# 为什么要报告与机器人的失败互动?!为了基于小插图的互动质量

Agnes Axelsson
TU Delft
Delft, Netherlands
a.axelsson@tudelft.nl

Merle Reimann
VU Amsterdam
Amsterdam, Netherlands
m.m.reimann@vu.nl

Ronald Cumbal
Uppsala University
Uppsala, Sweden
ronald.cumbal@it.uu.se

Hannah Pelikan
Linköping University
Linköping, Sweden
hannah.pelikan@liu.se

Divesh Lala
Osaka University
Osaka, Japan
lala.divesh.kanu.es@osaka-u.ac.jp

***Abstract*—尽管随着大型语言模型（LLM）的出现，人机交互的质量有所提高，但与人际交互相比，系统仍存在各种导致其次优化的因素。失败的性质和严重性通常依赖于交互的上下文，因此无法在已在 HRI 研究中实施的广泛场景和实验中进行概括。在这项工作中，我们建议采用一个在人机交互领域被忽视的技术，民族志小插曲，以清晰地突出这些失败，尤其是那些很少被记录的失败。我们描述了撰写小插曲背后的方法论，并根据我们在 HRI 系统中失败的个人经验创建了我们自己的小插曲。我们强调小插曲的优势在于能够从多学科的角度传达失败，促进对机器人能力的透明度，并记录原本会被研究报告遗漏的意外行为。我们鼓励使用小插曲来增强现有的交互评估方法。**

***Index Terms*—ethnographic vignettes, spoken interaction, dialogue, quality, HRI, HAI, reporting**

## I. 简介

与机器人进行高质量对话是许多人机交互（HRI）研究者的目标 [38] 。尽管技术不断进步，人机交互中的对话有时仍会失败。在本文中，我们提出采用小插图写作作为报告失败互动观察的方法。

虽然大型语言模型（LLMs）在模拟人类语言方面表现出色，激发了人们对生成有意义对话的兴趣和乐观态度，但其众所周知的缺陷仍然存在 [6, 9, 24] 。然而，在实现无瑕疵与机器人进行口语交互方面仍有很长的路要走 [45] 。为了实现这一目标，需要解决的挑战之一是在定义、描述和评估具体交互方面。在本文中，我们建议通过民族志方法描述对话中的失败时刻，这是理解、评估和定义人机交互的一条路径。我们借鉴了人机交互（HCI）中人机对话的悠久历史 [44, 52, 71] ，以及在语音识别机器人方面的研究 [45, 53, 63] 。

分析人机 [1] 语言交互具有挑战性。不受限的互动总会导致意想不到的边缘情况，这可能会以系统设计者未预料的方式影响用户体验。用户对机器人能力的期望，受到具象和多模态能力的影响，也可能难以预测。这导致了对互动质量 [41] 的不同期待。更重要的是，评估互动通常涉及定性 [7, 61] 或定量 [21, 39] 测量，这些测量可能仅捕捉到互动的特定方面，可能忽视了一些细微的重要因素。鉴于此，我们需要重新审视口语 HAI 的分析和报告。

故障，即“系统执行的行为或服务能力下降导致偏离理想、正常或正确功能的状态” [12, 64] ，在 HAI 评估中是一种特别有趣的现象。如果交互代理知道故障不是太大而无法解释或解释，则可以采取补救策略 [36] 来修复后果 [18] 。适当的策略取决于交互的性质 [14] 。故障会始终导致人类用户对机器人失去信任 [23, 54] 。用户可能对自己在交互中的贡献能力失去信任，或者对交互本身失去信任 [55] 。

在本文中，我们提出使用民族志小插图 [8, 25] —即对特定情境和背景下的互动的简短描述—来补充经典的互动评估指标，通过报告互动中的失败时刻。小插图让我们能够报告 HAI 中的意外事件和偏离计划场景。由于在人机互动中对于什么构成良好或不良的对话互动尚无共识 [22] ，像这些小插图这样的定性方法对于捕捉互动中的主观观察非常有价值。这些观察可以提供见解，使社区中的其他研究人员也认为值得进一步分析。

我们通过结合个人在机器人交互建造和观察中的经验，展示了使用民族志小插图的方法。我们说明了基于小插图的报告如何能够进一步提高对人机交互中挑战的认识，使研究人员能够记录被忽视或“无足轻重”的问题。我们的小插图反映了设计民族志，即通过讲述故事来影响和反映设计一个系统 [19] 的阶段，并将小插图写作视为从业者记录观察的工具，就像在教育科学中 [4, 20] 一样。

## II. 相关工作

### A. 人际互动中的质量

评估人类之间互动的一种视角是衡量互动伙伴如何实现他们的交流目标。使伙伴未能实现其共同目标的交流行为是较差的互动，而那些使伙伴迅速有效地实现其目标（即很少需要修正 [36] ）则表明是良好互动 [15] 。

在合作原则中，通过关于数量、质量、关系和方式的格言，Grice [32] 强调贡献应该简洁、真实、与互动相关并且对当前互动中的人易于理解。同样地，Clark 和 Allwood et al. 认为沟通是一种双方协商的合作活动，尽管关于这一过程的确切机制以及讲话者与听者之间的独立性未完全达成一致 [1, 15] 。因此，作为外部人员来评估一次互动本质上变得主观，因为互动的真正有效性只能从参与者个人和共享目标的角度进行分析。

### B. 评估人机交互

与 Grice [32] 所描述的人类不同，交互式机器并不会真正追随合作动态，也不会本能地适应用户的需求 [47] 。但是，用户却可以很容易地适应机器 [67] ——或者适应他们认为机器正在做的事情 [16] 。

评估人与代理之间的互动受到对话系统特征和与互动相关任务的显著影响 [22] 。尽管已经提出了许多量表和测量方法用于评估与人工社交代理的互动，但并没有标准的评估程序。相反，评估方法在不同研究中差异显著，因为

[1]在本文中，代理是指虚拟的或情境中的代理，而机器人则严格是指具有具体化身的情境中的代理。

每种方法通常依赖于任务和环境 [17, 65] 。Fitrianie et al. [26] 和 Bagchi et al. [2] 分别在 IVA 和 HRI 会议上批评了问卷的低重用率。

遵循第 II-A 节的原则，HRI 的用户体验也可以通过测量满意度 [10, 13, 40] 、挫折感 [68] 或用户对代理的总体感知来进行评估，从而深入了解交互 [29, 48] 。诸如 Godspeed 问卷 [3] 或广泛的 90 项人工社会代理问卷 [27] 等工具经常被使用。交互后的半结构化或开放式访谈 [51] 补充了更静态的问卷。

将目光从用户的主观体验转移开后，可以通过客观的对话次数 [42] 、交互持续时间 [43] 或对于基于任务的交互，任务成功率 [42] 来衡量质量。客观指标的优势是可以让机器人实时评估其自身交互的质量 [46] 。

主观和客观数据在评估交互质量方面都很有价值。这些测量使研究人员能够比较研究内或不同相关研究之间的交互。然而，它们通常只捕捉研究人员意图测量的结构，可能忽视许多对交互产生影响但未预期的元素。

### C. 观察人机交互

我们已经描述了在交互后测量主观和客观因素的方法，或在交互过程中测量客观因素。但是，关于将整个交互视为一个整体来记录其质量的研究较少。在一项早期研究中，Sabanovic et al. 专注于记录真实世界环境中的人机交互。作者推动了在自然环境中观察社交机器人，证明这种方法能够突出社交机器人设计中的重大问题，并对初始设计假设提出挑战，最终增强了机器人的交互质量 [57] 。

受民族学研究启发的方法对于描述完整的互动非常有用，但可能会被误解或被错误地认为缺乏科学有效性。Mutlu and Forlizzi 分析了一台在医院中自主配送机器人的人种学数据，揭示了个人在工作中如何将机器人整合进来的巨大差异。Forlizzi et al. 展示了用户在家庭环境中与清洁机器人互动的各种反应。研究表明，对机器人的社会归因促进了更快速的采用，并减少了与依赖技术相关的污名。Sabelli et al. 在老年护理中心观察了一台会话机器人长达 3.5 个月，深入了解了老年人如何与机器人互动。这些互动是在日常生活需求和活动的框架内进行的。[58]

许多这些研究被框定为长期研究项目中的探索性或初步步骤。需要强调的是，观察方法在所有评估和分析人机交互的研究中都可能是有价值的。正如这些研究所展示的，作者发现了一些意想不到的结果，从而显著增强了对交互的整体理解。虽然广泛的定性研究对于理解新兴系统对社会的广泛影响是重要的，但在本文中，我们希望引入一种方法，使开发人员能够在原型制作、测试和演示过程中记录观察。通过使用民族方法学对话分析来看待人机交互，有助于识别和分析视频记录中可见的不同类型的失败 [50, 56, 66, 69] 。

尽管对专注于人机交互失败的呼吁已有不少，它们在科学文献中并没有像预期或设计的行为那样被一致地报道。通过撰写情景短剧，我们希望更加关注那些在人机交互中从用户的视角出发，其交互出现问题或没有按设计意图演变的时刻。这符合会话分析的偏差案例分析传统，其中对典型互动模式的偏差给予特别关注。通过详细观察那些可被视为异常的案例，我们希望能更深入了解对话系统在人机交互中面临的持续问题。

我们以民族志小品的形式呈现我们的观察，其中 [8, 25] 侧重于展示互动未能达到预期质量的案例。系统地分析负面例子使我们能够以归纳的方式推断互动质量。我们并不声称已经找到了可能出现的所有问题，而是希望为批判性话语做出贡献，以便使 HRI 研究人员在做出明智的设计决策时，同时考虑人机交互领域的挑战和局限。

举例说明不是为了泛化和数量化；这些案例的发生提供了它们确实会发生的证据 [60] 。我们在多样化的研究团队中彻底讨论了每个小插曲，确保这些小插曲与口语 HRI 相关，并描述了轶事但真实、可操作和可推广的观察。

### D. 如何编写小插曲

撰写小插曲需要不同于撰写其他科学文本的写作风格和作者声音。它们描述了作者观察和经历的内容 [35] ，并总是由作者的视角所塑造。通过适当地构建小插曲，观察可以被映射到更普遍理解的社会世界和实践中。我们描述了三个一般步骤，借鉴自 [25] ，以帮助未接受人类学或民族方法学训练的人系统地描述他们的观察。

想象一下你正在向一个朋友描述这个事件。你会如何告诉他们这个事件？你会提供什么信息来介绍这个主题，让你的朋友想象他们自己在场？因为小插曲是对事件的个人表达以及从中获得的经验教训，因此对于不同的人来说，它们会有不同的表现——在第 **??** 节中的一些小插曲展示了这种在本节所描述范围内的变异性。

可以通过不同的方法来创建“页面上的场景”[25, p. 45] 。有些作者喜欢先写下一些笔记，然后再详细展开，而另一些作者则立即写出较长的文本部分。撰写小插曲是一个反复的过程，这意味着要一遍又一遍地修改你的小插曲。过几个小时后再回来补充遗漏的元素或删除次要的细节，以更好地反映事件的发生，这可能会有所帮助。计划一点时间让自己进入写作的状态，给自己空间详细回忆事件。

小插曲写作的特点是包含更多的细节、形容词和副词，尤其是在开始写作时。如 Goffman [31, p. 131] 在一次关于田野工作的演讲中所说：“尽管如此松散的副词化散文，它仍然是一个比那些被简化成几句话‘合理句子’的东西更丰富的起点。[···] 要在这个领域做到科学化，你必须首先相信自己，尽可能充分和华丽地写作。[···] 将自己置于你所写的情境中，以便后来你会看到如何修正你所说的话。你可以说，‘感觉是那样的，’‘我的感觉是，’‘我有这样的感觉’—这种事情。这是自律的一部分。”写作应该是反思性的，所以鼓励使用第一人称代词。通常在第一次迭代时添加过多细节会更好，然后逐步去除那些不需要的部分。我们撰写的适用于 HRI 的民族志片段的非正式指南位于附录 A 中。

我们提供了六个短文，这些短文描述了失败案例，并对这些短文所指向的人机对话当前面临的更广泛挑战进行了反思。

在开发一个系统时，开发者通常以一种临时的方式自行测试他们的系统。通过小插曲的形式记录这些观察，有助于报告那些未进一步探索的设计途径，或者概述在测试过程中遇到的问题，这些问题在实际部署中可能会引发麻烦。

*1) 插图 1：* LLM 生成的研究误导信息：

我们正在建立一个机器人系统，以探索其在公共交通中可以提供的功能。在早期测试中，我们研究了如何生成对乘客意外问题的回答。我们推测，一个大型语言模型（LLM）可以应用于基于乘客语音的文本转语音（TTS）输入快速生成此类问题的答案，从而减少生成响应的时间，

因为让远程操控我们机器人的“绿野仙踪”击打每个意外问题的答案会耗费不合理的时间。

在实验中，LLM 被提示了其应扮演角色的相关背景细节。在测试中，当被问到为什么它在车辆中时，由 LLM 驱动的机器人一贯地以编造的实验设置细节以及实验背后的研究人员信息作答。当被问及研究人员是谁时，该系统提供了并不存在的个人的电子邮件地址和名字，但这些人的隶属关系和身份对于不知情的参与者来说是可信的。同样，当被问及实验内容时，LLM 驱动的机器人总是在它被提示的信息之外，添加有关研究问题正在被探索的声明，尽管这并没有被要求，也不是它可以访问的信息的一部分。

*a) 反思：* 我们感到惊讶的是，系统如此轻易地准备说出一些我们可以分辨出的不真实的话，而乘客无法分辨其不真实，且系统本身没有任何验证的方法。我们本以为大语言模型（LLM）会用闲聊或无害的填充来填补空白，但相反，它的行为可能很容易损害到我们乘客的体验。最后，我们让操作员手动输入这些回应，但这带来了显著的延迟。在这种情况下，奥兹法师显然合理，因为它允许设计师在现地实时考虑系统应该具备的功能时，配合用户的想法进行即兴的头脑风暴。我们提示大语言模型不要编造实验细节，但大语言模型的角色明确是要处理那些意想不到的问题，我们无法为其提供预先写好的答案。处理这种问题将是一个无穷无尽的资格问题 [30] 。

无法识别用户的名字：为了实验室的开放，我正在准备一个病人面试机器人的演示。代码已经完成，其中的部分已经用通用答案进行了单独测试。然而，在最后测试期间，我尝试用不同的答案回应，多位国际同事也测试了这个系统。机器人在互动开始时会先自我介绍，然后询问用户的名字。当我或我的同事提供我们的名字时，机器人无法正确理解它们，要么误解为其他词汇，要么要求重复。我尝试通过以不同方式发音我的名字来让机器人理解，但机器人的语音识别一再失败。虽然一开始我们笑着并认为这是一个让机器人说出我们名字的挑战，但很快这种情况就变得令人沮丧，因为我们发现根本无法在对话中使用我们的真实名字。

这个小插曲说明了即使是看似简单的任务也可能会持续失败。单个模块的行为，比如在这里的 ASR，会影响整个交互。机器人无法理解用户的名字乍看并不是一个非常严重的问题，但它清楚地表明整个系统是不具有包容性的。尤其是当机器人从用户那里收集数据而不仅仅是闲聊时，确保数据的正确性和用户能够为机器人提供正确的信息是很重要的。

这些简短的陈述指出了对话中两个关键挑战——说出相关的内容和彼此倾听（见第 II-A 节）。插图 1 中描述的 LLM 幻觉违反了 Grice 的质量准则，即一个人不应该说没有证据支持的内容 [32] ——尽管尚不清楚是否可以说 LLM 持有某种信念 [47] 。

即使在开发机器人系统时投入了大量精力来解决交互问题，当用户实际开始与其交互时——通常是在研究中用来进行实验时，问题可能会在完全不同的层面上出现。我们展示了如何利用短文帮助描述在整个实验过程中普遍具有挑战性的方面——提供一种报告可以改进的内容但不使研究失效的方法。我们还展示了短文如何描述通常会被视为异常值的情况，从而提供一种讨论它们所指向的更大的挑战的方法。

*2) 场景三：* 提出过于冗长和复杂的问题：我们正在进行一项实验，实验对象是在疗养院中与一个专心聆听的机器人互动的老年参与者。参与者与机器人谈论他们曾经历过的有趣经历。一位参与者谈到了他们的海外假期，讲话大约持续了 30 秒。然后，机器人首先回应承认了该信息，解释说它也喜欢那个目的地，然后询问他们对那里的文化地标的看法。这个回应持续了大约 10 秒钟。我们期望参与者立即回答问题并继续在那个话题上谈话。然而，参与者却回应“啊？”并表现出困惑的样子。看起来他们似乎没抓住机器人在说什么。于是，他们继续讲述另一个话题，无视机器人的问题。

*a) 反思：* This vignette highlights that long turns, even ones meaningfully addressing the conversation, can be difficult to follow. This may be especially problematic for groups who may suffer more from hearing difficulties such as the elderly. While it is possible to modify an LLM prompt to mitigate this issue, this does not guarantee that the complexity of a question will be reduced. Instead of long and complex utterances, users require an empathetic response from the system, which could be very simple, such as asking for expansions (e.g. “You drove a car?”) [15] .

*3) 小插图 4：* 无法提供语境相关的后续信息：作为我进行的一个实验的一部分，我们配备对话功能的机器人被放置在超市的葡萄酒货架上，为顾客提供葡萄酒选购建议。一个海报告知顾客机器人的任务是提供葡萄酒建议。葡萄酒推荐基于价格、成分和葡萄酒类型等信息。在顾客与机器人成功互动并收到葡萄酒推荐后，他们会看到葡萄酒的一般信息和图片。然而，机器人并没有显示任何关于在货架上哪里可以找到该葡萄酒的信息。当顾客询问机器人在哪可以找到推荐的葡萄酒时，除了提到大概的葡萄酒类别外，它无法提供更多信息，这还是让顾客需要自己寻找两个货架。顾客表达了他们的沮丧，甚至没有开始寻找推荐的葡萄酒。他们没有使用提供的建议，而是决定自己选择一款葡萄酒，完全忽略了之前与机器人的互动。

这个小插曲强调了不仅仅是机器人的外观，机器人所处的具体环境也可能影响用户的期望。顾客希望能够提供建议的机器人也能够执行合乎逻辑的下一步操作（告诉推荐物品的位置）。如果一个人在商店里能够推荐一种酒，那么对人来说，顺便告诉顾客酒在货架上的位置是微不足道的。正如我们在这个小插曲中看到的，这种假设对机器人而言并不成立。系统设计师必须了解用户将基于系统的环境、位置和体现产生怎样的期望。在这种情况下，LLM 可能会给出更自然表达的对话，但不会帮助找到具体葡萄酒在过道中的位置信息。

转接设计、体现和多模态仍然是人机对话的挑战，如小插曲所示。如第 II-A 节所述，人们在进行人机交互时假设这些交互将遵循基本的沟通原则，无论是 Gricean [32] 或基于基础的 [1, 15] 。如果机器人不能符合用户的期望，用户就需要适应机器人的风格。在 场景三 中，这意味着要适应长篇大论的表达，导致互动质量差。在 小插图 4 中，对于希望机器人执行其无法执行任务的顾客来说，根本不可能产生一致性。

当系统被部署时，可以观察到许多有趣的交互。除了更深入的实地研究外，使用小插图来报告演示过程中的观察结果可以是一种阐明持续性问题的方法，这些问题可能指出了演示系统的弱点，或成为系统开发新迭代的动机（或

非正式的错误报告)。

*4) 场景 5：* 说出与情境不符的回应

我被大学邀请参加一个关于人工智能的小组讨论，另一位小组成员带来了一台通常在当地科技博物馆使用的机器人，以展示大型语言模型（LLM）如何应用于机器人。在为小组讨论做准备时，我与机器人进行了交流，向它打招呼并询问它能做些什么。由于环境非常动态且嘈杂，我意识到机器人可能无法应对这种情况。尽管我预计机器人可能不会回应，或者会说没有听清我说的话，我仍然感到惊讶，机器人在嘈杂的环境中产生了回应。当它回应时，很明显它没有正确地听到我说的话，因为它回应说“我也爱你”。尽管当时我对这个不合时宜的话语一笑置之，但当我后来回想这件事时，这个场景又浮现出来。如果旁边的程序员（就站在我身边）在这种情况下说出这句话，机器人的行为可能会被归类为骚扰。虽然机器人误听和产生不合时宜的回应是常见的，即使在有脚本的机器人中也是如此，但这个回应是由大型语言模型生成的事实引发了新的、不熟悉的感觉。在这种情况下，程序员是否应该受到指责？在我的话语和机器人的大型语言模型之间发生了什么，使得这被认为是一个合适的回应？

*a) 反射：* 这个小插曲说明，大型语言模型（LLMs）可能会呈现出情境上不合适的内容，而用户可能不明白为什么会发生这种情况。虽然一个编程好的博物馆机器人可能不会说“我爱你”，但 LLM 很容易产生一些违背其所在环境规范的回应。在 LLM 的情况下，责任归属完全不透明——是机器人的开发者，还是 LLM，或者说是机器人重现了其训练数据中存在的刻板印象？可以设定界限以防止出现可能不当的陈述，但我们希望通过这个例子说明，一个能够更加轻松和主动响应的机器人，也有更高的风险生成不合情理或可能被视为骚扰的言论，而这些言论无法在自动语音识别中轻易被解释为错误或识别出来。

*5) 场景 6：* 行为不符合情境: 在我参加的一次会议期间，一个配有显示屏的人形机器人被放置在大堂，以提供有关地点、会议和活动的信息。在休息期间，我决定与机器人互动，以获取一些关于当天会议活动的信息。

在接近机器人并受到欢迎后，我询问了有关会议的信息。作为回应，机器人展示了一长串符合该主题的新选项，同时也给出了口头回答。我决定在听取回答的同时先查看这些选项。然而，我没有注意到机器人在讲话时使用了手势。由于我站在机器人附近并触摸它的平板电脑浏览选项，机器人用其中一个手势打到了我。虽然这个动作力度不够大，不至于伤到我，但还是让我感到惊讶，因为我没有留意到机器人之前已经做过的一些小的不规则动作。在这之后，机器人没有承认刚刚发生的事情，而是继续进行下去。

这个小插曲表明，语音人机交互不仅仅是语言交流——它也是语言与其他模式的结合。机器人使用语言与用户交流，但对于选择预期用户在平板上按按钮。即使这些其他模式与语音完全断开使用，也不意味着它们不会对语音产生影响。在所描述的情况下，完全忽略对用户的打击使得整个互动显得很不真实，因为机器人显然打破了社会规范，先是打了我，然后又没有承认其不当行为。

*6) 要点：* 场景 5 和 场景 6 演示了在向公众展示系统时出现的各种问题。展示的系统必然是新的且开发程度较低，与潜在用户习惯使用的那些更成熟的系统不同。这意味着设计者和创作者没能预料的问题将会出现——就像我们在小插曲中看到的那样。这样的观察可能会激发未来的研究方向，因为它们唤起了关于互动应该如何展开的更普遍的问题。然而，目前这些问题可能不会被报告，因为很难控制演示环境、记录发生的情况（如果设计者一开始就看到了事件），并管理参与者的同意。形成小插曲可以成为一种捕捉事件核心并激发讨论和反思的方式。

## III. 讨论

民族志小插曲可以同时作为报告设计问题和系统测试时意料之外问题的一种方式。它们可以作为报告支持设计决定的经验的一种方式，以解决常见的反复出现的问题。多学科的视角尤其有助于找出一个旨在实现目的的系统为何未能实现，特别是由于设计者、用户和系统之间缺乏共同基础 [1, 15] 而导致的复杂故障。

半结构化访谈帮助我们获取参与者的非结构化想法，而无需事先知道这些想法是什么。同样，人种志小插图可以描述我们在实验交互过程中未准备好的情况。从开发的角度撰写的小插图可以让我们分享有趣的观察，这些观察本来不会成为计划评估的一部分。它们以描述性的方式提供信息，便于其他学科的研究人员理解。

展示系统交互场景需要付出努力、时间和占用出版空间。这需要权衡这些场景的益处，我们认为，许多像我们在第 **??** 节中展示的场景，并不像系统设计者认为的那样显而易见。人种志的片段不应取代定量分析，而应通过更细致的观察来补充。在系统的定性评估中，场景可能比调查中汇总的用户引用提供更多的洞察。

在有严格页数或字数限制的论文中发布插图是具有挑战性的。它们可能最终占据与一篇完整论文同样多的空间，但通常不被认为具有同等重要的贡献。我们建议将它们放在补充材料中，以增加研究目标的背景。有些出版机构已经开始引入额外的版块，系统行为插图本身可以成为一个很好的贡献。像 HRI 的视频版块或 CHI 的案例研究版块这样的版块，分别可以作为报告 HRI 和 HCI 实际经验的合适替代方案。

## IV. 设计建议

第 III 节为我们提供了未来在人机交互研究中的两个主要研究和工作流程设计建议。尽管最近的一些提议已经推动了更高标准的报告方法、结果、排除标准和招募程序在 HRI 中的应用 [2] ，我们的重点是改进研究设计和报告，包括那些目前尚未报告的内容。

### A. 关于系统能力的透明度

人际互动（HHI）的质量根据所执行的任务以不同方式进行评估，并可以被视为一种合作活动，其中的话语提供了必要的信息。这是可行的，因为对于许多互动，我们可以立即对目标作出假设。

我们场景中的不足之处凸显了 HHI 和 HRI 之间的差异。我们可以将糟糕的 HHI 归咎于某种程度上的共同基础的缺乏——例如缺乏共同的知识、性格差异或观察到的不同 [15] 。这些来自 HHI 的假设对于机器人来说并不成立，因为人类不能假定一个共同基础甚至存在 [47] 。在进行 HHI 的人类拥有即时解决和修复互动问题的机制 [36] ，而 HRI 中的问题可能源于错误的语音识别 [37, 55] 或对话模型 [33] 。

因为 HRI 涉及人类，所以交互总是不可预测的。这可能源于对机器人应如何响应、其能力，以及用户与机器人

[16, 70] 关系的先入为主的观念。我们的小插曲表明，这些期望往往没有被充分考虑，而且没有简单的方法去识别和处理它们。机器人的表现，包括外观和环境，极大地影响了对交互和机器人的能力的期望。因此，我们可以将这些用作基础机制。将用户的期望与机器人的能力相匹配是良好交互的前提，但如果能力非常有限，仍可能导致不良的交互。

用户期望可以通过实验前问卷调查来了解，类似于收集人口统计信息。在实验室环境中，我们建议研究更加关注定性指标，例如参与者访谈、共同设计甚至非正式讨论，作为与定量调查同等水平的分析形式。通过这些，我们可以更好地理解可能随时间变化的用户期望。

### *B.* 报告意外行为

正如在第 II-B 节中所强调的，在文献中已经使用了一些 HRI 评估标准。由于这些是经过验证的度量标准，可用于不同背景下的比较，因此在该领域是必需的。另一方面，许多定量指标无法捕捉第 **??** 节中描述的交互部分，因为它们可能不会频繁发生。这些情况可以用来描述机器人或交互设计中的缺陷，这些缺陷在 HRI 对话系统的进一步开发中仍需解决。因此，我们呼吁一种批判性视角，不仅呈现使用新技术（如 LLMs）的优势和成功之处，还要认真讨论其挑战。

我们发现撰写小插曲的方法对于讨论我们经历的系统意外行为非常有用。构建和叙述一个场景使我们能够更具体地讨论为何这种交互被如此感知，以及其潜在原因可能是什么。我们建议，HRI 和对话系统社区可以通过这种方法从更大规模地展示系统故障中受益，而不仅仅限于目前的做法。

## V. 结论

在未来，我们希望人与机器人的互动能够自然、有效、诚实，并明确其目的、目标和能力。这些并不是容易衡量或评估的互动方面，但通过将人机交互研究报告为机器人行为的研究，包括其失败，而不是作为功能完善系统的宣传材料，我们可以朝着这些目标前进。

## VI.

致谢 感谢 Felix Gervits 对本文早期版本提供的重要反馈。我们还要感谢在 场景 5 和 场景 6 中提到的系统的设计者，感谢他们允许我们在本文中将其系统用作例子。该研究由 混合智能中心 、一个为期 10 年、通过荷兰科学研究组织由荷兰教育、文化和科学部资助的项目资助，资助编号为 024.004.022。此外，该工作得到了 JST Moonshot R & D JPMJPS2011 支持。

## References

[1] Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. On the Semantics and Pragmatics of Linguistic Feedback. Journal of Semantics , 9(1):1–26, 01 1992. ISSN 0167-5133. doi: 10.1093/jos/9.1.1. URL https://doi.org/10.1093/jos/9.1.1.

[2] Shelly Bagchi, Patrick Holthaus, Gloria Beraldo, Emmanuel Senft, Daniel Hernandez Garcia, Zhao Han, Suresh Kumaar Jayaraman, Alessandra Rossi, Connor Esterwood, Antonio Andriella, and Paul Pridham. Towards Improved Replicability of Human Studies in Human-Robot Interaction: Recommendations for Formalized Reporting. In Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction , HRI '23, page 629–633, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399708. doi: 10.1145/3568294.3580162. URL https://doi.org/10.1145/3568294.3580162.

[3] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International journal of social robotics , 1(1):71–81, 2009.

[4] Stephanie Behm Cross. Whiteness in the academy: using vignettes to move beyond safe silences. Teaching in Higher Education , 22(7):879–887, October 2017. ISSN 1356-2517, 1470-1294. doi: 10.1080/13562517.2017.1340266. URL https://www.tandfonline.com/doi/full/10.1080/13562517.2017.1340266.

[5] Kathleen Belhassein, Víctor Fernández-Castro, Amandine Mayima, Aurélie Clodic, Elisabeth Pacherie, Michèle Guidetti, Rachid Alami, and Hélène Cochet. Addressing joint action challenges in HRI: Insights from psychology and philosophy. Acta Psychologica , 222:103476, 2022. ISSN 0001-6918. doi: https://doi.org/10.1016/j.actpsy.2021.103476. URL https://www.sciencedirect.com/science/article/pii/S0001691821002262.

[6] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency , FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

[7] James M. Berzuk and James E. Young. More Than Words: A Framework for Describing Human-Robot Dialog Designs. In Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction , HRI '22, page 393–401, New York, USA, 2022. IEEE Press.

[8] Anna Bloom-Christen and Hendrikje Grunow. What's (in) a Vignette? History, Functions, and Development of an Elusive Ethnographic Sub-genre. Ethnos , 0(0):1–19, March 2022. ISSN 0014-1844, 1469-588X. doi: 10.1080/00141844.2022.2052927. URL https://www.tandfonline.com/doi/full/10.1080/00141844.2022.2052927.

[9] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The Foundation Model Transparency Index, 2023.

[10] Simone Borsci, Alessio Malizia, Martin Schmettow, Frank Van Der Velde, Gunay Tariverdiyeva, Divyaa

Balaji, and Alan Chamberlain. The Chatbot Usability Scale: the design and pilot of a usability scale for interaction with AI-based conversational agents. Personal and Ubiquitous Computing , 26:95–119, 2022.

[11] C. Breazeal. Social interactions in HRI: the robot view. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) , 34(2): 181–186, 2004. doi: 10.1109/TSMCC.2004.826268.

[12] Daniel J Brooks. A human-centric approach to autonomous robot failures . PhD thesis, University of Massachusetts Lowell, 2017.

[13] Joana Campos, James Kennedy, and Jill F. Lehman. Challenges in Exploiting Conversational Memory in Human-Agent Interaction. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems , AAMAS '18, page 1649–1657, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

[14] Xiaoyu Chang, Yanheng Li, Sijia Liu, Ling Ma, and Ray Lc. "Sorry to Keep You Waiting": Recovering from Negative Consequences Resulting from Service Robot Unintended Rejection. In Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction , HRI '24, page 96–105, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703225. doi: 10.1145/3610977.3634959. URL https://doi.org/10.1145/3610977.3634959.

[15] Herbert H Clark. Using language . Cambridge University Press, Cambridge, UK, 1996. ISBN 9780521567459.

[16] Herbert H. Clark and Kerstin Fischer. Social robots as depictions of social agents. Behavioral and Brain Sciences , 46:e21, 2023. doi: 10.1017/S0140525X22000668.

[17] Enrique Coronado, Takuya Kiyokawa, Gustavo A Garcia Ricardez, Ixchel G Ramirez-Alpizar, Gentiane Venture, and Natsuki Yamanobe. Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an industry 5.0. Journal of Manufacturing Systems , 63: 392–410, 2022.

[18] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S. Melo, and Ana Paiva. Exploring the Impact of Fault Justification in Human-Robot Trust. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems , AAMAS '18, page 507–513, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

[19] Andy Crabtree, Mark Rouncefield, and Peter Tolmie. Doing design ethnography . Springer, London, 2012. ISBN 978-1-4471-2726-0.

[20] Viviene E. Cree. 'I' d Like to Call You My Mother.' Reflections on Supervising International PhD Students in Social Work. Social Work Education , 31(4):451–464, June 2012. ISSN 0261-5479, 1470-1227. doi: 10.1080/02615479.2011.562287. URL http://www.tandfonline.com/doi/abs/10.1080/02615479.2011.562287.

[21] Mirjam De Haas, Veerle Hobbelink, and Matthijs Smakman. Inclusive Dialogues: WokeBot Engaging Diversity Dilemmas. In Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction , HRI '24, page 379–382, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703232. doi: 10.1145/3610978.3640650. URL https://doi.org/10.1145/3610978.3640650.

[22] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. Survey on evaluation methods for dialogue systems. Artificial Intelligence Review , 54:755–810, 2021.

[23] Munjal Desai, Mikhail Medvedev, Marynel Vázquez, Sean McSheehy, Sofia Gadea-Omelchenko, Christian Bruggeman, Aaron Steinfeld, and Holly Yanco. Effects of changing reliability on trust of robot systems. In Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction , HRI '12, page 73–80, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310635. doi: 10.1145/2157689.2157702. URL https://doi.org/10.1145/2157689.2157702.

[24] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. Transactions of the Association for Computational Linguistics , 9: 1012–1031, 2021.

[25] Robert M Emerson, Rachel I Fretz, and Linda L Shaw. Writing ethnographic fieldnotes . University of Chicago press, Chicago 60637, 2011.

[26] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Amal Abdulrahman, and Willem-Paul Brinkman. What Are We Measuring Anyway? - A Literature Survey of Questionnaires Used in Studies Reported in the Intelligent Virtual Agent Conferences. In Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents , IVA '19, page 159–161, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366724. doi: 10.1145/3308532.3329421. URL https://doi.org/10.1145/3308532.3329421.

[27] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. The Artificial-Social-Agent Questionnaire: Establishing the Long and Short Questionnaire Versions. In Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents , IVA '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392488. doi: 10.1145/3514197.3549

612. URL https://doi.org/10.1145/3514197.3549612.

[28] Jodi Forlizzi, Carl DiSalvo, and Francine Gemperle. Assistive robotics and an ecology of elders living independently in their homes. Human–Computer Interaction , 19(1-2):25–59, 2004.

[29] Mary Ellen Foster, Andre Gaschler, Manuel Giuliani, Amy Isard, Maria Pateraki, and Ronald P.A. Petrick. Two People Walk into a Bar: Dynamic Multi-Party Social Interaction with a Robot Agent. In Proceedings of the 14th ACM International Conference on Multimodal Interaction , ICMI '12, page 3–10, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314671. doi: 10.1145/2388676.2388680. URL https://doi.org/10.1145/2388676.2388680.

[30] Matthew L. Ginsberg and David E. Smith. Reasoning about action II: The qualification problem. Artificial Intelligence , 35(3):311–342, 1988. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(88)90020-3. URL https://www.sciencedirect.com/science/article/pii/0004370288900203.

[31] Erving Goffman. On Fieldwork. Journal of Contemporary Ethnography , 18(2):123–132, July 1989. ISSN 0891-2416, 1552-5414. doi: 10.1177/089124189018002001. URL http://journals.sagepub.com/doi/10.1177/089124189018002001.

[32] H P Grice. Logic and Conversation. Syntax and Semantics , 3:41 – 58, 1975.

[33] Shanee Honig and Tal Oron-Gilad. Understanding and resolving failures in human-robot interaction: Literature review and model development. Frontiers in Psychology , 9, 2018. ISSN 1664-1078. doi: 10.3389/fpsyg.2018.00861. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00861.

[34] An Jacobs, Shirley A. Elprama, and Charlotte I. C. Jewell. Evaluating Human-Robot Interaction with Ethnography. In Céline Jost, Brigitte Le Pévédic, Tony Belpaeme, Cindy Bethel, Dimitrios Chrysostomou, Nigel Crook, Marine Grandgeorge, and Nicole Mirnig, editors, Human-Robot Interaction: Evaluation Methods and Their Standardization , pages 269–286. Springer International Publishing, Cham, Switzerland, 2020. ISBN 978-3-030-42307-0. doi: 10.1007/978-3-030-42307-0_11. URL https://doi.org/10.1007/978-3-030-42307-0_11.

[35] Alice Juel Jacobsen. Vignettes of interviews to enhance an ethnographic account. Ethnography and Education , 9(1):35–50, 2014. doi: 10.1080/17457823.2013.828475. URL https://doi.org/10.1080/17457823.2013.828475.

[36] Gail Jefferson. Repairing the Broken Surface of Talk: Managing Problems in Speaking, Hearing, and Understanding in Conversation . Oxford University Press, 198 Madison Avenue, New York, NY 10016, USA, 2018. ISBN 978-0-19-069796-9.

[37] James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction , HRI '17, page 82–90, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450343367. doi: 10.1145/2909824.3020229. URL https://doi.org/10.1145/2909824.3020229.

[38] Casey Kennington, Malihe Alikhani, Heather Pon-Barry, Katherine Atwell, Yonatan Bisk, Daniel Fried, Felix Gervits, Zhao Han, Mert Inan, Michael Johnston, Raj Korpan, Diane Litman, Matthew Marge, Cynthia Matuszek, Ross Mead, Shiwali Mohan, Raymond Mooney, Natalie Parde, Jivko Sinapov, Angela Stewart, Matthew Stone, Stefanie Tellex, and Tom Williams. Dialogue with Robots: Proposals for Broadening Participation and Research in the SLIVAR Community, 2024. URL https://arxiv.org/abs/2404.01158.

[39] Callie Y. Kim, Christine P. Lee, and Bilge Mutlu. Understanding Large-Language Model (LLM)-powered Human-Robot Interaction. In Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction , HRI '24, page 371–380, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703225. doi: 10.1145/3610977.3634966. URL https://doi.org/10.1145/3610977.3634966.

[40] A. Baki Kocaballi, Liliana Laranjo, and Enrico Coiera. Measuring User Experience in Conversational Interfaces: A Comparison of Six Questionnaires. In Proceedings of the 32nd International BCS Human Computer Interaction Conference , HCI '18, Swindon, GBR, 2018. BCS Learning & Development Ltd. doi: 10.14236/ewic/HCI2018.21. URL https://doi.org/10.14236/ewic/HCI2018.21.

[41] Laura Kunold, Nikolai Bock, and Astrid Rosenthal-von der Pütten. Not All Robots Are Evaluated Equally: The Impact of Morphological Features on Robots' Assessment through Capability Attributions. ACM Transactions on Human-Robot Interaction , 12(1):1–31, 2023.

[42] Cheongjae Lee, Sangkeun Jung, Kyungduk Kim, and Gary Geunbae Lee. Hybrid approach to robust dialog management using agenda and dialog examples. Computer Speech & Language , 24(4):609–631, 2010.

[43] Pierre Lison. A hybrid approach to dialogue management based on probabilistic rules. Computer Speech & Language , 34(1):232–255, 2015.

[44] Paul Luff, Nigel Gilbert, and David Frohlich, editors. Computers and Conversation . Computers and People Series. Academic Press, London, 1990. doi: https://doi.org/10.1016/C2009-0-21641-2. URL https://doi.org/10.1016/C2009-0-21641-2.

[45] Matthew Marge, Carol Espy-Wilson, Nigel G. Ward,

Abeer Alwan, Yoav Artzi, Mohit Bansal, Gil Blankenship, Joyce Chai, Hal Daumé, Debadeepta Dey, Mary Harper, Thomas Howard, Casey Kennington, Ivana Kruijff-Korbayová, Dinesh Manocha, Cynthia Matuszek, Ross Mead, Raymond Mooney, Roger K. Moore, Mari Ostendorf, Heather Pon-Barry, Alexander I. Rudnicky, Matthias Scheutz, Robert St. Amant, Tong Sun, Stefanie Tellex, David Traum, and Zhou Yu. Spoken language interaction with robots: Recommendations for future research. Computer Speech & Language , 71:101255, 2022. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2021.101255. URL https://www.sciencedirect.com/science/article/pii/S0885230821000620.

[46] Amandine Mayima, Aurélie Clodic, and Rachid Alami. Towards Robots able to Measure in Real-time the Quality of Interaction in HRI Contexts. International Journal of Social Robotics , 14(3):713–731, 2022.

[47] Erik Miehling, Manish Nagireddy, Prasanna Sattigeri, Elizabeth M. Daly, David Piorkowski, and John T. Richards. Language Models in Dialogue: Conversational Maxims for Human-AI Interactions, 2024. URL https://arxiv.org/abs/2403.15115.

[48] Pierrick Milhorat, Divesh Lala, Koji Inoue, Tianyu Zhao, Masanari Ishida, Katsuya Takanashi, Shizuka Nakamura, and Tatsuya Kawahara. A conversational dialogue manager for the humanoid robot ERICA. In Advanced Social Interaction with Agents , pages 119–131. Springer, Cham, Switzerland, 2019.

[49] Bilge Mutlu and Jodi Forlizzi. Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction. In Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction , HRI '08, page 287–294, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580173. doi: 10.1145/1349822.1349860. URL https://doi.org/10.1145/1349822.1349860.

[50] Hannah R.M. Pelikan and Mathias Broth. Why That Nao? How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems , CHI '16, page 4921–4932, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858478. URL https://doi.org/10.1145/2858036.2858478.

[51] Kathrin Pollmann, Wulf Loh, Nora Fronemann, and Daniel Ziegler. Entertainment vs. manipulation: Personalized human-robot interaction between user experience and ethical design. Technological Forecasting and Social Change , 189:122376, 2023. ISSN 0040-1625. doi: https://doi.org/10.1016/j.techfore.2023.122376. URL https://www.sciencedirect.com/science/article/pii/S0040162523000616.

[52] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. Voice Interfaces in Everyday Life. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems , CHI '18, pages 640:1——640:12, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3174214. URL http://doi.acm.org/10.1145/3173574.3174214.

[53] Merle M. Reimann, Florian A. Kunneman, Catharine Oertel, and Koen V. Hindriks. A Survey on Dialogue Management in Human-robot Interaction. J. Hum.-Robot Interact. , 13(2), jun 2024. doi: 10.1145/3648605. URL https://doi.org/10.1145/3648605.

[54] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario. In Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9 , pages 42–52, Cham, Switzerland, 2017. Springer.

[55] Damien Rudaz and Christian Licoppe. Public Speech Recognition Transcripts as a Configuring Parameter in Human-Agents Interactions. In Proceedings of the 2023 MP-COSIN Workshop , September 2023.

[56] Damien Rudaz, Karen Tatarian, Rebecca Stower, and Christian Licoppe. From inanimate object to agent: Impact of pre-beginnings on the emergence of greetings with a robot. ACM Transactions on Human-Robot Interaction , 12(3):1–31, 2023.

[57] Selma Sabanovic, Marek P Michalowski, and Reid Simmons. Robots in the wild: Observing human-robot social interaction outside the lab. In 9th IEEE International Workshop on Advanced Motion Control, 2006. , pages 596–601, Cham, Switzerland, 2006. IEEE, IEEE.

[58] Alessandra Maria Sabelli, Takayuki Kanda, and Norihiro Hagita. A conversational robot in an elderly care center: an ethnographic study. In Proceedings of the 6th International Conference on Human-Robot Interaction , HRI '11, page 37–44, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450305617. doi: 10.1145/1957656.1957669. URL https://doi.org/10.1145/1957656.1957669.

[59] Emanuel A Schegloff. Sequencing in Conversational Openings. American Anthropologist , 70(6):1075–1095, August 1968. ISSN 00027294, 15481433. URL http://www.jstor.org/stable/669510.

[60] Emanuel A. Schegloff. Reflections on Quantification in the Study of Conversation. Research on Language and Social Interaction , 26(1):99–128, 1993. doi: 10.1207/s15327973rlsi2601_5.

[61] Soyeon Shin, Dahyun Kang, and Sonya Kwak. User-centered Exploration of Robot Design for Hospitals in COVID-19 Pandemic. In Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction , HRI '22, page 1040–1044, New York, USA, 2022. IEEE Press.

[62] Jack Sidnell and Tanya Stivers. The handbook of conversation analysis , volume 121. John Wiley & Sons, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK, 2012.

[63] Gabriel Skantze. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. Computer Speech & Language , 67:101178, 2021. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2020.101178. URL https://www.sciencedirect.com/science/article/pii/S088523082030111X.

[64] Kavyaa Somasundaram, Andrey Kiselev, and Amy Loutfi. Intelligent Disobedience: A Novel Approach for Preventing Human Induced Interaction Failures in Robot Teleoperation. In Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction , HRI '23, page 142–145, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399708. doi: 10.1145/3568294.3580060. URL https://doi.org/10.1145/3568294.3580060.

[65] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. Common Metrics for Human-Robot Interaction. In Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction , HRI '06, page 33–40, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595932941. doi: 10.1145/1121241.1121249. URL https://doi.org/10.1145/1121241.1121249.

[66] Wyke Stommel, Lynn de Rijk, and Roel Boumans. 'Pepper, what do you mean?' Miscommunication and repair in robot-led survey interaction. In 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) , page 385–392, 2022. doi: 10.1109/RO-MAN53752.2022.9900528.

[67] Lucy A. Suchman. Human-machine reconfigurations: Plans and situated actions . Cambridge university press, New York, USA, 2007.

[68] Jesse Thomason, Shiqi Zhang, Raymond Mooney, and Peter Stone. Learning to Interpret Natural Language Commands through Human-Robot Dialog. In Proceedings of the 24th International Conference on Artificial Intelligence , IJCAI'15, page 1923–1929, Washington, DC, USA, 2015. AAAI Press. ISBN 9781577357384.

[69] Lucien Tisserand, Brooke Stephenson, Heike Baldauf-Quilliatre, Mathieu Lefort, and Frédéric Armetta. Unraveling the thread: Understanding and addressing sequential failures in human-robot interaction anonymous. Frontiers in Robotics and AI , 11, 2024. ISSN 2296-9144. doi: 10.3389/frobt.2024.1359782. URL https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2024.1359782.

[70] Sherry Turkle. Authenticity in the age of digital companions. Interaction studies , 8(3):501–517, 2007.

[71] Joseph Weizenbaum. ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM , 9(1):36–45, 1966.

## Appendix

### A. 小插图撰写说明

我们发现以下步骤有助于指导我们的写作过程。

*a) 设置场景。:* 试着回忆观察到的实例发生的背景。它发生在哪里？当时谁在场？你以何种角色参与事件？以一种能让人想象与你同在现场的方式描述场景。

*b) 描述动作序列的细节。:* 这将是你小插曲的核心，你现在需要聚焦于一个特定情况，选择一个特定的时间点。描述与机器人互动的具体人物，尽量详细地回忆事件的顺序。你可以在非常确定某件事情以特定方式表述时使用直接引语（她说 "你好吗?"）。你也可以使用间接引语（她问我好吗）和改述（我们互相问候）。虽然这里的重点可能在对话上，但你也需要详细描述人们所做的事情，某些动作或面部表情。如果你描述机器人所做的事情，尽量避免说 "机器人表达很高兴"，而是更具体地描述——例如机器人微笑，播放音调升高的声音等等。然后，你可以在下一步采用一种更具解释性的立场。

*c) 反思是什么使这变得值得报告 :* - 通常这涉及讨论感受。是什么让这个事件显得值得讨论？是什么让它令人惊讶、沮丧或其他值得铭记的？对于谁来说？这是你从第一人称视角表示你对事件体验的地方。试图找出机器人在该情况下违反了哪些社会规范，对在场的人而言。通常明确表述在这种情况下应有的期望以及实际发生的事形成对比会有帮助。你和/或在场的人预期会发生什么？实际发生了什么，你和/或在场的人是如何反应的？

尤其是在修改您的小插曲时，尝试捕捉事件发展的顺序（通常您只需将一些句子移来移去）。什么事情先发生，接下来是什么？将程序员的视角排除在小插曲之外——尽量从社交角度描述发生了什么。如果你愿意，你可以在最后增加第四点，如果你真的想从机器人的角度解释发生了什么。在那一部分，你应该纯粹专注于机器人/设计者的视角，把用户和开发者的视角分开 [67] 。