Journal of Intelligent Society

ISSN 2097-2091
9 772097 209239 07>
定价：45.00 元

JIS
第2卷
2023
第 4 期

智能社会研究

中华人民共和国工业和信息化部主管

智能社会研究
Journal of Intelligent Society
ISSN 2097-2091
CN 23-1615/C

第 2 卷
2023
第 4 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

智能社会研究
（双月刊）
2023 年
第 4 期
2023 年 7 月 10 日出版
2022 年 11 月 10 日创刊
总第 5 期

目　次

译　文

书　评

访　谈

CONTENTS

SPECIAL TOPICS ON INTELLIGENT SOCIETY AND POPULATION RESEARCH

RESEARCH REPORT

TRANSLATED TEXT

BOOK REVIEW

INTERVIEW

交际中人工智能对语言和社会关系的影响*

杰斯·海恩斯坦　勒内·克孜尔切克　多米尼克·迪弗朗佐

日拉·阿加贾里　汉娜·米奇科夫斯基　凯伦·利维

莫尔·纳曼　杰弗里·汉考克　马尔特·荣格　著**

李寒秋　陈典涵　译***

吕　鹏　校****

摘要:人工智能(AI)已广泛应用于日常交流中。尽管人们担心使用AI进行交流会给社会带来负面影响,但其用于交流的社会后果在很大程度上仍未得到探索。本文分析了AI最广泛的应用之一“算法回应建议”(智能回复)的广泛使用所带来的社会后果。两个随机实验结果表明,算法回应建议这类系统改变了人们与他人的互动和人际感知(以亲社会或反社会的方式)。也就是说,智能回复的使用改变了人际沟通的语言风格和社会关系。具体而言,智能回复系统加快了交流的速度,提高了积极情感语言的使用频率,加深了双方对彼此间亲密度和合作的感知。与人工智能负面影响的普遍假设一致,如果人们被怀疑使用了智能回复,则对他们的评价会更加负面。因此,虽然人工智能可以提高沟通效率,改善人际关系感知,但是如果公开使用人工智能,它普遍存在的反社

---

* 本文原题为“Artificial Intelligence in Communication Impacts Language and Social Relationships”,英文原文见 https://doi.org/10.1038/s41598-023-30938-9,经授权翻译发表。译文的顺利完成特别感谢在斯坦福大学攻读组织行为学和社会学博士学位的李子信(P. Lambert)师兄,他对译文的一些专业词汇给予了很多指导。

** 杰斯·海恩斯坦(J. Hohenstein),康奈尔大学信息科学系;勒内·克孜尔切克(R. Kizilcec),康奈尔大学信息科学系;多米尼克·迪弗朗佐(D. DiFranzo),利哈伊大学计算机科学与工程系;日拉·阿加贾里(Z. Aghajari),利哈伊大学计算机科学与工程系;汉娜·米奇科夫斯基(H. Mieczkowski),斯坦福大学传播系;凯伦·利维(K. Levy),康奈尔大学信息科学系;莫尔·纳曼(M. Naaman),康奈尔科技大学信息学系;杰弗里·汉考克(J. Hancock),斯坦福大学传播系;马尔特·荣格(M. Jung),康奈尔大学信息科学系。

*** 李寒秋,陕西科技大学经济与管理学院、北京大学武汉人工智能研究院;陈典涵,中南大学公共管理学院、北京大学武汉人工智能研究院。

**** 吕鹏,北京大学武汉人工智能研究院。

会内涵将会削弱这一潜在的优势。

关键词： 人工智能　社会交流　自动回复　社会关系

一、引言

沟通是人们形成对他人看法（Mairesse，Walker & Mehl et al.，2007）、建立和维护社会关系（Pennebaker，Mehl & Niederhoffer，2003）、成功实现合作（Zhang，Chang & Danescu-Niculescu-Mizil et al.，2018）的基本过程。从大语言模型（LLM）中提取的生成式人工智能，可能从根本上改变人们的交流方式。像 ChatGPT 这样的人工智能应用逐渐广泛用于生成各种类型的语言，比如短信、网络帖子、计算机程序以及演讲（Stone，Brooks & Brynjolfsson et al.，2016；Rahwan，Cebrian & Obradovich et al.，2019；Jakesch，Hancock & Naaman，2023）。

迄今为止，最普遍的人工智能应用之一是基于文本通信的个性化回复建议，通常被称为“智能回复”（Kannan，Kurach & Ravi et al.，2016）。截至 2017 年，智能回复占 Gmail 发送邮件总量的 12%（Bullock，2017），这就意味着每天约有 6 亿封电子邮件是由人工智能代写的（Kraus，2018）。智能回复系统利用一般的文本语料库来预测用户可能输入的内容，生成一个或多个回复建议，以供用户在回复信息时进行选择（图 1），从而提高了文本生成效率（Kannan，Kurach & Ravi et al.，2016）。关于生成算法回应的各种技术研究，促进了这一类型的人工智能在人际沟通中的迅速使用（Kannan，Kurach & Ravi et al.，2016；Henderson，Al-Rfou & Strope et al.，2017；Ritter，Cherry & Dolan，2011）。[①]

尽管人工智能应用在新产品和环境中得到快速推广，但是人们依然担心

① 亨德森（M. Henderson）等人提出了一种高效的机器学习自然语言以提供回复建议的方法：使用 N-gram 嵌入特征的前馈神经网络将信息编码到向量中，这些向量经过优化，使得“信息—回复建议”具有较高的点积值（两个向量之间的数量积），即让回复建议与信息内容相关度更高，从而优化了提供的回复建议。

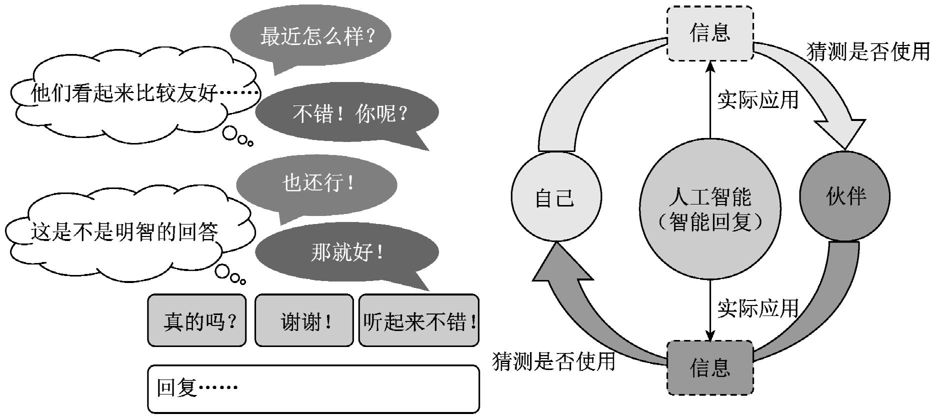

图 1　智能回复系统交互模式和影响过程

注：左图展示了使用人工智能回复的信息交互模式，方形条框（实际为橙色显示）是智能回复给出的回复示例以及呈现的形式。右图展示了人工智能对人际沟通影响的过程，任何一方或双方均可以获得人工智能的支持（选择智能回复建议进行交流），且当拥有人工智能支持时，参与者可以选择使用，也可以选择不使用。不过，用户对人工智能使用的感知与实际使用无关，并假设实际使用和感知使用都会影响整体信息交互和人们对彼此看法的形成。

它们给社会带来的负面影响（Shakirov，2016），科学界在很大程度上忽视了将人工智能生成的信息整合到人类交流中的潜在社会影响。现代人工智能研究所（AI Now Institute）在报告里将这种情况类比为“在不注意结果的情况下进行实验”（Crawford & Whittaker，2016），并一再指出目前对人工智能产生的社会影响研究不足，同时也呼吁增加其跨学科检测（Whittaker，Crawford & Dobbe et al.，2018）。

对此，计算机和社会科学交叉领域的研究越来越多地致力于探究人工智能系统如何影响人类行为（Rahwan，Cebrian & Obradovich et al.，2019；Lee，2018；Hancock，Naaman & Levy，2020）。[①] 初步研究发现，算法回应会影响人们的输入方式（Arnold，Chauncey & Gajos，2020）。用户认为，仅仅是智能回复的显示就影响了他们的交流方式，部分原因是智能回复的语言风格倾

① 李（M. Lee）研究发现，算法的公平性和可信度归因于用户的感知效率和客观性，而算法决策的负面情绪则归因于非人性化体验。汉考克（J. Hancock）等人探究了人工智能中介通信（AI-MC）的技术设计，以及将其引入人际交流后对人们心理、语言、关系、政策和伦理的影响。

向——与正常对话相比,智能回复倾向于表达过度的积极情绪(Hohenstein & Jung, 2020)。不过,让算法代表人们说话时,人们并不知道自己与他人的社会关系和对他人的看法会受到怎样的影响。

为了研究使用人工智能生成的信息对人际关系的影响,我们开发了一个定制的信息应用程序,并进行了两个随机实验,以研究在文本交流中实时显示和使用人工智能生成的智能回复将如何影响人们互动和人际感知的方式。我们发现,使用广泛部署的智能回复算法会以积极或者消极的方式影响人际沟通的各个方面,包括沟通速度、情绪语调以及人际评价。

二、结果

(一) 人工智能影响社会关系:消极评价但改善了人际感知

受到以计算机为中介的沟通如何影响亲密关系和关系维持理论的启发(Tong & Walther,2011),我们假设参与者看到人工智能生成的回复建议时会影响自己与对话伙伴的人际感知。为了测试人工智能中介对人际特质推断和合作式感知的影响,我们开发了一个新的信息传递应用程序(详见方法部分),它不仅可以控制智能回复内容的显示,还可以收集在沟通中有关智能回复使用的数据。

为了明确算法回应在对话中的影响和感知,我们随机分配了 219 对参与者("自己"和"伙伴"),让他们分别独立使用或不使用智能回复(谷歌回复 API[①] 生成的回复建议),共有四种信息传递场景:(1)两个参与者都可以使用智能回复;(2)只有自己可以使用智能回复;(3)只有对话伙伴可以使用智能回复;(4)两个参与者都不能使用智能回复。智能回复的可用性激励参与者

① 智能回复模型根据对话的完整上下文生成回复建议,模型会将信息与敏感主题列表进行比较,如果确定语言为英语,且并未检测到敏感主题,模型会提供三个建议的回复。历史记录不会占用内存,也不会被上传到服务器,以保障个人隐私安全。目前能够识别 103 种语言的文本,只需要给定几个单词,就能回传精准的判断结果,且只需要 1—2 毫秒的回应时间。

在交流中使用智能回复。为了更进一步评估实际使用智能回复而不是仅拥有其可用性对交流速度、情绪和人际感知的影响,我们使用了工具变量(IV)分析法。IV分析法是一种成熟的计量经济学方法。在实验过程中,当结果取决于个体选择采纳与否时,该方法用于评估个体选择与结果之间的因果关系效应(Angrist, Imbens & Rubin, 1996)。实验工具是对话伙伴智能回复的可用性,其可用性是随机分配的,参与者无法察觉收到的信息是否属于智能回复。这为满足IV分析法排除限制假设创造了理想的条件,因为工具(智能回复的可用性)对结果(如从属关系的评级)造成的所有影响完全来自其暴露(对话伙伴的信息中智能回复的比例)的情况。

参与者就政策问题进行交流,同时应用程序跟踪智能回复的呈现和使用情况。谈话结束后,向参与者介绍智能回复的定义[①],要求他们以此从1(从来没有)到5(总是)对对话伙伴使用智能回复的频率进行打分。同时,对参与者进行主导地位和从属关系感知的测试[使用修订的人际形容词量表(IAS-R)(Wiggins, Trapnell & Phillips,1988]。该测试向参与者展示了描述人们如何与他人互动的单词列表(如害羞、善良、外向等),要求参与者对描述对话伙伴的每个单词的准确程度按照从“极度不准确”(1)到“极度准确”(7)的等级进行打分。最后,参与者还需要完成一项合作式沟通测试(Lee, 1997),要求他们按照从“强烈不同意”(1)到“强烈同意”(7)的等级对“我们经常相互批评”等陈述进行打分。三项测试任务在参与者之间是随机的,以避免潜在的顺序效应。[②]

实验发现,算法回应的可用性一方面显著增加了人们在交流对话中使用智能回复的频次[第一阶段:$t(211)=13.800$, $p<0.0001$],智能回复平均占总信息发送量的14.3%;另一方面也有效提高了交流速度,每分钟信息发送条数提高10.2%[意向治疗估计:$t(198)=2.173$, $p=0.0309$]。智能回复提高了

① 在要求参与者评估感知到的对话伙伴智能回复使用情况之前,向参与者提供了“智能回复”的定义:智能回复是由人工智能(AI)生成的信息,您只需要点击即可发送。并提供如图1左图的图片示例。

② 顺序效应是指人们通常对最先出现的刺激或后面出现的刺激评价过大的倾向。

参与者的信息传递速度，因为对话伙伴使用智能回复并没有显著提高参与者自身的交流速度［IV 估计：$b=0.402$，$t(205)=0.825$，$p=0.410$］。尽管智能回复提高了交流速度，但它对人际关系感知的影响更为复杂。

参与者在一定程度上能识别出对话伙伴在使用智能回复：他们对对话伙伴使用智能回复频次的识别与实际使用情况有一定相关关系，但不显著［Pearson's $r=0.22$，$t(97)=3.62$，$p=0.0005$］。我们发现，参与者感知到对话伙伴使用智能回复与对他们的态度之间存在显著相关性，这与关于人工智能在社交互动中存在的负面影响的普遍看法一致（Jakesch, French & Ma et al., 2019）。如图 2 所示，即使在实际谈话过程中控制对话伙伴对智能回复的应用，参与者也认为对话伙伴使用智能回复越多，对他们的感知评价越为不易合作［$t(92)=-9.89$，$p<0.0001$］、越占主导地位［$t(92)=2.27$，$p=0.0256$］，且对他们的从属关系感知越少［$t(92)=-6.90$，$p<0.0001$］。这表明，在对话中被认为使用了智能回复的人付出了极大的人际关系代价，即使他们实际上并没有使用智能回复。但是，这一实验发现并没有证明人们的态度如何随着智能回复的使用而改变因果关系。

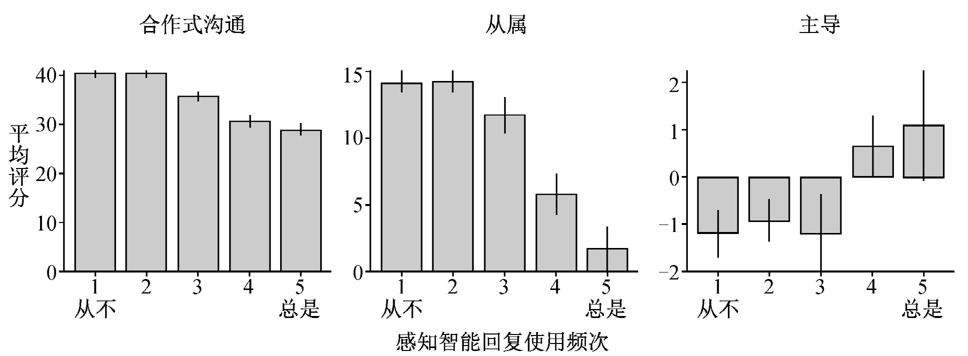

图 2　感知智能回复频次对对话伙伴评价的影响

注：图中展示了在感知对话伙伴使用智能回复的不同频次下，对对话伙伴的合作式沟通、从属关系和主导地位的平均评分（N=361），误差线表示该组聚类稳健标准误高于或低于平均值。

我们发现，对话伙伴使用智能回复频次的增加，实际上提高了参与者对他们的合作［IV 估计：$b=15.66$，$t(189)=2.39$，$p=0.018$］和从属关系［IV 估

计:$b=21.79$,$t(189)=2.75$,$p=0.007$]感知的评价,但没有提高对主导地位感知的评价[IV估计:$b=-0.53$,$t(189)=-0.13$,$p=0.900$]。虽然人们对智能回复的使用存在负面看法,但对话伙伴对智能回复的实际使用却产生了更积极的态度。不过,参与者自己使用算法回应不能显著影响合作和从属关系的感知[意向性治疗估计:合作 $b=0.397$,$t(188)=0.436$,$p=0.663$。从属关系:$b=-0.397$,$t(188)=-0.362$,$p=0.718$],而且只有当参与者自己使用算法回应时,他们对主导地位感知的评价才会降低。

本文还发现,对话伙伴增加使用智能回复频次,会导致参与者自己发送更积极情绪的信息[IV估计:$b=0.178$,$t(205)=2.02$,$p=0.045$]——即使智能回复信息被排除在情绪评分之外[$b=0.208$,$t(205)=2.17$,$p=0.031$]。如果参与者可以使用智能回复算法,那么他们发送的信息的情绪也会更积极[意向治疗估计:$b=0.026$,$t(198)=2.05$,$p=0.0422$],除非在计算信息情绪得分时将智能回复信息排除在外[$b=0.019$,$t(198)=1.35$,$p=0.1801$]。这表明,仅仅显示算法回应并不会影响人为发送的信息的情绪,而是通过使用倾向于具有积极情绪的智能回复信息来影响信息的情绪。综上所述,这些发现表明,人工智能中介对人际感知的影响与人工智能系统引入的语言变化有关。

(二)人工智能影响语言:智能回复信息情绪影响人们交流中的情感用语

为了更好地理解智能回复的情绪如何影响交流用语,我们进行了第二个实验,随机分配291对参与者使用实验提供的应用程序在以下四种条件之一中讨论政策问题:(1)谷歌智能回复(使用谷歌回复API生成);(2)积极的智能回复(由众包工人评定为具有积极情绪);(3)消极的智能回复(由众包工人评定为具有消极情绪);(4)没有向两个参与者提供可用于对话的智能回复。我们利用VADER测量对话内容的情绪——VADER是一个基于词汇和规则的情绪分析工具,适合分析简短的社交信息(Hutto & Gilbert, 2014);我们首先使用LIWC情感词典(Pennebaker, Francis & Booth, 2001)来判定对话中智能回复是否使用了更具情绪化的语言(参见方法部分);然后将VADER得分

汇总成一个情绪程度得分排名,从“最积极”(1)到“最消极”(-1),“中性”(0)在中间。平均而言,对话持续了6.33分钟(SD=2.67),信息条数包括智能回复在内共20条。

我们发现,使用消极的智能回复相比于使用积极的智能回复[t(127)=2.75,p=0.007,d=0.352]和谷歌智能回复[t(127)=2.40,p=0.018,d=0.323](图3),会导致参与者在交流过程中产生更多的消极情绪内容,这也说明了商业信息应用程序中智能回复明显的积极情绪偏好。谷歌智能回复对谈话情绪的影响与积极的智能回复相似[t(150)=0.51,p=0.61],但与没有使用智能回复相比,谷歌智能回复并没有引起更显著的积极情绪[t(137)=0.55,p=0.580]。此外,我们发现这些语言的变化是由人们真正使用智能回复引起的,如果只是展示智能回复建议并不会引发这些变化。通过在会话语料库中省略智能回复信息来重复分析,我们发现在使用智能回复的情况下谈话情感差异最小[F(3277)=0.360,p=0.782]。综上所述,这些发现证明了人工智能生成的情绪如何影响人们在对话中使用的情感语言。

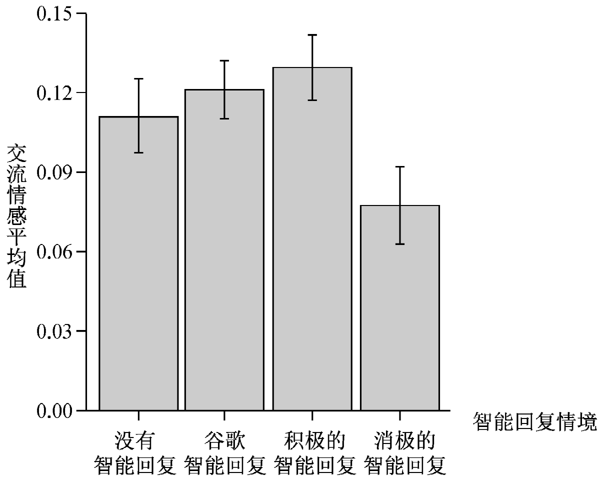

图3　智能回复情境

注:图中显示了参与者随机被分配到没有智能回复、谷歌智能回复、消极的智能回复或积极的智能回复的情况下,各自的对话情绪平均值。误差线表示该组聚类稳健标准误高于或低于平均值。

三、讨论

实验表明,生成式人工智能,包括商业部署的人工智能系统,可以对人们的沟通方式产生积极或消极的影响。我们发现,当人们有机会时,会在交流中尽可能地选择使用人工智能,这提高了交流的速度,并产生了更加积极的情感语言。但是当参与者认为对话伙伴频繁使用智能回复时,对他们的感知评价则为“越不易合作,越占主导地位”,且对他们的从属关系感知越少。这一发现可能与关于人工智能在社交互动中的负面影响的常见假设有关。例如,人们倾向于信任其他人而不是计算机(Promberger & Baron, 2006)。目前,大多数具有人工智能中介功能的通信系统对用户缺乏透明度(信息发送者知道信息由人工智能修改和生成,但接收者不知道),而用户倾向于减少交流中的这种不确定性(Berger & Calabrese, 1975),这就可能导致人们对人工智能产生负面看法。事实上,这些负面看法证实了最近的一些研究发现:人们认为人工智能回复往往不能捕捉到他们想说的话,可能会改变自己与其他人交流的方式(Hohenstein & Jung, 2018),认为被怀疑或被标记为人工智能生成的文本不太可信(Jakesch, French & Ma et al., 2019)。

尽管人们对人工智能在交流中所产生的影响存在负面评价,但算法回应的实际使用使得用户对对话伙伴的评价更积极。尽管人们对感知到的智能回复使用持有负面态度,但实际的智能回复使用会使交流者更容易产生合作和从属关系的感知。换言之,对交流中使用人工智能的负面看法与现实存在不符。

但是,这些实验结果与人们在交流过程中使用人工智能的具体情况相关,如果只是向用户提供人工智能回复建议而用户没有使用,就无法得出这些结果。也就是说,尽管还不知道接触但不使用智能回复会产生什么影响,但实验发现,智能回复的可用性鼓励人们使用它,人们由此调整了自己的交流用语和对话伙伴对自己的感知方式。

我们的研究对传播学和心理学有一定的启示。我们提供的证据表明,使

用人工智能可以塑造交流中的语言和人际感知。这一影响至关重要,因为语言与听众对说话者的性格(Mairesse, Walker & Mehl et al., 2007)、情绪(Pennebaker, Mehl & Niederhoffer, 2003)、态度(Hutto & Gilbert, 2014; Breck, Choi & Cardie, 2007)和主导地位(Rienks & Heylen, 2005)的描述密不可分。事实上,我们发现使用智能回复改变了人们对话中的情感表达。人工智能对人类情感交流的影响令人十分担忧,因为人工智能每天帮助人们撰写 6 亿封电子邮件(Kraus, 2018)。虽然其他形式的人工智能(如智能撰写)在日常交流中的日益普及,但研究者却很少关注人们如何允许人工智能辅助交流以及人工智能对人们交流潜在的长期影响。我们的研究表明,在人际交流中使用人工智能对人际关系可能产生积极的影响;但未来的研究需要调查这些变化的纵向影响,比如随着时间的推移,语言表达是否会变得同质化,导致个人交际风格丧失。

本研究对计算科学中有关人工智能开发的研究也产生了影响,因为我们强调了人工智能回复系统存在的机会和风险。我们展示了人工智能系统如何通过极其微妙的干预形式而积极地影响人际交流,如仅提供回复建议就可以改变交流中使用的语言,使得这些语言与算法建议的语言特性一致。此外,以前的研究表明,当交流出现错误时,人们会将一些原本归咎于人的责任归咎于人工智能(Hohenstein & Jung, 2020)。综上所述,这些发现表明,开发人员可能有机会通过刻意控制智能回复的语言来影响交流动态和结果(Sukumaran, Vezich & McHugh et al., 2011)。然而,人工智能在社会交互中被越来越广泛地使用,也给社会带来了潜在的风险。研究人员和从业人员在设计支持通信的算法时,需要考虑这些风险所产生的社会后果。

总的来说,我们展示了一个旨在帮助人们的人工智能系统是如何产生意想不到的社会后果的。人工智能有可能帮助人们在日常生活中更快捷地交流并改善人际感知,但这些好处和改变随着时间的推移对沟通模式产生的影响还需要进一步探讨。

四、方法

所有方法均按照相关的道德准则和规定进行,所有实验方案和材料均由康奈尔大学人类参与者研究机构审查委员会(IRB)批准(协议编号:1610006732)。实验得到了所有参与者的知情同意,且我们预先在AsPredict上注册了“实验一”(Hohenstein, Kizilcec & DiFranzo et al., 2020)。

(一) 实验一

我们随机选取219对参与者(“自己”和“伙伴”)参与实验,规定在进行有关政策问题的交流时可以或不可以使用智能回复,这导致出现四种模式:(1)两个参与者都可以使用智能回复;(2)只有自己可以使用智能回复;(3)只有对话伙伴可以使用智能回复;(4)两个参与者都不能使用智能回复。受到计算机中介沟通(利用计算机和互联网技术来辅助、促进或实现人与人之间的沟通和信息交流的方式)对亲密关系和关系维持的影响理论的启发(Tong & Walther, 2011),我们预计人工智能生成的回复建议会影响参与者对对话伙伴的感知以及交流用语。

1. 被试

在这项实验中,我们招募了438名土耳其机器人(Mechanical Turk)的众包工人参加这项实验,并给予被试一定的报酬。实验表明,参与者提供的数据通常达到甚至超过使用其他方式收集的数据设定的心理测量标准(Buhrmester, Talaifar & Gosling, 2018)。样本量与最近的其他实验相当,那些实验研究了算法介入沟通后的社会后果(Hohenstein & Jung, 2020; Jakesch, French & Ma et al., 2019)。

由于实验的重点是完整的对话,因此排除了总共交流少于10条信息的对话以及单个参与者发送少于3条信息(一对参与者中)的对话,还排除了没有意义的交流对话,得到了424名有效实验的参与者。对话平均持续6.81分钟(SD=2.31),平均包含21条信息(SD=7.55)。为了分析对话后的自我报告

结果，我们还排除了 63 名未完成完整调查的参与者，剩下的 361 名参与者（124 名女性，235 名男性，1 名“其他”）[①]的年龄从 18 岁到 68 岁不等（M＝34.07，SD＝10.1）。

2. 智能回复实验平台

我们开发了一种基于网络的实验工具 Moshi[②]，它能够在线招募参与者，并让他们获得智能回复支持以进行实时的人际沟通。

Moshi 是一种 Web 应用程序，允许两个参与者进行文本聊天。与现有具有智能回复功能的商业信息传递应用程序一样，参与者可以在标准文本框内手动输入信息，也可以直接选择智能回复点击发送。该研究工具为研究人员提供了一个实验平台，使研究人员能够完全控制显示的智能回复类型、显示方式和时间以及谁可以看到它们（请参阅补充材料以获取更多详细信息）。[③]

我们为“实验一”设置了两种模式，即不能使用智能回复和可以使用智能回复，每个模式都可以由参与者自由选择。在不能使用智能回复模式下，参与者必须手动输入想发送的信息；在可以使用智能回复模式下，参与者可以点击发送谷歌生成的智能回复。

3. 测量工具

为了评估智能回复对社会关系的影响，我们对主导和从属关系的感知进行了测量，对对话伙伴的合作沟通以及对话伙伴使用智能回复频次的感知进行了测量。为了评估智能回复对语言的影响，我们测量了交流互动过程中的沟通速度和消息情绪。

第一，主导地位和从属关系的感知。通过修订的人际形容词量表（IAS-

① 子数据之和小于总数，原文如此。——译者注

② Moshi 的设计灵感来自基于商业文本的 AI-MC 平台，例如谷歌已弃用的平台 Allo（法语“你好”），我们将平台命名为 Moshi（日语环境下接听电话时表示问候）。Moshi 是一个 Web 应用程序，可以在所有主流浏览器上运行，并且不需要参与者安装任何插件。该界面对设备类型做出反应，并调整自身大小以在台式机、平板电脑和移动设备上良好运行。除了用于发送信息的标准文本框外，参与者还可以点击发送智能回复。与现有的聊天应用类似，用户也可以随时滚动查看历史记录。Moshi 允许研究人员定义和构建他们的研究、平台的外观和感觉、实验室内的参与者数量等。设置完成后，研究人员只需要将参与者指向 Moshi，Moshi 即可自动分配并连接。

③ 补充材料可通过链接 https://github.com/Social-Design-Lab/moshi 获取。

R)对主导地位和从属关系的感知进行衡量,IAS-R 提供了人际交往中各个维度的实证衡量标准(Wiggins, Trapnell & Phillips, 1988)。为了简化测量步骤,根据维金斯(J. Wiggins)及其同事的分析,我们从每个人际关系的 8 个维度中选择了两个得分占比最高的形容词,从而对得出的 16 个项目进行排名,并进行说明:"下面是一系列描述人们如何与他人互动的陈述,根据你的直觉,请评估每个陈述对你的对话伙伴的描述准确性。"(Knutson, 1996)参与者根据"极度不准确"(1)、"有点准确"(4)和"极度准确"(7)的评分标准对每个陈述进行评分,研究者根据改编自维金斯等人(Wiggins, Trapnell & Phillips, 1988)的公式将这些评级组合在一起,以确定对从属关系和主导地位感知的评级(Knutson, 1996)。

第二,合作式交流的感知。我们通过一张 7 项量表(Lee, 1997)测试合作式交流,参与者根据与对话伙伴互动的结果,对量表提供的相关陈述进行同意度评分,并进行说明:"考虑你与对话伙伴的互动,请评价你对这些陈述的同意程度。"参与者根据"强烈反对"(1)和"非常同意"(7)的评分标准对每个陈述进行评分。

第三,智能回复频次的测量。参与者从"1 = 从不"到"5 = 总是"评价他们的对话伙伴可能使用智能回复的频次。所有调查内容的呈现在参与者之间是随机的,以避免参与者回复中潜在的顺序效应。

第四,沟通速度的测量。我们通过计算参与者平均每分钟发送的消息数量来计算沟通速度。

第五,消息情绪的测量。我们使用 VADER 测量消息情绪,VADER 是一种基于词汇和规则的情绪分析工具,专门分析社交媒体上表达的情绪(Hutto & Gilbert, 2014)。VADER 生成一个情绪指标,表明文本情绪的积极、消极或中立程度,其综合得分范围从"-1"(最消极)到"1"(最积极)。

4. 流程

参与者在在线调查平台 Qualtrics 上完成实验。参与者在对实验内容表示知情同意后,将使用该信息传递系统完成与匿名伙伴的讨论任务,即讨论一个与众包工人相关的问题——被不公平拒绝工作的问题(McInnis, Cosley

& Nam et al. , 2016)。具体而言,本研究要求参与双方就“土耳其机器人为了更好地处理被不公平拒绝工作的问题,可以做出的三项主要的改变”达成一致。在打开消息传递平台之后,参与者等待另一个参与者进入会话,如果 5 分钟后另一位参与者还没进入会话,该参与者可以提前退出调查并获得部分报酬。一旦在 5 分钟内另一个参与者进入会话,两人就有足够的时间进行商讨,从而就排名名单达成一致。在确认对话完成之后,参与者还需要完成任务后的测量。

5. 数据分析

按照工具变量(IV)评估的标准流程,我们计算了三种类型的估计:第一阶段效应、意向治疗效应和 IV 效应(Angrist, Imbens & Rubin, 1996)。在所有情况下,我们使用 clubSandwich R 软件包(Pustejovsky, 2017)中的 coef_test 函数计算组内聚类稳健标准误(CR^2)。第一阶段效应估计的是智能回复可用性的随机分配对参与者在对话中使用智能回复的频次的影响;意向治疗效应估计的是智能回复可用性的分配在多大程度上导致了结果的变化,如在调查后的评级、沟通速度或情绪等方面的变化;IV 效应评估对话伙伴增加智能回复的使用对自我评估结果的边际影响。具体来说,我们使用 IV 效应回归分析参与者自身的结果数据,通过对话伙伴随机分配条件来检测对话伙伴智能回复的使用情况,参与者自身的随机分配条件被当作协变量。我们在 AER 包(Kleiber, Zeileis & Zeileis, 2020)中使用了 ivreg 函数①,报告的估计值表示 IV 效应回归输出的系数、t 统计量和 p 值。

使用 IV 效应方法来评估智能回复使用(通过随机分配的可用性来衡量)对谈话速度、情绪和人际感知(主导地位、从属关系和合作式沟通)的影响。通过实验设计来排除限制假设是合理的,因为没有参与者被提前告知对话伙伴的智能回复可用性以及给定的消息是不是智能回复。

(二)实验二

为了更好地了解智能回复建议所传达的情绪如何影响交流语言,我们进

① R 语言计量包中的 ivreg 函数用以解决线性回归模型的内生性问题。

行了第二个实验,随机分配了 291 对参与者,使用以下四个应用程序之一来讨论政策问题:(1)两个参与者都收到谷歌的智能回复;(2)两个参与者都收到带有积极情绪的智能回复;(3)两个参与者都收到带有消极情绪的智能回复;(4)没有智能回复。

1. 被试

582 名土耳其机器人众包工人参加了这项实验,并从中获得了一定的金钱报酬:排除了总共交流少于 10 条消息的 13 对参与者;对话平均持续 6.33 分钟(SD = 2.67),平均包含 20.2 条消息(SD = 8.63);510 名有效参与者(92%)的年龄从 19 岁到 69 岁不等(M = 35.6,SD = 9.97);包括 206 名女性,275 名男性和 1 名“其他”。[①]

2. 测量工具

本实验使用了与“实验一”相同的研究平台,但扩展了两种额外的模式:积极情绪的智能回复和消极情绪的智能回复。例如,在积极情绪的智能回复条件下,参与者可能会看到诸如“我喜欢它”“我十分同意”的回复建议;在消极情绪的智能回复条件下,参与者可能会看到诸如“我不明白”“不,你不是”的回复建议。这些智能回复是从输入文件中随机选择的,不会过于重复(每次对话中推荐的三个备选回复不会立即被推荐在后续对话中)。智能回复使用的语句来自以前的交流习惯用语(Hohenstein & Jung, 2018)。实验要求参与者对智能回复的情绪进行评分——智能回复仅包括那些被评为具有明确积极情绪或消极情绪的建议。

为了评估智能回复对语言的影响,我们测量了消息传递情绪,该测量与“实验一”一样。

3. 流程

与“实验一”相似,只有具有智能回复可用性的参与者才被告知“使用人工智能介入的消息系统与你的对话伙伴进行对话。当你发送信息时,人工智能将提供智能回复建议,你只需要点击即可发送”;而控制条件下的参与者则

---

① 子数据之和小于总数,原文如此。——译者注

被告知“使用标准消息传递系统与你的交流伙伴进行对话”。

4. 数据分析

我们使用 estimatr R 包中的 lm_robust 函数,使用具有组内聚类稳健标准误的简单线性回归,在个体水平上分析了结果数据。因变量是个体语言测量(VADER 情绪),自变量是指定条件,没有协变量。报告的统计数据是相关系数的 t 统计数据和 p 值,以及手动计算的 Cohen’s d 值。

为确保实验发现的交流用语差异不是人口统计学差异的结果,实验检查了条件之间的人口构成(年龄、性别和种族),没有发现任何显著差异。

由于语言查询和单词计数(LIWC)是 VADER 情绪分析的先导,因此我们研究了 LIWC 提供的情感测量。LIWC 是一种基于词典的文本分析工具,用于确定反映一些语言过程、心理过程和个人关注的单词的百分比(Hutto & Gilbert, 2014; Pennebake, Francis & Booth, 2001)。我们利用 LIWC 情感分数来检查情感语言的使用是否随着智能回复的引入和使用而变化,情感的数值范围为[0,100],由 LIWC 中积极情绪分数和消极情绪分数的总和计算得出。

我们发现:积极智能回复和谷歌智能回复比没有智能回复对对话具有更大的影响[$t(124)=2.95$, $p<0.001$, $d=0.272$];积极智能回复和谷歌智能回复对情感的影响在统计学上相似[$t(150)=0.354$, $p=0.724$];与没有智能回复的情况相比,消极智能回复具有更强的负面影响[$t(123)=-3.50$, $p<0.001$, $d=0.454$]。综上所述,这些发现证明了人工智能生成的情绪语句如何影响人类对话中情感语言的使用。

参考文献

Angrist, J., G. Imbens & D. Rubin 1996, “Identification of Causal Effects Using Instrumental Variables.” *Journal of the American Statistical Association* 91.

Arnold, K., K. Chauncey & K. Gajos 2020, “Predictive Text Encourages Predictable Writing.” Proceedings of the 25th International Conference on Intelligent User Interfaces.

Berger, C. & R. Calabrese 1975, “Some Explorations in Initial Interaction and Beyond: Toward A Developmental Theory of Interpersonal Communication." *Human Communication Research* 1.

Breck, E., Y. Choi & C. Cardie 2007, "Identifying Expressions of Opinion in Context." *International Joint Conferences on Artificial Intelligence* 7.

Buhrmester, M., S. Talaifar & S. Gosling 2018, "An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use." *Perspectives on Psychological Science* 13.

Bullock, G. 2017, "Save Time with Smart Reply in Gmail." https://www.blog.google/products/gmail/save-time-with-smart-reply-in-gmail.

Crawford, K. & M. Whittaker 2016, *The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*, New York: AI Now Institute at New York University.

Hancock, J., M. Naaman & K. Levy 2020, "AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations." *Journal of Computer-Mediated Communication* 25(1).

Henderson, M., R. Al-Rfou & B. Strope et al. 2017, "Efficient Natural Language Response Suggestion for Smart Reply." arXiv preprint arXiv:1705.00652.

Hohenstein, J. & M. Jung 2018, "AI-Supported Messaging: An Investigation of Human-Human Text Conversation with AI Support." Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems: CHI'18, https://doi.org/10.1145/3170427.3188487.

Hohenstein, J. & M. Jung 2020, "AI as A Moral Crumple Zone: The Effects of AI-Mediated Communication on Attribution and Trust." *Computers in Human Behavior* 106.

Hohenstein, J., R. Kizilcec & D. DiFranzo et al. 2020, "As Predicted: Effect of Smart Reply Use on Language and Interpersonal Perceptions." https://aspredicted.org/4mi3z.pdf.

Hutto, C. & E. Gilbert 2014, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." Eighth International AAAI Conference on Weblogs and Social Media.

Jakesch, M., J. Hancock & M. Naaman 2023, "Human Heuristics for AI-Generated Language Are Flawed." *Proceedings of the National Academy of Sciences* 120(11).

Jakesch, M. , M. French & X. Ma et al. 2019, "AI-Mediated Communication: How Profile Generation by AI Affects Perceived Trustworthiness." Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems: CHI'19, https://doi. org/10. 1145/3290605. 3300469.

Kannan, A. , K. Kurach & S. Ravi et al. 2016, "Smart Reply: Automated Response Suggestion for Email." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Kleiber, C. , A. Zeileis & M. Zeileis 2020, "Package 'AER'." *R Package Version* 12(4).

Knutson, B. 1996, "Facial Expressions of Emotion Influence Interpersonal Trait Inferences." *Journal of Nonverbal Behavior* 20.

Kraus, R. 2018, "Gmail Smart Replies may Be Creepy, but They're Catching on Like Wildfire." https://mashable. com/article/gmail-smart-reply-growth.

Lee, J. 1997, "Leader-Member Exchange, the 'Pelz Effect', and Cooperative Communication Between Group Members." *Management Communication Quarterly* 11.

Lee, M. 2018, "Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management." *Big Data & Society* 5.

Mairesse, F. , M. Walker & M. Mehl et al. 2007, "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text." *Journal of Artificial Intelligence Research* 30.

McInnis, B. , D. Cosley & C. Nam et al. 2016, "Taking A HIT: Designing Around Rejection, Mistrust, Risk, and Workers' Experiences in Amazon Mechanical Turk." Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.

Pennebaker, J. , M. Francis & R. Booth 2001, "Linguistic Inquiry and Word Count: Liwc 2001." *Mahway: Lawrence Erlbaum Associates* 71.

Pennebaker, J. , M. Mehl & K. Niederhoffer, 2003, "Psychological Aspects of Natural Language Use: Our Words, Our Selves." *Annual Review of Psychology* 54.

Promberger, M. & J. Baron 2006, "Do Patients Trust Computers?" *Journal of Behavioral Decision Making* 19.

Pustejovsky, J. 2017, "Club Sandwich: Cluster-Robust (Sandwich) Variance Estimators with Amall-Sample Corrections. R Package Version 0. 2. 3." *The R Foundation for Statistical*

Computing.

Rahwan, I., C. Cebrian & N. Obradovich et al. 2019, "Machine Behaviour." *Nature* 568.

Rienks, R. & D. Heylen 2005, "Dominance Detection in Meetings Using Easily Obtainable Features." International Workshop on Machine Learning for Multimodal Interaction.

Ritter, A., C. Cherry & W. Dolan 2011, "Data-Driven Response Generation in Social Media." Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Shakirov, V. 2016, "Review of State-of-the-Arts in Artificial Intelligence with Application to AI Safety Problem." arXiv preprint arXiv:1605.04232.

Stone, P., R. Brooks & E. Brynjolfsson et al. 2016, "Artificial Intelligence and Life in 2030." in *One Hundred Year Study on Artificial Intelligence: Report of the 2015 - 2016 Study Panel*, Stanford: Stanford University Press.

Sukumaran, A., S. Vezich & M. McHugh et al. 2011, "Normative Influences on Thoughtful Online Participation." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI'11.

Tong, S. & J. Walther 2011, "Relational Maintenance and CMC." *Computermediated Communication in Personal Pelationships* 53.

Whittaker, M., K. Crawford & R. Dobbe et al. 2018, *AI Now Report 2018*, New York: AI Now Institute at New York University.

Wiggins, J., P. Trapnell & N. Phillips 1988, "Psychometric and Geometric Characteristics of the Revised Interpersonal Adjective Scales (IAS-R)." *Multivariate Behavioral Research* 23.

Zhang, J., J. Chang & C. Danescu-Niculescu-Mizil et al. 2018, "Conversations Gone Awry: Detecting Early Signs of Conversational Failure." arXiv preprint arXiv:1805.05345.

智能社会研究
Journal of Intelligent Society

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：2023 年 7 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：mailtojis@163.com

著作权使用说明