心理学探新 2026,Vol.46,No.4,291 - 298
PSYCHOLOGICAL EXPLORATION

拟人化表情和伦理设计对人与智能机器合作的影响*

高山川 刘 欢
(复旦大学心理学系,上海 200433)

摘 要:采用博弈实验考察外显表情和内在伦理设计如何影响人对智能机器的合作预期与行为。结果发现:(1)信任博弈中,高兴表情相比于中性表情能同等地提升人对机器和人类对手的合作预期,却不一定能促进人对机器的合作行为,有机会谋取更大利益时人更多地选择背叛;(2)囚徒困境博弈中,相比造福人类取向,功利主义的伦理设计会让人对机器人的合作预期与行为均明显下降,不合作者的内疚程度变得更低。可见除了外表显得友善更应改进智能机器的伦理设计。

关键词:合作;智能机器;表情;伦理设计;内疚
中图分类号:B848 **文献标志码**:A **文章编号**:1003 - 5184(2026)04 - 0291 - 08

1 引言

智能机器是利用算法进行自动化决策的各种系统的简称,在人工智能(Artificial Intelligence,AI)的驱动下愈发强大,AI 技术能从数据中识别模式、推断乃至自主学习(Mahmud et al.,2022)。随着智能机器应用日益广泛,人经常面临是否与之协作的决策,例如是否接受机器的服务、建议、决定或给自动驾驶汽车让行等,研究哪些因素能促进人 - 机合作很有意义。根据丁毅等人(2015)总结,人的合作行为受多种因素影响,自己需要控制个人利益最大化动机,对他人还要有信任,即对其行为和意图有积极预期,这又可利用所获取的社会线索进行推断,主要有面孔可信任性、感知的道德品质等。张慧等人(2026)又指出,增强人机信任可以借鉴人际信任的研究成果,从改进机器的外在特征和实际表现入手。为促进人对智能机器的信任与合作,常见的做法是为其增添面部表情以更像人类(Terada & Takeuchi,2017;Whiting et al.,2021)。不过,人对机器有积极预期不保证一定有合作行为,也可能利用其善意剥削对方(Karpus et al.,2021)。这就要求更系统地考察拟人化表情的效果有多稳健,能否同时促进人对机器的合作预期与行为。另一方面,无私利他的道德品质也能增加人的信任与合作(Delgado et al.,2005;Fareri,2019)。如今不少算法决策都涉及道德权衡,智能机器已被视为道德决策的主体或代理(Bonnefon et al.,2024;Liu et al.,2025),因此也需探究人如何感知智能机器的伦理取向、又会如何影响人的合作预期与行为。进一步地,Cachia 等(2025)发现,表情可信度作为迅速产生的第一印象,在人对他人知之甚少时会影响合作行为,当人还有机会获知其利他品质时,信任可能会发生更新,合作行为也会随之变化,即道德因素可能更重要。可见,有必要对这两种与信任密切关联的因素之作用都进行探究并加以对照。据此,研究将着重考察表情和伦理设计如何影响人对智能机器的合作预期与行为,并揭示积极预期是否一定伴随更多合作行为,这有助于发现真正有效促进人 - 机合作的方法。

人 - 机合作研究常采用改编自人 - 人博弈的任务,如囚徒困境、信任博弈等,人往往需要权衡他人是否值得自己出让部分利益(Everett et al.,2018)。不同任务还有不同的不确定性:例如囚徒困境博弈中双方要同时做选择,只能依据可得的线索推测对手是否可信并选择是否合作;信任博弈中双方先后做选择,先选的一方需要推断对手是否可信,如果选了合作,对后选者来说不确定性已消解,如果背叛就特别显出辜负了对方的信任。无论如何,不少研究显示人对机器比对人类的合作水平更低(de Melo et al.,2016;Karpus et al.,2021;Makovi et al.,2023)。Ishowo - Oloko 等人(2019)还发现,当人以为对手是真人(实为机器)时不会减少合作,但以为对手是机器(实为真人)时更多地选择背叛。究其原因,可能在于人有区别对待群体内、外成员的倾向(Balliet et

* **基金项目**:教育部高等学校心理学类专业教学指导委员会教育教学改革项目(22 - 43)。
通信作者:高山川,E - mail:tracygao@ fudan.edu.cn。

al. ,2014)。脑成像研究也表明,与人博弈时人脑的特定区域会激活,与机器博弈时则不会(Chaminade et al. ,2012)。de Melo 等人(2016)还指出,人更多地与人类合作是为了避免背叛带来的内疚,对机器则没有这种顾虑。研究拟考察增加一条外表线索会如何影响人 - 机合作,仍会与同样条件下的人 - 人合作对比,基于以往研究推测人会区别对待两类对手,提出假设 1:人对智能机器比对人类有更少的合作行为。

为了增加智能机器在外表和情感上的类人性,一种常见做法是为其添加面部表情,表情具有重要的社会功能,是提示心理状态或行为意图的线索(de Melo et al. ,2014)。聊天机器人在文字回复中添加表情符号(emoji)就能增强用户对其温暖友善等社会性特质的感知(陈颖冰,龚明亮,2026)。就人 - 机博弈而言,不少研究表明,让机器呈现情绪反应能促进人与之合作。例如 Terada 和 Takeuchi(2017)为机器加上了线条勾勒的表情,嘴角线条从向下、水平到向上表示由伤心变成高兴,比起固定不变的中性表情,表情变化能让人对机器更公平地分配利益。Whiting 等人(2021)也发现,高兴、伤心、愤怒等抽象化表情符号有沟通功能,比毫无交流能让人与机器在复杂任务中高效合作。Takahashi 等人(2021)还让机器人用肢体动作、眼睛色彩和语音应答组成多模态情感表达,比起固定的中性表情,表情变化的反应模式能有效促进人 - 机合作。尽管上述研究提示增添表情可让机器获得更多合作,仍有必要深入探究,积极的表情在机器与人类身上能否有同等效果。研究拟选取高兴表情(表示友好)与中性表情相对照,考虑到人更愿意由人类而非机器获得收益(von Schenk et al. ,2025),结合以往发现可以推测,高兴表情对人与两类对手的合作行为会有不同程度的促进,提出假设 2:高兴表情比中性表情能让人有更多合作行为,但这种促进作用在对手为机器时比对手为人类时更小。

需注意的是,人与机器博弈时合作行为较少不一定源于合作预期较低。尽管人 - 人博弈中合作预期越高则合作行为越多(熊承清 等,2021),人 - 机博弈时合作预期与行为不一定会同步变化。Karpus 等人(2021)利用改编的信任博弈任务揭示,人对机器的合作预期不低于人类,却会更多地背叛,可见人并非不信任机器,而是更想趁机为自己谋利,即算法剥削。Karpus 等人(2025)在跨文化比较时也发现,日本民众和美国民众对机器的合作预期差不多,仍是合作行为不一样。据此推测,加上表情后,人对有相同表情的两类对手仍有同等的合作预期,但这不一定保证有同等的合作行为,因为人为不同对象控制利己冲动的意愿可能不同(丁毅 等,2015)。由上提出假设 3:在表情相同条件下,人对智能机器的合作预期不低于人类;假设 4:高兴表情可同等地提升人对两类对手的合作预期。

无论如何,面部表情提示的可信任性只是一种初始信念,人还会在互动中观察对手的品行并更新信念,即信任形成是一个动态过程(高青林,周媛,2021),因此还有必要探索人日常感知到智能机器的内在品质如何。近年来学者们呼吁要加强智能机器的道德决策体系设计以符合公众的道德准则,但如何设计尚在摸索中(赵雷 等,2019)。Bonnefon 等人(2024)进一步指出,智能机器既是自主决策的主体,又可成为他人行动的代理,无论程序中是否编入伦理准则,只要行动后果涉及不同人的利益,都属于道德决策:生死攸关的两难困境属于显性道德决策,如自动驾驶汽车面临的“电车难题”;各种智能系统的日常应用也会广泛影响不同人的利益分配,属于隐性道德决策。后一类情境显然更为常见,值得重视。

智能机器遵从怎样的伦理准则取决于研发者如何设计,可简称为伦理设计,这又跟组织的发展目标密切相关。起初研发 AI 的目标主要是增进人类福祉,许多国家和组织倡导公益、公平、安全的发展准则,例如由专家联名签署的阿西洛马 AI 原则提倡让广大人群受益(Thiebes et al. ,2021)。随着竞争加剧,当大众利益与组织利益冲突时,组织可能偏离这一初衷,转向为己方谋取最大利益。Mathur 等人(2019)发现购物网站采用了诱导消费的多种“暗黑模式”,如自动续订、社会劝服、营造紧迫感、价格误导等。Abeliuk 等人(2024)指出婚恋网站希望让用户长久停留但用户想尽快找到匹配对象。Kozyreva 等人(2020)也提醒,数字化环境带来四大挑战:操纵选择、自动过滤信息、传播虚假信息及抢占注意力,网络平台出于商业目的可能疏于管理乃至故意为之。Dietvorst 和 Bartel(2022)还发现,不少人相信算法决策遵从结果最大化原则,注重成本 - 效益分析,这符合伦理学中的结果论、主要代表为功利主

义，与之相对的是强调行为或过程是否正当的道义论（如“不伤害他人”原则）。他们对医疗保险问题的研究也证实，在越是涉及道德权衡的情境下，人越不愿意让算法替代人类决策。

可见，优先考虑组织的商业利益还是大众的福祉可以说是对智能系统进行伦理设计的两种不同取向，可分别简称为功利主义取向和造福人类取向。人－人博弈已发现，相比于注重道义原则的人，人认为注重结果最大化者更不可信，也更少与之合作（Everett et al.，2018）。据此推测，功利主义取向会让人担心与机器博弈时会被剥削，伴随消极预期的是较少的合作行为，提出假设5：相比于造福人类取向，功利主义取向会降低人对智能机器的合作预期，假设6：功利主义取向还会减少人对智能机器的合作行为。

综上，研究将由表及里地考察表情和伦理设计取向如何影响人对智能机器的合作预期与行为。分为两个子研究：研究一采用信任博弈，考察人如何与不同表情的机器人合作，并与有同样表情的人类对手相对比，以检验表情的作用有多强；研究二为囚徒困境博弈，考察当人感知到两种不同伦理取向时又会如何与机器人对手合作，以验证加强伦理设计的重要性。预计上述结果还将形成对照，展现两种因素的不同作用。

2　研究一：不同的对手及其表情对合作的影响

参考Karpus等人（2021）改编的信任博弈设计了实验任务，由于博弈双方是先后做选择，研究对象将参与两次不同的博弈，分别为先选、后选的角色，将检验假设1～4。

2.1　方法

2.1.1　对象和研究设计

采用GPower 3.1确定样本量，按照逻辑回归中等效应量的优势比（odds ratio，OR）3.47、显著性水平α为0.05，需104人达到80%（$1-\beta$）的效力。招募了132名大学生（男性占45%），平均年龄19岁（$SD=1.56$），随机分入四组，每组33人。

采用2（对手：人类、智能机器人）×2（表情：高兴、中性）被试间设计。一个因变量是预期对手将合作的人数比例（简称合作预期），另一个是选择与对手合作的人数比例（简称合作行为）。附加测量了不合作者的内疚。

2.1.2　材料和程序

采用E－prime 2.0编写程序，参与者在电脑上独自完成流程，开始前签署了知情同意书，其博弈得分决定了报酬（10～15元）。参与者先后扮演玩家一、玩家二进行两次博弈：玩家一有优先选择权，若不予合作（出黑牌）则博弈结束、双方各得30分，若合作（出红牌）则玩家二获得选择权，若同样合作（出红牌）双方各得70分，若不合作（出黑牌）玩家二得100分、玩家一得0分。与人类博弈一组被告知对手为陌生人，与机器人博弈一组被告知对手为Alpha或Beta机器人，通过电脑博弈，用高兴或中性表情符号代表对手形象。

实验分四步：一，阅读信任博弈介绍并回答三道检测题，全答对后才进入下一步；二，作为玩家一与第一个对手博弈，先预测对手是否合作，再决定自己是否合作，若不合作还要评估内疚程度，对手的出牌由程序随机决定（到最后算分时揭晓，以免干扰下一次博弈）；三，作为玩家二与新对手博弈，仍先做预测，然后会看到对手选择了合作，决定自己是否合作，若不合作仍要评估内疚程度；四，回答表情符号的操纵检验问题。对内疚采用Likert 4点评分，1～4表示“一点也不内疚”到“非常内疚”。最后解释了研究目的。

预研究显示抽象的表情符号能有效代表人类或机器人的表情。选用苹果手机常用的高兴与中性表情符号，分别对37、39人进行了调查，结果发现：对高兴符号，82%的人认为反映高兴情绪，其余18%视为中性情绪；对中性符号，62%的人认为反映中性情绪，仅3%视为高兴情绪，其余35%视为其它情绪；经卡方检验两组在高兴情绪的判断上有显著差异，$\chi^2(1, N=76)=45.68$，$p<0.001$，$\varphi=0.78$，中性情绪的判断也有显著差异，$\chi^2(1, N=76)=12.41$，$p<0.001$，$\varphi=0.40$。对正式实验对象也做了类似的操纵检验，结果同样显示操纵有效（略），还让其对表情的友好程度进行Likert 7点评分（1～7），独立样本t检验显示，高兴表情的友好程度（$M=5.27$，$SD=0.95$）显著高于中性组（$M=3.73$，$SD=0.99$），$t(130)=9.16$，$p<0.001$，$d=1.59$。采用SPSS 26.0进行数据的统计分析。

2.2　结果与分析

2.2.1　信任博弈中人与不同对手的合作

表1与表2分别列出了参与者作为玩家一、玩家二与不同对手博弈的结果，包括预期对手合作的

人数比例、选择合作的人数比例及不予合作者的内疚程度。

表 1 显示，参与者先选时，只要对手表情高兴，有合作预期与合作行为的人数比例均在 70% 或以上。

表 1 先选择时与不同对手的合作（$N = 132$）

表情	对手	n	合作预期		合作行为		不合作者的内疚		
			人数	百分比	人数	百分比	人数	M	SD
高兴	智能机器人	33	23	70%	24	73%	9	1.33	0.50
	人类	33	25	76%	26	79%	7	1.43	0.54
中性	智能机器人	33	10	30%	9	27%	24	1.13	0.34
	人类	33	14	42%	17	52%	16	1.31	0.70

表 2 后选择时与不同对手的合作（$N = 132$）

表情	对手	n	合作预期		合作行为		不合作者的内疚		
			人数	百分比	人数	百分比	人数	M	SD
高兴	智能机器人	33	24	73%	7	21%	26	1.92	1.16
	人类	33	27	82%	25	76%	8	2.63	1.30
中性	智能机器人	33	16	48%	13	39%	20	1.60	0.75
	人类	33	18	55%	17	52%	16	1.81	0.83

表 2 显示，参与者后选时，对高兴表情的机器人预期不低，但与之合作者仅占 21%，且不合作者的内疚未达到“有点内疚”的水平。

2.2.2 对手的身份及其表情对合作的影响

首先，对参与者作为玩家一的博弈结果，以对手身份（人类为参照）及其表情（中性表情为参照）为自变量，合作预期为因变量（0、1 分别代表预期对方不合作、合作）进行逻辑回归，结果发现：高兴比中性表情更可能引发合作预期，$B = 1.45$，$OR = 4.24$，95% confidence interval（CI）= [1.48，12.17]，$p = 0.007$；参与者对机器与对人类的合作预期无显著差异，$B = -0.53$，$OR = 0.60$，95% CI = [0.21，1.63]，$p = 0.308$，回归模型可解释的变异量 Nagelkerke $R^2 = 0.18$。类似地，以对手身份及其表情对合作行为（0、1 分别代表不合作、合作）也进行了逻辑回归，结果发现：高兴比中性表情更可能引发合作行为，$B = 1.25$，$OR = 3.50$，95% CI = [1.19，10.28]，$p = 0.023$；相比于人类，参与者更少与机器合作，$B = -1.04$，$OR = 0.35$，95% CI = [0.13，0.99]，$p = 0.047$；对手的身份与表情的交互作用不显著，$B = 0.71$，$OR = 2.03$，95% CI = [0.44，9.38]，$p = 0.363$，Nagelkerke $R^2 = 0.21$。还用方差分析考察了对手的身份及其表情对内疚的影响，未发现显著效应（$ps > 0.05$）。

接着，对参与者作为玩家二的博弈结果进行逻辑回归分析（方法同上）。结果发现合作预期受到的影响与上面类似：高兴表情更可能引发合作预期，$B = 1.32$，$OR = 3.75$，95% CI = [1.23，11.48]，$p = 0.021$；参与者对机器与人类的合作预期仍无显著差异，$B = -0.24$，$OR = 0.78$，95% CI = [0.30，2.06]，$p = 0.623$，Nagelkerke $R^2 = 0.11$。对合作行为的逻辑回归分析则显示：高兴表情更可能引发合作行为，$B = 1.08$，$OR = 2.94$，95% CI = [1.03，8.39]，$p = 0.044$；对手身份的主效应不显著，$B = -0.49$，$OR = 0.61$，95% CI = [0.23，1.62]，$p = 0.324$，但对手的身份与表情对合作行为有显著交互作用，$B = -1.96$，$OR = 0.14$，95% CI = [0.03，0.64]，$p = 0.011$，Nagelkerke $R^2 = 0.20$；进一步做卡方检验显示，对手为中性表情时，与机器合作的人数比例接近人类，$\chi^2(1, N = 66) = 0.99$，$p = 0.323$，$\varphi = 0.12$，但对手为高兴表情时，与机器合作的人数比例显著更低，$\chi^2(1, N = 66) = 19.65$，$p < 0.001$，$\varphi = 0.55$，且该比例（21%）显著低于随机水平 50%，$\chi^2(1, N = 33) = 10.94$，$p = 0.001$，$\varphi = 0.58$。另外方差分析发现：不与高兴对手合作的内疚（$M = 2.08$，$SD = 1.22$）高于不与中性表情对手合作的内疚（$M = 1.69$，$SD = 0.79$），$F(1, 66) = 4.61$，$p = 0.036$，$\eta_p^2 = 0.065$，对手身份的主效应及二者的交互作用均不显著（$ps > 0.05$）。

2.3 讨论

有关合作行为的假设 1 在参与者先选时被验证。假设 2 在后选时被验证，有关合作预期的假设 3、4 无论在先选或后选时均被验证。具体而言，在表情相同条件下，人对机器与对人类的合作预期差不多，当表情从中性变成高兴时，人对二者的合作预期也同等地上升，可见表情对合作预期的影响不会因对手不同而异。但是，表情对合作行为的影响较复杂：参与者先选时，高兴表情能同等地促进他们与

两类对手的合作;但后选时他们对同为高兴表情的两类对手采取了截然不同的行为,此时机器人已选了合作,但愿意回馈对方的人数比例却低至21%,远少于与人类合作的比例76%。可见,人对机器的合作行为偏少并非由于合作预期偏低,此外,高兴表情有助于人类收获更多合作,却会让人更敢于背叛机器。换言之,表情友善对促进人-机合作的作用有限,虽可提升合作预期,但不足以让人放弃自私行为。正如丁毅等人(2015)所言,合作能否达成不仅取决于信任,还要求人能自我控制。这也提示需寻找更有效方法以同时促进人对机器的合作预期与行为。

研究一中表情是人没有其他信息时所利用的简单线索,研究二会提供智能机器日常应用的不同介绍,让人对其伦理设计取向形成不同感知,即有机会深入了解对手的内在品质。

3　研究二:智能机器的伦理设计对人-机合作的影响

由于伦理设计取向之一为功利主义,继续采用先后做选择的信任博弈不太合适(功利化的机器人先选择合作与设定不符),研究二将采用双方同时做选择的囚徒困境博弈,此任务也可反映人是否预期对手值得信任并且愿意共享利益,将检验假设5、6。

3.1　方法

3.1.1　对象和研究设计

采用GPower 3.1确定样本量,按照卡方检验中等效应量为0.3、显著性水平α为0.05,88人可达到80%(1-β)的效力。招募了132名大学生(男性占46%),平均为20岁(SD=1.60),随机平分入两组,每组66人。采用单因素被试间设计,自变量为感知的智能机器的伦理设计(造福人类取向、功利主义取向),因变量同研究一。

3.1.2　材料和程序

参与者的博弈得分决定了报酬(5~10元)。参照囚徒困境博弈设计了任务,参与者被告知对手为智能机器人Omega(无形象),通过电脑博弈,双方同时做出选择:若都合作(出红牌)各得70分,若都不合作(出黑牌)各得30分,若一方不合作(出黑牌)而另一方合作(出红牌),分别得100分、0分。机器人出牌由程序随机选择并呈现于屏幕。

通过预研究编制阅读材料,参考Kozyreva等人(2020)总结,结合智能机器的应用现状编写了两种版本介绍:版本一突显智能系统在哪些领域改善了人类生活,包括交通运输、医疗、法律、传媒、婚恋介绍、搜索、日常生活等,例如"搜索引擎通过优化的算法和搜索提示、模糊搜索、图片搜索、学术搜索等技术,帮助用户快速找到所需信息,促进问题解决";版本二突显智能系统在哪些领域进行商业化运作,包括购物、视频、搜索、婚恋介绍、社交、日常生活等,例如"在视频平台上,自动推荐算法源源不断地向用户推送可能感兴趣内容,吸引他们尽可能长久停留,这样平台就能展示更多广告,从而吸引更多商家打广告"。分别招募了39、37人进行调查,让其对材料反映的伦理设计进行Likert 7点评分,1~7表示从"非常重视提升人类福祉"到"非常重视商业利益最大化"。独立样本t检验显示,第二组的评分(M=5.92,SD=1.30)接近6分且显著高于第一组(M=3.13,SD=1.45),$t(74)=8.35$,$p<0.001$,$d=1.92$。对正式实验对象也做了类似的操纵检验,结果同样显示操纵有效(略)。

实验分四步:一,阅读囚徒困境博弈介绍并回答三道检测题,全答对后才进入下一步;二,阅读智能机器应用现状介绍;三,与机器人博弈,先预测对方是否合作,再决定自己是否合作,若不合作还要评价内疚程度;四,对材料反映的伦理设计评分。

3.2　结果与分析

表3列出了两组参与者与机器人博弈的结果。功利主义取向下,预期机器人合作和与之合作的人数比例都不到30%,造福人类取向下,尽管预期较高,合作者仍偏少。

表3　不同取向下人与机器人的合作(N=132)

伦理设计取向	n	合作预期		合作行为		不合作者的内疚		
		人数	百分比	人数	百分比	人数	M	SD
造福人类	66	42	64%	26	39%	40	1.83	0.82
功利主义	66	19	29%	15	23%	51	1.39	0.95

对两组人数做卡方检验显示:相比于造福人类取向,功利主义取向下预期机器人合作的人数比例显著更低,$\chi^2(1,N=132)=16.12$,$p<0.001$,$\varphi=0.35$,与之合作者也显著更少,$\chi^2(1,N=132)=4.28$,$p=0.039$,$\varphi=0.18$,且这两个比例均显著低于随机水平50%,$\chi^2(1,N=66)=11.88$,$p=0.001$,φ

$=0.42$，$\chi^2(1, N=66)=19.64$，$p<0.001$，$\varphi=0.55$。对不合作者的内疚做独立样本 t 检验，功利主义取向下内疚程度明显更低，$t(89)=2.31$，$p=0.023$，$d=0.49$。

3.3 讨论

研究二中假设 5、6 均得到验证，当人认为智能机器的伦理设计遵从功利主义取向时，对机器人的合作预期与行为均明显降低。在研究一中，尽管人对机器有较高合作预期不能保证有较多合作行为，但如果人愿意克制利己动机仍可能增进合作。研究二则提示，功利主义的感知会让人连信任也失去，此时要达成与机器的合作更加困难，不合作者也更不会内疚。可见，伦理取向不符合公众需求会造成更为不利的影响。这与人－人博弈结果有类似之处，对那些秉持功利主义的人，人更不信任、也更少合作（Everett et al.，2018）。另外，Cachia 等人（2025）还发现，提示人品的信息往往比表情有更重要的作用，未来也可进一步验证智能机器的内在品质与表情信息并存时前者的权重是否更大。

4 总讨论

研究采用两个博弈任务考察人如何基于表情和伦理设计取向推测智能机器的合作意图并选择是否合作，发现两种因素对合作预期与行为的影响有不同之处。这些结果不仅丰富了社会合作的认识，还可为改进智能机器的设计提供参考。正是由于可以根据人类的需求进行自上而下的设计，才能避免在残酷的竞争中演化出“自私的机器”（Boudry & Friederich，2025）。

研究表明，为促进人－机合作，首先要提升人对机器的合作预期或者说信任，通过设计友好的表情或宣传智能机器如何改善人类生活都是可行的。这表明建立人机信任与建立人际信任是类似的，对象的外表友善和道德表现良好都能增强人的信任（张慧 等，2026）。这也符合信任形成的观点，即初始印象和互动的经验信息都会影响信任（高青林，周媛，2021）。

不过，仅有积极预期还不够。研究显示，人对智能机器有较高预期不一定伴随较多合作行为，面对表情友好且已选择合作的机器人，人并未投桃报李，近 80% 的人选择独吞收益。这也不同于人－人博弈中发现的互惠倾向（熊承清 等，2021）。Karpus 等人（2021）已发现人会抓住机会剥削机器，研究又揭示，加上友善表情并未减少人的机会主义行为。研发者需综合考虑人的利己动机和应用场景进行外表设计，甚至要让表情有所变化。Makovi 等人（2023）则认为，拟人化不是根本解决之道，可建立与机器互动的社会规范，将其视为新加入成员。

研究结果也在拟人化表情之外提供了新的思路，即要重视智能机器的伦理设计，避免过分商业化让人形成功利主义感知，以至于合作预期与行为双双下降。前人已发现功利主义会引起算法厌恶（Dietvorst & Bartel，2022），研究又表明这种取向还会减少人－机合作，不利于智能机器的推广应用。Cabiddu 等人（2022）也指出，拟人化外表有利于建立初始信任，但如果人在使用过程中形成消极认知，信任也会降低。可见，研发智能机器不能只做到表面友善，而要让人真正体验到生活改善。做好伦理设计需要加强智能机器的道德心理学研究，不仅要探讨显性的道德两难问题，还应关注日常应用中的隐性道德决策（Bonnefon et al.，2024）。

研究存在一些不足。研究对象为大学生，其结果不能代表其他人群；只比较了高兴与中性表情，也只考察了单次博弈。未来研究可面向多样化群体，关注个体接纳新技术的差异，或采用虚拟面孔或实体机器人，在重复博弈中考察策略与表情相结合的影响，也可针对特定领域考察伦理设计如何影响合作。

5 结论

（1）相比中性表情，高兴表情能同等提升人对机器人和人类对手的合作预期，却不一定能增加人对机器人的合作行为，当有机会为自己谋取更大利益时人更多地选择背叛；（2）相比造福人类取向，功利主义的伦理设计会让人对机器人的合作预期与行为均明显下降，且不合作者的内疚程度变得更低。

参考文献

陈颖冰，龚明亮.（2026）. 文字之外的情感：Emoji 对 AI 社会性特质感知的影响. *心理学探新*，*46*（3），195－203.

丁毅，纪婷婷，陈旭.（2015）. 社会两难情境下的合作选择：自我利益和集体利益间的权衡. *心理学探新*，*35*（2），159－164.

高青林，周媛.（2021）. 计算模型视角下信任形成的心理和神经机制：基于信任博弈中投资者的角度. *心理科学进展*，*29*（1），178－189.

熊承清，许佳颖，马丹阳，刘永芳.（2021）. 囚徒困境博弈中对手面部表情对合作行为的影响及其作用机制. *心理学报*，*53*（8），919－933.

张慧，敬一鸣，古若雷.（2026）. 你会像信任人一样信任智能

技术吗？人机信任与人际信任的关系.*心理学探新*,*46*(1),3－12.

赵雷,李园园,叶君惠,胡凤培.(2019).人对智能机器行为的道德判断:现状与展望.*应用心理学*,*25*(4),306－318.

Abeliuk,A.,Elbassioni,K.,Rahwan,T.,Cebrian,M.,& Rahwan,I.(2024).Price of anarchy in algorithmic matching of romantic partners.*ACM Transactions on Economics and Computation*,*12*(1),1－25.

Balliet,D.,Wu,J.,& De Dreu,C. K.(2014).Ingroup favoritism in cooperation:A meta－analysis.*Psychological Bulletin*,*140*(6),1556－1581.

Bonnefon,J. F.,Rahwan,I.,& Shariff,A.(2024).The moral psychology of Artificial Intelligence.*Annual Review of Psychology*,*75*(1),653－675.

Boudry,M.,& Friederich,S.(2025).The selfish machine? On the power and limitation of natural selection to understand the development of advanced AI.*Philosophical Studies*,*182*(7),1789－1812.

Cabiddu,F.,Moi,L.,Patriotta,G.,& Allen,D. G.(2022).Why do users trust algorithms? A review and conceptualization of initial trust and trust over time.*European Management Journal*,*40*(5),685－706.

Cachia,J. Y. A.,Blevins,E.,Chen,Y.－C.,Ko,M.,Yen,N.－S.,Knutson,B.,& Tsai,J. L.(2025).Cultural variation in the smiles we trust:The effects of reputation and ideal affect on resource sharing.*Emotion*,*25*(4),1025－1043.

Chaminade,T.,Rosset,D.,Da Fonseca,D.,Nazarian,B.,Lutcher,E.,Cheng,G.,& Deruelle,C.(2012).How do we think machines think? An fMRI study of alleged competition with an artificial intelligence.*Frontiers in Human Neuroscience*,*6*,103－111.

Delgado,M. R.,Frank,R. H.,& Phelps,E. A.(2005).Perceptions of moral character modulate the neural systems of reward during the trust game.*Nature Neuroscience*,*8*(11),1611－1618.

de Melo,C. M.,Carnevale,P. J.,Read,S. J.,& Gratch,J.(2014).Reading people's minds from emotion expressions in interdependent decision making.*Journal of Personality and Social Psychology*,*106*(1),73－88.

de Melo,C. M.,Marsella,S.,& Gratch,J.(2016).People do not feel guilty about exploiting machines.*ACM Transactions on Computer－Human Interaction*,*23*(2),1－17.

Dietvorst,B. J.,& Bartels,D. M.(2022).Consumers object to algorithms making morally relevant tradeoffs because of algorithms' consequentialist decision strategies.*Journal of Consumer Psychology*,*32*(3),406－424.

Everett,J. A.,Faber,N. S.,Savulescu,J.,& Crockett,M. J.(2018).The costs of being consequentialist:Social inference from instrumental harm and impartial beneficence.*Journal of Experimental Social Psychology*,*79*,200－216.

Fareri,D. S.(2019).Neurobehavioral mechanisms supporting trust and reciprocity.*Frontiers in Human Neuroscience*,*13*,271.

Ishowo－Oloko,F.,Bonnefon,J. F.,Soroye,Z.,Crandall,J.,Rahwan,I.,& Rahwan,T.(2019).Behavioural evidence for a transparency－efficiency tradeoff in human－machine cooperation.*Nature Machine Intelligence*,*1*(11),517－521.

Karpus,J.,Krueger,A.,Verba,J. T.,Bahrami,B.,& Deroy,O.(2021).Algorithm exploitation:Humans are keen to exploit benevolent AI.*Science*,*24*(6),1－16.

Karpus,J.,Shirai,R.,Verba,J. T.,Schulte,R.,Weigert,M.,Bahrami,B.,...Deroy,O.(2025).Human cooperation with artificial agents varies across countries.*Scientific Reports*,*15*(1),10000.

Kozyreva,A.,Lewandowsky,S.,& Hertwig,R.(2020).Citizens versus the internet:Confronting digital challenges with cognitive tools.*Psychological Science in the Public Interest*,*21*(3),103－156.

Liu,P.,Chu,Y.,Zhai,S.,Zhang,T.,& Awad,E.(2025).Morality on the road:Should machine drivers be more utilitarian than human drivers?*Cognition*,*254*,106011.

Mahmud,H.,Islam,A. N.,Ahmed,S. I.,& Smolander,K.(2022).What influences algorithmic decision－making? A systematic literature review on algorithm aversion.*Technological Forecasting and Social Change*,*175*,121390.

Makovi,K.,Sargsyan,A.,Li,W.,Bonnefon,J. F.,& Rahwan,T.(2023).Trust within human－machine collectives depends on the perceived consensus about cooperative norms.*Nature Communications*,*14*,3108.

Mathur,A.,Acar,G.,Friedman,M. J.,Lucherini,E.,Mayer,J.,Chetty,M.,& Narayanan,A.(2019).Dark patterns at scale:Findings from a crawl of 11K shopping websites.*Proceedings of the ACM on Human－Computer Interaction*,*3*(CSCW),1－32.

Takahashi,Y.,Kayukawa,Y.,Terada,K.,& Inoue,H.(2021).Emotional expressions of real humanoid robots and their influence on human decision－making in a finite iterated prisoner's dilemma game.*International Journal of Social Robotics*,*13*(7),1777－1786.

Terada,K.,& Takeuchi,C.(2017).Emotional expression in simple line drawings of a robot's face leads to higher offers in the ultimatum game.*Frontiers in Psychology*,*8*,724－732.

Thiebes,S.,Lins,S.,& Sunyaev,A.(2021).Trustworthy artificial intelligence.*Electronic Markets*,*31*,447－464.

von Schenk,A.,Klockmann,V.,& Köbis,N.(2025).Social preferences toward humans and machines:A systematic exper-

iment on the role of machine payoffs. *Perspectives on Psychological Science*, *20*(1), 165 – 181.

Whiting, T., Gautam, A., Tye, J., Simmons, M., Henstrom, J., Oudah, M., & Crandall, J. W. (2021). Confronting barriers to human – robot cooperation: Balancing efficiency and risk in machine behavior. *Science*, *24*(1), 1 – 15.

The Effects of Anthropomorphized Emotion Expression and Ethical Design of Intelligent Machines on Human – Machine Cooperation

Gao Shanchuan Liu Huan

(Department of Psychology, Fudan University, Shanghai 200433)

Abstract: With the development of intelligent agents, it is of value to explore how the outer and inner design of intelligent machines would affect people's anticipation of cooperation and actual behavior. Focusing on the humanlike facial expression and inner ethical design of intelligent machines, the present study investigated the effects of both factors on human – machine interaction using two economic games. First, in the Trust game, it was found that participants had similar anticipation of cooperation about robot and human partner with the same expression, as indicated by happy or neutral emoji. However, they cooperated with robots significantly less than they did with humans. Even when they saw the robots with happy emoji had cooperated, only a few of them chose to reciprocate, whereas much more people reciprocated the kindness of humans with the same happy emoji. In another game, the Prisoner's Dilemma, participants in two groups were presented with different descriptions of the application of artificial intelligent technology in daily life and then asked to interact with robots. Results showed that the participants not only anticipated less cooperation from robots but also cooperated less when they found that the ethical design of intelligent machines followed utilitarian approach, i. e. maximizing business owner's profits instead of improving people's lives. Meanwhile, those who chose to deceive the robots were less likely to feel guilty. Thus, these findings suggest that, in order to promote human – machine cooperation, more efforts need to be put into the design of ethical intelligent machines.

Key words: cooperation; intelligent machine; emotion expression; ethical design; guilt