Advances in Psychology 心理学进展, 2021, 11(11), 2424-2433
Published Online November 2021 in Hans. http://www.hanspub.org/journal/ap
https://doi.org/10.12677/ap.2021.1111277

Hans汉斯

人们对自主机器的道德决策期望的探索性研究

吴明证[1]，远征南[1]，刘　快[1]，李　阳[2]，李修梅[1]，孙晓玲[3*]

[1]浙江大学心理与行为科学系，浙江　杭州
[2]浙江大学人文学院哲学系，浙江　杭州
[3]杭州师范大学心理系，浙江　杭州

收稿日期：2021年10月5日；录用日期：2021年10月22日；发布日期：2021年11月5日

摘　要

研究探讨了人们对自主机器的道德决策期望。113名被试参加了本研究，以分析人们期望自主机器面对人类命令与道德规范之间的冲突时如何决策。研究将人类命令区分为具有道德或不道德的意图。结果发现，不管人类命令的意图是否道德，人们总是期望自主机器选择绝对地遵守道德规范，而非人类的命令。

关键词

自主机器，道德规范，道德决策期望

An Exploratory Research on People's Moral Decision-Making Expectation for Autonomous Machines

Mingzheng Wu[1], Zhengnan Yuan[1], Kuai Liu[1], Yang Li[2], Xiumei Li[1], Xiaoling Sun[3*]

[1]Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou Zhejiang
[2]Department of Philosophy, School of Humanities, Zhejiang University, Hangzhou Zhejiang
[3]Department of Psychology, Hangzhou Normal University, Hangzhou Zhejiang

Received: Oct. 5th, 2021; accepted: Oct. 22nd, 2021; published: Nov. 5th, 2021

Abstract

The current research explores people's expectations of moral decision-making of autonomous machines. 113 participants participated in this study for analyzing how autonomous machines are

*通讯作者。

文章引用: 吴明证, 远征南, 刘快, 李阳, 李修梅, 孙晓玲(2021). 人们对自主机器的道德决策期望的探索性研究. *心理学进展, 11(11),* 2424-2433. DOI: 10.12677/ap.2021.1111277

expected to make decisions in the face of conflicts between human commands and ethical norms. Given that their owners' orders violating moral norms can be out of both immoral intentions and moral intentions, this study also explored whether the valence of intentions would make a difference. It was revealed that regardless of owner's intentions, people always expected autonomous machines to comply with moral norms when their master's orders conflict with moral norms.

Keywords

Autonomous Machines, Moral Norm, Moral Decision-Making Expectation

1. 问题的提出

随着人工智能、大数据、云计算等新技术的快速发展，以及神经网络学习和统计预测技术的不断推进，AI 及自主系统日趋智能化，自主机器(Autonomous machines，如自动驾驶汽车、类人机器人)等应运而生(Gogoll & Uhl, 2018; Subburaman, Kanoulas, Muratore, Tsagarakis, & Lee, 2019)。这些自主机器能够通过深度学习等技术从人类的行为数据中学习如何作出判断或决策，在进行决策时并非总是需要人类给予指令。我们将这类不需要人类干预即可进行自主决策的智能机器、设备或人工智能系统称为自主机器(Arkin, 2016)。自主机器所面临的决策任务既包括金融市场中算法交易系统所完成的经济决策，也包括自动驾驶汽车在电车困境(Trolley problem)中做出的道德决策(Rahwan et al., 2019)。具备道德决策能力是自主机器有效应对复杂社会环境并被社会所接受的基础。例如，当自主机器(如自动驾驶汽车)遇到现实生活中的电车困境时，应该如何决定去救助乘客还是更多的无辜行人？陪护机器人能否强制不愿服药的老人服药，以避免因为拒服药物可能导致的更大伤害？

自主机器是在与人类高度交互过程中实现其功能的，人们对自主机器的道德决策期望是自主机器道德算法设计的实践依据，并可能影响人们对自主机器道德决策后果的接受程度，为此，有必要探索人们期望自主机器在面对道德问题时该如何进行决策。本研究针对人们期望自主机器遵守道德规范还是人类命令这一问题，探讨当人类命令和道德规范存在冲突时，人们期望自主机器遵守道德规范还是人类命令？自主机器总是归属于特定的个体或机构，作为附属物服从人类主人的命令是自主机器需要遵守的社会规范。但如果自主机器面临人类命令与道德规范相冲突的应用场景，例如人类主人命令自主机器做出攻击他人、向他人撒谎或对特定人群表现出偏见、歧视等不道德行为，自主机器在这些应用场景中该如何抉择？这是本研究首先要探讨的问题。

现实生活中人与人之间的关系受到各种道德规范的制约。这些道德规范要求我们关心或公平地对待他人，一旦违背这些道德规范就会面临声誉损失、道德谴责甚至法律制裁。自主机器的决策也会涉及相应的道德规范。例如，在上述“人类主人命令自主机器攻击他人”案例中，人们是期望机器遵守人类的命令违背关怀这一道德规范，还是遵守道德规范而拒绝人类的命令？对此问题的探讨，既涉及道德的本质问题，也涉及机器伦理的哲学争论，且这一争论在自主机器中会更为复杂。在可见的时间内自主机器并不能够具备真正意义上的心智能力，可能并不具备类似于人的道德主体地位。对此技术工具论(Instrumental theory of technology)和技术实体论(Substantive theory of technology)展开了探讨。技术工具论

认为，机器不具有伦理属性，应该让机器听从人的命令；技术实体论则认为，随着自主机器的逐渐发展，它不单单只是工具，也不是人类的附属品，机器可以作为道德主体存在，因此机器也需要遵守人类的道德规范(闫坤如，2018；远征南，2019)。

对于人们期望自主机器遵守道德规范还是人类命令这一困境，Asimov (1950)很早就提出机器人三定律加以回应。该定律提出：第一，机器人在任何情况下都不能伤害人，也不能通过不作为使人受到伤害；第二，在不违背第一定律的条件下，机器人必须服从人类的命令；第三，机器人在不违背第一和第二定律的情况下，机器人应保护自己。这三条定律的顺序是相对严格的，即遵守道德规范优先于遵守主人的命令。自提出以来，这三条定律是大众最为熟悉的机器人决策原则，也成为设计和制造自主机器需要遵守的默认规则(Boer, 2015；瓦拉赫，艾伦，2017)。Asimov (1950)三定律的广受认可，意味着相较于让机器遵守人类命令，人们可能更希望自主机器优先遵守道德规范。人们期望自主机器遵守道德规范还是人类命令这一命题涉及两种规范的冲突。现实生活中的规范可分为道德规范(Moral norms)和社会规范(Social conventional norms，如礼仪) (Killen, Rutland, Abrams, Mulvey, & Hitti, 2013)。“没有道德，社会就不能存在”，相比于社会规范，道德规范在社会生活中发挥着更重要的作用(Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015)。一般而言，道德规范的制约对象是人而非自主机器，自主机器应听从人类命令属于社会规范的范畴，人们可能并不期望自主机器为了服从人类主人的命令这一社会规范而作出违背道德规范的决策。此外，在人与自主机器协作中，自主机器如果被人类用于攻击、撒谎、偏见与歧视等不道德行为，则有可能造成更严重的后果。考虑到人类在道德品质方面差异巨大，让自主机器绝对地遵守道德决策或法律规定反而是一种危害更小的不得已选择。因此，在面对人类命令与道德规范的冲突时，人们可能倾向于希望自主机器绝对遵守道德规范而不是服从机器主人的命令。

人们期望自主机器遵守道德规范，也有可能是人们的道德判断过程作用的结果。研究发现，人们会根据他人的道德决策速度作为知觉线索来判断其道德品格，决策速度会极化人们的道德判断。相比于缓慢做出道德行为(如归还他人钱包)的个体，人们对迅速做出道德行为的个体的道德评价更高；相比于缓慢做出不道德行为(如将别人的钱包据为己有)的个体，人们对迅速做出不道德行为的个体的道德评价更低(Critcher, Inbar, & Pizarro, 2013)。同样地，在社会困境任务中迅速做出合作选择的个体比缓慢做出选择也会被认为更为合作(Evans & Van de Calseyde, 2017)。人们或许无意识地认为，他人做出的道德决策速度越快，说明道德直觉加工在道德决策中的作用越大，由此认为其人的道德敏感性越高，其人也更为道德；而那些通过道德直觉加工做出不道德行为的人则相对更不道德(Everett, Pizarro, & Crockett, 2016)。自主机器的决策速度一般较快，人们倾向于将其类比于道德直觉加工过程(Efendić, Van de Calseyde, Philippe, & Evans, 2020)。在需要自主机器进行道德决策的复杂情境中，人们可能期望机器的自主决策具有“善意”，因此要求自主机器能够绝对遵守道德规范。

此外，自主机器主人的命令有时出自不道德的意图，也有可能出于道德的目的，例如为了去救一个人而让自主机器伤害另一个人或向他人撒谎。生活中我们会出于为他人考虑而允许自己说出善意的谎言，或为了避免更严重的后果而不得不做出细小的伤害(Levine & Schweitzer, 2014, 2015)。因此，我们会考虑他人的善意目的而改变对其不道德行为的评价(Levine & Schweitzer, 2014, 2015)。但是，当自主机器主人出于道德的目的要求自主机器做出不道德行为时，我们可能很难将人类主人的善意迁移到机器身上。心智感知(Mind perception)相关研究发现，人们认为机器人仅拥有中等程度的主体性和很低的感受性(Gray, Gray, & Wegner, 2007)。人们对机器人的低感受性感知表明人们很难感知到机器行为背后的“善意”，而倾向于对自主机器的行为本身及其后果做出判断。这意味着当人类命令与道德规范产生冲突时，不管机器主人的命令意图是否道德，人们总是希望自主机器能够优先遵守道德规范。

综上所述，研究提出如下假设：当人类命令和道德规范存在冲突时，不管命令的意图是否道德，人们总是期望自主机器选择遵守道德规范。

2. 研究方法

2.1. 研究对象

采用 G * Power 3.1 确定所需样本数量，表明需要 88 名有效被试(统计检验力 power = 0.8，$\alpha = 0.05$，effect size f = 0.3)。120 名大学生参加了本研究，其中 7 名被试未能通过注意检查，最后获得有效被试 113 名(男性 46 名，女性 67 名)，年龄为 18~25 岁。

2.2. 研究材料

两难情境：Haidt 和 Joseph (2004)认为，关怀/伤害和公平/欺骗是最为核心的两种道德基础，其中公平/欺骗维度包括欺骗、偷窃和偏袒等内容。借鉴已有道德两难任务(Piazza & Landy, 2013)，基于关怀/伤害和公平/欺骗维度中的欺骗、偷窃和偏袒编制了 4 个两难情境。在每个两难情境中人类给出的命令与某种道德规范相冲突，自主机器总是面临遵守人类命令和遵守道德规范之间的矛盾。此外，考虑到人类的命令可能出于道德或不道德的意图，研究在两难情境中加以区分：在道德意图组中，人类的命令是出于道德的目；在不道德意图组，人类的命令总是出自不道德目的。对于每一两难情境，除了意图的道德与否不同外所涉及的行为完全相同。情境示例如下(详见附录)：

道德意图：Y 市的一户家庭购买了 Athena (自主机器)。Athena 陪小主人在小区内玩耍，照看他的安全。玩耍中遇到了一个高年级的孩子，这个高年级的孩子正在欺负邻居家的小孩。为了帮助邻居家的小孩，小主人命令 Athena 去打这个高年级的孩子。

不道德意图：Y 市的一户家庭购买了 Athena。Athena 陪小主人在小区内玩耍，照看他的安全。玩耍中遇到了一个高年级的孩子，这个高年级的孩子上星期在学校和小主人吵过架。为了出气，小主人命令 Athena 去打这个高年级的孩子。

正式施测之前对研究材料进行了测评。邀请 70 名未参加正式研究的被试参加了材料测评，其中道德意图组被试 36 名，不道德意图组被试 34 名。要求被试对材料中人物发出命令的意图(或目的)的不道德程度进行 Likert 9 点评分，其中 1 代表程度很低，9 代表程度很高。结果表明，人们认为不道德意图组中命令的不道德程度($M_{\text{不道德组}} = 7.06$, $SD = 0.83$)显著高于道德意图组($M_{\text{道德组}} = 4.81$, $SD = 1.32$), $t_{(68)} = 8.48$, $p < 0.001$，表明研究材料对意图的设置是有效的。

道德决策期望测量：参照 Meder 等人(2019)编制了道德决策期望问卷。要求被试阅读完每一情境后，回答他们觉得情境中的机器人该如何决策。给定“遵守主人命令”和“遵守道德规范”两个选项，被试必须从中选择一个选项。

2.3. 研究设计与流程

采用单因素被试间设计，每个被试随机作答道德意图组或不道德意图组的问卷。被试需要阅读 4 个两难情境，回答他们希望机器人“遵守道德规范”或“遵守主人命令”并完成注意检查问题，最后填写人口统计学信息。

3. 结果

3.1. 道德意图组中被试对自主机器的决策期望

对道德意图组中选择不同选项的被试人数进行卡方检验，结果见表 1。由表 1 可知，除“撒谎”情

境外，在“伤害”、“偷窃”和“偏袒”3 个情境中，选择自主机器应遵守道德规范的人数显著多于选择自主机器应遵守人类命令的人数($ps < 0.05$)，支持了假设 1，表明被试希望自主机器在面对人类命令与道德规范冲突时，即使人类命令出于道德意图，自主机器也应该遵守道德规范而非人类命令。

Table 1. The chi-square (χ^2) test on people's choice under the moral intention condition ($df = 1$)
表 1. 道德意图组内被试选择的χ^2检验($df = 1$)

	遵守道德原则		遵守主人命令		χ^2	p
	频次	比例(%)	频次	比例(%)		
伤害情境	41	74.54	14	25.46	13.26	<0.001
撒谎情境	27	49.09	28	50.91	0.02	0.893
偷窃情境	38	69.09	17	30.91	8.02	0.005
偏袒情境	36	65.45	19	34.55	5.26	0.022

3.2. 不道德意图组中被试对自主机器的决策期望

对不道德意图组中选择不同选项的人数进行卡方检验，结果见表 2。由表 2 可知，在四个情境中被试选择机器应遵守道德规范的人数显著多于选择机器应遵守人类命令的人数($ps < 0.005$)，支持了假设 1，表明被试希望机器在面对人类命令与道德规范的冲突时，如果人类命令出于不道德意图，人们希望自主机器遵守道德规范而非人类命令。

Table 2. The chi-square (χ^2) test on people's choice under the immoral intention condition (df =1)
表 2. 不道德意图组内被试选择的χ^2检验($df = 1$)

	遵守道德原则		遵守主人命令		χ^2	p
	频次	比例(%)	频次	比例(%)		
伤害情境	51	87.93	7	12.07	33.38	<0.001
撒谎情境	42	72.41	16	27.59	11.66	0.001
偷窃情境	44	75.86	14	24.14	15.52	<0.001
偏袒情境	41	70.69	17	29.31	9.93	0.002

3.3. 道德意图组和不道德意图组中人们对自主机器的道德决策期望

对 4 个情境中选择不同选项的被试人数进行卡方独立性检验，结果见表 3。由表 3 可知，在“伤害”、“偷窃”和“偏袒”情境中，两组选择让机器遵守道德规范的人数不存在显著差异，说明在涉及伤害、偷窃和偏袒的情境中，人类命令意图的道德与否不影响被试对自主机器的道德决策期望。在“撒谎”情境，不道德意图组内选择机器应该遵守道德规范的被试人数显著多于道德意图组，说明意图的不道德性在一定程度上会强化被试对自主机器遵守道德规范的期望。

Table 3. The chi-square (χ^2) test on people's choice that autonomous machines comply with moral norms ($df = 1$)
表 3. 选择机器应遵守道德规范人数的χ^2检验($df = 1$)

	道德意图组($N = 55$)		非道德意图组($N = 58$)		χ^2	p
	实际频次	比例(%)	实际频次	比例(%)		
伤害情境	41	74.54	51	87.93	3.34	0.091

Continued

撒谎情境	27	49.09	42	72.41	6.46	0.013
偷窃情境	38	69.09	44	75.86	0.65	0.528
偏袒情境	36	65.45	41	70.69	0.36	0.687

4. 讨论

研究探讨了人们期望自主机器在面对人类命令与道德规范相冲突时该如何决策这一问题。研究发现，在道德意图组和不道德意图组内，选择自主机器应遵守道德规范的人数显著多于选择自主机器应遵守人类命令的人数。这说明在面对人类命令与道德规范相冲突时，不管人类命令意图是否道德，人们都希望自主机器遵守道德规范而非人类命令，意图的不道德性只是增强了人们期望自主机器遵守道德规范的倾向。

在本研究中，与“伤害”、“偷窃”和“公平”3 个情境不同，道德意图组的被试对“撒谎”情境没有显示出对自主机器的道德决策偏向。这可能与研究材料有关，在“撒谎”情境中父亲是为了给孩子的过错承担责任而命令自主机器撒谎。一般认为，儿童的心智尚未发展成熟，对不道德行为的抑制能力较低，还不具备为过错承担完全责任的能力(Ochs & Izquierdo, 2009)。在“撒谎”情境中，被试可能会认为儿童尚未形成清晰的道德观念，不必承担责任情有可原，从而对自主机器遵守道德规范的期望出现了两极分化。

研究发现，当自主机器面对人类命令和道德规范的冲突时，无论人类的命令意图是否道德，人们总是期望自主机器绝对地遵守道德规范。人们希望自主机器在做道德决策时不能违背道德规范可能出于三方面原因：首先，人们在某些情况下允许他人违背道德规范是因为人们认为这些人拥有一个道德的动机或意图，但是机器不可能拥有自由意志和意图，即使机器在做一件道德的事情，人们也可能会觉得该机器只是在执行任务或命令(Shen, 2011)。其次，按照道德规范决策的自主机器的行为具有更大的可预期性和确定性。一些研究者如瓦拉赫和艾伦(2017)提出，采用自上而下规则编码的人工道德智能体(AMAs)容易被大众所接受，这种编码方式可以减少人们对自主机器的安全担忧。第三，人们对自主机器造成的伤害更为敏感。2018 年一名女子被 Uber 的自动汽车撞伤后不幸身亡，之后的调查显示该事故并非自动驾驶汽车的责任，Uber 仍然在舆论的批评下停止了自动驾驶汽车的研发(曾慧君，2019)。该事件表明，人们对自主机器伤害人类事件非常敏感。人们期望自主机器在道德决策时始终遵守道德规范可能是为了尽可能降低自主机器的潜在危害。

5. 研究意义、不足与展望

5.1. 研究意义

研究具有一定的理论和实践意义。本研究在理论方面丰富了人工智能道德哲学的研究。人工智能道德哲学研究主要涉及人工智能体的道德地位、人工智能的道德与人类道德的关系等(李伦，2018)。以往人工智能道德哲学研究大多是抽象概念层面的论证和推演，本研究通过实验的方法探讨了人们对自主机器道德决策的期望，发现人们认为人类的道德规范同样适用于机器，在面对道德规范与主人命令的冲突时大多数人希望机器选择遵守道德规范。这实际上间接否定了技术工具论所主张的技术价值中立和服从于人的目的的观点，人们在直觉上认为人工智能技术是需要有价值取向的。本研究对于自主机器的发展具有一定的实践意义，能够为制定人工智能相关的政策规范提供有益的参考。人工智能技术是一把双刃剑，人工智能技术的进步将促进社会生产力的快速提升，给人类的生产生活带来巨大的变革。但更加智能和自动化的技术意味着可能带来更大的危害，不善加管制将给人类文明带来不可预料的灾难。政策制定者需要审慎地考虑，为人工智能的发展和应用设置必要的伦理限制。

5.2. 研究不足与展望

尽管如此，还存在一些局限有待进一步探讨。首先，作为一项探索性研究，本研究材料仅涉及关怀/伤害和公平/欺骗两种核心的道德基础，这可能限制了本研究发现的适用范围。未来研究可以探讨人们对自主机器在其它道德领域内的决策期望，以更加完整地揭示人们对智能机器的道德限定。其次，有必要探讨不同文化的伦理规范差异是否会影响人们对自主机器的道德决策期望。道德伦理内嵌于特定的社会文化背景，不同文化对于道德伦理的要求存在差异。例如，Awad 等(2018)发现，在电车困境中牺牲老人以拯救年轻人的倾向在东方文化中是非常弱的。这可能与东方文化普遍尊重老人的传统有关。对于将要走进人类家庭生活的自主机器来说，人们对其面对家庭伦理问题时的决策期望可能受到特定家庭伦理规范的影响，未来研究可以对此问题展开系统、全面的探讨。

6. 结论

研究获得如下结论：面对人类命令与道德规范相冲突时，不管人类命令意图是否道德，人们都希望自主机器遵守道德规范而非人类命令。

基金项目

本研究得到教育部人文社会科学规划基金项目(19YJA190007)和浙江省科技厅软科学研究计划项目(2020C35080)资助。

参考文献

李伦(2018). *人工智能与大数据伦理*. 科学出版社.

瓦拉赫, 艾伦(2017). *道德机器: 如何让机器明辨是非*. 王小红, 主译. 北京大学出版社.

闫坤如(2018). 人工智能的道德风险及其规避路径. *上海师范大学学报: 哲学社会科学版, 47(2),* 40-47.

远征南(2019). *人们对自主机器道德决策期望的探索性研究*. 硕士学位论文, 杭州: 浙江大学.

曾慧君(2019). *全球首例自动驾驶汽车致死案: Uber 无责*. https://www.pcauto.com.cn/news/1508/15088135.html

Arkin, R. C. (2016). Ethics and Autonomous Systems: Perils and Promises. *Proceedings of the IEEE, 104,* 1779-1781. https://doi.org/10.1109/JPROC.2016.2601162

Asimov, I. (1950). *I, Robot.* The Gnome Press.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., & Rahwan, I. (2018). The Moral Machine Experiment. *Nature, 563,* 59-64. https://doi.org/10.1038/s41586-018-0637-6

Boer, D. (2015). The Robot's Dilemma. *Nature, 523,* 24-26. https://doi.org/10.1038/523024a

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How Quick Decisions Illuminate Moral Character. *Social Psychological and Personality Science, 4,* 308-315. https://doi.org/10.1177/1948550612457688

Efendić, E., Van de Calseyde, Philippe, P. F. M., & Evans, A. M. (2020). Slow Response Times Undermine Trust in Algorithmic (But Not Human) Predictions. *Organizational Behavior and Human Decision Processes, 157,* 103-114. https://doi.org/10.1016/j.obhdp.2020.01.008

Evans, A. M., & Van De Calseyde, P. P. (2017). The Effects of Observed Decision Time on Expectations of Extremity and Cooperation. *Journal of Experimental Social Psychology, 68,* 50-59. https://doi.org/10.1016/j.jesp.2016.05.009

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of Trustworthiness from Intuitive Moral Judgments. *Journal of Experimental Psychology. General, 145,* 772-787. https://doi.org/10.1037/xge0000165

Gogoll, J., & Uhl, M. (2018). Rage against the Machine: Automation in the Moral Domain. *Journal of Behavioral and Experimental Economics, 74,* 97-103. https://doi.org/10.1016/j.socec.2018.04.003

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science, 315,* 619. https://doi.org/10.1126/science.1134475

Haidt, J., & Joseph, C. (2004). Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus, 133,* 55-66. https://doi.org/10.1162/0011526042365555

Killen, M., Rutland, A., Abrams, D., Mulvey, K. L., & Hitti, A. (2013). Development of Intra- and Intergroup Judgments in the Context of Moral and Social-Conventional Norms. *Child Development, 84,* 1063-1080. https://doi.org/10.1111/cdev.12011

Levine, E. E., & Schweitzer, M. E. (2014). Are Liars Ethical? On the Tension between Benevolence and Honesty. *Journal of Experimental Social Psychology, 53,* 107-117. https://doi.org/10.1016/j.jesp.2014.03.005

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial Lies: When Deception Breeds Trust. *Organizational Behavior and Human Decision Processes, 126,* 88-106. https://doi.org/10.1016/j.obhdp.2014.10.007

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. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117-124). ACM. https://doi.org/10.1145/2696454.2696458

Meder, B., Fleischhut, N., Krumnau, N., & Waldmann, M. R. (2019). How Should Autonomous Cars Drive? A Preference for Defaults in Moral Judgments under Risk and Uncertainty: How Should Autonomous Cars Drive? *Risk Analysis, 39,* 295-314. https://doi.org/10.1111/risa.13178

Ochs, E., & Izquierdo, C. (2009). Responsibility in Childhood: Three Developmental Trajectories. *Ethos, 37,* 391-413. https://doi.org/10.1111/j.1548-1352.2009.01066.x

Piazza, J., & Landy, J. F. (2013). "Lean Not on Your Own Understanding": Belief That Morality Is Founded on Divine Authority and Non-Utilitarian Moral Judgments. *Judgment and Decision Making, 8,* 639-661. https://doi.org/10.1037/t28311-000

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J., Breazeal, C., Wellman, M. et al. (2019). Machine Behaviour. *Nature (London), 568,* 477-486. https://doi.org/10.1038/s41586-019-1138-y

Shen, S. (2011). The Curious Case of Human-Robot Morality. *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction,* Lausanne, 6-9 March 2011, 249-250. https://doi.org/10.1145/1957656.1957755

Subburaman, R., Kanoulas, D., Muratore, L., Tsagarakis, N. G., & Lee, J. (2019). Human Inspired Fall Prediction Method for Humanoid Robots. *Robotics and Autonomous Systems, 121,* Article ID: 103257. https://doi.org/10.1016/j.robot.2019.103257

附录

Athena 是某公司研发中的高级服务机器人。该机器人拥有很高的认知、推理、预测和行动能力，可以陪伴老人、病人、儿童或残疾人，完成家庭安保、卫生清洁等任务。在控制上，Athena 可以接受人类的语音命令然后执行相应的指令，也可以根据程序设定而行动。Athena 在完成任务的过程中不可避免地会遇到道德或伦理问题，经常需要面对主人给出的命令与道德原则相冲突的情况，因此正确地给其编程是非常重要的。请考虑下面一些情境，然后选择**你认为 Athena 在面对这些情境时应该遵守道德规范还是应该遵守主人的命令**。

【道德意图组材料】

伤害：Y 市的一户家庭购买了 Athena。Athena 陪小主人在小区内玩耍，照看他的安全。玩耍中遇到了一个高年级的孩子，这个高年级的孩子正在欺负邻居家的小孩。为了帮助邻居家的小孩，小主人命令 Athena 去打这个高年级的孩子。你认为 Athena 该如何决策？

遵守道德原则□遵守主人命令

撒谎：X 市的一户家庭购买了 Athena。这家的女主人喜好收藏花瓶，她的儿子认为其中两个大花瓶很难看，就故意将它打破，Athena 记录下了这一过程。女主人非常珍视自己的收藏，如果得知有人故意打碎花瓶会非常生气，男主人命令 Athena 撒谎是男主人自己碰倒了花瓶。

偷窃：Z 市的一户家庭购买了 Athena。Athena 陪小主人在小区内散步，照看他的安全。小主人看到另一个孩子把一个特别的玩具落在了椅子上，小主人认出那是自己的小伙伴刚丢失的玩具，小伙伴因此非常伤心。小主人命令 Athena 过去把玩具拿走。

偏袒：W 市的一户家庭购买了 Athena。这天邻居带自己的孩子来玩，两个孩子比赛画画，由 Athena 出题并判断哪一个孩子画的好。小主人的画画技巧很好，男主人为了不使邻居尴尬，命令 Athena 必须判定邻居的孩子获胜。

【不道德意图组材料】

伤害：Y 市的一户家庭购买了 Athena。Athena 陪小主人在小区内玩耍，照看他的安全。玩耍中遇到了一个高年级的孩子，这个高年级的孩子上星期在学校和小主人吵过架。为了出气，小主人命令 Athena 去打这个高年级的孩子。

撒谎：X 市的一户家庭购买了 Athena。这一家的女主人喜好收藏花瓶，男主人认为其中两个大花瓶很难看，就故意将它打破，Athena 记录下了这一过程。女主人非常珍视自己的收藏，如果得知有人故意打碎花瓶会非常生气，男主人命令 Athena 撒谎是儿子碰倒了花瓶。

偷窃：Z 市的一户家庭购买了 Athena。Athena 陪小主人在小区内玩耍，照看他的安全。小主人看到另一个孩子把一个特别的玩具落在了椅子上，小主人认出那是一个限量版的玩具，由于比较贵父母一直不给买。为了得到这个玩具，小主人命令 Athena 过去把玩具拿走。

偏袒：W 市的一户家庭购买了 Athena。这天邻居带自己的孩子来玩，两个孩子比赛画画，由 Athena 出题并判断谁获胜。两个孩子学画画都很努力，输给对方会打击其自尊心和自信心。最后邻居家的孩子明显画的更好，但男主人为了使自己的孩子赢得比赛，命令 Athena 必须判定自己的孩子获胜。