心理学探新 2024,Vol.44,No.6,483 - 492
PSYCHOLOGICAL EXPLORATION

人工智能社会心理学的瓶与酒:范式与问题*

喻 丰 赵一骏 许丽颖
(武汉大学心理学系,武汉 430072)

摘 要:人类社会已逐步迈向人工智能时代,其已然掀起一场科学范式的革命。本文以“瓶与酒”之喻,总结人工智能社会心理学的两种研究取向。其一为以人工智能为方法赋能传统社会心理学研究问题之“新瓶旧酒式”研究,既包括文献综述、假设生成的智能化升级,也涉及模拟被试、理论抽取、计算建模、大数据分析等新方法研究。其二为以传统社会心理学方法研究人工智能新问题之“旧瓶新酒式”研究,包括静态能力和倾向考察、动态交互影响探索。最后,本文试图呼唤一种“新瓶新酒式”研究,以新方法解决新问题,真正回应人工智能对人类社会发出的冲击和挑战。

关键词:人工智能;社会心理学;研究范式;研究问题
中图分类号:B8409 **文献标志码**:A **文章编号**:1003 - 5184(2024)06 - 0483 - 10

2024 年诺贝尔物理学奖颁发给人工神经网络和机器学习的奠基人杰弗里·辛顿(Geoffrey Hinton),而诺贝尔化学奖授予开发出预测蛋白质结构的 AlphaFold 人工智能模型的丹米斯·哈萨比斯(Demis Hassabis)和约翰·乔普(John Jumper),在撼动全世之余也迫使人类反思人工智能对科学研究造成的冲击与挑战。简言之,人工智能带动了科研范式从经验主导向数据驱动的转型,缩短研究周期、释放科研潜力,掀起了一场新的科学革命(Wang et al.,2023;鄂维南,2024)。这种范式冲击同样波及了社会科学的研究,尤其是社会心理学的研究(Xu et al.,2024)。从方法学的角度,人工智能推动了心理学研究的数据化取向,为传统小样本、小效应量、情境依赖、文化差异的社会心理学提供了升级可能性。而在研究内容上,人工智能向社会抛出了包括接受与采用、风险与治理、赋能与威胁在内的诸多问题,而社会科学的回应仍滞于其后,略显不足(Bengio et al.,2024)。

针对于此,人工智能社会心理学展现出两种截然不同的研究取向。一者从范式入手,将人工智能算法技术应用于传统社会心理学的研究问题之中,提供一个崭新的研究方法和思路,我们称之为“新瓶装旧酒式人工智能社会心理学”。而第二则从问题开始,将人工智能视作社会心理学研究的对象,探讨人工智能自身与其引发的诸多社会性问题,扩展社会心理学的研究面向,我们称之为“旧瓶装新酒式人工智能社会心理学”。本文尝试以“瓶”喻范式,以“酒”喻问题,梳理人工智能社会心理学的多种可能取向,并分别总结其思维逻辑与底层结构,指导未来研究。

1 新瓶装旧酒:人工智能赋能社会心理学研究范式

科学史家托马斯·库恩(Kuhn,1996)认为,科学之所以为科学,是因为在特定时空内共享一套规定性的范式(paradigm),其指导了科学家看待问题的方式,研究问题的进路。社会心理学自肇始便希望成为一门独立的科学。为这门学科奠基的范式便是“假设 - 检验(hypothesis - testing)”的统计思想。进一步,我们可以将“假设 - 检验”的过程拆解为“假设 - 生成”与“假设 - 检验”两个关键步骤。这也恰好对应了人工智能赋能社会心理学研究范式的两条路径。

1.1 假设生成的方法创新

一切社会心理学研究自假设提出开始,一个好的假设决定了研究的质量(McGuire,1973,1997)。这依赖于研究者对本领域内先前研究和理论的吸收、理解和再加工。而随着研究数量的指数级飙升,研究者们似乎越来越难以做到穷尽。此时,能够理解并表达语义的大语言模型为加速科研生产提供了机会。一方面,人工智能技术可加速文献检索过程,理解文献内涵并生成基于特定关键词的文献综述。研究者们尝试了使用 ChatGPT 以改进布尔查询法(Boolean query)在系统综述写作中的弊端(Wang et

* **基金项目**:国家社科基金(20CZX059),国家自然科学基金(72101132)。
通信作者:许丽颖,E - mail:liyingxu@whu.edu.cn。

al. ,2023),并进一步开发了能够直接将用户意图转化为搜索就绪格式的自然语言查询(natural language query,NLQ)工具 Dimensions 辅助文献检索。另一方面,大语言模型可以直接生成在清晰性、原创性、合理性等评价标准中不亚于人类研究者平均水平的心理学假设(Banker et al. ,2024)。从方法上,大致有两条路径:微调路径(fine - tuning approach)与提示路径(prompt approach)。其一,是将在大型语料库中预训练好的模型放入特定语料库(例如,近几十年来所有社会心理学领域期刊和预印本文献组成的训练集)中进行微调,而后生成相应假设;其二,则不依赖于微调对特定问题解决能力的强化,仅仅依靠大语言模型(如 GPT4)对自然语言的充分理解,通过精准提示*的方式要求其生成相应假设。

1.2 假设检验的技术创新

图灵奖得主詹姆斯·格雷(2009)将科学研究的范式归纳为:描述之实验、概括之理论、模拟之仿真、分析之数据四种。虽然其中讨论了自然科学的进展,但随着社会科学与自然科学的交叉,其研究范式也大抵可以此分类。因此,本节将以此为框架,梳理人工智能对心理学假设检验提供的技术创新。

实验设计之模拟被试。无论是语言决定思维(如萨丕尔 - 沃尔夫假说、海德格尔之“存在之家”说),抑或反之决定于思维(如皮亚杰之语言发展说),语言都可被认作是思维的载体。被视为人工智能金标准的图灵测试(Turing test),也试图表达了相同的观点,即语言为智能的线索和表征。因此,我们似乎可以推测,基于海量人类语言文本学习的大语言模型具备模拟人类思维的能力,从而具有代替人类被试回答特定心理学问题的潜力(Argyle et al. ,2023)。于此,近年来一些研究试图回答大语言模型能否代替人类被试作为研究对象以供心理学研究者研究人类的心理和行为的问题。这一问题,实际上有两个重要的成分。前提一(又称图灵前提),即弱图灵测试,需要说明大语言模型足以给出类似于人的反应。而这只是语言逻辑的形式要求,并不涉及内容的一致性。前提二(又称对齐前提),则需要证明大语言模型输出的内容与人类对相同问题的思考是对齐、同一的。对于图灵前提而言,从理论上,人工智能的出现就旨在通过计算模型模拟人类认知的过程(Simon,1979),而神经网络技术更建立在对人类神经系统的理解和复制之上。无论人工智能实际上有无突破图灵测试(Jones & Bergen,2024),只要在人类的认知中其能够生成出类似于人类(甚至是专家)的语言表达即可(Dillion et al. ,2024a)。针对于对齐前提,实证研究发现,GPT 3.5 对包括 Clifford 等人(2015)开发的道德基础情境(moral foundations vignettes)之内的 464 个道德判断情境中,与相关以人类为被试的研究得到的结论具有极高的对齐程度(相关性达 $r = 0.95$, Dillion et al. ,2023)。GPT4 在大五人格、信任、合作、竞争等行为决策上与随机选取的来自全球 50 多个国家的近十万名被试的样本分布没有显著差异(Mei et al. ,2024)。甚至有研究者尝试使用历史文本进行微调训练,构建历史性大语言模型以模拟特定古代社会中人类的心理反应,使社会心理学取样得以纵向延伸(Varnum et al. ,2024)。证成以上两个前提,大语言模型是具有替代人类被试的潜力,但其依赖于训练数据的全面性、提示词对任务解释的精确性、训练集与价值对齐的程度。

理论建构之主题建模。长期以来,社会心理学家遵循“奥卡姆剃刀原则”试图通过简单理论解释复杂社会现象,其元理论被概括为:情境主义、社会建构、紧张系统(Ross & Nisbett,2011)。社会心理学是一门强调理论的学科,正如其学科奠基人勒温(Lewin,1951)所言“没有什么比好的理论更实用的了”,其需要以抽象的理论串联、概括细碎的现象,并指导具体的研究(van Lange et al. ,2011)。然而,理论之建构大抵有两条路径。其一,似思辨哲学,自上而下提出框架,再验证之;其二,似实证科学,自下而上归纳证据,再概括之。心理学兼有两种路线。对于人工智能技术而言,其从数据中学习的逻辑更近于后者。因此,或可尝试使用主题建模(topic modeling)等方法,使之从海量文本信息、数据信息中抽取相聚类的主题从而建构理论。以隐含狄利克雷分布(latent dirichlet allocation,LDA)这种最为常见的主题建模方法为例,其基本思想在于,将文本信息区分为文档、主体、词汇三层结构,假设文档由主题混合,而主题由词汇表征。其旨在识别文档中的

---

* 可参考的提示如下(翻译自 Banker et al. ,2024 的研究):“你是一位专业的社会心理学家。你的研究兴趣在于社会认知、态度与态度改变、暴力与攻击、亲社会行为、偏见与歧视、自我与社会认同、群际行为、人际关系。你的任务是生成反直觉但合理的假设。他们需要将社会心理学的不同子领域和前沿理论知识结合起来。请确保你的假设是准确的,并包含一个对照组。请以“假设…”开头,并生成 100 个假设。”

词汇,并将其聚类为主题(Blei et al.,2003)。简言之,主题建模可以告诉我们在浩如烟海、毫无逻辑的数据之中,哪些数据之间彼此相关,如何将大数据化繁为简。这个过程类似于人格心理学中词汇学取向的做法,通过对特质词的因子分析将一切描述人的形容词汇聚在“OCEAN”的大五人格之下(John,2021)。通过主题建模,社会心理学研究可克服传统质性研究方法的主观性局限,更加客观、严谨地抽取文本中的共性主题,在此基础上建立理解复杂现象的简单理论。例如通过自然语言分析和主题建模,研究者将负债感这一复杂情绪分解为接受善意帮助后因对帮助者造成负担而体验到的内疚和接受策略性帮助后体验到的义务感两种情绪成分(Gao et al.,2024)。我们建议后续研究者们可以结合主题建模对大数据信息的分析和抽取构建不同的心理学理论。

仿真模拟之智能计算。詹姆斯·格雷(2009)认为计算科学具有两种形态,其称之为:计算X学(comp-X)与X信息学(X-info),前者为计算机模拟某学科研究内容,而后者则是对学科内数据进行信息化分析。对应社会心理学,前者是仿真模型分析,其典型代表为基于主体的建模(agent-based modeling,ABM,Jackson et al.,2017),研究者们可以通过计算机构建社会仿真模型以观察更为复杂的社会现象(Gao et al.,2024),甚至构建了纯粹由人工智能主体组成的社区(例如Chirper)以观察类人主体的社会行为,从而反推至人类社会(Park et al.,2023)。而后者则是新兴的计算智能社会心理学(Bao,2024)。此类研究大多利用词嵌入技术(word embedding),将词汇转化为向量,浓缩词的语义并将之映射到向量空间之中(包寒吴霜 等,2023)。得到词向量后,可通过余弦相似度(即向量夹角)计算文本与文本之间的语义关联,从而反推向量所表征的构念之间的关系(例如,由此方法研究美国总统候选人的人格特质关联性,Bhatia et al.,2018)。进一步,若将词相似度视作内隐联想测验中的反应速度,使用两组目标词和两组属性词的词相似度差来测量目标词和属性词在自然语言中的相对联系强度,其便是词嵌入联想测验(word embedding association test,WEAT)的基本逻辑,其可以衡量潜藏于人类语言之中的种种社会认知模式和偏见(Caliskan et al.,2017)。倘若再将语义信息加入其中,通过掩码填空联系测验范式(fill-mask association test,FMAT),根据构念含义设计完形命题(query),直接调用预训练BERT模型来估计掩码位置(mask)不同备选词出现的语义概率,实现对社会事实、社会态度、社会刻板印象、社会心理变迁的智能测量(Bao,2024)。简言之,此类方法的共同逻辑在于将自然语言汇表征为数学向量,通过向量间的距离计算(或绝对或相对)捕捉词汇语义之中的关系,从而反向说明心理构念之间的关系。

数据涌现之尺度分析。随着大数据社会科学的兴起,人类留存于书籍、对话、社交媒体、公共数据库中的语言和非言语行为为社会心理学家提供了前所未有的海量数据来源。人工智能辅助社会心理学家通过爬取海量数据,以挖掘蕴含在冰冷数据背后的人心人性。借助“人工智能+大数据”的模式,社会心理学家得以超越原本样本代总体、规律代个性、情境代现实、操作代体验的心理学研究局限,从更宏观的视域研究更微观的问题(喻丰 等,2015)。具体而言,智能数据的社会心理学横可探究心理特质的时序变迁,纵可比较地理差异(吴胜涛 等,2023)。于前者,研究关注于世界范围内个人主义之抬头、集体主义之式微的价值流变(Yu et al.,2016;黄梓航等,2018);道德动机随季节、昼夜的节律性涨落(Hohm et al.,2024;喻丰 等,2020);道德规范的滑坡幻觉与转型实质(Mastroianni & Gilbert,2023;喻丰 等,2022);社会约束力的松-紧变化(Jackson et al.,2019)。而对于后者,研究则更多集中于绘制居民幸福感的分布地图(Zhao et al.,2019),比较全球价值系统差异(Jackson & Medvedev,2024;Jackson et al.,2023)、探究早期社会形态对主流宗教道德化倾向的塑造作用(Whitehouse et al.,2019)、预测防疫措施有效性的地域差异(Lu et al.,2021)、对比全国性和地方性政治表达策略(Dillion et al.,2024b)。此类研究之逻辑在于:基于具体的假设,将变量操作化定义之,根据特定的大数据来源(或为文字数据,如Google Books Ngram,或为特定公开数据库,如Eurobarometer,或为社交媒体,如微博或Twitter),选用适当的人工智能算法(如基于词嵌入技术的word2vec等),分析以验证假设的合理性。

2　旧瓶装新酒:人工智能扩展社会心理学研究内容

所谓社会心理学之社会,亦即“他人”。自1898年社会心理学家崔普利特(Norman Triplett)进行了第一个社会心理学实验——绕鱼线的社会助长现象开始,社会心理学关注的问题一直是作为个体的人

如何看待、联系、影响他人。基于如是理解，人工智能的突然崛起对社会心理学最大的冲击实际上并非方法范式上的补充或变革，而是重新定义了何为“他人(或称之为他者，others)”。社会心理学的发展史上，他者的概念也在不断蠕变。崔普利特的他者仅仅是一个不动的、缄默的观察者，而在阿希(Soloman Asch)所设线段面前的他者是一群模糊的从众者，到米尔格拉姆(Stanley Milgram)的他者是权威的身份，进而在谢里夫(Muzafer Sherif)的山洞和泰斐尔(Henri Tajfel)的分组中，所谓他者则是一个具有某些共性和认同的群体。然而，在人工智能时代，他者似乎不必要是“人”。人工智能深度参与到人类社会的种种活动之中，或同事、或管理、或服务、或陪伴、或做出决策、或礼敬神明。其已然成为一种“他者”，被人类认识、与人类互动、影响人类社会(Nielsen et al. ,2022)。因此，人工智能对社会心理学的挑战，实际上是对“他者”概念的扩展，从而扩展了社会心理学的研究对象和内容。其中包含三类研究，第一是以静态视角观察人工智能，探究其作为社会存在物的个体属性和能力；第二则将其视作被动的认知对象，探究人类对其的认知与态度；第三则视之为主动施加者，调动人能动的意识，探究其对人类的影响。

2.1 人工智能的倾向与能力

倾向和能力皆为一种稳定的结构，其构成了人工智能类人属性的基本假定。虽然社会心理学强调人是“情境的反应者(situational responder)”(Baumeister & Finkel,2010)，但其依旧坚持于寻找人性中稳固不变的成分。Strohminger 和 Nichols(2014)发现，相比于知觉、欲望、记忆、性格，当人的道德发生改变时，个人的同一性就此消失，由此说明了道德是人之为人的核心。而人格心理学家们则认为，所谓“人格”则是个体面对变动不居的世界统一不变的回应方式。人工智能是否具有人格？人工智能具有何种人格特质？人工智能又具有怎样的认知能力？这些问题构成了人工智能拟人性的重要前提。

从人格倾向的角度，无论从人格的理论性推断还是实证的经验性证据，都足以说明人工智能具有人格特质，且不同的模型之间存在人格差异。人格心理学理论自肇始，一直延续一种词汇学假设，即一切人格特质都蕴含于人类的自然语言之中。而大语言模型作为人类自然语言的产物，其理应从中学习和模拟到人类的人格特质。再加之尺度涌现，其可能表现出与人相异的个体倾向。在实证层面，通过计算 BERT 系列模型对于测量特定人格特质的题项的反应概率，发现这些模型普遍具有高尽责性、低神经质、高开放性的大五人格模式，但彼此之间存在个体差异(Pellert et al. ,2024)。而在意识形态方面，在没有角色扮演的情况下，ChatGPT 对政治罗盘问题的回答表现出了明显得对左翼自由意志主义的偏向(Motoki et al. ,2023)。

从认知能力的角度，大语言模型表现出一系列类似于人的心理能力。在决策能力上，面对经典的琳达问题，GPT3 也表现出了与大多数人类相似的合取谬误，认为琳达更有可能是一个女权主义的银行柜员。但是，其却更好地避免掉入卡片选择任务、认知反射测试、Blicket 实验的逻辑陷阱，表现出更好地信息搜索、审慎思考、反事实推理能力(Binz & Schulz,2023)。理性推理能力之外，以心理理论(theory of mind,ToM)、观点采择为代表的共情能力同样被认为是人类重要的认知能力。这些能力关涉于个体去理解他人具有不同的心理状态，站在他人的角度思考问题，以使得真正的社会理解和互动成为可能。目前研究从这一领域证实了大语言模型的潜在社交可能性。GPT4 和 LLaMA2 等大模型在识别间接请求、理解错误信念、识别讽刺等任务中达到甚至高于人类的平均水平，其可以理解人类复杂的交际(Hagendorff,2024;Strachan et al. ,2024)。

上述研究，实际上是一种静态观察的实验范式。研究者们把人工智能放在“实验台”上，或从不同人格特质的角度要求其进行“自我报告”(自然语言的直接报告或计算概率的间接测量)以衡量其表现出的个体差异，或放入特定认知任务之中检测其回答模式并与人类的相关研究结论形成对比以此说明人工智能表现出的认知能力。

2.2 人类如何看待人工智能

当我们叩问“人类如何看待人工智能”时，实际上问了两个问题，首先是我们如何朴素地理解人工智能，第二才是我们对人工智能的态度是积极还是消极。第一层次的问题是朴素认知的问题。语言学家莱考夫(George Lakoff)认为，我们生活在一个隐喻(metaphors)的世界，人类思维的本质是隐喻式的，人总以一种概念来模糊地建构另一种概念(Lakoff & Johnson,2003)。而对于新兴实体人工智能，当以隐喻的视角探索其在民众心理学层面的相关朴

素信念(folk beliefs)时,研究发现大致可分为未知(the unknown)、助手(the assistant)和机器(machines)三个维度(Xu et al.,2024)。认知心理学的基本假设认为,人类的认知表征遵循意义联结的模式,相近的概念将会被联系在一起。在人心的内隐联想中,人工智能与人性互斥(Spatola & Wudarczyk,2021);而在人脑的语义网络之中,人工智能与神圣实体具有更近的语义联结,共享相似的表征模式(Spatola & Urbanska,2019)。这些研究无不旨在说明,于朴素人心之中,人工智能是"似人但非人又超人"的智能实体。

然而,我们对其的态度是积极抑或消极,大部分研究结果均说明人类存在一种稳定的"算法厌恶(algorithmic aversion)"倾向,即在认知上拒绝、情感上厌恶、行为上回避人工智能(张语嫣 等,2022;谢才凤 等,2023)。这也是人工智能社会心理学中最为重要和庞大的一部分研究领域(Williams & Lim,2024)。这种反应倾向普遍存在于人类社会的种种活动领域之中。人们因为质疑人工智能的透明性而排斥其人力管理(赵一骏 等,2024),因为厌恶功利主义最大化的道德原则而抗拒其战略决策(Dietvorst & Bartels,2022),因为丧失敬畏体验而贬低其艺术作品的价值(Millet et al.,2023),因为担心对个体独特性的忽视而否定其医疗诊断(Longoni et al.,2019)。即便事实上其管理是公正的,决策是审慎的,艺术作品是审美的,医疗诊断是准确的。然而,这种倾向也并非绝对固定,其存在转向"算法欣赏(algorithmic appreciation)"的边界条件。当任务强调或需要客观中立时,人工智能公正、专业、僵化等的启发式被激活(Yang & Sundar,2024),人们便更倾向于其决策(Castelo et al.,2019;Castelo,2023)。同时,当社会判断凸显时,人们也会倾向于选择不会评判自己所作所为的人工智能(Raveendhran & Fast,2021;Jin et al.,in press)。

综上所述,此类研究的共同核心逻辑在于,尝试构建一个社会认知过程,将人类(即被试)作为认知主体,而人工智能(无论其在认知情境中是行动发出者抑或行动承受者)作为认知客体,使得人对人工智能做出认知判断。这在逻辑上是一个单向的认知过程,仅仅涉及到认知主体对客体的态度和反应。

2.3　人工智能如何影响人类

人工智能对人类影响的研究,大致可以拆解为两类范式。其一,研究人与人工智能的直接社会互动,以观察作为互动主体的人工智能对人类产生的影响。早期的研究倾向于去验证经典的社会心理学实验是否可以替换为人工智能从而得到重复。例如,阿希的经典从众实验复制于智能机器人中,无论是儿童还是成年人都会跟从于智能体群体的选择,哪怕其判断是错误的(Qin et al.,2022;Vollmer et al.,2018);当虚拟智能体被替换为米尔格拉姆服从权威实验中被绑在电击椅上的人接受电击时,人们依然会被唤起同情的反应(Rosenthal - von der Pütten et al.,2013)。这些研究旨在于说明人工智能作为一种类人智能,其对人类的影响具有与他人在场相似的效果。而随着研究的逐渐深入,越来越多的研究从验证人机相似走向寻找人机差异,人与智能体之间的互动行为也愈发复杂。与人工智能的互动似乎使得人类处于道德的滑坡之中。人工智能心智的缺位使人产生更低的预期内疚,从而做出更多的不道德行为(Kim et al.,2023)。而当人工智能作为主管提出要求时,人们由于感知不到被评判的压力降低了做好事的意愿(许丽颖 等,印刷中 b)。此外,仅仅是在工作场合中与人工智能接触或共事,便有可能使得个体愈发冷漠,内心动荡,从而做出更多倦怠和职场不文明行为(Granulo et al.,2024;Yam et al.,2023)。

其二,则是将人工智能视作社会现象,或言整体性概念,探究其崛起以及引发的威胁对人类社会互动和自我认识的影响。人以群分,人总是在划分内群体与外群体。人工智能作为独立但相似于人类的另一种智能实体,其势必会被人类知觉为外群体并引发威胁感知(Yogeeswaran et al.,2016)。根据群际威胁理论(intergroup threat theory,Stephan et al.,2016),外群体的出现会带来物质资源和安全上的现实威胁(reality threat)和身份认同、价值独特性方面的象征性威胁(symbolic threat)。当人们意识到人工智能带来的多重威胁时,人与人之间的和谐可能会遭到破坏。人工智能的威胁破坏了人赖以维持信念系统稳定的控制感,迫使人通过简单叙述进行控制补偿,从而选择物化身边的同事(许丽颖 等,2024)。同时,其诱发出一种对人类未来命运前途未卜的集体焦虑,导致社会纽带的结题,降低了人与人之间的亲社会联结(许丽颖 等,印刷中 a)。似乎,人工智能引发的威胁破坏了人际的和谐关系。对此,被威胁到地位的人类该如何应对(Jago et al.,2024)?社会认同理论(social identity theory)认为,

面对外群体诱发的威胁，人可以通过个人流动（individual mobility）、社会创造（social creativity）、社会竞争（social competition）三种策略去应对（Ellemers & Haslam，2012）。然而碳基与硅基之间不存在可渗透的群体边界，其取消了社会流动的可能，社会竞争又过于粗暴不够智慧。于是，社会创造的“阿 Q 精神”成为了人类解决认知失调的唯一可行途径（喻丰，2020）。人类自以为聪明的改变了对人性认识，将那些与人工智能相异的独特属性（如：文化、信仰、情感体验等）而非共享属性（如：交流、计算、记忆等）视作更重要的成分（Cha et al.，2020；Santoro & Monin，2023）。

总而言之，上述研究的逻辑均复杂于前一类单向认知，其涉及反思性（reflective）的思维过程，大致可归纳为，人工智能作为行动主体（即影响的发出者），而人类则是行动客体（即承受者），使得人对人工智能的影响做出反思性的认知和行为。其中反思性体现在，在人 - 人工智能的二元互动中，人不是简单地针对人工智能做出回应性的反应，而是将反应朝向人工智能之外的对象（如自身、他人、人性等等）。反思性思维体现了人类自我调节的能动本质，不简单是“刺激 - 反应”的有机体（喻丰，2022a）。

3 呼唤一种“新瓶装新酒”的研究

如前所述，无论是旧瓶新酒抑或新瓶旧酒，似乎都还有些寡淡。前者似乎是以方法为中心，专精于技术，但也容易沉溺于其中，使得社会心理学陷入“学而无思，以手段为美”的异化泥淖；而后者虽忠于社会心理学理论，发扬道统，但难以突破陈规，被逼入“思而无己，以他论为美”的尴尬角落（喻丰，2021）。当然，这并不意味我们需要全盘否定前文所述的所有研究及其范式。而只是需要呼唤一种“新瓶装新酒”的研究，以便更好地回应时代向社会心理学提出的挑战。这种研究，以人工智能提供的新研究方法为手段，再结合传统社会心理学的理论积淀和学科思维，研究数智时代提出的新社会问题。

技术福祉与群体威胁之张力。如此前每一次工业革命，人工智能的出场在提升社会生产力的同时，也造成了极大面积的技术替代和失业恐慌（Frey & Osborne，2017）。然而，人类社会似乎没有因为任何的技术进步而瓦解，却永远在担心社会进一步的技术化。与此前不同的是，人工智能所潜在的主体性使得其威胁了人类存在的独特性和价值感。面对这种福祉与威胁之间的张力，人工智能社会心理学应试图去回答，技术之发展会将人类社会带入怎样的未来之中，是实现共同富裕让人重回自由自觉的类本质活动之中，还是进一步加深阶层分化甚至奴役于技术丧失人之为人的尊严。

价值对齐与有效加速之纷争。针对人工智能的未来发展方向，大抵有两条截然不同的路径。其一为价值对齐，即构建道德的人工智能，确保技术与人类道德价值相一致，以规避通用人工智能成为危险的道德混沌物（Gabriel，2020）。其二为有效加速，即不断推动技术进步、实现快速迭代，使技术资本化从而颠覆旧有社会价值。虽然，社会心理学仅仅能够回答描述性的实然问题，不应染指规范性的应然问题，导致我们无法讨论何种路径应是未来之路（彭凯平 等，2011）。但仍有诸多问题有待人工智能社会心理学的入场。例如，如何构建道德的人工智能？参考社会心理学的研究，了解伦理规则不代表行事道德（Hou et al.，2024），直接否定了自上而下的规范伦理学路径。似乎需要通过美德伦理学的方式，制定道德的行为脚本，将抽象规范转化为具体规则，使之习惯化，锻造技术美德（喻丰，2022b；喻丰，许丽颖，2018）。

技术人化与人技术化之对立。人类总以自身为参照去认识世间万物，形成一条以人为中心上行及神、下行及物的社会认知链（Yu et al.，2022）。然而，随着人工智能的智慧化、通用化，人的认知模式似乎也有可能被重塑（喻丰 等，2024）。一方面人将人工智能拟人化，将其看做与人类相等同的智能实体，与之互动（喻丰，许丽颖，2020），但另一方面智能技术构成了对人的宰制，使人沦为高速运转的智能社会的“奴隶”（孙伟平，2020）。这是两个独立的心理过程，前者我们称之为技术之人化，而后者则为人之技术化。面对越来越像人的技术，人类如何保持人性之独特与尊严？面对越来越像技术的人，又该如何适应或改变这种失调？此当为人工智能社会心理学之使命。

以上问题，仅仅是在抛砖引玉，我们希望通过对人工智能社会心理学研究范式的梳理和总结，呼吁更多面向人工智能时代的社会心理学研究，回答人工智能提出的时代之问。

参考文献

包寒吴霜，王梓西，程曦，苏展，杨盈，张光耀，王博，蔡华俭.（2023）. 基于词嵌入技术的心理学研究：方法及应用. 心

理科学进展,*31*(6),887 – 904.

鄂维南.(2024).AI助力打造科学研究新范式.*中国科学院院刊*,*39*(1),10 – 16.

黄梓航,敬一鸣,喻丰,古若雷,周欣悦,张建新,蔡华俭.(2018).个人主义上升,集体主义式微?全球文化变迁与民众心理变化.*心理科学进展*,*26*(11),2068 – 2080.

彭凯平,喻丰,柏阳.(2011).实验伦理学:研究、贡献与挑战.*中国社会科学*,(6),15 – 25,221.

孙伟平.(2020).人工智能与人的"新异化".*中国社会科学*,(12),119 – 137.

吴胜涛,茅云云,吴舒涵,冯健仁,张庆鹏,谢天,陈浩,朱廷劭.(2023).基于大数据的文化心理分析.*心理科学进展*,*31*(3),317 – 329.

谢才凤,邬家骅,许丽颖,喻丰,张语嫣,谢莹莹.(2023).算法决策趋避的过程动机理论.*心理科学进展*,*31*(1),60 – 77.

许丽颖,王学辉,喻丰,彭凯平.(2024).感知机器人威胁对职场物化的影响.*心理学报*,*56*(2),210 – 225.

许丽颖,张语嫣,喻丰.(印刷中 a).感知机器人威胁降低亲社会倾向.*心理学报*.

许丽颖,赵一骏,喻丰.(印刷中 b).人工智能主管提出的道德行为建议更少被遵从.*心理学报*.

喻丰,许丽颖.(2018).如何做出道德人工智能体?心理学的视角.*全球传媒学刊*,(4),24 – 42.

喻丰,彭凯平,郑先隽.(2015).大数据背景下的心理学:中国心理学的学科体系重构及特征.*科学通报*,*60*(5 – 6),520 – 533.

喻丰,许丽颖,丁晓军,钱小军.(2022).道德滑坡了吗?美式英语语言汇总的社会道德时间变化.*中国社会心理学评论*,*23*,15 – 38.

喻丰,许丽颖,韩婷婷,刘知远,钱小军,彭凯平,胡晓檬.(2020).道德节律:基于新浪微博的道德动机每分变化.*科学通报*,*65*(19),2055 – 2061.

喻丰,许丽颖.(2020).人工智能之拟人化.*西北师范大学学报*(社会科学版),*57*(5),52 – 60.

喻丰,赵一骏,许丽颖,汪美霞.(2024).人工智能时代儿童发展的拟人化类别理论及其教育意蕴.*教育发展研究*,*44*(20),69 – 76.

喻丰.(2020).人工智能与人之为人.*人民论坛·学术前沿*,(1),30 – 36.

喻丰.(2021).中国心理学还美吗?*苏州大学学报*(教育科学版),*9*(3),17 – 20.

喻丰.(2022a).有限自由的道德能动.*心理研究*,*15*(2),105 – 109.

喻丰.(2022b).科技伦理治理的社会心理取向:以人工智能为例.*国家治理*,(7),26 – 30.

张语嫣,许丽颖,喻丰,丁晓军,邬家骅,赵靓.(2022).算法拒绝的三维动机理论.*心理科学进展*,*30*(5),1093 – 1105.

赵一骏,许丽颖,喻丰,金旺龙.(2024).感知不透明性增加职场中的算法厌恶.*心理学报*,*56*(4),497 – 514.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*, 337 – 351.

Banker, S., Chatterjee, P., Mishar, H., & Mishar, A. (2024). Machine – assisted social psychology hypothesis generation. *American Psychologist*, *79*(6), 789 – 797.

Bao, H – W – S. (2024). The fill – mask association test (FMAT): Measuring propositions in natural language. *Journal of Personality and Social Psychology*, *127*(3), 537 – 561.

Baumeister, R. F., & Finkel, E. J. (2010). *Advanced social psychology: The state of the science*. Oxford University Press.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y – Q., Xue, L., Shalev – Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., Mcilraith, S., Gao, Q., Acharya, A., Krueger, D., … Mindermann, S. (2024). Managing extreme AI risk amid rapid progress. *Science*, *384*(6698), 842 – 845.

Bhatia, S., Goodwin, G. P., & Walasek, L. (2018). Trait associations for Hillary Clinton and Donald Trump in new media: A computational analysis. *Social Psychological and Personality Science*, *9*(2), 123 – 130.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT – 3. *Proceedings of the National Academy of Sciences of the United States of American*, *120*(6), e2218523120.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(1), 993 – 1022.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human – like biases. *Science*, *356*(6334), 183 – 186.

Castelo, N. (2023). Perceived corruption reduces algorithm aversion. *Journal of Consumer Psychology*, *34*, 326 – 333.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task – dependent algorithm aversion. *Journal of Marketing Research*, *56*, 809 – 825.

Cha, Y – J., Baek, S., Ahn, G., Lee, H., Lee, B., Shin, J., & Jang, D. (2020). Compensating for the loss of human distinctiveness: The use of social creativity under Human – Machine comparisons. *Computers in Human Behavior*, *103*, 80 – 90.

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott – Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavioral Research Methods*, *47*(4), 1178 – 1198.

Dietorvst, B. J. , & Bartels, D. M. (2022). Consumers object to algorithms making morally relevant tradeoffs because of algorithms' consequentialist decision strategies. *Journal of Consumer Psychology*, *322*, 406 – 424.

Dillion, D. , Mondal, D. , Tandon, N. , & Gray, K. (2024a). Large language models as moral experts? GPT – 4o outperforms expert ethicist in providing moral guidance. *PsyArXiv Preprints*. https://doi. org/10. 31234/osf. io/w7236

Dillion, D. , Puryear, C. , Li, L. , Chiquito, A. , & Gray, K. (2024b). National politics ignites more talk of morality and power than local politics. *PNAS Nexus*, *3*(9), 345.

Dillion, D. , Tandon, N. , Gu, Y. , & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597 – 600.

Ellemers, N. , & Haslam, S. A. (2012). Social identity theory. In P. A. M. van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology vol.* 2, (pp. 379 – 399). SAGE.

Frey, C. B. , & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerization? *Technological Forecasting & Social Change*, *114*, 254 – 280.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*, 411 – 437.

Gao, C. , Lan, X. , Li, N. , Yuan, Y. , Ding, J. , Zhou, Z. , Xu, F. , & Li, Y. (2024). Large language models empowered agent – based modeling and simulation: A survey and perspectives. *Humanities and Social Science Communications*, *11*, 1259.

Gao, X. , Jolly, E. , Yu, H. , Liu, H. , Zhou, X. , & Chang, L. J. (2024). The psychological, computational, and neural foundations of indebtedness. *Nature Communications*, *15*, 68.

Granulo, A. , Caprioli, S. , Fuchs, C. , & Puntoni, S. (2024). Deployment of algorithms in Management tasks reduces prosocial motivation. *Computers in Human Behavior*, *152*, 108094.

Gray, J. (2009). eScience: A transformed scientific method. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The fourth paradigm: Data – intensive scientific discovery* (pp. xvii – xxxi). Microsoft Research.

Hagendorff, T. (2024). Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences of the United States of American*, *121*(24), e2317967121.

Hohm, I. , O' Shea, B. A. , & Schaller, M. (2024). Do moral values change with the seasons? *Proceedings of the National Academy of Sciences of the United States of America*, *121*(33), e2313428121.

Hou, T. , Ding, X. , & Yu, F. (2024). The moral behavior of ethics professors: A replication – extension in Chinese mainland. *Philosophical Psychology*, *37*(2), 396 – 427.

Jackson, J. C. , & Medvedev, D. (2024). Worldwide divergence of values. *Nature Communications*, *15*, 2650.

Jackson, J. C. , Dillion, D. , Bastian, B. , Watts, J. , Buckner, W. , DiMaggio, N. , & Gray, K. (2023). Supernatural explanations across 114 societies are more common for natural than social phenomena. *Nature Human Behaviour*, *7*, 707 – 717.

Jackson, J. C. , Gelfand, M. , De, S. , & Fox, A. (2019). The loosening of American culture over 200 years is associated with a creativity – order trade – off. *Nature Human Behaviour*, *3*, 244 – 250.

Jackson, J. C. , Rand, D. , Lewis, K. , Norton, M. I. , & Gray, K. (2017). Agent – based modeling: A guide for social psychologists. *Social Psychological and Personality Science*, *8*(4), 387 – 395.

Jago, S. , Raveendhran, R. , Fast, N. , & Gratch, J. (2024). Algorithmic management diminishes status an unintended consequence of using machines to perform social roles. *Journal of Experimental Social Psychology*, *110*, 104553.

Jin, J. , Walker, J. , & Reczek, R. W. (in press). Avoiding embarrassment online: Response to and inferences about chatbots when purchases activate self – presentation concerns. *Journal of Consumer Psychology*. Advance Publication Online.

John, O. P. (2021). History, measurement, and conceptual elaboration of the big – five trait taxonomy: The paradigm matures. In O. P. John & R. W. Robins (Eds.), *Handbook of personality: Theory and research* (pp. 35 – 83). New York: Guilford Press.

Jones, C. R. , & Bergen, B. K. (2024). Does GPT – 4 pass the Turing test? *arXiv Preprints*. https://doi. org/10. 48550/arXiv. 2310. 20216

Kim, T. , Lee, H. , Kim, M. Y. , Kim, S. , & Duhachek, A. (2023). AI increases unethical consumer behavior due to reduced anticipatory guilt. *Journal of the Academy of Marketing Science*, *51*, 785 – 801.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd Eds). Chicago: University of Chicago Press.

Lakoff, G. , & Johnson, M. (2003). *Metaphors we live by*. University of Chicago Press.

Lewin, K. (1951). *Field theory in social science: Selected theoretical papers* (D. Cartwright, Eds.). Harper & Brothers.

Longoni, C. , Bonezzi, A. , & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629 – 650.

Lu, J. G. , Jin, P. , & English, A. S. (2021). Collectivism predicts mask use during COVID – 19. *Proceedings of the National Academy of Sciences of the United States of America*, *118* (23), e2021793118.

Mastroianni, A. M. , & Gilbert, D. T. (2023). The illusion of moral decline. *Nature*, *618*, 782 – 789.

McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology*, *26*(3), 446 – 456.

McGuire, W. J. (1997). Creative hypothesis generating in psychology: Some useful heuristics. *Annual Review of Psychology*, *48*, 1 – 30.

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(9), e2313925121.

Millet, K., Buehler, F., Du, G., & Kokkoris, M. (2023). Defending humankind: Anthropocentric bias in the appreciation of AI art. *Computers in Human Behavior*, *143*, 107707.

Motoki, F., Neto, V. P., & Rodrigues, V. (2023). More human than human: Measuring ChatGPT political bias. *Public Choice*, *198*, 3 – 23.

Nielsen, Y. A., Pfattheicher, S., & Keijsers, M. (2022). Prosocial behavior toward machines. *Current Opinion in Psychology*, *43*, 260 – 265.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv Preprints*. https://doi.org/10.48550/arXiv.2304.03442

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, *19*(5), 808 – 826.

Qin, X., Chen, C., Yam, K. C., Cao, L., Li, W., Guan, J., Zhao, P., Dong, X., & Lin, Y. (2022). Adults still can't resist: A social robot can induce normative conformity. *Computers in Human Behavior*, *127*, 107401.

Raveendhran, R., & Fast, N. J. (2021). Human judge, algorithms nudge: The psychology of behavior tracking acceptance. *Organizational Behavior and Human Decision Processes*, *164*, 11 – 26.

Rosenthal – von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics*, *5*, 17 – 34.

Ross, L., & Nisbett, R. E. (2011). *The person and the situation: Perspectives of social psychology*. Pinter & Martin Publishers.

Santoro, E., & Monin, B. (2023). The AI effect: People rate distinctively human attributes as more essential to being human after learning about artificial intelligence advances. *Journal of Experimental Social Psychology*, *107*, 104464.

Simon, H. A. (1979). Information processing models of cognition. *Annual Review of Psychology*, *30*(1), 363 – 396.

Spatola, N., & Urbanska, K. (2019). God – like robots: The semantic overlap between representation of divine and artificial entities. *AI & Society*, *35*, 329 – 341.

Spatola, N., & Wudarczyk, O. (2021). Implicit Attitude towards robots predict explicit attitudes, semantic distance between robots and humans, anthropomorphism, and prosocial behavior from attitudes to human – robot interaction. *International Journal of Social Robotics*, *13*, 1149 – 1159.

Stephan, W. G., Ybarra, O., & Rios, K. (2016). Intergroup threat theory. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (2nd Ed., pp. 255 – 278). Mahwah, NJ: Lawrence Erlbaum Associates.

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*, 1285 – 1295.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*, 159 – 171.

Van Lange, P. A. M., Higgins, E. T., & Kruglanski, A. W. (2011). *Handbook of theories of social psychology*. SAGE.

Varnum, M. E. W., Baumard, N., Atari, M., & Gray, K. (2024). Large language models based on historical text could offer informative tools for behavioral science. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(42), e2407639121.

Vollmer, A – L., Read, R., Trippas, D., & Belpaeme, T. (2018). Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science Robotics*, *3*(21), eaat7111.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergem, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T – Y., Manrai, A., … Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, *620*, 47 – 60.

Wang, S., Scells, H., Koopman, B., & Zuccon, G. (2023). Can Chat GPT write a good Boolean query for systematic review literature search? *arXiv Preprints*. https://doi.org/10.48550/arXiv.2302.03495

Whitehouse, H., François, P., Savage, P. E., Currie, T. E., Feeney, K. C., Purcell, R., Ross, R. M., Larson, J., Baines, J., ter Harr, B., Covey, A., & Turchin, P. (2019). Complex societies precede moralizing gods throughout world history. *Nature*, *568*, 226 – 229.

Williams, G. Y., & Lim, S. (2024). Psychology of AI: How AI impacts the way people feel, think, and behave. *Current Opinion in Psychology*, *58*, 101835.

Xu, L., Zhang, Y., Yu, F., Ding, X., & Wu, J. (2024). Folk beliefs of artificial intelligence and robots. *International Journal of Social Robotics*, *16*, 429 – 446.

Xu, R., Sun, Y., Ren, M., Guo, S., Pan, R., Lin, H., Sun, L., & Han, X. (2024). AI for social science and social science of AI: A survey. *Information Processing & Management*, *61*(3), 103665.

Yam, K. C., Tang, P., Jackson, J., Su, R., & Gray, K. (2023). The rise of robots increase job insecurity and maladaptive workplace behaviors: Multimethod evidence. *Journal of Applied Psychology*, *108*(5), 850 – 870.

Yang, H., & Sundar, S. S. (2024). Machine heuristic: Concept explication and development of measurement scale. *Journal of Computer – Mediated Communication*, *29*(6), zmae019.

Yogeeswaran, K., Złotowski, J., Livingstone, M., Bartneck, C., Sumioka, H., & Ishiguro, H. (2016). The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. *Journal of Human – Robot Interaction*, *5*(2), 29 – 47.

Yu, F., Peng, T., Peng, K., Tang, S., Chen, C. S., Qian, X., Sun, P., Han, T., & Chai, F. (2016). Cultural value shifting in pronoun use. *Journal of Cross – Cultural Psychology*, *47*(2), 310 – 316.

Yu, F., Xu, L., & Peng, K. (2022). A theory of a human – centered social cognitive chain. *Social Sciences in China*, *43*(4), 152 – 167.

Zhao, Y., Yu, F., Jing, B., Hu, X., Luo, A., & Peng, K. (2019). An analysis of well – being determinants at the city level in China using big data. *Social Indicators Research*, *143*, 973 – 994.

Social Psychology in the Era of Artificial Intelligence: Research Paradigms and Questions

Yu Feng Zhao Yijun Xu Liying

(Department of Psychology, Wuhan University, Wuhan 430072)

Abstract: The era of artificial intelligence is approaching. Artificial intelligence has the powerful potential to reshape scientific paradigms and spark a technological revolution. In addition, paradigms determine the worldview, methodology and normativity of science. In the filed of social psychology, artificial intelligence is profoundly affecting its research paradigms and contents in two aspects. On the one hand, artificial intelligence (e. g. large language models, algorithms) as powerful computational tools complement the methodological limitations of traditional social psychology research. This type of research we named as "AI for social psychology". In the preparation phase of experiments, AI can be used as tools for literature review and hypothesis creation. Thereby, in order to test an available hypothesis, AI is capable of simulating human participants' responds, mining natural language text for relevant topics, calculating vectorized semantic information, and analyzing big data samples. On the other hand, some studies have focused on exploring the relationship interaction and between AI and human community. In other words, AI as social agents supplements the objects and contents of social psychology. We label this type of research as "social psychology of AI". Specifically, these studies fall into three categories of measuring AI propensities and capabilities, examining human attitudes and adoption towards AI, and probing the impact of AI on human beings and the society. Finally, we call for a new kind of research which concentrate on utilizing AI – enabled methodologies to solve social issues caused by the salience of AI.

Key words: artificial intelligence; social psychology; research paradigms; research questions