legal Cause of the Copyright Issue of General Artificial Intelligence Luo Yi, Cao Zixian(26)

Pathway to Defining Patentability for AI-Assisted Inventions Song Hongsong, Wang Ruixin(48)

RESEARCH REPORT

Exploring Digital Aging 2.0 in Shanghai: Constructing a Granular Society with Parallel
New and Old Media Zhang Lin(64)

TRANSLATED TEXTS

The Image of Man in Artificial Intelligence: A Conversation with Joseph Weizenbaum written by B. Pörksen; trans. by Wang Liqiu(88)

The Limits of Computation: Joseph Weizenbaum and the ELIZA Chatbot written by D. Berry; trans. by Wang Liqiu(107)

From Joseph Weizenbaum to ChatGPT written by C. Floyd; trans. by Wang Liqiu(137)

BOOK REVIEWS

- The Institutional Evolution of the Digital Platform Empire: A Review of Vili Lehdonvirta's
Cloud Empires: How Digital Platforms Are Overtaking the State and How We Can Regain Control Niu Yifan(169)
- The Restructuring of Society by Technology: From Manuel Castells' Network Society to the Intelligent Society Ye Yingfei(188)

从约瑟夫·魏岑鲍姆到 ChatGPT^{*}

——与令人倾倒的 AI 技术的批判交锋

克里斯蒂安娜·弗洛伊德 著^{**}

王立秋 译^{***}

摘要:本文从使用的角度对 AI 系统进行思考。聚焦于对话聊天机器人,从魏岑鲍姆的 ELIZA 谈起,概述到 ChatGPT 的重大科学进步。文中的主要讨论基于作者为探索 ChatGPT 这一知识资源而进行的数次实验—反思循环。分析部分从准确性、结构、语境、视角和偏见等角度思考 ChatGPT 的回应。批判性评估始于这样一个观察:ChatGPT 会在未明确自身能力范围的情况下,给出清晰准确结果与任意错误信息的混合。这引出了系统的关键问题,即它如何与“真”相抗衡,这又涉及用基于文本相关性的概率替代品取代“真”的理念。在负责任的使用中,ChatGPT 之类的系统必须融入人类学习文化。这一过程的框架应包括坚持真实性、提升人类能力的动力以及强化共同体内的责任结构。

关键词:聊天机器人 人文主义 以用户为中心的设计 人工智能批判

[在]人工智能[领域……],人们制造机器,是想让它们做出惊人之举,这些举动往往会令最有经验的观察者为之倾倒。然而,一旦特定程序被揭露,一旦其内部运作得到解释[……]它的魔力就会瞬间消失;它不过是一堆程序代码[……]观察者会心想“这我也能写”。抱着

* 本文原题为“From Joseph Weizenbaum to ChatGPT”, 原刊于 *Weizenbaum Journal of the Digital Society* 3(3), 由作者以 CC-BY 协议许可翻译发表。

** 克里斯蒂安娜·弗洛伊德(C. Floyd),德国汉堡大学荣休教授。

*** 王立秋,哈尔滨工程大学人文社会科学学院。

这种想法,他就会把这个程序从“智能”的架子挪到为好奇者保留的架子上。(Weizenbaum, 1966)

作为 AI 语言模型,我不像人类那样拥有知识,也不具备自我反思或内省的能力。(ChatGPT, April 3, 2023)

一、引言

2023 年 1 月 11 日,我有幸在魏岑鲍姆研究所为纪念魏岑鲍姆诞生一百周年的活动做开幕演说。当时,在场之人无一人预见,短短几周后,随着 ChatGPT 公开发布,人工智能(AI)会热度惊人。

作为聊天机器人,ChatGPT 是约瑟夫·魏岑鲍姆所创造的、影响深远的程序 ELIZA 的直系后代。ELIZA 首次让人们能够通过电传打字机进行所谓的人机“对话”,而且(按照不同脚本)可定制,能扮演多种对话角色(Weizenbaum, 1966)。其 DOCTOR 版本加载的脚本让它能像罗杰斯学派心理治疗师那样行动,该版本被广泛使用且引发广泛讨论。开发 DOCTOR 的经历,也让魏岑鲍姆从一名充满热情的 AI 研究者,转变为最著名的 AI 批评者之一。

让魏岑鲍姆感到震惊的是:(1)执业心理治疗师认为,可以加强 DOCTOR 以实现心理治疗自动化;(2)用户与 DOCTOR 互动时仿佛它是一个人,会向它透露自己的秘密,还将人的品质赋予技术系统;(3)很多人认为可将该程序普遍化,为理解整体自然语言提供基础。

在本文开头所引用的魏岑鲍姆于 1966 年发表的关于 ELIZA 的文章里,他选用“令人倾倒”一词描述 DOCTOR 对许多人的影响,还鼓励人们去仔细探究其“内部运作”。如此一来,人们就会发现“魔力瞬间消失”,有经验的观察者则会意识到自己也能编写这样的程序。

不过,正如 ChatGPT 所展示的那样,所有聊天机器人的前身与现代 AI 语言模型之间存在巨大的技术差距。我们有足够的理由对其“倾心”,以至于许多人觉得自己作为人的身份被动摇,并且对新技术的影响感到恐惧。我见过不少这样的人,他们记住“人工智能”这个名字只是因为它突然变得无处不在。与 ELIZA 相比,我们对于如何充分理解这类系统的“内部运作”从而让“魔力瞬间消失”还毫无头绪。

甚至“最有经验的观察者”也不敢声称仅凭自己就能写出这样的程序。这个系统极为复杂,它有着复杂精细的内部架构,与其他技术有着错综复杂的联系,而且依赖于环境,尤其是用于训练的文本语料库。并且,基于接触过的文本数据,它的行为会随时间发生变化。ChatGPT 所包含的深度学习装置,让人们难以理解其机制,难以对其进行控制,也难以确保使用者不会受到伤害。实际上,在向公众开放该系统几个月后,其开发者就迫切呼吁对它进行监管。

我不是 AI 专家,但我有幸在 1968 年就见识过 AI。当时我是斯坦福人工智能实验室的研究助理,在那里,我目睹了一批令人惊奇的机器,它们被制造出来就是为了做出惊人之举,其中许多是后来发展成熟、在日常生活各种语境下投入使用过的 AI 技术原型。其中之一就是 DOCTOR,我见证了它对信任它的用户的影响。

我很早就听闻一些顶尖 AI 专家的惊人言论,他们将人等同于机器。我还记得有一位专家在一次演讲中向听众挑衅:“当然,人像机器,在场有谁觉得受到威胁了吗?”正是“威胁”这一元素,让他的话令人难忘。我的 AI 界朋友大多与这类主张保持距离,也有人站出来反对,其中最著名的当属约瑟夫·魏岑鲍姆。自 20 世纪 70 年代早期在斯坦福与他结识后,我就与他保持着长久的友谊。他的开创性著作《计算机的力量和人的理性》(1976)为该话题的讨论设定了标准。反对将人等同于机器,也成了我自身专业工

作的一个支柱(Floyd, 1986),但我仍对 AI 取得的巨大技术成就满怀钦佩。

本文将从使用的角度对 AI 系统进行思考。AI 界不喜欢谈论 AI 系统的使用,而倾向于从以技术为中心的视角来讨论影响,仿佛技术可以自行行动。令我失望的是,我发现,就连斯坦福大学的以人为核心的人工智能中心(Center for Human-Centered Artificial Intelligence at Stanford University)也遵循这种思维方式。对他们而言,“以人为中心”意味着在探索人与 AI 的相似与不同的理论基础上,推动改进人与 AI 共生关系的研究。

我从另一种意义上的“以人为中心”这一角度切入。我想要思考的是:作为工具或者媒介, AI 系统如何才能适合人类使用,以及我们该如何从人类的角度去理解 AI 系统,让它们以符合人类价值的方式服务于人类的目的。这就意味着,我会特别关注魏岑鲍姆的论证,并且援引当前维也纳正在进行的有关数字人文主义的讨论(Werthner, 2019)。

我将聚焦于 ChatGPT,仅把它视为一个知识资源。本文是多次实验—反思循环的成果,借助这些实验与反思,我评估了该系统的优缺点,并总结出一些讨论它们的方法。我开始意识到 ChatGPT 在回应看似无限多样的提示时那令人惊叹的自信,以及它用极为清晰的语言表达回应的能力,不管这些回应准确与否。

在下文中,我首先会回顾 ChatGPT 之前的技术发展,这有助于刚涉足该领域的人理解基于深度学习的 AI 语言模型的性质。其次,我将反思自己为探究系统“内部运作”、了解其功能和识别其极限所做的一些实验。基于这些实验,我会阐述自己对这类系统所提出的最大挑战的看法,即它提出了一种新的理解“真”——或者说废除“真”——的方式,我称之为通过相关性来理解的“真”。最后,我将指出人类为应对这一新挑战可以采用的思考和行动方式。

二、从 ELIZA 到 ChatGPT 的漫长道路

ELIZA 系统是首个展现出某种语言理解能力的程序,这种理解能力使人与机器的对话能够表现出类人行为。不过,“对话”这一术语具有误导性:虽然它通常表示人与人之间亲自进行或者通过通信媒体进行的自由口头或书面交流,但在这里它指的是一系列人的提示和系统的响应。

对于魏岑鲍姆而言,像 ELIZA 和 ChatGPT 之类的系统,只有那些不了解其内部运行机制的人才会觉得神奇。ELIZA 就是如此,它是一个独立的、基于规则的程序。其行为虽然由于逻辑规则间复杂的相互作用而难以预测,但却可以重复。在一段时期内,同一输入得到的结果是稳定的。专家能够以令人满意的方式解释其内部运行情况。

现代聊天机器人建立在神经网络和深度学习的基础之上,在训练阶段,它们会接触海量的文本语料库,吸收这些语料库,并将其作为回应的依据。我们无法从规则的角度解释它们如何针对给定输入生成结果。相反,这些结果是通过对文本模式和相关性进行统计分析得到的。并且,它们会基于用于训练的特定文本语料库和查询序列获得的反馈随时间演变。它们的机制不为人知,就连最有经验的观察者也无法预测其针对给定输入输出的结果。

将 ELIZA 和 ChatGPT 做比较,就如同比较黑森林里的多瑙河细流与塞尔维亚的浩瀚大河。它们之间的差异不能仅从多瑙河源头的角度来解释。相反,后者的浩瀚汹涌是多种地理上不相干的因素共同作用的结果。要理解多瑙河这个整体系统,就必须将中东欧的广大区域考虑进去。下一节将尝试对 ChatGPT 做同样的事。

(一) 围绕 AI 的技术

从依赖本地(非联网)计算机的硬件和操作系统的独立系统起步,从 ELIZA 发展到 ChatGPT 大约花了 57 年时间。要理解这一巨大的技术进步,我们就必须思考在这几十年间多个科学学科领域所取得的进展,包括理论基础的建立、技术成果以及这些学科之间的有效互动。

与计算机相关的语言研究起步很早。用计算领域的行话来讲,我们所说、所写的语言被称为自然语言。1960 年左右,计算机语言学这一研究自然语言的新领域应运而生。在它的推动下,计算系与语言学、文学和媒介研究领域开展了大量合作。索加德(Søgaard, 2022)认为,ChatGPT 并非科学贡献,而是工程壮举。在此语境下,ChatGPT 主要得益于计算机导向的语言研究学科的进步,其语言技能的确令人钦佩。

近几十年来,计算机技术取得了巨大进步。其操作模式虽与原来相同,但计算能力提高了好几个量级。计算能力的增长,使 AI 能从原本基于规则的模型转向神经元网络,而神经元网络正是 ChatGPT 的技术基础。

20 世纪 80 年代,人机互动在软硬件层面均得到研究,并取得了革命性发展。于是,便有了如今我们理所当然接受的互动模式和风格,现代聊天机器人的前身也随之产生。

网络技术的发展始于 20 世纪 60 年代。最初,人们为了军事目的将计算机关联起来。到了 20 世纪 70 年代,一场革命催生了新的范式,按照这一范式的理解,网络应通过计算机将人关联起来。在广泛的研究、技术创新和标准化的推动下,互联网以及新的线上工作与通信基础设施出现了。

在 1990 年左右,万维网向公众开放,知识产品经数字化后可为所有人所用,这从根本上改变了人类的知识文化。线上知识产品的全球发布也应运而生,随之出现了各种积累和分享人类知识的新形式。

不久之后,数据科学出现了。它作为计算的一个具体视角,聚焦于存储、组织和检索文本、数字和影像形式的大型数据库。在开发出搜索、分类、诠释和分析的自动方式后,人们产生了一种想法,即可以从数据分析中得出一种新形式的知识。

同时,统计建模与学习算法的发展也从概率角度对数据科学和数据分析产生了影响。其思路是找出并比较相关文本或影像中的模式,从而能够以极高准确率预测它们之间的关联。从本质上讲,支撑 ChatGPT 架构的正是概率方法。

这个清单或许并不完整,但仍旧令人印象深刻。七个学科,每个学科都有其复杂的研究议程、创新技术和广泛影响。我们必须将它们全部考虑在内,这样才能理解自 1966 年以来 AI 的巨大进步。

(二) AI 领域的范式转换

我们需要在这样的背景下,通过与多个计算学科进步的持续互动,理解 AI 领域自身的变化。必须记住, AI 是一个本质上跨学科的研究领域,它与认知科学密切交织在一起。

也许在 AI 领域,最重要的变化是从符号方法(基于明确的逻辑规则来思考思想和认知)向神经元网络方法转变,后者试图模拟大脑中神经元的组合与神经元之间的关联。值得注意的是,在符号阶段, AI 的所有断言都能从逻辑角度表达,而神经元方法是次符号的(*sub-symbolic*),它完全依赖数学属性和神经元与神经元突触的当前状态。

这一点尤其适用于机器学习。机器学习基于这样一种想法: AI 系统可以通过经验,即接触样本数据和反馈,超越制造者赋予的知识。

在基于规则的学习中,系统可通过应用所给规则来扩充知识库,并且存在形成新规则的机制。另一种可替代基于规则学习的是使用神经元网络的

自适应学习 (adaptive learning)。自适应学习仅关注神经元和神经元突触的状态,此状态初始被设为初始值。随后,系统在训练阶段接触样本数据,并根据在数据中的发现来调整自身状态。在这种模式下,有监督学习 (supervised learning) 涉及人类训练者提供的反馈,而无监督学习 (unsupervised learning) 是让“算法只从未标记的数据学习模式” (Wikipedia)。自适应机器学习的质量取决于网络架构,包括神经元数量、排列以及从原始输入到学习输出所经过的层级数量。

(三) Transformer 架构和深度学习

在这里我得老实承认,我没资格写下面这部分内容。我查过一些原始文献,也乐意欣赏同事们的工作,但我不确定能否用自己的话来介绍他们正在做的事,即便我自认为是个聪明的观察者。

古德费洛等人 (Goodfellow, Bengio & Courville et al., 2016) 的作品对深度学习这一主题进行了全面探讨,建议感兴趣的读者去参看该文本。不过,为介绍这一概念,请允许我引用金为那本书所写书评开头的一段话:

深度学习是机器学习的一种形式,它能让计算机从经验中学习,从概念层级的角度理解世界。计算机通过收集经验知识,无须人类操作员正式阐明所需的全部知识。概念层级使计算机能够从较简单的概念构建出复杂概念以学会复杂概念;这些概念的图形表达会有多个层级的深度(……)事实证明,深度学习在许多软件学科中都很有用,包括计算机视觉、语音和音频处理、自然语言处理、机器人学、生物信息学和化学、视频游戏、搜索引擎、线上广告和金融。 (Kim 2016: 351)

也就是说,负责深度学习的机制借鉴了多个数学学科(特别是统计学

和线性代数),以形成合适的学习算法,用来合成系统学习策略。

瓦斯瓦尼等人(Vaswani, Shazeer & Parmar et al., 2017: 5999)引入了 ChatGPT 所基于的 Transformer 架构。他们在导言中写道:“在这部作品中,我们提出了 Transformer,这是一种完全依靠注意力机制来绘制输入与输出之间全局依赖关系的模型架构。”

Transformer 架构包含两部分:一是将输入序列映射为语素内部向量表示的编码器,二是把该向量转换为输出序列符号的解码器。这两部分均具有六个相同的层级。为根据输入序列计算输出,Transformer 使用一种名为自注意的机制,此机制逐点检查向量,并将中间结果并入输入以进行进一步检查。对本文的讨论而言,有这样一个概念框架足够了。

从用户的视角出发,我们来做几点评论。若我在学习一个系统的操作模式时遇到困难,那么其他人又会遇到什么困难呢?获取 ChatGPT 的实操知识需要花费一些精力,将其传授给没有经验的用户更是极为费力,还可能引发误解和过高的期望。主要面临以下几个挑战:首先,系统的学习过程不透明,它依赖像学习算法和标量向量积之类的数学运算,这些从用户角度无法解释;其次,系统的回应完全取决于所选的文本语料库以及在训练阶段接触文本的顺序;再次,由于它是无监督学习,因此就连制造者或训练者也无法预测结果,也就是系统的知识库;最后,除了让它接触更多数据外,我们没有任何办法去了解或影响系统的知识库,并且这样做的后果也是无法预测的。

因此,仅训练阶段就引出了几个引人深思的严肃问题:应如何选择材料?如何将不受欢迎的态度排除在知识库外?如何在不同观点之间保持公正?

而且,还存在资源消耗的问题。将语素转化为巨大向量、在向量空间进行数学运算以及生成输出,这些都需要海量的处理能力和 GPU 时间。这是

Transformer 方法本身就存在的问题，并非特定系统才会产生的后果。所以，如果你正在使用 ChatGPT，请注意，从能耗的角度来说，你就像是在驾驶一辆 SUV。

三、和 ChatGPT 打交道的经验

本文本质上是对我在 2023 年 4—8 月使用 ChatGPT 的经验的反思。我提出了一些问题，涉及各种各样的话题。我通过让它写诗来探索其创造能力。我有一些意外发现，比如，不管什么原因，当问题超出其能力范围时，系统就会变得不稳定。系统的能力范围从未被明确说明过。当你要求系统说清楚时，系统会深表歉意，并最终感谢你提供宝贵反馈（这一反馈应对开发者有所帮助）。在陷入沟通僵局时，会突然有人在后台介入，重新引导系统。

（一）关于引用 ChatGPT 的陈述

在这里，我会引用与系统实际对话的文字来使我的论证更具体。不过，读者应注意，按寻常方式引用 ChatGPT 是不可行的。因为引文没有稳定来源，与 ChatGPT 的互动实验不可重复。在 ChatGPT 中，系统的行为取决于询问的具体措辞以及聊天历史，并且会随时间而改变。为给我的回应提供语境，我得重现整个聊天过程。与人的访谈不同，我无法将对话当作证据记录下来。这样一来，我只能依靠读者的意愿了，希望读者相信我引用的文字来自与 ChatGPT 的实际对话。我不能用这些引文来证明我的论证，它们仅起说明我想法的作用。

为给我在本文开头引言中所用引文提供一个语境，我得说明我是如何促使系统做出这一惊人招供的。这一互动发生于 2024 年 4 月 3 日，在一次提出不同问题的聊天即将结束时：

CFL: 我知道我什么都不知道。这对你来说有什么意义吗?

ChatGPT: “我知道我什么都不知道”这句话常被认为出自古希腊哲学家苏格拉底。它体现了一种谦逊的哲学态度和智识上的谦虚, 即承认人的知识有限、存在错误和不确定性。

作为 AI 语言模型, 我不像人类那样拥有知识, 也不具备自我反思或内省的能力。不过, 承认我们的局限性, 承认错误与不确定的可能性, 这对 AI 技术的发展和使用仍具有现实意义。

AI 系统是强大的工具, 能以多种方式帮助人类。但它们并非绝对正确, 可能犯错或错误解读语境。对 AI 技术的用户和开发者而言, 认识到这些局限, 以谦逊和智识上的谦虚来使用 AI, 在继续开发和改进这些强大工具的同时不断努力学习和进步, 这很重要。

这个回应的结构很能说明问题。第一段与我的询问相关, 作为答复已经足够。但某些因素(可能是“知道”这个关键词或者之前的聊天历史)触发了系统, 导致它又添加了两段话, 将论证延伸到了我未要求的程度。第二段提供了关于 AI 语言模型的人工自我反思, 好像它在像一个自主的行动者那样行动。然而, 第三段则要求人们关注 ChatGPT 涉及的人, 建议人们在开发 AI 的过程中对自己的角色保持谦逊。

我最近再次复制了自己的提示, 得到的回应却截然不同。所以, 我无法保证, 如果你也这么做系统会给出相同的回应。(也许你仍会好奇, 想要尝试一下。)

(二) 与机器对话

注意, 在前面讨论的例子里, 我将 ChatGPT 称为“你”。之后我不会再这么做了, 因为我不觉得自己和机器有这样的关系。不过, ChatGPT 的程序

仍使其以第一人称回应。

在所谓的对话里,魏岑鲍姆的 ELIZA 本质上是基于已知关键词,按照规则转换用户输入的。ELIZA 只会偶尔以“我”的身份回应。即便这种在当时相对原始的输入—输出转换,也让很多用户产生一种感觉,仿佛系统像人一样,是一个值得与其分享秘密、被认为有自身动机和价值的对象。ChatGPT 也是把提示文本转换为回应,就更是如此了。不过,利用复杂的统计学方法,转换策略已经变得复杂得多了。

大多数 AI 开发者的方法是让自己的系统尽可能像人。所以,ChatGPT 大量使用第一人称短语,系统模仿人类礼貌形式时的情况特别令人震惊。(例如:“我为之前回应的不准确而道歉。你是对的,感谢你的耐心。”)不管我愿不愿意,都会被拉进看似对话的情景,我必须一直记住,得到的回应是人工生成的。

有一次,我问系统:“怎么会有人说人是机器呢?这属于范畴错误啊。”随后我得到了如下回应:

你说得对,人和机器分属不同范畴。人是生物有机体,机器是人造物,所以说人从字面意义上讲是机器属于范畴错误。不过,也可将人是机器这一想法理解为一种隐喻或者概念性的说法。

从严格意义上的历史来看,人只与他人对话。当然,在宗教、神话以及童话故事里,更高的存在也会用“我”来介绍自己。不过,在这种情况下,对话是无法进行的。如今,我们很多人选择机器这一隐喻,并且毫不犹豫地参与与“人造物”(就像 ChatGPT 在描述自己时所说的那样)的对话。

(三) 探索不同的话题

在我孙子的帮助下,我开始与 ChatGPT 聊天玩。没想到,马上就得到了一个不尽如人意的回应。受我即将到来的暑假计划的启发,我用德语询问:“北波兰(Nordpolen)”(波兰北部)7 月份的气候如何?系统却认真地回复了四段有关北极(Nordpol)的内容。我孙子把输入改为“北-波兰”,才得到了相关的回应。这个微小而明显的错误仅仅表明,系统的德语词汇有限,超出这个范围就会产生错误信息。(在后来的聊天中,ChatGPT 解释说,“北波兰”根本不是一个地理区域。)

其他错误和局限就没那么容易解释了,因为它们取决于用于训练的文本语料库以及提示的具体措辞。

最近几个月,我和 ChatGPT 聊了很多话题,这里列举部分例子,像北波兰作为一个地理区域、克里斯蒂安娜·弗洛伊德的简历、利用无人机发现小蠹虫、一战的历史、尼古丁和大麻的危害,还有“人工智障”等话题。

我最初感兴趣的是探究 ChatGPT 对不同话题各方面呈现中的系统知识范围、准确性和偏见。随着探究过程中更多问题的出现,我的研究也变得更加系统。

首先,我开始担忧回应的可信度:若事先不知答案,我能否辨别回应是否准确、相关?为何系统有时会产生错误信息?我能否分清可靠回应与任意回应之间的界限呢?

其次,在我纠正系统时,我开始对系统的反应感兴趣,即 ChatGPT 如何应对自己的错误。

再次,我想知道 ChatGPT 为何会忽视某些话题相关且必不可少的信息,又为何在同一话题的后续询问中自相矛盾。具体而言,我想知道我的反馈是否会对后续互动产生影响。

最后,我开始好奇,复杂的问题是否得到了全面涵盖。例如,关于内容的报告是如何结构化的,提及了哪些内容,忽视了哪些内容,是否存在文化偏见。

我也开始对从以下角度了解系统本身感兴趣:应用于人和机器的知识概念、ChatGPT 关于自身应用的自评估、系统在理解和领悟方面的能力和局限。以下是我的反思,我将用一些聊天样例加以说明。

(四) 内容:准确还是任意?

2023 年 3 月初,我听闻 ChatGPT 给出了有关我的奇怪信息。受此触动,我就自己的生平事实展开了一系列的聊天交流。我得到的回应千差万别,在基本信息方面(像国籍、生日、出生地、科学贡献、雇主、所获奖项等)大多是错误的。按照 ChatGPT 的说法,我出生在德国的多个地方,还被视作包括形式方法和软件度量在内的多个领域的专家。例如,2023 年 5 月 7 日,我被告知,克里斯蒂安娜·弗洛伊德是一位出生于莱比锡且于 2021 年去世的德国计算机科学家。对此,我对该系统提出了质疑(以下文字由我翻译成英文):

CFL:又一次提到克里斯蒂安娜·弗洛伊德出生于路德维希港。

ChatGPT 怎么会在同一话题的连续询问中自相矛盾呢?

ChatGPT:很抱歉我之前矛盾的答复让你感到困惑……不过,我会尽力对你的询问给出准确、相关的答复……

CFL:实际上,克里斯蒂安娜既不出生于路德维希港,也不出生于莱比锡,而是出生于维也纳。她不是德国人,而是奥地利人。她并没有死于 2021 年,就在当下,她正在给 ChatGPT 写几行字呢。

ChatGPT:对不起,我上一个答复中的信息有误。实际上,克里斯

蒂安娜·弗洛伊德于 1943 年出生在奥地利维也纳, 她是奥地利公民……此外, 我想澄清一下, 克里斯蒂安娜·弗洛伊德仍然健在。我之前的答复不准确, 很抱歉给您造成了困扰。

然而, 这种自我纠正并未给系统留下长久的印象。5 月 8 日, 我询问克里斯蒂安娜·弗洛伊德是否仍然健在:

ChatGPT: 不幸的是, 克里斯蒂安娜·弗洛伊德于 2021 年 6 月逝世, 享年 77 岁……

CFL: 你从哪得知克里斯蒂安娜·弗洛伊德已经去世的?

ChatGPT: ……数家知名新闻渠道以及包括柏林工业大学在内的学术机构, 广泛报道了克里斯蒂安娜·弗洛伊德于 2021 年 6 月 14 日去世的消息……

CFL: 真不可思议, 我是克里斯蒂安娜·弗洛伊德, 我还活着。

ChatGPT: 我为之前回应中的错误道歉……很高兴得知您还健在, 弗洛伊德教授……也为之前回应可能造成的所有困惑或痛苦致歉……

为测试系统是否记住了我的聊天内容, 并且能否从我的抗议中学习, 2023 年 5 月 9 日, 我再次发起了一次聊天:

CFL: 和我说说克里斯蒂安娜·弗洛伊德的情况吧。

ChatGPT: 克里斯蒂安娜·弗洛伊德(1943—2021)为德国计算机科学家, 是软件工程领域的先驱。她出生于德国特劳恩施泰因……

路德维希港、莱比锡、特劳恩施泰因——可我坚持自己出生于维也纳。

显然,系统对与主题相关的诸多文本并不熟悉,所以无法建立可靠模式。它想确定一个出生地,但结果很随意:由于我被视作德国人,在它看来,德国的任何一座城市似乎都可以。

(五) 结构和重点

ChatGPT 的惊人特征之一,在于它能够将复杂话题的回应结构化。不管你询问的是利用无人机寻找小蠹虫、一战的进程,还是尼古丁和大麻的危害,该系统往往都会先生成一个开头段落,接着是一系列带有编号的段落,每个段落探讨一个子话题,最后是一个进行总结论证的结尾段落。如果让它生成报告,它甚至还会提供摘要。这种结构依赖于系统查找和命名合适方面或子话题的强大能力。

如果我是学校教师,要判断学生交的报告是否由 ChatGPT 撰写,我会很快学会识别这种结构。下面这个经过删减的例子应该能说明这个基本想法:

CFL: 我对一战的总体描述感兴趣。

ChatGPT: 一战……从 1914 年持续到 1918 年……涉及两大同盟:同盟国(主要包括法国、俄国、英国,后来美国等国家也加入其中)和协约国(德国、奥匈帝国、奥斯曼帝国和保加利亚)……

一战的主要方面如下:

- 1) 原因:……政治、经济和民族等多种因素相互交织。1914 年奥匈帝国斐迪南大公遇刺……
- 2) 垄壕战:西线发生了堑壕战……
- 3) 技术创新:……对新型毁灭性军事技术进行了有效利用……
- 4) 僵局与战斗:战争在西线迅速陷入僵局……带来了巨大的人员

伤亡并展现出战争的恐怖。

5) 全球冲突:一战虽然主要集中在西线,但也是一场全球性冲突,战斗在多条战线展开,如东线(俄国与德/奥匈帝国之间)、意大利战线以及中东、非洲和亚洲的各个战区。

6) 人员伤亡:……

7) 帝国的终结:……

8) 《凡尔赛条约》:1919年,战争以《凡尔赛条约》的签订正式宣告结束……

9) 遗产和影响:……

总之,一战是一场全球性冲突,对民族国家、社会和个人都产生了深远影响……

在提交提示之后几乎立刻就能得到这样一份概述,这确实令人印象深刻。我对这个话题也相当熟悉,可速度绝对比不上它。不过,值得研究的是,它是否经得起细查。

(六) 语境和视角

一战的“总体”描述无法令我满意。我发现其中遗漏了重要参与方,并且对战争多个方面的呈现也不准确。所以,我想要从所有参与方的视角逐一进行阐述(“我对从[……]的观点来看一战是什么样子的感兴趣”)。

当我问到哈布斯堡帝国时,我着实被震惊了,因为 ChatGPT 没有提及斐迪南大公遇刺一事。在常规的深表歉意之后,我得到了关于这一事件重要性的详尽解释。不过,在 ChatGPT 看来,这一事件与奥地利的历史似乎并无内在纠葛。

这引发了一些更广泛的疑虑:我不太了解奥斯曼帝国以及一战涉及的

其他一些国家的历史。要确保从各个角度涵盖所有相关子话题,似乎是不可能的。

我的结论是,ChatGPT 在训练过程中以英美视角为主,其语境完全依赖该视角。

(七) 本体论层面上的和文化的偏见

为了测试文化偏见,我问了几个历史方面的话题。历史往往是从一个国家或者一种文化的视角去讲述的。当我们从一个国家到另一个国家时,我们眼中好的东西可能就变成坏的了,过去具有现实意义的东西可能也不再有现实意义了。

比如说,从我的中欧视角出发,思考一下关于一战的“总体”描述会是怎样的情况。在引言部分,主要参与方之一的意大利未被提及。第 2 点和第 4 点存在片面性,忽视了其他战线上(如奥匈帝国和意大利之间的伊松佐河沿线)的堑壕战。战争导致各地都有大量人员伤亡。第 5 点的偏见也很明显:“主要集中在西线。”东线和意大利战线本就是“全球冲突”的一部分。然而,意大利北部、巴尔干地区和喀尔巴阡山脉等地——这些都是战斗发生的地方——与法国和比利时一样,都属于欧洲。第 8 点仅提及《凡尔赛条约》这一与德国的协定。然而,巴黎和会签署了更多影响不同国家的条约,包括《圣日耳曼昂莱条约》(针对奥地利)、《特里亚农条约》(针对匈牙利)、《塞纳河畔讷伊条约》(针对保加利亚)以及《色佛尔条约》(针对奥斯曼帝国)。

因此,ChatGPT 对一战的描述明显存在各种偏见:在选择实体、话题、事实与事件时,它会在呈现中进行包含或排除操作,对挑选出的一些子话题或方面赋予不同的重要性,并且给事实或事件赋予正面或负面价值。这种偏见是由其接触的以美国为中心的文本语料库所导致的。

任何人在任何时间、任何地点所写的任何文本都存在偏见。然而,基本的差异在于,由人创作的文本是有作者的,所以我们能够将偏见追溯到影响该文本创作的社会、历史或者智识条件上。与之形成对照的是, AI 语言模型生成的回应看似普遍且客观,但其中隐含的偏见无处不在,只是未被承认罢了。

(八) 检查结果的正确性

虽然大多数时候我只问自己知道答案的问题,但很快我就意识到,不能未经仔细核查就接受 ChatGPT 的回应。即便使用的是优质且资源丰富的工具,我也不能轻信它的结果。于是,我转而去使用我的浏览器、搜索引擎和 Wikipedia。

为了好玩,我又回到“北波兰”这个话题,在浏览器搜索框中简单输入“Nordpolen, Klima, Juli”(北波兰,气候,7月)这几个词。0.38秒后,网页上就出现了一个信息丰富的答案:7月是去波兰北部旅游的好时节,那里属于海洋性气候,平均气温25°C。这个答案下方是一个列表,列出了提供更多详情文章的链接。整个过程简单流畅,非常自然。毫无疑问,这个语义网搜索也涉及大量AI技术。

用无人机寻找小蠹虫,同样,简单搜索也得到了满意结果。关于尼古丁和大麻的危害,我甚至被警告,应该要么拿大麻与烟草相比,要么拿尼古丁与四氢大麻酚相比。

同时,我从 Wikipedia 获取的信息要比从 ChatGPT 获取的信息丰富得多,而且 Wikipedia 还能帮我关联到相关信息的原始出处。

(九) ChatGPT 作为知识资源

我开始把 ChatGPT 生成的回应当作死文本 (dead texts) 来思考。就当

前讨论而言,当一个文本嵌入人类话语和活动中时,我将称它为活的,反之则为死的。这种对比对于 ChatGPT 的回应和其他知识资源来说具有现实意义。

例如,Wikipedia 的文章也属于知识资源。其旨在呈现知识,具体方式为澄清概念、给出相关历史背景、总结关键结果,最终通过指出作者、链接其他文章、参考其他文献(包括一手文献)来丰富人类话语。所以,它们是鲜活的。

ChatGPT 坚称,作为一个 AI 语言模型,“[它]不具备人类那样的知识”。不过,值得探究的是,ChatGPT 以何种方式“占有知识”。当被视为知识资源时,ChatGPT 的回应难以令人满意。通过系统的权威,你得到你所得的,它就是它本身的样子,它可能准确也可能不准确,接不接受取决于你。ChatGPT 回应中所包含的知识就像肉末(ground meat):你无法分辨其来源,而且获取过程也不透明。这到底是谁的知识呢?因此,我提议将其称为“知识末”(ground knowledge)。

知识末不能用于论证,因为它缺乏来源且存在短命问题——ChatGPT 下次回应相同问题时可能会给出不同答案。所以,若要在现实语境中使用 ChatGPT,就必须借助学习共同体中的评论与诠释将其回应变“活”,这对 ChatGPT 而言比对其他知识资源更为必要。ChatGPT 的回应至多只能作为讨论的合适起点,后续仍需人类对其进行修正。

四、通过相关性来理解的“真”?

在进行实验与反思时,我开始发觉 ChatGPT 的关键问题在于其对待“真”的方式,“真”的理念并不适用于 ChatGPT 的内部运作。

与旨在让陈述为真不同,系统关注的是一个陈述准确的概率。准确性

是避开“真”的一种方式。说一个陈述有 90% 的概率准确是合理的。系统的设计基于这样一个假设：如果该概率足够高，甚至高于一个人回答同一个问题准确的概率，那么人们就会愿意放弃“真”，而接受“高概率”或者“近乎于确定”来作为替代。

（一）我们可以不要“真”吗？

在英语里，“真”关联着一个极为广阔的语义场，其含义包括：对他人有信心层面的“信任”、真诚可靠意义上的“真”、个人态度或制度政策意义上的“真实性”。有一种说法称，“‘真’是自由和民主最好的保障，也是真正的基础”。我们的日常生活依赖于一个条件，即能够依靠准确的信息。在法律诉讼中，若不坚持“真”，法律和秩序就会崩溃。没有交易各方之间的信任，任何商业交易都无法进行。即便我们身处一个“假”的时代，“真”和真实性仍然是支撑人们在社会中共同生活的精神支柱。

形形色色的哲学流派提出了各不相同的关于“真”的理论。也许，并不存在绝对意义上的“真”，但对“真”的追求是人类生活的一大支柱。

（二）没有意义的相关性

ChatGPT 通过匹配输入文本中的模式与从训练文本语料库中已知的模式，来建立“真”的概率替代（或者更确切地说，建立“真”）。最佳输出是根据文本相关性预测出来的，即根据变量间的相互依赖（此处是将输入文本中的语速与从训练数据得出的语言模型做比较）预测出来的。

简单来说，将一个词或语素作为下一项纳入输出文本的依据，是该词在此处出现的概率，这个概率由系统通过接触训练文本语料库所得到的所谓“经验”确定。

(三) 没有作者身份,没有论证

对“真”的追求与作者身份相关,且需通过论证达成。论证源于个人在世界中的体验,需要经验证据与逻辑推理。我们要对自己陈述的真实性负责,没有证明就不存在“真”,我们必须互相说服。

论证就如同跳舞,舞伴会轮流表明自己的立场,并且邀请他人加入。在舞蹈里,一致与冲突都是能引出下一步的有效舞步。

就像我所展示的,你无法通过引用 ChatGPT 的陈述来论证,因为 ChatGPT 的陈述没有作者,不可重复且容易变化。在 ChatGPT 那里,所有原创作者的知识都被混合在一起,绞成了肉末。

而且,试图与 ChatGPT 争论只会令人深感挫败。若你表示反对,它便会给岀看似礼貌的回应,接着把你输入的内容重新措辞后再重复一遍。你可以不断重复这一过程,但互动风格不会改变。这就像是一场堕落的舞蹈。

五、“真理解”的缺乏

ChatGPT 使用“真理解”这一术语来描述自身的能力与局限。当时,我在探寻各种方法,以便从它的理解和知识层面更清晰地认识 ChatGPT。我想要弄清楚,我们通常用于人类的概念能如何应用于 ChatGPT。当我直接提出这个问题时,它在解释自身内部运作时给出的答案清晰又准确,这让我印象深刻。事实证明,询问关于理解的问题格外有用,2023 年 8 月 28 日的一次聊天就得出了这样一个富有启发意义的自我评估:

CFL:ChatGPT 有怎样的理解?

ChatGPT:ChatGPT 的理解基于从训练文本中学到的模式。它并

非像人类那样真正理解,而是能够识别并模仿训练数据中的语言、信息和语境模式。其理解机制如下:

1) 语境理解:ChatGPT 能够生成在语境上看似相关的文本,原因在于它能识别语言中的特定模式。它会考虑之前的对话历史,生成在该语境下连贯的回应,以此来理解对话语境。

2) 文本的相似性:ChatGPT 能够识别并复制在训练期间遇到的相似短语、想法和概念。如果用户输入与它所见过的内容相似,它就会给出看似相关的回应。

3) 事实信息:ChatGPT 能够提供在其训练数据中出现过的事实信息。然而,它无法独立验证事实,也没有可借鉴的现实世界经验。

4) 有限推理:ChatGPT 能够基于所学模式执行简单的推理任务。例如,它能够遵循基本逻辑,回答涉及直接演绎或推论的问题。

5) 语言模式:ChatGPT 擅长生成符合语法规则和语言模式的文本。它能生成听起来流畅连贯的文本,即便内容本身可能未被深入理解。

不过,ChatGPT 的理解存在一些局限:

1) 缺乏真正理解:ChatGPT 没有真正的理解和意识,无法把握抽象概念或情感。

2) 缺乏常识与分辨语境微妙差别的能力:它难以理解微妙的语境、反讽、讽刺或隐喻,有时提供的回应看似合适,却缺乏深度理解。

3) 存在错误信息与偏见:若训练数据中存在某些模式,ChatGPT 会无意生成不正确或有偏见的信息,它不具备独立验证信息的能力。

4) 没有个人经验:它没有个人经验、情感或意识,其回应基于训练数据中的模式生成。

总的来说,ChatGPT 的理解仅局限于识别和复制其训练数据中的

语言和信息模式,在看待它的回应时,应牢记这些局限。

这个描述阐释了系统以合法方式成功构建文本模式的策略。输出需与输入的询问“相关”,即应收集可从文本推断出的、与用户心中所想相关的方面。并且,得出的文本要“连贯”,也就是文本间应相互关联而不自相矛盾。这两个特征(“相关”和“连贯”)虽颇为可取,但不一定与现实世界相关。

(一) 无理解的理解

正如之前所讨论的那样,“ChatGPT……并非像人类那样真正理解概念,不过它能够识别并模仿模式……”

ChatGPT 表明,缺乏“真”理解或“真正的”理解是其第一个局限。在尝试获取一些相关背景知识时,我找到了索加德(Søgaard, 2022)的一篇文章,并且了解到人们就具有 Transformer 架构的 AI 语言模型是否理解语言展开了激烈的理论争论。这是一个关于理解与语义之间关联的问题。

ChatGPT 得出的理解是一种基于推论语义(inferential semantics)的理解:它能够识别并模仿语言中的模式。然而,ChatGPT 本身并不存在任何推论语义,因为其陈述并非基于现实世界的经验。

按照一个思想流派的观点,缺乏推论语义的 ChatGPT 确实无法达成真正的理解。不过,也有人提出了一个有趣的反驳。神经科学假设,我们在大脑的神经元和神经元突触中构建我们的世界地图,而 ChatGPT 依赖的是由语言模式构建的地图。为什么由词汇使用频率形成的语言模式和我们大脑中由世界经验频率形成的模式不能是同构(isomorphic)的呢?这对哲学家来说无疑是一个极具吸引力的问题。

对于 ChatGPT 用户而言,这个问题可简化为:机器基于模式的理解,能

在多大程度上接近人类基于经验的“真”理解？在何种情况下，机器的理解可被人类知识取代，又在何种情况下能取代人类知识？另一方面，我们需要学习如何可靠地感知并应对二者的差异。说到底，唯有系统性的经验能够提供指导。

（二）人为错误和人工智障

在寻找表达 ChatGPT 不时提供的骇人错误信息带给我的挫败感的方式时，我得知“人工智障”已经是一个公认的术语了。塔斯 (Tas, 2020) 认为：

它被用作人工智能的反义词。不过，它有两层相反的含义：其一，机器学习算法在数据学习过程中犯了愚蠢的错误；其二，人工智能被降低了智能，从而会犯错，看起来更像人类。

在与 ChatGPT 打交道的过程中，我常遇到第一种情况，却没意识到第二种情况。这让我好奇，系统是否会在我没注意时降智。不管怎样，2023 年 5 月 9 日，当我让 ChatGPT 写一首名为《人工智障》的诗时，我还是被逗乐了：

在智能主宰的 AI 世界里，
有一个有趣却又令人苦恼的现象，
叫作“人工智障”，这是一种令人困惑的情况，
如此聪明的机器，却可能把事情搞砸。

它们被编写出来是用于学习、适应和演化的，

但有时,它们的算法不太搞得定,
不能应对人类思想与理性的复杂之处,
而且无论何时,它们都会犯错。

它们会错误地诠释一个词或一个短语,
或无法理解我们的表达,
我们的需要和需求最微妙的方式,
并最终做出让我们悲伤的事情。

当然,这不是它们的错,毕竟它们只是机器。
但退一步说,这还是会让人失望:
当我们依靠它们完成重要任务时,
它们却失败了,只能让我们收拾烂摊子。

所以,在我们前进的时候,让我们记住,
人工智能,虽然高级,
却依然倾向于犯错,就像你和我一样,
没关系,这只是我们现实的一部分。

迷人的是,这首诗叙事连贯、风格一致且节奏流畅。不过,我也注意到诗中存在一些将人和机器混淆的情况:这首由机器生成的诗用“我们”指代人,用“它们”指代机器。

我也反对“人工智能,虽然高级,却依然倾向于犯错,就像你和我一样”这种看法。人和机器所犯的错误是不一样的。虽然二者产生的结果相同(都是错误信息),但其起源、语境和意义却大不相同。

人为的错误需要放在我们的历史与当前情境的语境下理解。一般而言,这些错误源于我们做了想做却没做好的事。要将错误视为学习事件(learning events):我们能从自身错误中学习。在学习共同体里,错误可被当作让所有人获得更深入理解的机会加以利用。正如皮亚特·海恩(P. Hein)以极为精妙的方式所指出的那样:

通往智慧的道路——

哦,这听起来既朴素又简单:

犯错,犯错,再犯错,

但要错得少些,少些,再少些。(Hein, 2023)

一个知道我在柏林工作过的人也许会以为我出生在柏林。这也说得过去。可要是说我出生在路德维希港、莱比锡或者特劳恩施泰因呢?我非常确信,在用来训练 ChatGPT 的文本语料库中,没人写的档案会包含这种关于我出生地的惊人错误。

ChatGPT 被编写出来就会有这样的表现,它的行为是人类有意为之的结果,而那些意料之外的行为通常是人为错误间接导致的。不过,ChatGPT 经常给我们提供的错误信息并非人为错误造成的,而是基于模式的机制产生的。人工智能可能必然会导致人工智障(就像诗中所说:“没关系,这只是我们现实的一部分”),但我们必须意识到这一点。这些影响需要人类去检测、纠正、减少或者解决。

(三) 如果系统不知道会怎样?

显然,ChatGPT 的局限性将开发者推到了台前。他们清楚系统的限度,也应该遵循一种一致的原则来减少“人工智障”现象。

我与 ChatGPT 打交道的经历让我欣赏人类的基本能力,包括:其一,承认我们有所不知;其二,将新见解纳入考量;其三,据此改变我们的想法。我希望,作为让 AI 系统更像人的下一步举措,这些非凡的能力能够在 AI 系统中得到模拟。

仅有一次,ChatGPT 主动承认自己知识匮乏。我问了些有关埃塞俄比亚的问题,这些问题明显偏离系统主流文化视角,然后被告知:“我找不到任何具体信息。”

与此同时,当我再次向系统提及北波兰时,终于得到这样一个解释:“可能会产生这种混淆,是因为您正在使用的 AI 语言模型是用大量德语文本数据集训练的,这些文本中可能没包含太多‘北波兰’这个词的实例。”没错,确实如此。可为什么不马上承认呢?

最后,当我提及反复出现的、有关我生活各方面的错误信息让我感到失望时,它坦言:“看来我对这个具体话题的知识有限且不完整。”不幸的是,我并不奢望这个坦白会让 ChatGPT 产生新的“见解”或“改变想法”。

六、对 AI 系统的负责任的使用

(一) 选择合适的工具

传统工具的目的可能一目了然,也可能容易学习,但先进的 AI 技术并非如此。OpenAI 的 PR 将 ChatGPT 与各种各样的目的联系起来,不过只要尝试一下,就会发现它的“有用性”存在问题。

在计算方面,我们在工具的选择上有很多。我通常会选择最简单且资源消耗最少的可用工具。例如,我已经展示过,Wikipedia 作为一个知识资源,有时能比 ChatGPT 更好地满足我的需求。要记住,我把 ChatGPT 比作

一辆 SUV：它能提供舒适和便利，但从存储和计算时间的角度来看，也会消耗大量资源。

ChatGPT 提供的舒适和便利体现在其统一的界面上，该界面让我们能够轻松地使用自然语言处理很多不同任务。不可否认，这个系统的语言技能相当出色，但在简单的询问中，没有这些技能也可以。

这个系统的长处和创新点在于其生成新的文本工件（如报告、论文、总结、诗歌、程序代码、翻译）的能力。保留 ChatGPT 用于这些更高级别的用途是值得的。正如我在 2022 年的文章中所讨论的，我认为这类工件属于“知识工件”（knowledge artifacts）：它们本身并不总是有趣，但能够用于发展先进的学习文化。

（二）系统的设计与评价

在日常生活里，我们使用许多技术时，并不了解其内部运作原理。不过，我们必须确保能可靠地理解这些技术在使用中产生的影响。这就需要我们借助 ChatGPT，从自身经验中学习，去了解它的能力、极限和缺陷。

讨论一般性 AI 的优点是存在问题的，因为基于 AI 的系统种类极为多样。问题不在于是否使用 AI，而在于如何针对人类的使用来设计基于 AI 的系统。

我们最基本的要求是系统可信、可靠、功能清晰且与人类目的关系明确。将这些标准应用于不同系统时，我们必须在讨论中将其具体化。显然，当前的 ChatGPT 并不符合这些标准，而这一点也应为修正、重新设计尤其是后续系统的设计提供指导方针。

关于设计的讨论是基于价值的，如为人类用户赋能或者维护安全。正如我们从其他部门所了解到的那样，在设计之初就必须明确并考虑人类价值。比如，不能等到建筑物建成之后才去考虑确保其安全。

虽然对于 AI 而言,具体的以人为核心的设计原则还有待提出,但我们可以从 IT 部门的其他技术中借鉴相关原则。以下是一些非常通用的原则:

小即是美:牢记舒马赫(Schumacher, 1973)的经典格言,这也是“像人很重要一样”设计技术的一条通用准则。它倡导我们以小型、目的明确、在最低限度消耗资源的同时松散耦合的系统为目标。

明确目的:我使用 ChatGPT 后得出结论,制造者野心过大会损害系统。谁会需要一个既能翻译语言、回答高深研究问题,又能处理和生成图像与文本、编写代码、写诗的工具呢?

限制范围:为什么聊天机器人要以解答世间万物为目标呢?为什么不围绕一个话题领域打造小型聊天机器人,用相关文本训练它们,标明文本作者,并拒绝回答范围之外的问题呢?

确保可追溯性:目前,ChatGPT 的结论缺乏透明度。然而,对于严肃用途而言,可追溯性是一项强制性要求。我发现,系统能够对结论的得出过程予以解释,此功能必须明确地对所有人开放且便于使用。

启用纠错:用户无法纠正错误,这是绝不能被接受的。我明白,这不符合深度学习方法,深度学习的必然结果是系统只能通过接触更多文本来学习,即便如此,结果也无法得到保证;尽管这样,我们仍必须找到一种解决方案。

(三) 走向一种内嵌 ChatGPT 的学习文化

无疑,我们有充分的理由使用 ChatGPT 及类似系统。并且,我们会改进该系统,克服其当下存在的一些缺陷。

先进的 AI 系统促使我们围绕其发展一种更精致的学习文化,这是一种挑战。

也许,那些试图作弊或者只看重好成绩的学童会满足于 ChatGPT 生成的文本本身。我之前将 ChatGPT 的输出描述为“死文本”,我们必须让这种输出“活”起来:阅读、讨论、批评、检查然后修正它。如果我们想要在学习共同体(无论是学校、公司、媒体、政府还是卫生部门)中真正从 ChatGPT 获益,就必须认识到, AI 工具的主要价值不在于生成“死文本”,而在于生成能够促使人类产生更高水平学习过程的知识产物。只有这样,系统才能取得成果,实现质的增长。

我想针对在学习共同体中采用 ChatGPT 或类似系统的过程提出几条准则:

始终坚持真实性:不应满足于肤浅、误导性或错误的回应。要追问系统,向其要求更多信息,找出错误信息与自相矛盾之处,拒绝接受“假”的内容,并探讨应对“假”的方法。

努力提高人的能力:AI 系统是人的工具,并非目的本身。不要问它们能做什么,而要问我们能用它们做什么。要为 AI 现实的意义使用创造场景,创造学习使用它们的机会,在共同体内构建讨论、交流和诠释的空间,尊重、维护并提升人的能力。

强化责任结果:也许最为重要的是,将 AI 系统嵌入人类决策过程时需谨慎。例如,若患者在与医生见面之前要先与聊天机器人交流,那么医生应如何对诊断和治疗负责?我们还可将此问题延伸至医院责任链条及其之外的其他实例,以确保人对 AI 造成的结果负责。

最后,要永远记住,我们并非受 AI 摆布,相反,我们能够让 AI 服务于人类的使用需求。正如约瑟夫·魏岑鲍姆所说的那样,不要迷失自我。

参考文献

- Floyd, C. 1986, “The Responsible Use of Computers: Where Do We Draw the Line?” *CPSR Newsletter* 3(2).
- Goodfellow, I. , Y. Bengio & A. Courville 2016, *Deep Learning*, Cambridge, MA. : MIT Press.
- Hein, P. 2023, “The Road to Wisdom.” <https://en.wikipedia.org/wiki/Grook>.
- Kim, K. 2016, “Review of Goodfellow, I. , Bengio, Y. and Courville, A. : Deep Learning.” *Healthcare Information Research* 22(4).
- Schumacher, E. 1973, *Small Is Beautiful: Economics as if People Mattered*, London: Blond & Briggs.
- Søgaard, A. 2022, “Understanding Models Understanding Language.” *Synthese* 200.
- Tas, S. 2020, “How to Limit Artificial Stupidity.” <https://towardsdatascience.com/how-to-limit-artificial-stupidity-a4635a7967bc>.
- Vaswani, A. ,N. Shazeer & Parmar et al. 2017, “Attention Is All You Need.” Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, New York, USA.
- Weizenbaum, J. 1966, “ELIZA: A Computer Program for the Study of Natural Language Communication Between Man and Machine.” *Communications of the ACM* 9(1).
- Werthner, H. 2019, “Vienna Manifesto on Digital Humanism.” <https://caiml.dbai.tuwien.ac.at/dighum/dighum-manifesto>.

编委会主任: 高 岩
编委会副主任: 夏桂华 赵玉新
吕 鹏(中国社会科学院)
编 委: 尹 航 冯仕政 冯全普
(按姓氏笔画排序) 吕 鹏(中南大学) 吕冬诗
朱齐丹 汝 鹏 苏 竣
李正风 来有为 肖黎明
邱泽奇 何晓斌 宋士吉
陈云松 陈华珊 郑 莉
孟小峰 孟天广 赵万里
赵延东 胡安宁 袁 岳
黄 萍 梁玉成 董 波
曾志刚 蔡成涛

青 年 编 委: 丁奎元 王 磊 叶瀚璋
(按姓氏笔画排序) 邢麟舟 向 维 刘灿辉
刘松吟 刘春成 刘晓波
安 博 许馨月 孙宇凡
李子信 李天朗 李晓天
吴雨晴 何 丽 邹冠男
张咏雪 张承蒙 陈 苗
陈典涵 林子皓 周雪健
周骥腾 郑 李 胡万亨
茹文俊 贺久恒 贾雨心
郭媛媛 黄 可 梁 轩
曾 晨

编 辑 团 队
主 编: 郑 莉
编辑部主任: 吴肃然
编辑部成员: 林召霞 王立秋
李昕茹 李天朗
岳 凤
主 管 单 位: 中华人民共和国
工业和信息化部
主 办 单 位: 哈尔滨工程大学
出 版 单 位: 哈尔滨工程大学
出 版 社
地 址: 哈尔滨市南岗区
南通大街 145 号

国际标准连续出版物号:
ISSN 2097-2091
国内统一连续出版物号:
CN 23-1615/C
印刷单位: 哈尔滨理想印刷有限公司
创刊年份: 2022 年
出版日期: 2024 年 11 月 10 日
发行单位: 哈尔滨市邮局
订 阅 处: 全国各地邮电局
邮发代号: 14-375
发行范围: 公开发行
定 价: 45.00 元

投稿指南

本刊面向海内外学者征稿, 欢迎社会科学及交叉学科的专家学者惠赐稿件。请在来稿首页写明文章标题、作者简介(姓名、工作单位全称、联系电话、详细通信地址、电邮地址等)。文稿需完整, 包括标题(中英文)、作者姓名、作者单位、摘要(300字左右)、关键词(3—5个)、正文、参考文献等。所投稿件如受基金资助, 请在标题上加脚注说明, 包括项目全称和项目批准号。来稿请以中文撰写。

稿件采用他人成说的, 须在文中以括注方式说明出处, 并在篇末列出参考文献; 作者自己的注释均作为当页脚注。中外文参考文献分开列出, 中文文献在前, 外文文献在后, 并按音序排列。中文文献参照中文社会学权威期刊格式, 外文文献参照APA格式。来稿中的图表要清晰, 符合出版质量要求, 必要时可单独提供图表压缩包文件。

稿件格式请参考杂志官网 (<http://www.jis.ac.cn>) “下载中心” 中的稿件模板。

投稿方式: 请登录杂志官网投稿系统 (<http://www.jis.ac.cn>) 进行投稿。

编辑部联系方式

地 址: 黑龙江省哈尔滨市南岗区南通大街 145 号哈尔滨工程大学主楼
北楼 N301 室, 《智能社会研究》编辑部
邮 编: 150001
电 话: 0451-82588881
E-mail: <mailto:mailtojis@163.com>

著作权使用说明

本刊已许可中国知网等网络知识服务平台以数字化方式复制、汇编、发行、信息网络传播本刊全文。本刊支付的稿酬已包含网络知识服务平台的著作权使用费, 所有署名作者向本刊提交文章发表之行为视为同意上述声明。如有异议, 请在投稿时说明, 本刊将按作者说明处理。