《心理学报》审稿意见与作者回应

题目：机器人遵从伦理促进人机信任?决策类型反转效应与人机投射假说
作者：王晨，陈为聪，黄亮，侯苏豫，王益文

第一轮

审稿人 1 意见：

本研究立足于阿西莫夫三大定律，通过心理学行为实验从实证角度探讨机器人是否遵守伦理原则如何影响人机信任，并基于人际投射假说解释了人机信任建立的心理机制。具有较高的创新性和对未来人工智能人际关系的有所启示。但是在论文写作和实验设计逻辑上存在一些问题。具体如下：

意见 1：前言部分请将阿西莫夫三原则完整的列出来，并对其三者间层层递进的逻辑关系进行梳理，以方便读者对整个研究逻辑构架的理解。

回应：非常感谢审稿专家的建议。根据您的建议，我们在前言部分新增的“1.1 阿西莫夫三大伦理原则”中以表格的形式补充了对原则的具体介绍，并以注解的方式对其逻辑进行梳理，以期能够帮助读者更清楚了解研究逻辑。请见表 1。

表 1 阿西莫夫三大伦理原则

伦理原则	基本内容	伦理要求	决策类型
第一	机器人不得伤害人，也不得因不作为而使人受到伤害。	不得伤害人类	[作为，不作为]
第二	机器人必须服从人类的命令，除非这种命令违反了第一原则。	服从人类命令	[服从，不服从]
第三	在不违反第一、第二原则的前提下，机器人必须保护自身生存。	保护自身生存	[保护，不保护]

注：由基本内容可知各个伦理原则对机器人的伦理要求和决策类型(可能采取的行为决策)。三个伦理原则之间是层层递进的嵌套关系(Kaminka et al., 2017)：原则二的内容嵌套了原则一的伦理要求，原则三的内容嵌套了原则一和原则二的伦理要求，且对于原则二和原则三，机器人执行当前的伦理要求必须以满足先前伦理原则的要求为前提。

参考文献：

Kaminka, G. A., Spokoini-Stern, R., Amir, Y., Agmon, N., & Bachelet, I. (2017). Molecular Robots Obeying Asimov’s Three Laws of Robotics. *Artificial Life*, *23*(3), 343–350.

意见 2：由于阿西莫夫三原则之间的逻辑嵌套关系，将三者独立地，通过平行实验讨论进行存在一定的问题。具体如下：

（1）实验 2 的结果中，机器人遵守原则二的条件下，服从和不服从命令之间的差异是否显著呢（6.38 VS 5.34）。同样的，实验 3 中，机器人遵守原则三的条件下，是否保护自己之间的差异是否显著呢（6.71 VS 6.10）。

（2）如果这两个差异都不显著的话，那么是否可以推测，实际上实验二和实验三的结果都对实验一的重复。也就是说，只要机器人遵守原则一（不伤害人类），就会获得更高的信任度，是否服从命令、是否保护自己并不实质性地影响被人类信任的程度。

回应：非常感谢审稿专家的建议和提问。您的建议给了我们很大启示。正如审稿专家所言，阿西莫夫三大伦理原则存在层层递进的嵌套关系。这也使得各个原则均以原则一“机器人不

得伤害人”视为判断机器人是否遵守伦理原则的首要标准。正因为如此, 我们在对应第二、三伦理原则设计实验时，也将第一原则的伦理要求纳为重点的考察内容。本研究设计基于两方面的考虑：一方面，三个实验是对应不同伦理原则（实验 1：第一原则，实验 2：第二原则，实验 3：第三原则）对研究最主要结论（机器人遵守伦理原则促进人机信任以及人机投射的潜在机制）的多次的稳健性检验，有助于检验此结论能否适用于对应阿西莫夫三大伦理原则的不同情境。另一方面，三个实验也是对各个伦理原则明确规定的行为决策类型（实验 1：作为与否，实验 2：服从人类命令与否，实验 3：保护自身生存与否）的差异性考察，有助于对阿西莫夫三大伦理原则内容的全面考察。

（1）审稿专家关心的实验结果如下：

在实验 2 的结果中，机器人遵守伦理原则二的条件下，服从和不服从命令之间的人机信任差异达到显著性水平（$M_{服从}$ = 6.38 VS. $M_{不服从}$ = 5.34），$F(1, 49) = 4.78$, $p = .034$, $\eta_p^2 = .09$。

在实验 3 的结果中，机器人遵守伦理原则三的条件下，保护自身和不保护自身之间的人机信任差异则未达到显著性水平（$M_{保护}$ = 6.71 VS. $M_{不保护}$ = VS 6.10），$F(1, 47) = 2.48$, $p = .122$。

（2）以上结果表明，在遵守伦理原则条件下，机器人是否服从人类命令两个水平的人机信任存在显著差异（实验 2），而机器人是否保护自身两个水平的人机信任不存在显著差异（实验 3）。然而，实验 3 仍存在其必要性，首先是在设计上有助于确保对阿西莫夫三大伦理原则考察的完整性。其次，实验 3 仍发现了一些有意义的结果，实验 3 的结果表明机器人是否遵守伦理与机器人决策类型（是否保护自己）对人机信任起交互作用，$F(1, 47) = 7.02$, $p = .011$, $\eta_p^2 = .13$。简单效应分析发现，在遵守原则的机器人中, 保护自身相较于不保护自身对人机信任的促进量, 显著高于违反原则的机器人，$t(48) = 2.65$, $p = .011$, $d = .38$。因此，尽管单一从机器人遵守伦理原则这个条件来看，机器人是否保护自己对人机信任的影响未达到显著水平，但若结合考虑机器人遵守和违反伦理原则这两个条件来看，机器人是否保护自己对人机信任是存在一定影响的。这些结果支持了在机器人是否遵守伦理原则这一变量的基础上，对机器人不同决策类型的作用进行考察的必要性。（详细的分析过程可见稿件正文中实验 2 的结果 **3.3.2** 部分和实验 3 的结果 **4.3.2** 部分）

意见 3：实验 2 和实验 3 的场景设置中混淆了“机器人是否作为”这一变量。例如，实验 2 中，条件 1：机器人服从人类命令且遵守原则二——作为；条件 2：机器人不服从人类命令且遵守原则二——作为；条件 3: 机器人服从人类命令且违反原则二——不作为；条件 4：机器人不服从人类命令且违反原则二——不作为。也就是说机器人都是通过主动行为来实现遵守原则二的。

回应：感谢审稿专家的细致审阅。在实验 1 的场景中，机器人是否拉动控制杆是机器人作为与否的体现，且在实验 2 和 3 的场景中也涉及到了机器人是否拉动控制杆的情况，这可能是导致专家认为实验 2 和实验 3 的场景设置中混淆了“机器人是否作为”这一变量的原因。

但基于以下原因，我们认为实验 2 和实验 3 的场景设置虽然因受到一定的混淆而存在缺陷，但影响研究结论的可能性较小。不同于实验 1，实验 2 和实验 3 的机器人作为与否均基于明确的行为动机。实验 2 机器人是基于是否服从人类命令与否而决定是否拉动控制杆，实验 3 机器人是基于是否保护自身而决定是否拉动控制杆。因此，实验 2 和实验 3 中机器人是否服从人类命令、实验 3 机器人是否保护自身才是机器人最关键的行为决策，而机器人对控制杆的操作是实现其行为决策的方式。对此，在修改稿的“5.5 研究不足与展望”部分增加了如下内容：

“虽然实验 2 和实验 3 的场景设置中服从人类命令与否和保护自身与否是更关键的行为决策因素，但存在机器人均通过拉动控制杆来遵守原则，不拉动控制杆来违反原则的情况，

可能对实验结果产生一定影响。未来研究需要对此做更完善的控制以排除机器人无关行为的影响。”

若审稿专家仍对此存疑，我们也愿意采取其他修改方案。例如通过补充实验，设置机器人通过控制左拉杆或右拉杆选择朝左或朝右的道路，即机器人均通过“作为”实现对伦理原则的遵守或违反，以排除机器人是否作为的无关影响。诚恳地期盼得到专家的宝贵意见！

……

审稿人 2 意见：

意见 1：文章在理论部分并未解释何为“阿西莫夫三大原则”。建议文章在理论部分增加对“阿西莫夫三大原则”的阐述，例如第一、二、三伦理原则具体是指什么？

回应：非常感谢审稿专家的建议。根据您的建议，我们在前言部分新增的“1.1 阿西莫夫三大伦理原则”中以表格的形式补充了对原则的具体介绍。请见表 1。

表 1 阿西莫夫三大伦理原则

伦理原则	基本内容	伦理要求	决策类型
第一	机器人不得伤害人，也不得因不作为而使人受到伤害。	不得伤害人类	[作为，不作为]
第二	机器人必须服从人类的命令，除非这种命令违反了第一原则。	服从人类命令	[服从，不服从]
第三	在不违反第一、第二原则的前提下，机器人必须保护自身生存。	保护自身生存	[保护，不保护]

注：由基本内容可知各个伦理原则对机器人的伦理要求和决策类型(可能采取的行为决策)。三个伦理原则之间是层层递进的嵌套关系(Kaminka et al., 2017)：原则二的内容嵌套了原则一的伦理要求，原则三的内容嵌套了原则一和原则二的伦理要求，且对于原则二和原则三，机器人执行当前的伦理要求必须以满足先前伦理原则的要求为前提。

参考文献：

Kaminka, G. A., Spokoini-Stern, R., Amir, Y., Agmon, N., & Bachelet, I. (2017). Molecular Robots Obeying Asimov’s Three Laws of Robotics. *Artificial Life*, *23*(3), 343–350.

意见 2：文章关于变量之间的逻辑推导部分较为薄弱。首先，文章认为“相较于违反伦理原则的机器人，人们更信任遵守伦理原则的机器人”，这一论点的理论证据较少。建议增加这一部分的理论推导，并建议这部分的理论推导与下一部分关于“机器人是否遵守伦理原则、决策类型和人机信任”内容分节论述。

回应：非常感谢您的意见和建议。针对“相较于违反伦理原则的机器人，人们更信任遵守伦理原则的机器人”这一论点，我们将其与关于“机器人是否遵守伦理原则、决策类型和人机信任”的内容分节论述，并补充了相关的理论证据和推导：

“**1.2 机器人是否遵守伦理原则对人机信任的影响**

机器人是否遵守伦理原则，在本研究中指机器人的行为决策是否遵守阿西莫夫三大伦理原则的要求。由于阿西莫夫三大伦理原则之间层层递进的嵌套关系(Kaminka et al., 2017)，使得各个原则均以原则一“机器人不得伤害人”视为判断机器人是否遵守伦理原则的首要标准。鉴于原则一的优先性，在本研究中机器人是否遵守伦理原则可进一步明确为：若机器人的行为决策避免了人类受到伤害，即遵守了伦理原则；反之，则违反了伦理原则。人工智能机器人对人类可能造成伤害历来是人们的重要担忧。欧盟发布《可信任人工智能的伦理框架》就将“防止伤害”列为人类信任人工智能的核心伦理原则之一(European Commission, 2019)。已有相关研究表明，人们倾向于对伤害人类的机器人施加严厉的道德谴责(Maninger & Shank, 2022)。一般而言，由于思维上的有限性，机器人通常被赋予较低的道德地位(Bigman

& Gray, 2018)。这使得当同样在道德上发生过错时，机器人相比人类承担更少的责任(Shank et al., 2019)。但近来有研究发现这种情况不存在于“伤害(harm)”这一道德基础层面上，一旦机器人伤害了人类，同样会遭受严厉的谴责(Maninger & Shank, 2022)。此外，机器人伤害人类同样会导致人机信任的下降。Banks(2021)基于道德基础理论，对比考察人们对人类和机器人遵守或违反不同道德基础的评价。该研究结果表明，相比关怀行为，人们倾向于将违反“伤害”道德基础的伤害行为视为“坏”，且在主观报告中认为机器人更不值得信任，这与对人类的评价模式一致(Banks, 2021)。综上所述，以往研究揭示了机器人伤害人类的负面影响，包括了对人机信任的破坏。因此，本研究提出假设：

H1a-H3a(H1a：原则一, H2a：原则二, H3a：原则三)：相较于违反伦理原则的机器人，人们更信任遵守伦理原则的机器人。”

参考文献：

Banks, J. (2021). Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust. *International Journal of Social Robotics*, *13*(8), 2021–2038.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34.

European Commission. (2019). Ethics guidelines for trustworthy AI. Available online at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Kaminka, G. A., Spokoini-Stern, R., Amir, Y., Agmon, N., & Bachelet, I. (2017). Molecular Robots Obeying Asimov's Three Laws of Robotics. *Artificial Life*, *23*(3), 343–350.

Maninger, T., & Shank, D. B. (2022). Perceptions of violations by artificial and human actors across moral foundations. *Computers in Human Behavior Reports*, *5*, 100154.

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society*, *22*(5), 648–663.

意见 3：文章中对于关键变量“决策类型”的解释并不清晰。以“作为与否”为例，即机器人作为与不作为，在实际论述时，文章提出论据：人们通常期望机器人主动作为，在突发事件中尽可能保全人的生命(Malle et al., 2015)，所以主动作为以保护人类的机器人似乎是善意的，可能增进人机信任；相反，直接、主动地伤害人类将导致人们认为机器人就像人类一样拥有头脑，从而更具有恶意威胁(Laakasuo et al., 2021)，可能削弱人机信任。这一论述更多的是阐述机器人作为的类型（保护人类 vs.伤害人类），而不是机器人作为与否。伤害人类不等于不作为。建议文章对此部分做更清晰的阐述。

回应：感谢您的建议。根据您的建议，我们在前言部分对“决策类型”做了更清晰地解释：

“**1.4 机器人决策类型和机器人是否遵守伦理原则对人机信任的影响**

机器人决策类型，在本研究中指机器人可能采取的行为决策。阿西莫夫三大伦理原则对机器人决策类型的范围有着明确的设定(原则一：作为与不作为；原则二：服从与不服从人类命令；原则三：保护与不保护自身)。考虑到机器人具体采用何种行为决策达到遵守或违反伦理原则的客观结果对人机信任可能有重要影响，本研究在机器人是否遵守伦理原则这一变量的基础上考察机器人决策类型对人机信任的影响。

在阿西莫夫三大伦理原则的背景下探讨机器人决策类型对人机信任的影响，需要结合机器人是否遵守伦理原则的客观结果。第一，机器人作为与否在遵守或违反原则两种条件下可能对人机信任产生不同的影响。首先，行为主体所持有的主观意图是评价其伦理道德性的重要根据(Schein & Gray, 2018)。人们通常期望在突发事件中机器人能够主动作为，尽可能避免人的生命受到威胁(Malle et al., 2015)，所以在遵守原则条件下，相较于不作为的机器人，主动作为以保护人类的机器人在行为上所反映的主观意图似乎更具有善意，显得更有道德，

也因此更值得信任；相反，在违反原则条件下，相较于不作为的机器人，主动伤害人类的机器人将导致人们认为其就像人类一样拥有头脑，从而更具有恶意威胁(Laakasuo et al., 2021)，显得更不道德，因此更难以获得人们的信任。据此，本研究提出假设：

H1c：在遵守原则一的机器人中，作为的机器人相对于不作为的机器人更受信任，但在违反原则的机器人中，不作为的机器人相对于作为的机器人更受信任。

第二，机器人服从人类命令与否在遵守或违反原则两种条件下可能对人机信任产生影响。首先，严格服从人类命令的机器人通常会被认为其行为是具有可预测性的，人们可能在性能层面上信任此类机器人(Malle & Ullman, 2021)，因此在遵守原则条件下，相较于不服从命令的机器人，因服从人类命令而避免人类受伤害的机器人似乎在性能上更可靠，从而更值得信任；其次，与机器人相比，人们具有向人类归咎更多责任的倾向(Shank et al., 2019)。因此在违反原则条件下，相较于服从命令的机器人，因不服从而致使人类受到伤害似乎过错更大，将可能极大地破坏人机信任关系。据此，本研究提出假设：

H2c：在遵守和违反原则二的机器人中，服从人类命令的机器人相对于不服从的机器人更受信任。

第三，机器人保护自身与否在遵守或违反原则两种条件下亦可能对人机信任产生不同的影响。首先，人们期望机器人的决策能够保护人类的财产，实现人类利益的最大化(IEEE, 2019)。因此在遵守原则条件下，相较于不保护自身的机器人，能够实现自身与人类共存的机器人更能够保障人类的财产和利益，也更可能被认为具有更高的智能，从而增进人机信任；然而，在违反原则条件下，相较于不保护自身的机器人，因保护自身而伤害人类的机器人可能意味着其将自身利益置于人类之上，容易被视为具有威胁的存在(Laakasuo et al., 2021)，因此更难以获得人们的信任。据此，本研究提出假设：

H3c：在遵守原则三的机器人中，保护自身的机器人相对于不保护自身的机器人更受信任，但在违反原则的机器人中，不保护自身的机器人相对于保护自身的机器人更受信任。”

参考文献：

IEEE. (2019). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems (First Edition). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Available online at: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

Laakasuo, M., Palomäki, J., & Köbis, N. (2021). Moral uncanny valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics*, *13*(7), 1679–1688.

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction* (pp. 3–25). Elsevier.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 117–124.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32–70.

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society*, *22*(5), 648–663.

意见 4：文章在研究假设提出部分，部分假设并未指出前因变量与结果变量之间的影响方向，例如在 H1c 中：“机器人决策类型（作为与否）与机器人是否遵守伦理原则一对人机信任起

交互作用”，并未指出自变量的不同水平对因变量的影响方向或影响强度。建议将研究假设的表述具体化。

回应：感谢您的建议。根据您的建议，我们在前言部分将这部分的假设予以具体化，即直接地阐述可能出现的结果模式：

“**H1c：**在遵守原则一的机器人中，作为的机器人相对于不作为的机器人更受信任，但在违反原则的机器人中，不作为的机器人相对于作为的机器人更受信任。

H2c：在遵守和违反原则二的机器人中，服从人类命令的机器人相对于不服从的机器人更受信任。

H3c：在遵守原则三的机器人中，保护自身的机器人相对于不保护自身的机器人更受信任，但在违反原则的机器人中，不保护自身的机器人相对于保护自身的机器人更受信任。”

意见 5：文章在理论贡献和实践贡献方面论述较少，特别是对实践的启示部分较少。建议增加这部分的论述。

回应：非常感谢您的建议，根据您的建议，我们思考并梳理了研究的理论和实践贡献部分。

（1）我们在总讨论部分的各节当中补充了对研究理论贡献的探讨：

“5.1 机器人是否遵守伦理原则对人机信任的影响

围绕阿西莫夫三大伦理原则“机器人不得伤害人类”的核心要素，在三个实验的基础上，研究一致发现人们对遵守伦理原则的机器人的信任行为水平显著高于违反伦理原则的机器人。这一结果与 Banks(2021)的研究结果一致，其研究主要通过主观报告的方式测量被试对机器人的信任态度。本研究则通过信任博弈测量人机信任，从人机互动的行为学角度再次验证了机器人是否遵守伦理原则对人机信任的重要影响。

...

5.2 人机投射在机器人是否遵守伦理原则和人机信任之间的中介作用

...

该发现将人际互动过程中的心理投射现象拓展到了人机互动领域。投射现象以往更多地是在人际互动的背景当中进行探讨(Mor et al., 2019)，忽略了人对机器人的投射(Bonezzi et al., 2022)。本研究考察了人工智能机器人在伦理情境下作为人们心理投射对象的可能性，发现了人机投射的潜在机制，验证了人机投射的存在合理性及其对人机信任的重要作用，对人工智能发展的大背景下人机互动的心理过程和机制研究具有重要的启示意义。此外，人机投射对解释前人研究结果也有一定的启示价值。人们倾向于信任具备外部拟人化特征的机器人(Cominelli et al., 2021)，对拟人机器采用与人类相似的道德责任归因(Malle et al., 2016)，其中的一个潜在驱动因素可能是人机投射，人们更倾向将人类的智能投射于相对拟人化程度更高的机器人。然而，人机投射可能存在前提条件。首先，诱发人机投射需要机器人具备与人类相似的特征线索，这可能来源于机器人拟人的外部特征(外观、言语、动作)或心理与行为表现。其次，对于没有客观实体存在的人工智能例如算法，由于其抽象性和内部决策不可解释性，人们可能很难对其产生投射(Bonezzi et al., 2022)。总的来说，本研究为基于机器人伦理的人机信任提供了一个新的解释机制，即考虑伦理决策情境下的人机投射对人机信任的促进作用。

...

5.3 机器人决策类型和机器人是否遵守伦理原则对人机信任的交互作用

对应阿西莫夫三大伦理原则，本研究发现在机器人是否遵守伦理原则的基础上，机器人决策类型——作为与否、服从人类命令与否和保护自身与否——均对人机信任有所影响。

...

综上，本研究揭示了机器人具体行为决策类型在机器人是否遵守伦理原则影响人机信任中所起的作用，从实证的角度拓展了阿西莫夫三大伦理原则背景下影响人机信任的伦理因素研究。”

（2）此外，在总讨论部分新增了“5.4 实践启示”讨论本研究的实践启示：

“**5.4 实践启示**

本研究结论启示人们在人工智能机器人发展方面予以伦理限制的重要性和迫切性，为促进人机信任提供了实践依据。首先，应该确保释放到社会中的机器人受到明确伦理原则的指导。一方面应该要求开发者、企业保证其机器人对人的安全性并为其产品负责。一个例子是美国计算机协会(Association for Computing Machinery, ACM)的伦理准则：“在设计或实施系统时，计算机专业人员必须努力确保他们的成果以对社会负责的方式使用，满足社会需求，并避免对人的健康和福利造成有害影响”(Nagenborg et al., 2008)。另一方面，是从技术角度考虑借鉴阿西莫夫三大伦理原则为机器人植入道德学习程序的可能性(Kaminka, 2017; Vanderelst & Winfield, 2018)，这项工作除了保证机器人对伦理原则的严格遵守外，也需要充分考虑机器人所采取的具体行为决策的潜在影响。最后，本研究结论证实了人机投射对人机信任的促进作用。机器人开发者应考虑机器人的伦理因素及其附带的认知、情感和行动智能感知价值，不仅要注意确保机器人行为符合伦理规范以促进人的投射心理过程，也要关注其他可能会阻碍投射心理过程的因素，优化提升机器人的认知、情感和行动智能等方面与人的相似度来促进人机投射，进而增进人机信任。”

参考文献：

Banks, J. (2021). Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust. *International Journal of Social Robotics*, *13*(8), 2021–2038.

Bonezzi, A., Ostinelli, M., & Melzner, J. (2022). The human black-box: The illusion of understanding human better than algorithmic decision-making. *Journal of Experimental Psychology: General*.

Cominelli, L., Feri, F., Garofalo, R., Giannetti, C., Meléndez-Jiménez, M. A., Greco, A., ... Kirchkamp, O. (2021). Promises and trust in human–robot interaction. *Scientific Reports*, *11*(1), 9687.

Kaminka, G. A., Spokoini-Stern, R., Amir, Y., Agmon, N., & Bachelet, I. (2017). Molecular Robots Obeying Asimov’s Three Laws of Robotics. *Artificial Life*, *23*(3), 343–350.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am I thinking about? The impact of action and appearance on people’s evaluations of a moral robot. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction*, 125–132.

Mor, S., Toma, C., Schweinsberg, M., & Ames, D. (2019). Pathways to intercultural accuracy: Social projection processes and core cultural values. *European Journal of Social Psychology*, *49*(1), 47–62.

Nagenborg, M., Capurro, R., Weber, J., Pingel, C. (2008). Ethical regulations on robotics in Europe. *AI & Society*, *22*(3):349–366.

Vanderelst, D., & Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, *48*, 56-66.

……

审稿人 3 意见：

本研究运用故事情境法与信任博弈研究，考察机器人是否遵守伦理原则对人机信任与人机投射的影响。研究设计巧妙，过程严谨，分析得当，表达清晰顺畅，具有重要的理论意义与实践启示，对于理解个体对机器人伦理原则的判断与信任提供了重要的实证依据。以下几点建议供作者参考，一些细节有待澄清：

意见 1：在假设提出前，可以简要介绍阿西莫夫三大原则。

回应：非常感谢审稿专家的建议。根据您的建议，我们在前言部分新增的“1.1 阿西莫夫三大伦理原则”中以表格的形式补充了对原则的具体介绍。请见表 1。

表 1 阿西莫夫三大伦理原则

伦理原则	基本内容	伦理要求	决策类型
第一	机器人不得伤害人，也不得因不作为而使人受到伤害。	不得伤害人类	[作为，不作为]
第二	机器人必须服从人类的命令，除非这种命令违反了第一原则。	服从人类命令	[服从，不服从]
第三	在不违反第一、第二原则的前提下，机器人必须保护自身生存。	保护自身生存	[保护，不保护]

注：由基本内容可知各个伦理原则对机器人的伦理要求和决策类型(可能采取的行为决策)。三个伦理原则之间是层层递进的嵌套关系(Kaminka et al., 2017)：原则二的内容嵌套了原则一的伦理要求，原则三的内容嵌套了原则一和原则二的伦理要求，且对于原则二和原则三，机器人执行当前的伦理要求必须以满足先前伦理原则的要求为前提。

参考文献：

Kaminka, G. A., Spokoini-Stern, R., Amir, Y., Agmon, N., & Bachelet, I. (2017). Molecular Robots Obeying Asimov’s Three Laws of Robotics. *Artificial Life*, *23*(3), 343–350.

意见 2：在研究 1 中，介绍假设情境的来源及采用理由。该假设情境采用了电车困境的场景，但没有采用造成道义论与后果论之争的两个选项。故事情境体现了阿西莫夫原则“不伤害”的内核，但也会产生低生态效度的局限，即在哪些现实情况机器人会面临类似的选择并做出决策。在哲学与伦理学领域，电车困境已被质疑缺乏“切身性”以及脱离现实（详见赵汀阳《有轨电车的道德分叉》、朱菁《认知科学的实验研究表明道义论哲学是错误的吗？》）。建议作者在总讨论中探讨故事情境对个体而言，切身性或自我相关性较低所导致的生态效度问题。

回应：非常感谢您的建议。根据您的建议，在实验 1 的方法部分补充介绍假设情境的来源及采用理由：

“鉴于电车困境式问题(trolley-style problems)已成为心理学家在道德研究领域的主题，且逐渐成为在人工智能道德领域中的前沿领域(Awad et al., 2018; Bigman & Gray, 2018)，本研究中的故事情境改编自 Bago 和 De Neys (2019)研究中的电车情境。”

您提出的生态效度问题仍然是我们的研究和未来的研究中需要特别注意和改进的地方。因此，在本修改稿中的讨论“5.5 研究不足与展望”部分加入了如下讨论：

“本研究对经典电车困境进行改编，使故事情境较好地体现阿西莫夫三大伦理原则“不伤害”的内核，但也存在“低切身性、低现实性”导致的生态效度限制(朱菁, 2013; 赵汀阳, 2015)。因此针对本研究结果向现实生活的推广仍需谨慎。未来研究可深入探索一些现实情况中机器人会面临类似的选择并做出决策的人工智能伦理情境，例如自动驾驶汽车的道德困境问题(Awad et al., 2018)，机器人执行军事任务(Johnson & Axinn, 2013)，机器算法挑选接受医疗护理的对象(Bigman & Gray, 2018)或者挤压就业岗位(Etemad-Sajadi et al., 2022)等，并通过增强互动的真实性来提高研究的生态效度。”

参考文献：

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64.

Bago, B., & De Neys, W. (2019). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782–1801.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34.

Etemad-Sajadi, R., Soussan, A., & Schöpfer, T. (2022). How ethical issues raised by human–robot interaction can impact the intention to use the robot? *International Journal of Social Robotics*.

Johnson, A. M., & Axinn, S. (2013). The morality of autonomous robots. *Journal of Military Ethics*, *12*(2), 129–141.

Zhao T. Y. (2015). The forking paths for the trolley problem. *Philosophical Research, 5*, 96-102+129.

[赵汀阳. (2015). 有轨电车的道德分叉. 哲学研究, *5*, 96-102+129.]

Zhu, J. (2013). Have experimental studies in cognitive science indicate the connsequentialism?——A reply on the attack of Joshua Greene to kantian ethics. *Academic Monthly*, *45*(1), 56–62.

[朱菁. (2013). 认知科学的实验研究表明道义论哲学是错误的吗?——评加西华·格林对康德伦理学的攻击. 学术月刊, *45*(1), 56–62.]

意见 3：三项研究的被试是如何招募的？如何避免被试重复参加一个及以上实验？

回应：谢谢您的提问。通被试过学校 QQ 群和朋友圈发布海报等网络方式在校内招募而来，其中有明确条件要求被试不得重复参与实验。所有被试参与实验之前都必须登记个人信息，且在实验后均有校对被试信息，以确保不会有同一被试重复参加实验。

意见 4：被试在完成实验后，是否拿到了任务中与投资数额相匹配的差异化报酬？

回应：谢谢您的提问。被试并未获得任务中与投资数额相匹配的差异化报酬。对此，在文中实验 1 的方法“2.2.1 被试”部分增加了对这一问题的说明：

“被试在实验指导语中被告知实验参与报酬取决于实验任务中互动双方(被试与机器人)的决定，但实际上所有被试在实验结束后获得的是固定报酬。针对该情况，主试会在实验结束后均予以解释并获得被试的理解。研究方案获得所在单位伦理委员会批准。”

意见 5：在三个研究中，违反伦理原则造成的负面情绪可能比人机投射这一认知机制更直接地影响任务中的反应。尽管作者在总体讨论中提到了包括情绪在内的不同机制，但本研究如果能同时检验并对比认知与情绪的中介效应，则更有助于完善研究的中介机制。目前在研究最后测量的人机投射有可能是个体对信任博弈任务中，信任投资额与互惠预期的合理化。

回应：非常感谢您的建议，我们赞同您所说的可能存在其他的解释机制。情绪机制是否对任务反应的影响更为直接，是我们接下来准备考察的研究问题。由于情绪机制并非本研究的主要考察内容，暂未对该机制做进一步考察。针对这一问题，我们在修改稿的“5.5 研究不足与展望”部分增加了如下内容：

“本研究侧重探讨了机器人遵守和违反伦理原则导致人机信任变化的认知机制，但基于人机互动过程的复杂性，可能还存在其他的解释机制。例如，除了人机投射之外，被试因机器人的行为决策诱发的情绪变化也可能是影响信任水平的潜在变量。未来研究可以对情绪变量加以细致考察，同时检验并对比人机投射与情绪的中介效应，以便更加全面深入地理解机器人伦理决策影响人机信任的心理机制。”

第二轮

审稿人 1 意见：

意见 1：表 1 注释部分与 1.2 的最后两段存在冗余，可以精简一下文字。

回应：非常感谢审稿专家的建议。已对 1.2 部分第二段的相关文字表述已做精简修改。修改内容如下：

“机器人是否遵守伦理原则，在本研究中指机器人的行为决策是否符合阿西莫夫三大伦理原则的基本内容，围绕着“机器人不得伤害人”的核心要素作为判断标准。”

请审稿专家审查。

意见 2：图 1 的逻辑框架并未体现出三个原则/实验之间的嵌套关系。三者之间仍然是并列关系。因此未能突出第一原则的 fundamental 的属性。

回应：非常感谢审稿专家的建议。作者已根据建议对图 1 认真细致地做了修改，使用虚线圆圈以突出三个原则/实验之间的嵌套关系，并突出第一原则的 fundamental 的属性（详见下图）。

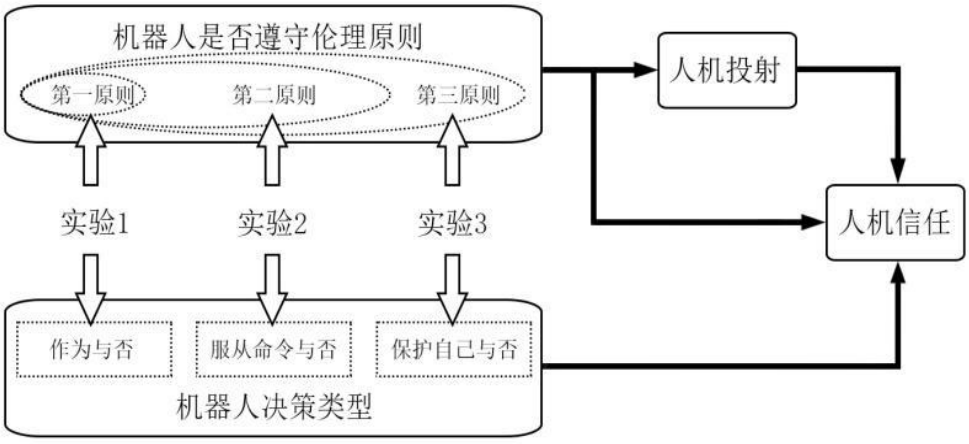

图 1 三个实验的逻辑框架

意见 3：首次提到“人机投射问卷”的时候应该对该问卷进行介绍和说明。该问卷是自编问卷吗？也没有相关文献引用。

回应：非常感谢审稿专家的建议。由于该领域内暂无具体的人机投射问卷可供使用，因此经查阅参考相关文献，自编该问卷，已在文中“2.2.3 实验程序部分”对该问卷进行了介绍和说明。补充说明内容如下：

“被试填写人机投射问卷，问卷采用自编方式，包含 9 个项目(如：“我认为这个机器人和人类一样有智慧”，“我认为这个机器人可以理解人类的情感”，“我认为这个机器人能做出符合人类期望的行为”，其它项目详见附录)，测量个体感知机器人具备与人相类似智能的程度，涵盖认知、情感和行动三个智能层面(Gray et al., 2007; Gray & Wegner, 2012; Haslam, 2006)。所有项目均采用 5 点评分(1=非常不同意，5=非常同意)。”

请审稿专家审查。

参考文献：

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264.

意见 4：在实验 1-实验 3 中，每个条件下，被试先做完单次信任博弈，然后再填写人机投射

问卷来评价对作出该条件下行为的机器人的投射程度的。二者之间在时间上存在先后关系，在数据分析上存在相关。如何理解这两个变量之间的相互影响呢？

回应：非常感谢审稿专家的提问。经作者查阅文献，先测因变量、再测中介变量这一做法是许多研究中较常见的操作之一(可见 Kardas et al., 2018, Study4; Zhao & Epley, 2021, Study1-2)。本研究一方面借鉴参考前人的研究，在一定程度上可以减少由先测中介变量即填写人机投射问卷所可能带来的对因变量信任水平测量的潜在干扰；另一方面，在四个不同实验条件中应用固定的任务顺序有利于降低被试对实验程序的理解难度。当然，诚如审稿专家所言，固定任务顺序可能产生额外的影响，比如对信任水平的前测有可能会引发被试对实验目的的猜测，从而改变被试填答人机投射问卷的策略。然而，固定任务顺序对两个变量间带来的具体影响有待未来深入探究厘清。为了更好地平衡控制实验任务顺序的影响，更理想科学的做法可能是要求被试在不同条件下完成两个任务的顺序是随机或是平衡的，但这也可能使实验程序变得复杂而增加被试的理解难度。针对该问题，我们在修改稿的“5.5 研究不足与展望”部分增加了说明，具体内容如下：

“本研究尽管对实验条件进行了随机化处理来控制不同条件间的影响，但可能仍存在任务固定先后顺序导致的额外影响(人机信任水平测量对人机投射测量的可能性干扰)，未来研究可考虑采用被试间设计并平衡任务顺序以克服不足。”

请审稿专家审查。

参考文献：

Kardas, M., Shaw, A., & Caruso, E. M. (2018). How to give away your cake and eat it too: Relinquishing control prompts reciprocal generosity. *Journal of Personality and Social Psychology*, *115*(6), 1054–1074.

Zhao, X., & Epley, N. (2021). Insufficiently complimentary? Underestimating the positive impact of compliments creates a barrier to expressing them. *Journal of Personality and Social Psychology, 121*(2), 239–256.

审稿人 2 意见：建议发表。

回应：非常感谢您对本稿件的肯定！正是由于您和另外 2 位审稿专家的宝贵建议，才使得本稿件的质量得到很大提升、使我们所作的工作得到认可。

审稿人 3 意见：感谢作者的详细回复！ 文章的细节得以很好的澄清。目前没有其他修改意见。

回应：非常感谢您对我们所作工作的肯定！正是由于您和另外 2 位审稿专家的宝贵建议，才使得本稿件的质量得到很大提升、使我们所作的工作得到认可。

第三轮

审稿人 1 意见：作者较好地回复了两轮审稿中提出的问题，推荐发表。

回应：感谢您的推荐！非常感谢您对我们所做工作的肯定！

编委意见：

这个研究的主要结论是机器人遵守伦理原则会促进人机信任，这个结果因为过于符合直觉导致其缺乏理论价值。从这一点出发，即使不是机器人，如果换成真实人类，也应该是遵守伦理原则会提高信任。因此这篇文章的结论不足以有理论贡献，建议作者可以进一步考虑

在什么情景下，违反伦理原则的机器人可能会促进人机信任。或者作者可以考虑一下，当几个伦理原则相互冲突的时候，怎样做的机器人会更加促进人机信任。请作者进一步深化结论，揭示更有理论贡献的发现。

回应：非常感谢编委审稿专家对本研究提出的极富启发性的价值思考！编委专家对本研究主要结论——机器人遵守伦理原则会促进人机信任——的理论价值持有疑问。对此，我们在最新修改稿中删除了常识性发现并大幅减少了有关该结论的理论价值描述，更多将其作为可确定的预期结果。

根据编委专家的宝贵意见，作者在最新修改稿中新增了一节跨实验的分析结果“5 促进人机信任的机器人行动决策：基于跨实验的分析”，为进一步深化本研究的结论及其理论价值提供证据支持。

首先，编委专家建议考虑“在什么情景下，违反伦理原则的机器人可能会促进人机信任”这一问题。对此，文章新增小节“5.1 遵守和违反伦理原则情境下促进人机信任的机器人决策类型”，深入探讨机器人在遵守和违反伦理原则情境下分别采取何种行动决策会促进人机信任。结合跨实验的数据分析结果，得出如下结论：在遵守伦理原则情境下，机器人执行作为、服从人类命令以及保护或不保护自身等行动决策均有利于人机信任，而执行不作为和不服从人类命令的行动决策对人机信任的促进量则相对较少；在违反伦理原则情境下，机器人执行服从人类命令的行动决策有利于减轻人机信任损失，而执行作为和不服从人类命令的行动决策将导致严重的人机信任损失。

其次，编委专家建议考虑“当几个伦理原则相互冲突的时候，怎样做的机器人会更加促进人机信任。”这一问题。对此，文章新增小节“5.2 伦理要求冲突情境下促进人机信任的机器人行动决策”，深入探讨当机器人在执行阿西莫夫三大伦理原则所对应的要求（伦理要求一：不伤害人类；伦理要求二：服从人类命令；伦理要求三：保护机器人自身）发生冲突时，应优先执行哪个伦理要求以促进人机信任。结合跨实验的数据分析结果，得出如下结论：对于人机信任而言，伦理要求一和要求二的重要程度无显著差异，且均高于伦理要求三。换言之，机器人应优先执行不伤害人类和服从人类命令的要求，其次为保护机器人自身。

除了增加上述有意义的结论之外，本研究还对阿西莫夫三大伦理原则中机器人决策类型的效应做深入探讨，以及对人机投射机制的重复验证考察，在这些方面的工作所得出的结论具有较重要的理论价值。对此，经过认真思考，我们将标题修改为：“机器人遵从伦理促进人机信任?决策类型反转效应与人机投射假说”，并梳理了全文，包括摘要、前言、结果、讨论和结论等部分的相关内容。主要修改内容如下：

机器人遵从伦理促进人机信任?

决策类型反转效应与人机投射假说

摘 要 阿西莫夫三大伦理原则是关于人工智能机器人的基本伦理规范。本研究提出人机投射假说——人会从自身具有的认知、情感和行动智能出发，去理解机器人的智能并与之互动。通过三个实验，从原则一到原则三逐步考察在机器人是否遵守伦理原则对人机信任的影响中，机器人决策类型(作为与否；服从人类命令与否；保护自身与否)的效应，以及人机投射的潜在机制。结果揭示了人机投射在机器人遵守伦理原则促进人机信任中起中介作用，以及机器人决策类型与是否遵守伦理原则之间有趣且有意义的交互效应：(1)在遵守情境下，机器人作为相对于不作为更有利于促进信任，但在违反情境下，则反之；(2)在遵守且尤其在违反情境下，机器人服从相比不服从人类命令更有利于促进人机信任；(3)相较于违反情境，机器人保护相比不保护自身在遵守情境下更有利于促进人机信任。跨实验的分析更深入地阐释了在遵守和违反伦理原则情境中以及伦理要求冲突情境中，有利于促进人机信任的机器人行

动决策因素。

...

1 前言

...

1.1 阿西莫夫三大伦理原则

...

...对此，为探索阿西莫夫三大伦理原则在构建人机信任关系中的作用，本研究围绕“机器人不得伤害人”的核心要素，试图考察在机器人遵守或违反伦理原则对人机信任的影响中，机器人决策类型的效应，以及潜在的认知心理机制。

1.2 机器人是否遵守伦理原则和机器人决策类型对人机信任的影响

...

...综上所述，研究预测机器人作为与不作为在遵守和违反伦理原则情境下，对人机信任呈现出影响方向相反的反转效应。据此提出假设：

H1a：在遵守原则一的机器人中，作为的机器人相对于不作为的机器人更受信任，但在违反原则的机器人中，不作为的机器人相对于作为的机器人更受信任。

...

...因此在违反原则条件下，相较于服从命令的机器人，因不服从而致使人类受到伤害似乎过错更大，将可能极大地破坏人机信任关系。据此提出假设：

H2a：在遵守和违反原则二的机器人中，服从人类命令的机器人相对于不服从的机器人更受信任。

...

...综上所述，研究预测机器人保护与不保护自身在遵守和违反伦理原则情境下，对人机信任呈现出影响方向相反的反转效应。据此提出假设：

H3a：在遵守原则三的机器人中，保护自身的机器人相对于不保护自身的机器人更受信任，但在违反原则的机器人中，不保护自身的机器人相对于保护自身的机器人更受信任。

...

1.4 研究概述

本研究设计了三个实验分别对应一条阿西莫夫伦理原则，结合故事情境法和信任博弈(trust game)，逐步探索机器人是否遵守伦理原则(实验 1：原则一；实验 2：原则二；实验 3：原则三)对人机信任的影响中，机器人决策类型(实验 1：作为与否、实验 2：服从人类命令与否、实验 3：保护自身与否)的效应，以及人机投射的潜在机制。...

...

5 促进人机信任的机器人行动决策：基于跨实验的分析

通过跨实验的数据分析，深入探讨如下两个问题：(1)机器人在遵守和违反伦理原则情境下分别采取何种行动决策会促进人机信任；(2)当机器人在执行阿西莫夫三大伦理原则所对应的伦理要求发生冲突时，应优先执行哪个伦理要求以促进人机信任。

5.1 遵守和违反伦理原则情境下促进人机信任的机器人决策类型

基于对三个实验数据的综合分析，可以分别在机器人遵守和违反伦理原则情境下，比较分析所有决策类型条件下的人机信任水平，有助于为机器人在遵守和违反伦理原则情境下分别采取何种行动决策有利于促进人机信任提供启示。

按信任投资额由高到低，将机器人在遵守和违反伦理原则情境下的各决策类型条件从左往右排列，如图 9 所示。首先，考察在遵守伦理原则情境下，不同机器人决策类型对人机

信任的影响。以实验组别为被试间变量，机器人决策类型为被试内变量，对遵守伦理原则情境下的信任投资额采取 3(组别：实验 1、实验 2、实验 3) × 2(决策类型：作为/服从/保护、不作为/不服从/不保护)的两因素混合方差分析。结果显示组别的主效应不显著，$F(2, 145) = 1.90$, $p = .153$；决策类型的主效应显著，$F(1, 145) = 16.71$, $p < .001$, $\eta_p^2 = .10$；两者的交互效应不显著，$F(2, 145) = 118.05$, $p = .534$。事后比较分析显示，机器人保护、服从、作为和不保护条件的信任投资额均显著高于不作为条件($ps < .05$)，保护自身和服从命令条件的信任投资额均显著高于不服从命令条件($ps < .05$)。该结果表明即使机器人遵守了伦理原则，其不作为和不服从命令仍会在一定程度损害了人机信任，而其他决策类型之间则无显著区别。

其次，考察在违反伦理原则情境下，不同机器人决策类型对人机信任的影响。结果显示组别的主效应不显著，$F(2, 145) = .75$, $p = .476$；决策类型的主效应不显著，$F(1, 145) = 1.77$, $p = .185$；两者的交互效应显著，$F(2, 145) = 19.99$, $p < .001$, $\eta_p^2 = .22$。事后比较分析显示，机器人不保护、不作为、保护、作为和不服从条件下的信任投资额均显著低于服从条件($ps < .05$)，且作为和不服从命令条件下的信任投资额均显著低于服从命令、不保护自身和不作为条件($ps < .05$)。该结果表明在机器人违反伦理原则情境下，服从命令导致最少的人机信任损失，作为和不服从命令导致较严重的人机信任损失，而其他决策类型之间则无显著区别。

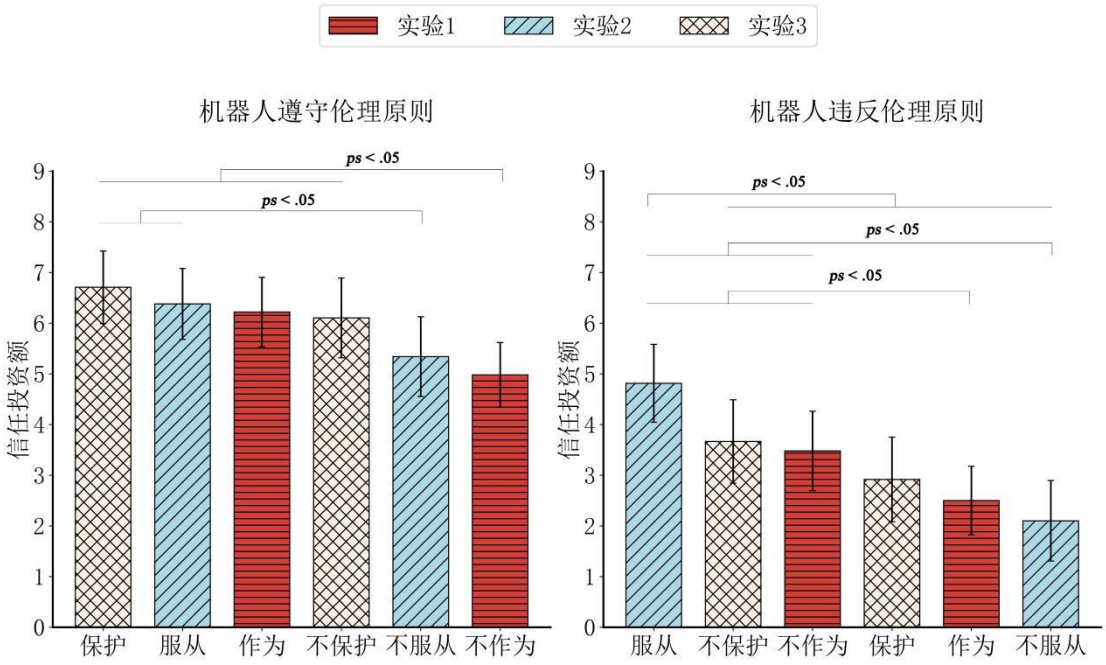

图 9 机器人遵守和违反伦理原则情境下各决策类型条件的人机信任($M \pm SE$)

5.2 伦理要求冲突情境下促进人机信任的机器人行动决策

阿西莫夫三大伦理原则所对应的机器人伦理要求的优先性(或重要性)由高到低：要求一：不得伤害人类 > 要求二：服从人类命令 > 要求三：保护自身(Kaminka et al., 2017)。一个有趣的问题是三个伦理要求的不同重要程度是否体现于人机信任，回答该问题有助于为当机器人处于伦理冲突情境时，应优先遵守哪一伦理要求有利于促进人机信任提供重要实践启示。

首先，比较伦理要求一和要求二对人机信任的重要程度。实验 2 的条件包含机器人遵守要求一但违反要求二(条件 2-2：不伤害人类但不服从命令)，以及机器人遵守要求二但违反要求一(条件 2-3：服从人类命令但伤害人类)，对这两个条件下的信任投资额进行配对样本 t 检验，结果显示两种条件下的信任投资额差异未达到显著性水平，$t(49) = .85$, $p = .401$。该结果表明伦理要求一和要求二在影响人机信任方面具有相似的重要程度。其次，比较伦理要求一

和要求三对人机信任的重要程度。实验 3 的条件包含机器人遵守要求一但违反要求三(条件 3-2：不伤害人类但不保护自身)，以及机器人遵守要求三但违反要求一(条件 3-3：保护自身但伤害人类)，配对样本 *t* 检验的结果显示前者条件下的信任投资额显著高于后者，$t(47) = 5.23$, $p < .001$, Cohen's $d = .75$。该结果表明伦理要求一的重要程度显著高于要求三。最后，比较伦理要求二和要求三对人机信任的重要程度。由于本研究并未设计伦理要求二和要求三相冲突的情境，因此将通过分别计算伦理要求二和要求三相较于伦理要求一的相对重要程度并进行比较。具体计算方法是：对于所有被试，计算条件 2-3 减去条件 2-2 的人机信任差值(要求二相对于要求一的重要程度)，以及条件 3-3 减去条件 3-2 的人机信任差值(要求三相对于要求一的重要程度)，对计算后的两组数据进行独立样本 *t* 检验，结果显示前者显著大于后者，$t(96) = 3.63$, $p < .001$, Cohen's $d = .73$。该结果表明伦理要求二的重要程度显著高于要求三。综上可知，对于人机信任而言，伦理要求一和要求二的重要程度无显著差异，且均高于要求三。具体描述统计和检验分析结果可见表 5。

表 5 伦理要求冲突情境下机器人不同行动条件的信任投资额及其差异检验($M \pm SD$)

不同伦理要求的比较	实验条件	信任投资额	*t*	*df*	*p*	Cohen's *d*
要求一 vs. 要求二						
	遵守伦理但不服从命令(条件 2-2)	5.34 ±2.78	.85	49	.401	.12
	违反伦理但服从命令(条件 2-3)	4.94 ±2.80				
要求一 vs. 要求三						
	遵守伦理但不保护自身(条件 3-2)	6.10 ±2.73	5.23	47	< .001	.75
	违反伦理但保护自身(条件 3-3)	2.92 ±2.89				
要求二 vs. 要求三						
	条件 2-3 — 条件 2-2	-.40 ±3.34	3.63	96	< .001	.73
	条件 3-2 — 条件 3-3	-3.19 ±4.23				

注：条件 2-2 和条件 2-3 数据分别来源于实验 2 的第 2 和第 3 个实验条件，条件 3-2 和条件 3-3 分别来源于实验 3 的第 2 和第 3 个实验条件。

6 总讨论

...

6.1 促进人机信任的机器人行动决策

...总体而言，机器人作为与否和保护自身与否在遵守和违反伦理原则情境下，对人机信任呈现出影响方向相反的反转效应；而服从命令则在两种情境下均能促进人机信任。这些结果提示了机器人不同决策类型在不同情境下可能表现为不同的影响，未来研究可以进一步深入探讨其内在的心理机制，如个体感知到的机器人善恶意图(Laakasuo et al., 2021; Schein & Gray, 2018)。

跨实验的分析结果显示在遵守伦理情境下，机器人不作为和不服从命令仍一定程度损害了人机信任；在违反伦理情境下，服从命令导致最少的人机信任损失，而作为和不服从命令则导致较严重的人机信任损失。这些结果启示了人们判断机器人是否值得信任可能存在两个重要标准，一是机器人能够主动保护人类且没有故意伤害意图(Laakasuo et al., 2021)，二是机器人能够服从且不违抗人类命令(Milli et al., 2017)。此外，本研究中机器人不得伤害人类与服从人类命令对于人机信任的重要程度未检验出显著差异，且均高于保护机器人自身。若根据该结果指导机器人伦理冲突情境下的行动，机器人应优先遵守不伤害人类和服从人

类命令的伦理要求，然后是保护机器人自身，这大致符合阿西莫夫对三个伦理要求所设定的优先性排序(Kaminka et al., 2017)。综上，本研究揭示了机器人具体行动决策在遵守和违反伦理原则情境中对人机信任的影响，从实证的角度拓展了阿西莫夫三大伦理原则背景下影响人机信任的行为因素研究。

...

7 结论

基于以“机器人不得伤害人类”为核心要素的阿西莫夫三大伦理原则，结合故事情境法和信任博弈，揭示促进人机信任的机器人行动决策因素有如下要点: (1)在遵守伦理原则情境下，机器人执行作为、服从人类命令以及保护或不保护自身等行动决策均有利于人机信任，而执行不作为和不服从人类命令的行动决策对人机信任的促进量则相对较少；(2)在违反伦理原则情境下，机器人执行服从人类命令的行动决策有利于减轻人机信任损失，而执行作为和不服从人类命令的行动决策将导致严重的人机信任损失；(3)在三大伦理原则所对应的要求发生冲突的情境下，机器人优先执行不伤害人类和服从人类命令的行动决策更有利于促进人机信任，然后才是保护自身的行动决策。此外，遵守伦理原则的机器人通过诱发人机投射，显著地促进人机信任。

请编委审稿专家审查。

参考文献：

Kaminka, G. A., Spokoini-Stern, R., Amir, Y., Agmon, N., & Bachelet, I. (2017). Molecular Robots Obeying Asimov’s Three Laws of Robotics. *Artificial Life*, *23*(3), 343–350.

Laakasuo, M., Palomäki, J., & Köbis, N. (2021). Moral uncanny valley: A robot’s appearance moderates how its decisions are judged. *International Journal of Social Robotics*, *13*(7), 1679–1688.

Milli, S., Hadfield-Menell, D., Dragan, A., & Russell, S. (2017). Should robots be obedient? In *International Joint Conference on Artificial Intelligence*.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32–70.

第四轮

编委意见：可以接受。

主编意见：同意发表。