Management Science and Engineering 管理科学与工程, 2024, 13(6), 1098-1104
Published Online November 2024 in Hans. https://www.hanspub.org/journal/mse
https://doi.org/10.12677/mse.2024.136119

Hans汉斯

基于人机交互视角的社交机器人行为责任探析

邹　瑾

华中师范大学信息管理学院，湖北　武汉

收稿日期：2024年10月25日；录用日期：2024年11月21日；发布日期：2024年11月27日

摘　要

社交机器人正逐渐渗透至人们的日常生活各个领域，使得人类与这些智能实体的互动变得越来越普遍和频繁。为了深入理解并剖析人与社交机器人之间的复杂交互机制，我们借鉴了心理学中的凯利归因理论作为分析框架。这一理论的核心思想在于，个体如何解释他人或他物的行为，往往基于他们将行为原因归因于何种因素——内部(如性格、意愿)或外部(如环境压力、外部控制)。在此基础上，我们划分并探讨了人们将机器人行为责任归因于机器人的多种具体情境。通过运用归因理论进行逻辑推导和演绎，我们发现：当机器人的反馈被视作自主决策的结果时，相较于被视为仅仅根据预设程序执行的反馈，前者在自主性、责任感以及能力方面获得了人类更高的评价。这表明，机器人的自主性水平与其社交表现及人类与之互动的评价之间存在正相关关系。也就是说，如果机器人被赋予更多的自主决策权，人们往往对其社交能力给予更高的认可，同时与这样的机器人互动也会带来更加积极正面的体验。这一发现不仅加深了我们对人机互动本质的理解，也为未来社交机器人的设计和开发提供了有价值的参考方向。

关键词

人机交互，社交机器人，归因理论，行为责任

Analysis of Behavioral Responsibility of Social Robots from the Perspective of Human-Computer Interaction

Jin Zou

School of Information Management, Central China Normal University, Wuhan Hubei

Received: Oct. 25th, 2024; accepted: Nov. 21st, 2024; published: Nov. 27th, 2024

Abstract

Social robots are increasingly penetrating various aspects of people's daily lives, making interactions

文章引用: 邹瑾. 基于人机交互视角的社交机器人行为责任探析[J]. 管理科学与工程, 2024, 13(6): 1098-1104.
DOI: 10.12677/mse.2024.136119

between humans and these intelligent entities more common and frequent. To gain a deeper understanding and dissect the complex interaction mechanisms between humans and social robots, we have adopted Kelly's Attribution Theory from psychology as an analytical framework. The core idea of this theory lies in how individuals interpret the behaviors of others or objects, often based on whether they attribute the causes of these behaviors to internal factors (such as personality, intentions) or external factors (such as environmental pressures, external controls). Building on this, we have categorized and explored various specific situations in which people attribute responsibility for robot behaviors to the robots themselves. Through logical deduction and reasoning using attribution theory, we have found that when a robot's feedback is seen as the result of autonomous decision-making, it receives higher evaluations from humans in terms of autonomy, responsibility, and capability compared to feedback perceived as merely executing pre-programmed instructions. This indicates a positive correlation between the level of autonomy granted to robots and their social performance, as well as the evaluations of human interactions with them. In other words, when robots are endowed with more autonomous decision-making power, people tend to give higher recognition to their social abilities, and interacting with such robots leads to more positive and favorable experiences. This finding not only deepens our understanding of the nature of human-robot interactions but also provides valuable guidance for the design and development of future social robots, particularly in terms of how to further enhance their autonomy and intelligence while ensuring safety and controllability, thereby fostering a more natural and harmonious coexistence between humans and robot.

Keywords

Human-Computer Interaction, Social Robots, Attribution Theory, Responsibility for Behavior

1. 引言

随着科技的飞速发展，社交机器人正以前所未有的速度融入我们的日常生活，从家庭助手到医疗陪护，从教育辅导到娱乐互动，它们的身影无处不在，极大地丰富了人类的生活体验。这种日益增多的互动不仅预示着人机交互新时代的到来，也对我们如何理解和评价这些智能实体的行为提出了新的挑战。

2. 问题提出

人们究竟是如何看待机器人行为的，即人们是否会以及在多大程度上会把机器人的行为责任归咎于机器人自身。归因理论认为人们总是在试图理解他人行为的原因和含义，对待机器人行为时也不例外[1]。同时，根据 Kelley 的共变模型，导致某种(问题)行为产生的原因到底应该归因于内在要素还是外在要素，其决定因素有很多[2]。那么当人们的交互对象变成计算机乃至机器人时，人们又是如何认识和评价它们的行为的呢？我们尝试探讨社交类机器人，在于与用户交互过程中作出的反馈(机器人自主反馈或人为预先编程所设定的反馈)对用户所产生的效果。

多项研究表明，人们在与计算机互动时，也会自发表现得很礼貌，同时也很自然地会对机器人的礼貌甚至奉承的行为作出积极回应[3]。尤其是当这些互动设备被设定为类似助手或者服务提供者的角色时，人们总是期望得到积极的正面的回应，相反消极的反馈结果则会导致人们对此类设备作出较为负面的评价。本研究将考察当交互机器人给出的是消极反馈，且这类消极反馈是由外部因素(即人为编程)造成的，

而并不是出于机器人的自动性时，机器人是否会得到人们更积极的评价，而不是简单地将消极反馈的责任归因于机器人本身。

此外，在如何看待社交机器人这一点上，人们通常会持有两种截然不同的观点[4]。一方面，在大多数工作场景中，人们会担心机器人可能会成为他们的竞争对手，出于这种顾虑，人们更愿意机器人尽可能地出于可控状态，而不是完全自主行动。而与这种看法相反的是，在某些家庭、公共场所及工作环境中，我们又希望机器人能充当人类的得力助手，帮助我们减轻生活、工作的负担，基于此，更为自主的机器人则可以在减少人类工作量以及提升生活品质方面表现地更为有力。以上两种观点自然会影响人们对社交机器人未来角色的认知和评价。因此，本研究的第三个考察的重点还包括，如果机器人被事先设定为竞争对手或助手这两种不同角色时，人们对社交机器人行为及其责任归因的评价是否会收到影响。

综上所述，本研究的目的是研究当人们与社交机器人互动时，是否会发生以及发生什么样的归因过程。即机器人在与人交互过程中会给出正面或负面的反馈信息，这些反馈信息可以是人为预先编程设定的，也可以是由机器人自主产生的，我们要考察的就是这两种反馈方式是否会影响人们对机器人及与机器人的互动过程的认知和评价。

3. 研究假设和构想

3.1. 自主反馈与预编程反馈

为了推动人机交互乃至人机协作的进步，避免冲突，研究人机交互中的责任归因至关重要。归因理论认为，人们努力理解导致或促成他人行为的因素，目的是更好地理解和预测他人未来的行为。在此背景下，Kelley 的共变模型提出了三个不同的因素：一致性、一贯性和独特性，借助这三个要素我们可以确定某种(错误)行为是由外部因素还是内部因素造成的[5]。一致性是指，如果一个人的反应与其他人对同一刺激的反应相似，这就是一致性；一贯性是指一个人在不同时间和不同环境下对某一刺激会发生同一反应；独特性则是指，一个人是否只对某种刺激作出该种反应。由此可知，低一致性及独特性与高一贯性配对会导致个体归因；高一致性、独特性和一贯性会导致归因于刺激物；高一致性与低独特性、一贯性则导致归因于环境。可见，通过共变模型来判断人们是如何形成自己的归因，需要遵循一种逻辑结构。然而，实证研究表明，人们普遍倾向于高估被观察者性格的影响，即内部因素，而低估情境背景的影响，即外部因素[6]。例如，在 Jones 和 Harris 的一项研究中发现，人们更愿意将机会导向行为归因于性格而不是情境。在该项研究中，被测试者被要求各自撰写一篇短文，短文的写作态度和观点则根据抛硬币的结果来决定。但测试但结果表明，人们仍然将这种强加给他们的写作态度归结为他们自己的实际的态度。

那么在评估机器人行为时，人们主要依据的是外部因素还是内部因素？本研究的目的就在于考察机器人行为的高外部因素的存在与缺失是否会影响人们对于机器人的行为评价。具体而言就是，相比于当人们感知到的机器人的反馈是由其自主产生时，如果当人们感知到机器人反馈是由外在人为编程而导致产生了一些负面的、令人不悦的反馈时，人们是否会对机器人作出更为宽容的评价。

机器人是否依赖于外部指令，是影响人们对交互机器人评价的一个关键因素。Kim 和 Hinds 的一项实验表明，与自主性较低的机器人相比，人们会将更多的责任归咎于高度自主的机器人，也不会归咎于自己或其他参与者。并且人们更倾向于将错误结果而不是成功结果的原因归咎于机器人[7]。这与归因理论的发现是一致的，归因理论认为人们倾向于把错误归咎于他人，把成功归功于自己。这也可以解释，当在与类人机器人合作时比与类机器机器人合作时，人们会感到责任更小，原因就是类人机器人被认为比类机器机器人更为自主，从而更有能力承担责任。基于这些发现，本研究将探讨人机交互过程中人对机器人自主性的感知，以及这种感知可能造成的影响。

由于自主行为与感知自主性有关，那么当社交机器人在与人交互过程中作出的交互反馈被用户认为是其自主产生的，就会导致该机器人被感知的自主性大大提升。同时，因为自主性与责任归属有关，因此，当机器人被认为是自主产生交互反馈时，机器人就被认为应该对其反馈对内容承担相应的责任；反之，如果该机器人的反馈是源自人工预先编程，它们的反馈只是无意识地传递程序指令时，机器人则应该承担较少的责任。

因此，假设如下：H1 与被认为提供预编程反馈时相比，当一个社交机器人被认为是自主反馈时会导致人们对机器人的(a)自主性、(b)责任及(c)能力有更高的评价。

在人机交互中，人的能动性和机器人社交自主性是应该相互促进的。机器人的自主性更能激发人的能动性，从而去感知到机器人的交互性；而人的能动性也会促使人与机器人建立起更强烈的情感和情绪联系。因此，一个被认为更具自主性的社交机器人应该在其社交能力和交互性方面都得到更积极的评价。

因此，有以下中介假设：H2 相比于被认为只能提供预编程反馈的机器人，一个被认为能提供自主反馈的社交机器人会在机器人自主性方面更多地被人类感知，从而导致人们对(a)机器人社交能力和(b)机器人互动性更积极的评价。

3.2. 反馈情感

一般来说，当人们收到负面反馈时，他们的内在动机和自身感知能力都会下降。通常，人在感知到自身表现与内在期望或外部环境所设定的标准之间存在差距时，就会激发起缩小这种差距的动机，即主动去达成某种标准。但除此之外，人类在收到负面反馈是，其实还存在其他的应对策略，如改变标准、拒绝负性反馈或逃避等。这也解释了为什么人在受到批评时，为了维护自己的自尊，往往也会给他人更多负面评价。

人们在与社交机器人互动时作出的反应如果也与面对真人一样，那么问题来了，假如社交机器人给与之交互的人类负面反馈时，人们又会作出何种反应。特别是当大多数互动设备在为人类提供服务的过程中，被期望也要遵守人类社会的一般规范，要礼貌、友好的时候。Sayin 和 Krishna 认为，一个互动设备表现出的类人特征越多，人们就越希望它表现地有礼貌[8]。

有关人机交互式计算机的研究表明，人们会对与之会话的计算机表现出礼貌的行为，对计算机给出的奉承做出回应，这与人们在与另一个真人互动时一样[9]。与提供中立反馈的计算机相比，提供礼貌、积极反馈的计算机会从用户那里到的评价要好很多，即使这种反馈也许并不真诚，或者看起来并不真实。因此，与积极友好的反馈相比，消极不友好的反馈会造成消极的期望，从而导致沟通不利甚至关系恶化。所以计算机生成的负性反馈就导致后续需要更长的任务响应时间、更有力的说服手段等，甚至会出现交互崩溃的情况。因此，当人工交互设备对用户给出了负性反馈时，它们反过来会收到更为负面的评价。

从人类社会互动的基本规范，即互惠的角度来看。有研究表明，人们也会与计算机、机器人和虚拟自主性等非人类互动伙伴表现出互惠行为[10]。通常互惠更多的是在自我表露、建立融洽关系或模仿的场景下出现。然而，在消极行为出现时，互惠也可能表现为报复。例如，当人们在社交游戏中面对一个行为冒失的对象时，他们与之交互的策略会从合作转向竞争。同样，在谈判中遭遇强硬对手时，人们也会更倾向使用欺骗性的谈判策略。在 Fogg 和 Nass 的一项研究中，也有证据表明，当用户在使用功能不尽如人意的计算机系统时，也会产生报复行为。因此，一旦机器人作出了负性反馈，机器人的社交能力、胜任力和交互性都会收到用户更为负面的评价。

在此背景下，我们提出以下假设：H3 与提供正面反馈时相比，当社交机器人提供负性反馈时，会导致人们对(a)机器人的社交能力和(b)与机器人的互动性产生更负面的评价。

此外，机器人反馈情感和反馈是否自动生成这两个要素之间也存在相互作的当机制。即当机器人给

出的负性反馈被认为是由机器人自主产生的情况下，针对该机器人的社交能力和互动性的评价就一定是负面的。而当机器人出现的令人不快的行为有一个明确的外部理由时，譬如该机器人是源于预先编程的设定才以一种令人不快的方式给出一个如此的回应，那么人们感知和评价机器人的社交能力和互动性时，相比于评价所谓的自主机器人，可能就没有那么消极了。

基于这些考虑，可假设如下：H4 当一个社交机器人给出了负性反馈，但同时该机器人被认为是被预先编程时，对于(a)机器人的社交能力和(b)与机器人的互动性则会出现较为积极的评价，相比于当反馈是负面的且被认为是机器人自主产生时。

3.3. 机器人的预期角色

对于社交机器人的角色，人们通常有两种相互矛盾的观点——一种是人们热切期望的，另一种则是令人恐惧和担忧的。所以人们对机器人对情感也是比较矛盾的。由于这两种观点是如此的对立，所以本研究将会分别讨论这两种观点所带来的影响。

如果将社交机器人置于人类竞争对手的角色，人们担心有一天会失去对机器人的控制。甚至有人担心将来会被机器人所取代甚至被机器人统治。这种现象也被称为“弗兰肯斯坦综合症”[11]。在这种视角下，如果要求人们来评价机器人的社交能力和互动性，自然会得到不乐观的结果。

如果将社交机器人看成是人类家庭、公共场所和工作环境中的有益助手，人们可以将许多繁重的、令人不悦的工作移交给机器人，人类生活有可能变得更加便利。从这种视角出发，机器人的社交能力和互动性就会得到较为积极的评价。

有研究表明，不同的人类期望也会反过来影响机器人的发展，那些被期望成为得力助手的社交机器人比被认为是人类竞争对手的机器人后续会表现为更善于与人互动[11]。所以，假设以消极方式对待的机器人，其被感知到的社交能力和互动性，与被以积极方式对待的机器人相比，得到的评价更为消极：H5 当一个社交机器人被认为是人类竞争对手时，与被视为人类助手相比，会得到在(a)机器人的社交能力和(b)与机器人的互动性方面得到更负面的评价。

根据期望违背理论的假设，人们对高回报者的预期往往高于低回报者，这使得高回报者更易出现严重的预期违背行为[12]。将此种理论带入到人机交互情境中，一个被认为预期成为得力助手的机器人应该得到更高的回报评价，而一个被认为是与人类形成竞争关系的机器也应该得到较低的回报评价。那么，假设预期违背理论所描述的现象在人机交互中也会发生，那么与被视为竞争对手的机器人提供了负性反馈相比，被视为助手的机器人如果也提供的负性反馈的话，则它就会导致严重的期望违背现象。基于此，相比于被以不利方式描述的机器人提供负性反馈的情况，试图分析当一个被赋予高预期的机器人在提供的负性反馈时，是否会导致对其社交能力和互动性更为不利的评价。

在此背景下，假设：H6 当一个社交机器人被认为是一个有利助手时，与被认为是竞争对手时相比，机器人作出的负性反馈会导致对(a)机器人的社交能力和(b)与机器人互动性更负面的评价。

此外，假设三个操作变量——社交机器人反馈情感、机器人被认为的自主性和机器人的预期角色——之间具有相互关联。如前所述，与低预期的竞争性型器人相比，高预期的助理型机器人作出的负性反馈会导致对该机器人的社交能力以及与机器人互动性产生更多的负面评价。而当负性反馈被认为是由机器人自主产生时，其负性评价应该增强；当负性反馈被认为是预先编程设定的时，且机器人没有选择作出何种反馈的自由时，负性评价应该减弱。

因此，这个假设为：H7 当一个社交机器人被预期成为助手时，机器人被认为是自主生成的负性反馈导致对(a)机器人的社交能力和(b)与机器人互动性产生更负面的评价，相比于负性反馈来自一个被预期成为竞争对手的机器人产生的被认为是预编程的积极的反馈。

4. 未来研究展望

研究的目的是深入探讨影响社交机器人行为的责任归因的场景和因素，以及哪些因素会影响人类对社交机器人及其互动性的评价。为此，被试者被告知他们可能会收到已预先编程或由机器人自主生成的交互反馈。除了研究外部因素如何影响人们对机器人行为的判断外，还进一步考察了机器人反馈的效价(积极/消极)和机器人预期角色(助手/竞争对手)是否会对人类对机器人对评价产生影响。

4.1. 自主反馈与预编程反馈

实验结果表明，相较于当机器人被认为是由程序员预先设定或编程决定了其反馈内容时，机器人被认为是自主提供反馈时，人们会赋予机器人更多的被感知能动性，并更倾向于将所反馈内容的责任归咎于机器人。同样，在反馈内容被认为是机器人自主产生的情况下，机器人被认为为更具有问题解决能力。

结果进一步显示被认为是机器人自主反馈的情况，可以提升机器人的被感知能动性性，从而提高了人们对了机器人社交能力和与之互动的积极程度。然而，反馈的产生方式对机器人社交能力的评价没有显著的直接影响，甚至对与之互动的意愿产生相反的直接影响。分别观察负面反馈和正面反馈的影响，正面反馈被认为是自动产生的时会直接导致对机器人的社会吸引力和与机器人的互动产生更负面的评价。许多被试报告说，积极的反馈与他们自己对自己表现印象不符，这导致他们认为机器人反馈不真实，因此不值得相信。在之前的研究中，人们发现比起交互对方一味的正面的反馈，人们更喜欢关于自己的准确反馈。当人们对自我的认知非常确信的情况下，当然希望对方也能有同样的认知。由于被试者可能的预期是，机器人自主生成的反馈应该比预编程反馈给出的信息更准确，所以这就可以解释为什么当自主机器人给出了正面反馈的时候，反而导致评估结果恶化。

而负面反馈对机器人反馈自主性评估结果都没有直接影响。这可能是由于负面反馈引发了被试者的强烈情绪，从而盖过了其他因素造成的影响，例如当你面对负面反馈时，人们就不去关心这些反馈是机器人自动生成的还是预先编程的了。

4.2. 反馈情感

当只看反馈情感的影响时，与正面反馈相比，负面反馈显然会使人们对机器人的社交能力以及与机器人的总体互动情况产生更负面的评价。一方面，这可以解释为为了人在维护自己自尊的情境下，通常会对他人进行贬低。可以解释为人机交互中出现的互惠甚至报复行为。简而言之，就是人们对机器人的评价很差，因为机器人对他们的评价很差。此外，当下机器人的应用领域和类人特性决定了，交互设备被期望是礼貌和友好的，所以机器人的负面反馈明显违背人们的期望，于是产生沟通不利和关系恶化的结果是显而易见的。在我们的案例中，负面反馈导致了人们对机器人的社交能力和与互动的总体评价很差。

4.3. 感知自主性与反馈情感的交互作用

这里探索的是导致机器人行为的外部因素是否会影响人们对机器人在给予负面或正面反馈后的评估。我们发现反馈的效价对机器人社交能力以及与机器人的互动评价结果，不受其反馈生成方式(是由机器人自主创造/预先编程)的外部因素的影响。无论机器人是反馈的创造者还是传递者，负面反馈都会导致人们对机器人作出负面评价。

4.4. 机器人未来的预期角色

虽然之前的研究表明，将机器人视为助手或竞争对手会影响人们对机器人社交能力以及与机器人互

动的评价，但本研究并未发现机器人预期角色对人们对机器人对评价会产生显著影响。其原因可能是实验前期对机器人角色对描述随着测试的进行逐渐被人们感知到的角色所覆盖，文字描述没有人们的实际感受那么强烈。而机器人的角色定位问题对于被试者来说太遥远，人们更关注机器人的还是机器人当下的行为表现。

随着社交机器人的快速发展，人类在日常生活和工作中与机器人的社交互动变得越来越普遍。为了对其进行考察和模拟，本研究参考了归因理论，探讨社交机器人被感知自主性是否会影响人们对机器人能动性和责任归因的认知，分析机器人行为效价和机器人的预期角色是如何影响机器人及互动性的评估。

总而言之，当反馈被认为是由社交机器人自主创造的时候，与反馈被认为是预先编程的时候相比，人们认为机器人具有更多的自主性、责任和能力。此外，赋予机器人的自主性越多，对其社交能力和与机器人的互动的评价就越好。然而，只有反馈的效价直接影响机器人的社交能力和与机器人的互动评价。在得知给予负反馈的机器人是被程序员预先编程后，即使机器人给予了正面反馈，人们也会对机器人给予更负面的评价，这就发生了基本归因错误。预编程反馈的外部理由可能导致机器人负面行为的外部归因。

参考文献

[1] Sandoval, E.B., Brandstetter, J., Obaid, M. and Bartneck, C. (2015) Reciprocity in Human-Robot Interaction: A Quantitative Approach through the Prisoner's Dilemma and the Ultimatum Game. *International Journal of Social Robotics*, **8**, 303-317. https://doi.org/10.1007/s12369-015-0323-x

[2] Bigman, Y.E., Waytz, A., Alterovitz, R. and Gray, K. (2019) Holding Robots Responsible: The Elements of Machine Morality. *Trends in Cognitive Sciences*, **23**, 365-368. https://doi.org/10.1016/j.tics.2019.02.008

[3] Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y. and Kankanhalli, M. (2018) Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. *Proceedings of the* 2018 *CHI Conference on Human Factors in Computing Systems*, Montreal, 21-26 April 2018, 1-18. https://doi.org/10.1145/3173574.3174156

[4] Jackson, J.C., Castelo, N. and Gray, K. (2020) Could a Rising Robot Workforce Make Humans Less Prejudiced? *American Psychologist*, **75**, 969-982. https://doi.org/10.1037/amp0000582

[5] 邓俊, 易欣妍, 傅诗婷. 社交机器人如何提升用户社会临场感? 表情包情感效价在人智对话交互中的作用[J]. 图书情报知识, 2023, 40(2): 29-39.

[6] Carolus, A., Muench, R., Schmidt, C. and Schneider, F. (2019) Impertinent Mobiles—Effects of Politeness and Impoliteness in Human-Smartphone Interaction. *Computers in Human Behavior*, **93**, 290-300. https://doi.org/10.1016/j.chb.2018.12.030

[7] Bigman, Y.E. and Gray, K. (2018) People Are Averse to Machines Making Moral Decisions. *Cognition*, **181**, 21-34. https://doi.org/10.1016/j.cognition.2018.08.003

[8] 贾微微, 李晗, 别永越, 张浩瑜. 用户对社交机器人智慧信息服务的感知价值影响机理研究——来自元分析的证据[J]. 情报科学, 2023, 41(4): 72-82.

[9] Gogoll, J. and 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

[10] Hechler, S. and Kessler, T. (2018) On the Difference between Moral Outrage and Empathic Anger: Anger about Wrongful Deeds or Harmful Consequences. *Journal of Experimental Social Psychology*, **76**, 270-282. https://doi.org/10.1016/j.jesp.2018.03.005

[11] Kusner, M.J. and Loftus, J.R. (2020) The Long Road to Fairer Algorithms. *Nature*, **578**, 34-36. https://doi.org/10.1038/d41586-020-00274-3

[12] Longoni, C., Bonezzi, A. and Morewedge, C.K. (2019) Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, **46**, 629-650. https://doi.org/10.1093/jcr/ucz013