自主性是一种后天习得的偏好：探索开发者对 GitHub 机器人的偏好

Amir Ghorbani*、Nathan Cassee†、Derek Robinson*、Adam Alami‡、
Neil A. Ernst*、Alexander Serebrenik†、Andrzej W.asowski§
* 维多利亚大学，加拿大 {amirrezaghorbani, drobinson, nernst}@uvic.ca
† 埃因霍温理工大学，荷兰 {n.w.cassee, a.serebrenik}@tue.nl
‡ 奥尔堡大学，丹麦 adal@cs.aau.dk§ 哥本哈根信息
技术大学，丹麦 wasowski@itu.dk

arXiv:2302.05048v1 [cs.SE] 10 Feb 2023

摘要——**软件机器人在协作式软件开发中发挥着重要作用，开发者对其采用有望提高生产力。既有研究发现，沟通过于频繁的机器人会令开发者感到恼怒，从而影响机器人的效用。然而，人机协作的其他哪些属性会影响开发者的偏好，以及这些属性可能产生何种影响，目前尚不清晰。本文的主要目标，是在 GitHub 拉取请求的背景下，探索影响开发者对人类与机器人交互偏好的特征。我们开展了一项探索性序贯研究，先进行访谈，随后开展基于情境短文的调查。我们发现，开发者通常偏好亲和型但自主性较低的机器人；不过，经验更丰富的开发者往往更偏好自主性更高的机器人。基于这些实证证据，我们建议机器人开发者增加机器人的配置选项，使个体开发者和项目能够根据自身偏好与项目文化对机器人进行最佳配置。**

索引词—**软件机器人、拉取请求、人的因素**

一、引言

软件机器人可以提高开发者生产力 [34], [52]。这或许就是机器人在协作式软件开发中日益普及的原因 [38]。例如，机器人可作为静态分析工具的前端，用于安全筛查和开发流程管理，比如关闭长期未处理的问题。然而，由于软件机器人相对较新，我们尚不了解哪些行为模式会让开发者接受或拒绝与机器人协作。现有研究侧重于理解特定机器人 [29], 、描述机器人类型 [45], [51] 和机器人行为 [34], [52] ，而非关注人机协作过程。虽然已知过于频繁的通知会惹恼人类 [53], ，但机器人通信的其他哪些特征会影响开发者，目前尚不清楚。

在这项工作中，我们探讨**特定机器人行为如何影响开发者对机器人的感知**，旨在**为设计行为良好的软件机器人提供依据**。我们首先开展一项范围较广的探索性访谈研究（第三节），以探讨以下研究问题：

研究问题 *1*：哪些机器人特征会影响人类对机器人行为的感知？

我们的访谈确定了受访者所熟悉的机器人。随后，我们深入探究了受访者偏好的机器人行为方面。受访者表示，即使对于执行极其有用任务的机器人，其某些行为方面也会影响开发者对机器人的看法。我们的访谈促成了一组主题，突出了若干重要因素，包括机器人角色（任务）、机器人自主行动的程度（自主性），以及机器人呈现的人格 。由于 Erlenhov [15], Wessel [51], 和其他人已经对机器人角色（任务）进行了充分研究，我们将本研究的第二阶段重点放在自主性和人格上。遵循探索性序贯设计 [11], ，我们提出了第二个研究问题：

研究问题*2*：自主性（或人格）的程度如何影响开发者在拉取请求讨论中对机器人操作的偏好？

为回答这一问题，我们采用定制的情境短文式工具开展了一项调查，以评估两个自变量——人格和自主性——对用户偏好的影响（第四节）。调查中的每段情境短文都向受访者并列呈现两个几乎相同的情境。这两个情境展示了一个机器人在 GitHub 软件项目的拉取请求讨论中进行交互。关键在于，我们改变机器人的行为，以对比自主性或人格的不同方面，使情境短文的两侧分别代表该构念的不同极端。随后，我们询问开发者更偏好哪一侧。

本研究的主要贡献包括：

- 一组源自从业者访谈的主题，揭示了影响人类与机器人交互感知的关键方面。
- 对机器人自主性和人格构念的识别与定义，并对这些构念进行了具体的操作化。
- 评估自主性和人格的影响，采用定制设计的情境短文式工具，以提高机器人行为研究的真实性并控制相关变量。
- 一组经过统计验证的模型，为开发者对特定机器人行为的偏好提供了定量证据。研究结果表明，开发者通常更偏好反应式、亲和型机器人。
- 基于我们的实证证据提出可行的建议。我们的建议为未来在软件开发活动中使用机器人指明了方向。我们建议，机器人的自主性和人格

应保持灵活，并由开发者根据自己的偏好和当前任务进行调整。

我们的研究发现为后续研究开辟了新方向，可进一步探索如何为特定项目设计机器人角色形象，以及如何针对不同开发者人口统计群体进行调整。这些发现表明，研究应聚焦于可接受的自主性。

II. 背景与相关研究

Storey 和 Zagalsky [45] 将软件机器人定义为“用户与服务之间的通道或接口，通常通过对话式用户界面（UI）实现”。交互和智能的概念是区分机器人与其他开发工具的主要主题 [15], [55]。事实上，Erlenhov等人发现，开发者对机器人的定义有三种不同理解 [15], ，具体取决于他们是否认为机器人主要用于聊天、执行智能操作，或自主行动。

机器人已得到广泛采用，并被认为很有用：Wessel等人以及 Peng 和 Ma 均发现，开源软件项目采用机器人后，合并的拉取请求数量会增加，从而提高工作分工效率 [52],[34]。然而，引入机器人也并非没有风险，因为 Wessel等人发现，开发者在描述机器人时，常常将噪声视为一个核心且反复出现的问题 [54]。为了解决这一问题，Wessel等人提议使用自动化审核器来过滤并汇总机器人行为 [50]。Wessel等人的研究聚焦于机器人在社交编程平台上的恼人行为，而我们的研究范围更具体，关注开发者偏好机器人展现哪些特定的机器人构念。

自主性和人格是访谈研究中提炼出的两个特征，并在实验调查中进一步探讨。

自主性。 虽然有些人可能希望软件机器人具有高度自主性，但他们在软件机器人用户中占少数[15]。Seiffer 等人 发现，过于自主的机器人会引发用户的怀疑，因为用户无法追踪和管理机器人的工作 [42]。Liao等人 在软件工程领域之外也发现了这种怀疑：曾使用自动化个人助理的专业人士不愿接受更主动的智能体，因为这存在打扰风险 [24]。在评估自动化助理时，Schaffer 等人 发现，专业人士自我报告的经验会影响参与者接受自动化代理协助的意愿 [40]。鉴于软件工程中已有高度主动式机器人的实例（例如，Dependabot），我们旨在了解开发者更偏好反应式还是主动式机器人，因为现有文献大多研究软件工程之外的自主性概念。与此同时，Erlenhov等人 [15] 和 Seiffer等人 [42] 的研究均未在实验设置中评估开发者的偏好。

人格。尽管许多软件机器人通过基于文本的对话式用户界面进行交流，但它们并不一定要采用科幻机器人典型的枯燥且分析型的人格，因为 Nass等人发现，人类会对计算机生成的线索表现出社会行为 [31]。Farah等人发现，软件机器人

表一 受访者人口统计信息

社区	ID	Role	经验	国家
FOSSASIA	P1	维护者	4年	印度
	P2	新成员	3 年	印度
	P3	贡献者	6年	印度
	P4	维护者	1.5年	印度
	P5	贡献者	5年	德国
	P6	维护者	6年	新加坡
	P7	维护者	4年	斯里兰卡
ROS	P8	贡献者	12年	USA
Coala	P9	贡献者	4年	印度
	P10	贡献者	2年	印度
	P11	贡献者	5年	日本
RTEMS 社区	P12	贡献者	3 年	印度

以双关语的形式运用幽默时，显得更加友好，也更富有个性 [16]。此外，表现出更高程度拟人化的 软件机器人通过营造熟悉感，能够与其人类对应者建立更多信任。这一点在软件工程领域 [42],[23], [35] 内外 [22], [9], [8]均已得到证实。然而，机器人可能并不总是需要具备更高程度的拟人化，因为 Clark等人发现，人类仍会将机器人以及与机器人的交互视为与人类交互在根本上不同的事物 [10]。尽管我们知道，拟人化机器人能够建立更多信任，也显得更加友好，但开发者是否偏好使用更多亲和语言的机器人，尚未得到研究。尤其是在模拟 GitHub 界面的实验设置中，让开发者直接比较亲和型机器人和事实型机器人，并要求他们表达偏好的研究，更是如此。

III. 第一阶段：对机器人行为的认知

我们采用了探索性序贯设计 [11]来回答两个研究问题。定性阶段（访谈）产生了在定量阶段（第四节，调查）中进行研究的主题。本节介绍的第一阶段通过访谈初步探讨了人类对机器人的特征和行为的认知。

第一阶段的研究结果指导了第二阶段的设计（第四节）。这一设计（定性 → 定量）使我们能够在第一阶段得出有依据且可靠的研究结果。在第二阶段，我们使用调查工具和更大的样本来拓展实证覆盖范围、控制无关变量，并检验前期提出的一些论断。由于目前尚无关于软件机器人行为和开发者感知的清晰理论，我们将本研究的目标定位为探索性研究。

A. 访谈开发者

为了了解开发者对开源社区中机器人行为的感知并回答研究问题 *1*，我们访谈了来自四个社区的十二名开源软件开发者。

抽样与招募： 我们通过便利抽样 [33]招募了访谈研究的参与者。我们请三位来自 FOSSASIA、ROS 和 Coala 的维护者协助我们吸引各自社区的贡献者和维护者

表 II 最重要的访谈问题。Q1 和 Q2 是简短的引导性问题，此处未列出。

核心问题
Q3 - 您在社区中使用机器人吗？以前使用过机器人吗？出于什么目的？如何使用？
Q4 - 在为社区作贡献时，与机器人或人类打交道对您来说有区别吗？有什么区别？
Q5 - 如果机器人立即拒绝了你的拉取请求，你会有什么感受？
Q6 - 如果机器人立即接受了你的拉取请求，你会有什么感受？
Q7 - 机器人提交拉取请求时，你会有什么感受？
Q8 - 具体来说，你希望机器人如何行动？
Q9 - 你认为机器人在你的项目/社区中有哪些发展机会？
追问问题示例
Q10 - 您能否分享一个您所在社区中机器人的示例 []？ Q11 - 如果机器人要接受拉取请求 [，例如，您在审查流程中的态度], 会改变吗？ Q12 - 您是否遇到过类似情况 [，即机器人拒绝或接受了拉取请求]？您能否与我们分享？ Q13 - 机器人应该怎样做才不会让您感到厌烦？

社区。我们与所选社区保持着长期联系，并请联系人协助招募参与者。我们准备了一份参与邀请，并通过两位联系人（FOSSASIA 和 Coala）发送给一组贡献者和维护者。相比之下，我们在 ROS 的联系人推荐了一些可能的候选人。

我们联系了四位可能参与 ROS 访谈的参与者，而我们在 FOSSASIA 和 Coala 的联系人分别代表我们向社区贡献者和维护者发出了十二份和十份邀请。我们成功访谈了一位 ROS 参与者、七位 FOSSASIA 参与者和三位 Coala 参与者。有一位受访者同时活跃于 Coala 和 RTEMS 社区，但选择认同后者。

参与者人口统计信息： 表 I 展示了受访者的人口统计信息。第一列是参与者所贡献的社区。第三列是他们在各自社区中的角色（贡献者或维护者），“经验”是他们为社区作出贡献的累计年数。“国家”是他们居住的国家。

数据收集：为了兼顾获取丰富数据的需求和访谈过程中的专注度，我们选择采用半结构化访谈收集第一阶段的数据。我们使用预先制定的访谈指南对访谈进行半结构化设计（表 II），并准备了追问问题（例如“不会让你感到厌烦”），以促使参与者进一步详细讲述他们的经验。

这项访谈研究旨在收集数据，了解贡献者和维护者在各自社区的拉取请求（PR）流程中使用机器人的经验。这是一个相关的数据来源，而这种方法使我们能够获得一种本位 视角，即人们赋予事件、关系、行为和经验的意义所体现出的观点。通过半结构化访谈收集的数据，能够提供关于实践中发生了什么以及人们如何看待周围事件的内部信息或知识，而这些信息或知识很难通过其他方式获得 [47]。

表 II 列出了访谈中的关键问题。为避免社会期许性回答（即提供研究人员想听的回答） [18] 和抽象的回答，我们设计了旨在获取细致、具体答案的问题（例如 Q5–Q8），以促使受访者分享与其经验相关的事例。详细的访谈指南可在复现包中查阅（第七节）。我们使用虚拟会议工具 Zoom 开展访谈。每次访谈结束后，我们使用在线转录工具 Temi[1] 转录录音。如果逐字稿内容不清晰，我们会对照录音进行人工核查。访谈平均持续 60 分钟，访谈转录稿平均长 20 页。我们获得了许可，可将访谈的匿名版本作为复现包的一部分公开。访谈于 2020年十月至十二月期间进行。

B. 访谈记录编码

内容分析是对语言或行为数据进行分类，以便对数据进行归类、总结和制表的过程。我们选择了归纳分析方法 [28]。当人们对研究现象了解有限时，通常会采用这种方法。归纳分析旨在从收集到的数据中提炼含义，以识别模式和关系，并据此构建理论。根据建议 [28]，本研究采用了两步分析过程。

步骤*1*：构建编码：在编码的初始阶段，我们逐行对数据进行编码，并以研究问题 *1* 为分析视角，对访谈文本进行解释性阅读。同时，我们为文本片段赋予有意义的编码。第一步编码的成果是一本编码手册，供后续编码阶段使用。我们使用 ATLAS.ti[2]导入访谈转录稿并管理编码手册。完整的编码列表见我们的复现包（第七节）。

步骤*2*：编码分类：我们在前一阶段识别出的编码中发现了模式，并将其归类为不同主题。本文的两位作者使用 ATLAS.ti，按照上述归纳式编码方法开展工作。总共进行了六轮编码构建（步骤1）。前三轮中，两位作者各自对一份访谈转录稿进行编码，之后举行讨论会。第四轮没有对访谈转录稿进行编码，但作者们合并了编码并修订了编码列表。第五轮中，一位作者完成了剩余访谈的编码；另一位作者则在第六轮完成了相同工作。第六轮结束后，两位作者共同对编码进行分类，形成最终的编码列表。

由于在这一编码过程的每一轮中，都会组织编码者之间的复盘会议，以讨论

[1]https://www.temi.com [2] https://atlasti.com

主题（频次）	定义	数据中的证据
态度 (104)	人类如何看待机器人行为	“所以，我完全不介意 *[*机器人*]*进行审查，也不介意机器人进行审查并说，你知道，这样不行”(P10).
自主性 (41)	是否必须由人类来调用/管理 bot	“... 如果 *[*机器人*]* 只是直接告诉你出了问题，那还不如它能够比如，替你修好它”(P8).
		“机器人的好坏，取决于它为你提供的自动化程度”(P8).
人格 (67)	人类感知到的机器人性格由人类进行	“它 *[*机器人*]* 应该有点酷。带点讽刺，讲些有趣的笑话，这样才能保持开发中的活力和热情”(P3).
任务 (144)	机器人可以执行的任务	“机器人已配置为自动运行各种测试。比如说，你创建了一个拉取请求，那么接下来会根据依赖项以及你的特定项目的 *CI/CD* 工作流来处理，无论有人提交了什么拉取请求，他们可能会使用一些自定义测试，例如 *Travis CI*。因此，自定义机器人可以运行专门为此构建的一些测试.....”那个特定的组织”（P1）。
感受 （28）	人类产生的情绪反应可能由机器人或人类引发	“.....如果是机器人 *[*在审查拉取请求*]*.....而且他 *[*原文如此*]*只是拒绝了我的拉取请求，那我会有点难过，而且我对那个机器人的看法也会因此改变。”（P3）。
项目规范(32)	开源软件项目中的隐性实践	“*RTEMS* 实际上是一个通过电子邮件开展工作的社区，目前我们不使用机器人”（P12）
角色 (14)	机器人或人类在项目	“那么，这个机器人可以进行 *[*风格检查*]*。向审查者请求审查很耗时..... 我们可以说，机器人已经成为审查者之一。” (P2).

表 III 主题、定义及其在数据中的依据

分歧、解决分歧，并就低层代码达成共识，因此评审者间信度测量不适用于这一建构主义和涌现过程 [1]。

C. 第一阶段的发现

访谈深入揭示了软件开发者如何看待软件机器人，并归纳出七个主题：态度、自主性、人格、任务、感受、项目规范和角色。表 III-B 对这些主题进行了定义。

受访者指出，机器人在他们的项目中发挥着重要的作用，主要被视为助手。机器人会极大地影响社区观感和项目规范，使人们产生各种感受，例如恼怒、气馁或开心等。探究这些感受背后的态度，最突出的因素是机器人的自主程度，以及它给人的印象。自主性和人格都会影响受访者对特定机器人的感受和态度。

例如，“我见过很多机器人都非常不近人情。机器人似乎就是被设计成不近人情的。我个人不喜欢这样”(P6)。无论机器人能完成什么任务，要发挥作用，就必须得到项目社区的接纳：“尽管 *[*这个机器人*]* 拒绝了我的拉取请求，但这说明它能够理解我的拉取请求有什么问题；可它还不够智能，无法在评论区写出问题所在，并等待整个开发过程完成”(P3)。

在第一阶段的数据中，**自主性** 和**人格** 这两个主题似乎在我们的访谈对象中引发了最强烈的反应。例如，机器人行为之后缺乏理性且基于事实的解释，可能会影响开发者对机器人行为的感受。如果机器人毫无理由地拒绝他的拉取请求，P3 会对此感到失望。他表示：“如果是机器人在执行任务，而机器人没有评论说明我犯了什么错误，只是简单地拒绝我的拉取请求，那么我会感到有些难过，而我对那个机器人的看法也会真正改变”（P3）。P9 表达了更强烈的反应：“我认为，作为一名新成员，这种感觉会糟糕得多，因为如果这是我的第一次贡献，而机器人拒绝了我，我就不会再为这个项目做贡献了。我与这个项目就此结束（P9）。

同样，过高的自主性会削弱参与贡献过程的乐趣。例如，P2 和 P6 解释说，如果他们的拉取请求被机器人接受，他们不会感受到同样的“喜悦”。他们解释道：“拉取请求的妙处在于，经过审查者提出许多建议后，最终获得合并时会让人感到喜悦。而这种喜悦确实无法与机器人在没有任何审查的情况下直接合并拉取请求相提并论。” (P2) 以及 “如果是人类接受了我的拉取请求，我可能会更高兴一些 *[sic]* ”(P6)。

我们得出结论：自主性和人格对塑造开发者对机器人的感知影响更大。因此，在第二阶段，我们决定聚焦这两个变量，并据此提出研究问题*2*，以进一步了解自主性和人格对开发者在拉取请求流程中对机器人的感知和预期所产生的影响。

IV. 第二阶段：测试对机器人的感知

我们设计了一项随机情景问卷，以探究自主性和人格如何影响更广泛开发者群体对机器人的看法，即它们如何影响机器人行为偏好这一因变量。情境短文描述了虚构场景，其中机器人在 GitHub 上与用户进行交互，而这种交互（交替地）体现了机器人的自主性和人格。该设计回应了研究问题2，并总结于图1中。调查能够更好地控制所提问题，并覆盖更广泛的受访者群体。我们选择将情景实验嵌入调查中，这种方法也称为

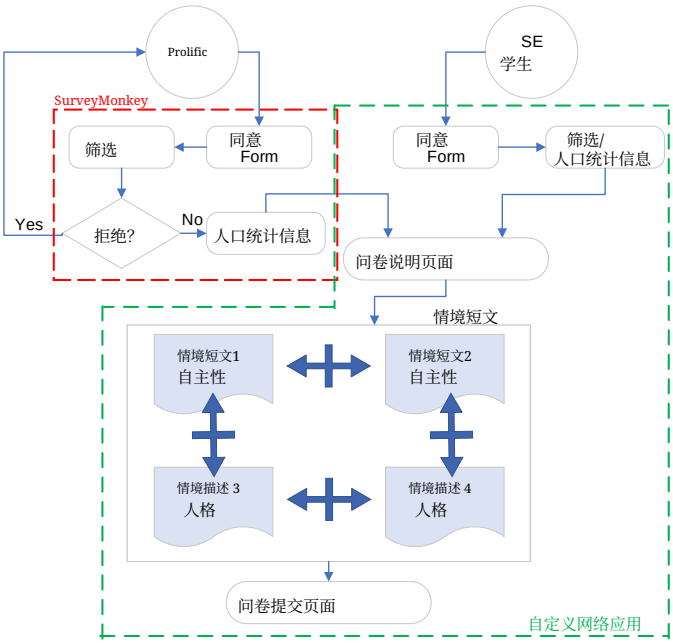


图1. 调查工具中的参与者流程。

析因调查，而非传统调查，因为它支持更高程度的真实性，允许系统地改变描述，并且感觉不那么单调 [44]。情境短文调查是一种成熟的社会学研究工具，由 Rossi 等人于20世纪70年代提出 [37]，近年来也被用于软件工程研究 [25], [27], [32], [39]。下面我们将讨论该调查的设计、实施和结果。

A. 目标总体

目标总体是了解拉取请求工作流程的软件开发者。显然不存在一份包含此类群体的抽样框的名单 [4]。我们采用非概率目的性抽样方法。虽然这会影响研究结果的普遍适用性，但在我们的探索性研究背景下，我们可以利用受访者特征进行推断，以探索差异。我们的目标对象是维多利亚大学软件工程专业三年级本科生（该校是一所位于北美的中型大学），以及 Prolific，[3] 这是一个在线参与者招募平台，借助内部筛选机制，让研究人员能够接触到世界各地的参与者。

B. 调查设计

调查以筛选问题和人口统计问题开篇，以确保 Prolific 的受访者属于目标总体，这也是 Danilova 等人和 Ebert 等人 [12], [13]的建议。最重要的问题见表 IV，筛选问题则收录于复现包中（第七节）。

此外，我们还筛选了受访者对编程和基于拉取请求的开发（例如，GitHub 账户、拉取请求定义）的了解情况。任何一题筛选问题答错，都会导致我们终止该受访者的调查。对于 Prolific，我们会将不符合条件的受访者返回 Prolific Study 页面以完成调查，这是 Prolific 对补偿的要求。随后，系统会随机排列四段简短的 情境短文 [27]，展示 GitHub 拉取请求界面中的机器人行为示例（如图 2 所示）。尽管访谈数据中频繁出现自主性和人格这两个主题，但在编写调查情境短文时，我们需要确定一种具体的操作化方式，以构建能够体现这些主题的构念。

构念——自主性：基于第一阶段访谈研究中提炼出的编码，我们将自主性构念界定为两个极端：主动式及其反面——反应式。主动式机器人能够在代码仓库内独立采取行动，无需人类触发。P7 解释道：“……但说到轻量级任务，比如做一些非常小的代码更改或文档修改……我们可以用机器人将这些任务自动化，而不是手动完成。”（P7）。相反，反应式机器人会先由人类提出协助请求，例如，由维护者手动触发机器人。Repairnator 机器人 [30] 能够在代码仓库中独立创建拉取请求，这就是主动型机器人的一个例子。一项近期关于基准测试机器人的研究发现了四种反应式机器人和五种主动式机器人 [26]。

构念——人格：我们只关注机器人通过文本表达体现出的人格，不考虑机器人的视觉外观（即，我们希望机器人使用中性的头像和名称）。我们将人格构念界定为两个极端。事实型机器人回复不包含任何拟人化细节，只是传达信息中的事实内容。P10 解释道：“……所以，首先，如果某件事是正确的，而机器人批准了它，它就应该使用友好的语言，比如说：‘嘿，你做得对，没问题。’如果某件事不对，而它拒绝了拉取请求，理想情况下，它会告诉我拒绝拉取请求的原因。”（P10）。P5 期望机器人完全客观；他们表示：“机器人没有任何感情之类的东西。它只是一段告诉你该做什么的软件。”（P5）。与之相反，我们将展现出类似人类人格的机器人称为亲和型机器人。这种机器人使用非正式且非常热情的文本。例如，P6 表示：“有些机器人甚至也有自己的个性，对吧？比如在末尾加个小表情符号，这种做法会带来很大的不同……它能让机器人多一些人性化因素。”（P6）。例如，在情境描述 3 中，亲和型机器人首先感谢贡献者提交拉取请求，然后称其“很有价值”。

我们为调查设定的最后一个构念，是受访者**对偏好**在给定情境短文中机器人行为的偏好。我们构建情境短文，以涵盖两个构念的两极（即事实型—亲和型和主动型—被动型）。我们将偏好的解释留给参与者，并通过开放式回答来描述他们如何说明自己的偏好。我们将其编码为一个三值变量：偏好左侧、中性、偏好右侧。

调查情境短文：在简答筛选和人口统计问题之后，我们向参与者展示了情境短文。情境短文是虚构场景，其中机器人会与用户在

[3]https://www.prolific.co/

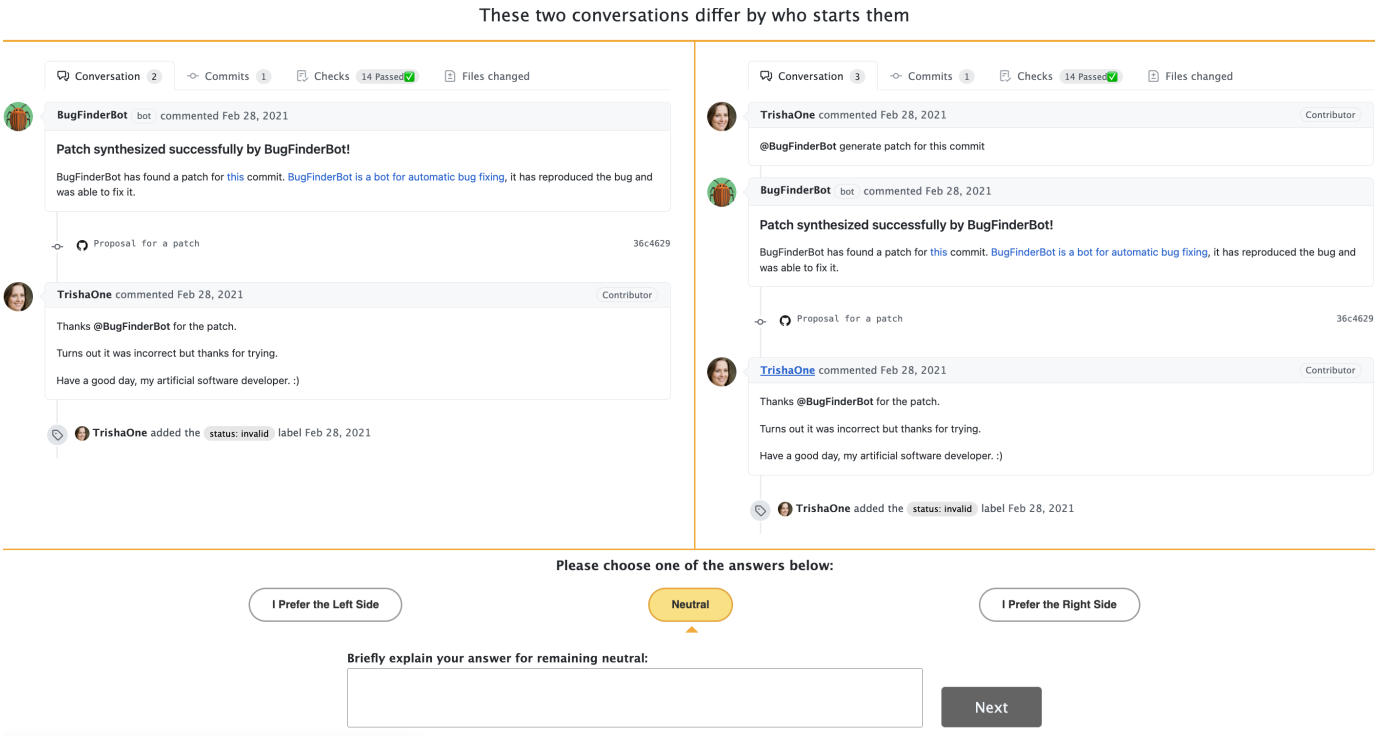


图 2. 展示情境短文 #1 的网络应用程序截图。为保护隐私，用户头像已遮挡。底部显示选项输入框和用于填写理由的文本输入框。中性为默认选项。

GitHub 上交互，交互内容交替体现机器人的自主性和人格。每篇情境短文都呈现了一段拉取请求讨论，为了增强真实感，这些内容取自现有的 GitHub 示例。

我们将机器人的头像设为中性，并更改 GitHub 贡献者的姓名以保护其隐私。我们还调整情境短文中的讨论顺序和/或机器人的消息，以突出正在测试的构念（自主性或人格）。因此，每个情境短文包含两个不同的情境，如图 2 所示：左侧对应构念的一个极端（例如主动型机器人），右侧对应另一个极端（例如响应型情形）。

我们会向受访者说明情境短文两侧有何不同。我们写道：“这两段对话的区别在于由谁发起”（自主性），以及“这两段对话的区别在于 [机器人]使用的文本”（人格）。我们尽量采用中性的措辞，以避免使受访者倾向于其中一方。这样的提示可以节省受访者寻找差异的时间，因为这些差异可能只是偶然出现的，同时也能让他们专注于正在研究的构念。我们请受访者仔细查看情境短文中的每种情境，然后指出他们更喜欢哪一种：左侧、都不喜欢、右侧。受访者还必须说明自己为何更喜欢所选的一方。

为确保情境短文中的两个自变量不会发生交互，每个情境短文只在一个构念上有所不同。例如，在图 2 最右侧的情境短文中，左侧和右侧情景分别展示了主动式机器人和响应式机器人的交互。两侧的机器人人格均固定为事实型。在这种情况下，只需确保机器人在每种情景中使用的文本完全相同即可轻松做到这一点。表 IV 下半部分对此进行了详细说明。

C. 部署调查

我们按照 Ebert等人 [13] 的建议，在数天内分批、于不同时间向参与者发送了这项调查的邀请。参与者完成调查后可获得每小时 7.5 英镑的报酬，即使未通过筛选也是如此。

我们的调查共收到 $N = 56$ 份有效答复（维多利亚大学 30 份，Prolific 26 份）。我们于 2022 年 2 月下旬发布了第一版调查。在受邀的 69 名维多利亚大学学生中，有 30 人选择参与调查。我们分 4 轮，在大约一个月内邀请了 300 名 Prolific 用户参与调查。在受邀的 300 名 Prolific 用户中，我们筛除了 274 名未能成功完成筛选问题的参与者，最终收到 26 份有效的 Prolific 答复。技术性较强的目标总体研究中，筛选成功率为 8.7% 并不少见。这有助于确保参与者符合要求。Danilova等人在一道较简单的技术问题中发现，成功率为 25% [12]。

34 名（61%）受访者是学生。19 人认同为女性，34 人认同为男性，2 人认同为非二元性别，另有 1 人未透露。34 名受访者年龄为18至24岁，18 人为25至34岁，4 人为35岁及以上。27 人有 1 至 3 年的软件开发经验，19 人有 3 年以上经验，10 人的经验不足一年。10 人曾与机器人交互。只有 6 人无法定义软件机器人。

表 IV 调查问题与情境短文

人口统计信息	**答案**
您最认同哪种性别认同？	以下选项之一：男性，女性，非二元性别，其他，倾向于不透露。
你多大了？	以下选项之一：未满18岁，18至24岁，25至34岁，35-44岁，45-54岁，55-64岁，65+
您的职业是什么？	打开输入框。
您有多少年工作经验从事您的职业？	以下选项之一：不到一年，一到三年，三到五年，超过五年。
你曾向任何开源项目提交过拉取请求吗？	以下选项之一：是，否。
你曾审查过任何开源项目的拉取请求吗？	以下选项之一：是，否。
您与代码仓库协作的频率是——GitHub 上的仓库，GitLab，Bitbucket，etc？	以下选项之一：至少每天一次，至少每周一次，At 至少每月一次，至少每六个月一次，至少每年至少一次，从未。
你是否曾在某个开源项目中与机器人互动？如果是，请填写至少一个此类机器人的名称。	打开输入框。
什么是机器人？请简要描述或提供一个定义。	打开输入框。
筛选	
筛选问题包含在随附的复现包（第七节）。	
情境短文	**变量**
情境短文 #1：在此情境短文中，BugFinderBot 为一次提交生成补丁，该提交位于一个拉取请求中。在左侧情境中，它生成了在合适的情境下自主提交由开发者手动触发。	变量：自主性，个性设定为事实型。
情境短文 #2：此情境短文中的情况与情境短文 #1 相似，唯一的不同之处在于机器人的表达较为非正式（亲和型）而情境短文 #1 中的表达则更为事实型的机器人写作。	变量：自主性，人格固定为亲和型。
情境短文 #3：一个 CLA 机器人回复一位新用户在发起拉取请求时，被要求签署一份贡献者许可协议。在左侧的情境中，机器人是事实型的，而在右侧则是亲和型的。	变量：人格，自主性固定为反应式。
情境示例 #4：在这个情境短文中，一位管家机器人回应用户发起拉取请求，告知他们代码仓库的规则，并提醒他们该拉取请求指向了错误的分支。	变量：人格、自主性——固定为主动式。

D. 描述性分析与结果

为了了解自主性和人格如何影响偏好（研究问题*2*），我们首先对结果进行描述性统计。随后，我们根据可能的预测变量（包括人格、自主性和 GitHub 活动）的不同表述形式，对偏好这一自变量的模型进行比较。

图 3 展示了总体选择分布。我们的受访者相较于主动式、事实型机器人，参与者更偏好反应式、亲和型机器人。若像表 V 所示，将各个情境短文分别分析，呈现出的情况略有不同。这一趋势仍然适用于

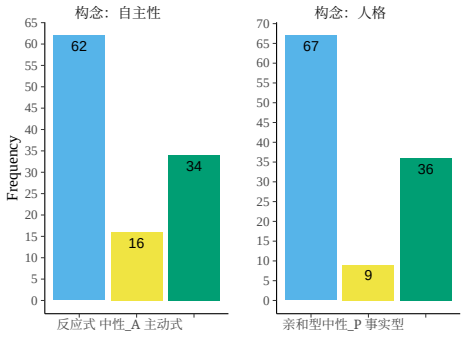


图 3. S　两种情境短文中受访者选择的汇总，按　　构念。

表 V 情境短文构念选择。

情境短文	构念	数量	情境短文	构念	数量
1	反应式	31	3	事实型	13
1	中性	6	3	中性	4
1	主动式	19	3	亲和型	39
2	反应式	31	4	事实型	23
2	中性	10	4	中性	5
2	主动式	15	4	亲和型	28

反应式机器人：在情境短文 1 和情境短文 2 中，绝大多数人都偏好反应式机器人。然而，在情境短文 3 中，人们明显更偏好亲和型机器人（39 人），而非事实型机器人（13 人）；在情境短文 4 中，这一优势则有所减弱（分别为 28 人和 23 人）。

我们还考察了编程经验的替代指标。编程经验有多种操作化方式 [43]，我们考察了几种不同的替代指标，它们可能为经验如何影响偏好提供互补见解：GitHub 活动水平、拉取请求经验，以及以编程年数衡量的总体经验。

图 4 通过按 GitHub 活动情况进行分组，进一步细化了分析。我们将每月使用 GitHub 超过一次的用户定义为有经验的GitHub 用户。不频繁用户虽然拥有 GitHub 账户，但每年使用 GitHub 不到一次。中等用户则介于这两种情况之间。有趣的是，专家对响应式和主动式机器人（情境短文 1 和 2）的偏好程度相同，而新手和中级用户则更倾向于响应式机器人（费舍尔精确检验，p=0.015, α=0.05）。Liao等人 [24]同样发现，专家用户较不愿接受自动化代理的输入，并且担心受到打扰，因此对主动式自动化代理持谨慎态度。

机器人使用经验也与此相关。我们要求参与者定义机器人；如果他们曾在开源项目中与至少一个机器人互动过，还要求他们至少列举一个，结果见表 IV。10 位受访者给出了一些机器人示例，44 位未回答这一非必答问题。6 位受访者给出的机器人定义表明他们并不确定。如果参与者能清晰地定义机器人，我们便将其归类为对机器人有中等了解（40/56）。如果他们还列出了曾与之互动过的机器人，我们则将其归类为有机器人使用经验（10）。其余

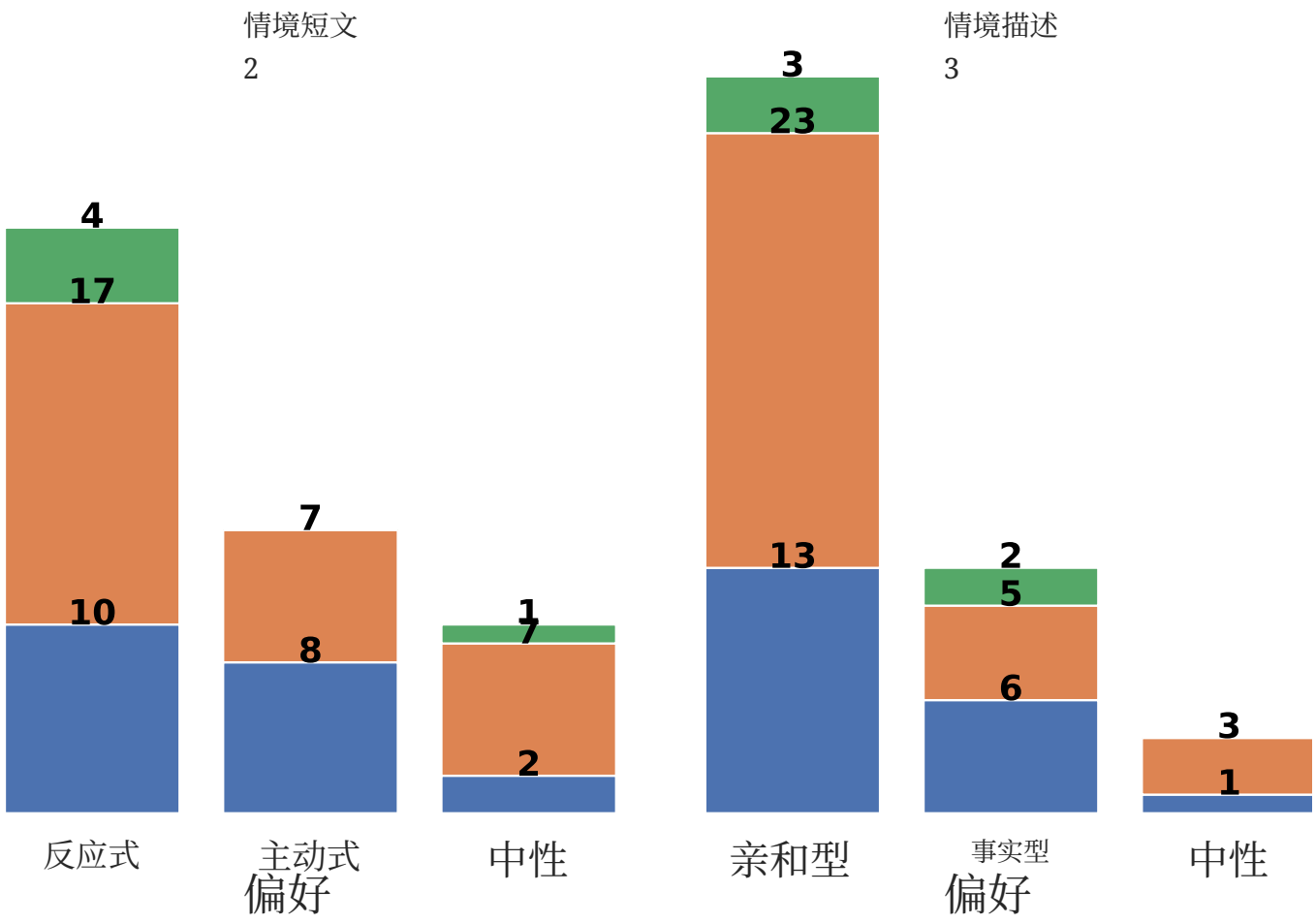


图 4. 基于 GitHub 活动的情境短文选择。

表 VI 关于理由的开放式问题的开放式编码结果。

Code	极性（频次）	定义
嘈杂度	嘈杂（4）/ 清晰（41）	机器人通过发送过多消息，使对话变得杂乱或频繁向开发者发送提醒，或者消息清晰、简短且易于理解。
人格	礼貌（40）/ 粗鲁（7）	该机器人很友好，其用语与人类对话相似，或者该机器人在消息和行动中不顾及人类情感。
控制	Bot-（30）/ 人类发起（41）	机器人处于控制之下，会发起操作，并且对这些操作负责，或开发者控制机器人、发起机器人操作，并且对机器人的行为负责
生产力提升	(2) / 损害 (1)	机器人高效且能改善工作流程通过节省时间和资源，或者机器人效率低下，消耗资源并浪费时间
信息内容	丰富（19）/ 贫乏（1）	该机器人经过配置，能够为分配的任务提供足够的信息，或者该机器人会..... 未产生有用信息

我们将其归类为缺乏经验的受访者（6）。总体模式在不同的机器人使用经验群体中均成立（偏好反应式和亲和型）。在我们的数据中，机器人使用经验似乎不会影响自主性偏好（p=0.27，α=0.05）。根据受访者的拉取请求活动，也似乎不存在任何特定模式。

E. 通过开放式回答解释选择

为了更好地理解受访者如何评价自主性和人格，我们对他们偏好的理由（图 2 底部）进行了开放式编码。论文的三位作者分别对受访者的开放式回答进行了三轮编码；每轮结束后，编码者都会讨论彼此的分歧，以增进对编码任务的共同理解，并改进编码指南。在最后一轮结束时，我们达到了$\kappa = 0.723$ 的较高评分者间一致性。最终形成了五个高层级编码，每个编码包含两个代表相反极性的子编码（表 VI）。括号内显示了编码频次。我们还为表示自己没有偏好的评论，以及噪声／不可用数据设置了一个编码。

受访者解释其偏好的原因，最常见的是机器人传达的信息清晰（41 次）、机器人有礼貌（40 次），或信息丰富（19 次）。对于机器人自主性的偏好，机器人发起（30 次）与人工发起（41 次）的控制方式数量相当。开放式回答表明，参与者的反应确实与我们的构念相符，也呼应了第一阶段访谈中的主题。

F. 定量数据的贝叶斯分析

我们采用贝叶斯数据分析 [19] ，并参照 Furia等人 [17] 的研究实例和 Torkar等人 [46]的指南，探究可能影响受试者偏好的因素。在频率学派框架下，样本量为 56 时，我们无法发现可能较小的效应。不过，贝叶斯分析即使在数据较少的情况下，也能得出有效的预测，但需要加入一个先验概率分布（在我们的研究中，该分布提供的信息较弱） [20]。

因果模型及相关统计模型：我们构建了因果图 [36] ，用于模拟访谈主题（第三节 B 部分）之间的关系。因果模型是一种有向无环图（DAG），其中的箭头表示变量之间的因果关系（影响）。我们的因果图见复现包。模型表明，机器人经验、GitHub 活动和PR 活动 会影响人们对自主程度和角色形象风格的偏好。一个隐藏节点表示其他未建模的变异来源。

我们用变量 情境短文选择_自主性 和 情境短文选择_人格来表示偏好。导致这些选择的处理效应是自主性程度和角色形象程度。也就是说，我们对受访者在图 2 的情境短文中作出某种选择的概率进行建模（在我们的测量工具中，选项为“偏好左侧”“中性”或“偏好右侧”）。

我们根据因果图推导出序数回归模型 [7] ，并在表 VII 中列出三个实例。在表中，波浪号表示“因变量（左侧）由自变量交互项（右侧）解释”。我们针对四个情境短文分别建立 3 个模型（共 12 个模型）。表中展示的是自主性情境短文（1 和 2）对应模型的表述方式，人格情境短文（3 和 4）也采用类似的表述。

模型 $\mathcal{M}_1$ 最简单。它假设，仅凭每个情境短文中的 GitHub 活动 水平就足以作为预测变量。模型 $\mathcal{M}_2$ 加入了 PR 经验、机器人以及总体软件开发经验，以检验这些因素是否能增强模型的解释力。模型 $\mathcal{M}_3$ 检验不同情境短文中独立构念的回答是否相关（例如，某人对自主性的选择是否会影响其对人格的选择，反之亦然）。

分析的目标是找出一个能够用最少预测变量最好地解释数据的模型。向模型中添加更多预测变量会增加对所收集数据过拟合（产生偏差）的风险，并降低我们解释更一般现象的能力。随后，我们比较了不同模型的解释力。模型能够解释的数据越多，就越能表明其中可能存在重要影响。

G. 定量分析结果

我们遵循 Bürkner 和 Vuorre [7] 的指南进行推断。潜在参数 Y_{autonomy} （分别为 Y_{persona}）的累积／连续响应序数模型 ~ ~用于对用户在机器人行为某一方面的一般偏好进行建模。~潜在连续变量 Y 被划分为三种

表 VII 自主性偏好候选模型

模型编号	统计模型
$\mathcal{M}_1$	情境短文选择_自主性 ~ GitHub 活动
$\mathcal{M}_2$	情境短文选择_自主性 ~ 机器人经验 + GitHub 活动 + PR 活动 + 经验
$\mathcal{M}_3$	情境短文选择_自主性 ~ 情境短文选择_人格

响应类别：“反应式”“中性”“主动式”（分别为“事实型”“中性”“亲和型”）。其基本思路是，这种划分能够将数据恰当地视为有序类别，而非连续度量响应，并通过贝叶斯推断找出最可能的划分方式。

后验分布的条件效应： 三个模型中的后验推断揭示了回归参数的分布（用于分析的 R 脚本见复现包，第七节）。遵循 Gelman 等人 [21], 为软件研究提出的贝叶斯工作流程，我们采用模型比较方法，重点是构建对观测数据的充分解释，从而有助于回答问题、进行决策分析或作出预测。这并不意味着它是最佳模型，只是给定因果模型中的各种因素后，能够最佳解释（“最不意外于”）数据的模型。

由推断过程生成的后验分布，使我们能够评估在给定某个自变量的条件下，情境短文中某一特定选择的概率。我们的目标是探究自变量是否会影响因变量，以及如何影响。

推断过程会生成后验预测分布，这是它与其他推断方法相比最有用的差异之一。在这里，我们可以利用后验分布来针对模型提出问题。图6中的条件效应图展示了后验分布中的一个样本。x 轴表示模型中的一个预测变量，在本例中是受访者 GitHub 活动的程度（有经验、不频繁、中等）。y 轴显示响应类别的概率，即在情境短文1中选择主动式（蓝色圆点）、中性（绿色圆点）或反应式（红色圆点）的概率。误差条表示模型的 95% 可信区间（即从后验分布中抽取的样本有 95% 落在误差条范围内）。圆点表示所取样本的均值。

在此情境短文中， $\mathcal{M}_1$ 显示，中等和不频繁使用 GitHub 的用户更可能选择反应式选项（蓝色／最右侧的均值），而有经验的用户更可能在情境短文中选择主动式机器人行为。不过，误差线重叠表明，我们对有经验用户选择概率的判断信心较低。

我们还可以使用信息准则和留一法验证（LOOIC, [48]），根据观测数据近似留出数据的似然，并量化哪个模型提供的信息最多。

较低的 LOO 信息准则分数（LOOIC）表明模型对数据的拟合效果更好 [17]。可以比较同一数据集不同模型的分数相对差异。当 LOOIC 分数低于标准误的整数倍时，领域知识等其他因素（即预测变量影响结果的可能性）以及模型简约性（相比更多预测变量，更倾向于使用较少的预测变量）也很重要。

如图5所示，我们关于软件开发或机器人相关经验（$\mathcal{M}_2$）的模型信息量较少，误差范围也更大。角色构念上的选择（情境短文3和4）更可能受到以下因素影响：

模型	情境短文	LOOIC	SE	情境短文	LOOIC	SE
$\mathcal{M}_1$		**101.4**	8.7		**117.8**	9.1
$\mathcal{M}_2$	V1	116.2	12.1	V2	120.5	12.2
$\mathcal{M}_3$		111.9	8.7		119.4	7.6
$\mathcal{M}_1$		95.5	11.2		112.3	10.8
$\mathcal{M}_2$	V3	115.7	14.1	V4	127.6	7.3
$\mathcal{M}_3$		**87.6**	10.6		**111.1**	7.7

图5。使用留一信息准则比较模型。数值越低越好。SE 表示标准误。对于情境短文1和2（自主性），模型1更受偏好；对于情境短文3和4（人格），模型3更受偏好。

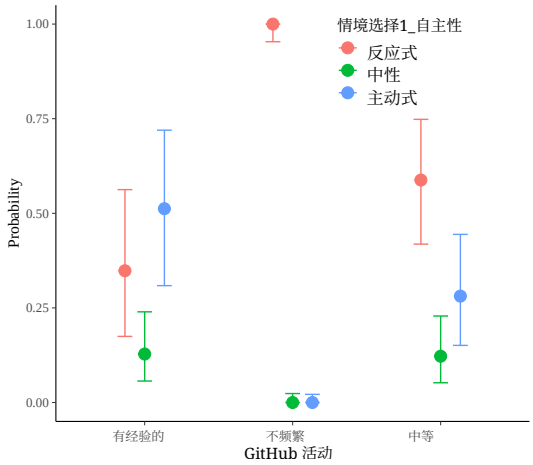


图6. 条件效应图展示了在不同 GitHub 活动条件下选择情境短文1的概率。从左到右依次为：反应式（红色）、中性（绿色）和主动式（蓝色）。

情境短文1中的选择（即，受访者偏好反应式还是主动式机器人行为）。

事实上，回到研究问题*2*，推断出的模型证实，缺乏经验的开发者偏好反应式、亲和型机器人。这与其他研究的报告一致：缺乏经验的开发者将机器人视为自动化的智能程序（Erlenhov的Sam人格 [14]）。这些模型为我们的观点提供了证据：有经验的开发者更偏好主动式、事实型机器人。

五、讨论

A. 自主机器人与开发者认知

用户和专业人士似乎对主动式代理持谨慎态度 [24], [42], [40]。Wessel 等人指出，在开源领域，过于热心的机器人会被开发者视为噪声 [54]。我们的研究发现，与经验较少的开发者相比，有经验的开发者更倾向于选择主动式机器人。参与者给出的理由很少提及 Wessel 等人所说的机器人噪声或打扰。相反，随着开发者经验的增长，他们对主动式机器人的信心（或容忍度）似乎也会提高。反过来，经验较少的开发者可能会觉得主动式机器人过于令人生畏或咄咄逼人。我们的研究表明，机器人创建者不仅应决定如何配置机器人，也提醒我们，开发者

是且始终是开发过程的核心。假设并非如此的工具可能很难被采用。因此，我们建议在机器人设计中考虑根据开发者的经验水平调整机器人自主性的程度。因此，我们提出：

> **建议 #1**：应向开发者提供选项，让他们调整参与其开发流程的机器人的自主性。例如，他们可以在主动式和反应式之间进行选择。

开发者还可以选择使用自己觉得舒适的机器人（例如 Dependabot，例如，P10 认为它“非常有用”），也可以选择将哪些机器人设为静音。不过，这种调整应以开发者偏好为依据，正如我们的调查和 Erlenhov 等人 [15]的调查所揭示的那样。因此，

> **建议 #2**：为了调整机器人行为，项目必须有更好的方式来收集有关机器人行为和开发者偏好的反馈。

一种做法是收集开发者对机器人评论的反应，例如使用点赞表情符号。

B. 机器人角色形象的重要性

机器人使用的语言会影响用户对机器人的看法 [10]。Volkel 等人发现，用户可能更愿意接受那些体现自身个性的机器人 [49]。我们的研究结果表明，开发者对机器人角色形象的偏好取决于情境短文的具体内容。实际启示是，机器人角色形象很重要，不应被视为事后才考虑的因素。应向开发者提供选项，让他们能够影响参与其开发活动的机器人的角色形象，尤其是因为已有研究表明，展现出更高拟人化程度的机器人能够与开发者建立更强的信任 [42], [23], [35]。例如，开发者可以从一系列预设的角色形象和沟通风格中进行选择。一些受访者表示，他们更希望机器人只保留消息中的信息，并省去没有信息价值的话语。基于这些见解，我们提出：

> **建议 #3：** 开发者应能够根据自己的偏好选择和更改机器人的角色形象。

未来的研究应探讨其他因素如何影响开发者对机器人角色形象的偏好。例如，可以利用开源项目的特点，按照社区偏好来指导预定义角色形象的设计。Alami等人 [2], [3] 发现，开源社区的拉取请求治理流程通常符合三种不同但并非互斥的治理风格：宽松型、公平型和保护型。可以研究与社区风格相符的不同机器人角色形象风格，作为贡献者的潜在默认选项。

C. 机器人交互中的更广泛主题

我们的研究聚焦于自主性和人格这两个关键主题，并指出应提升机器人的可配置性，并增强对开发者认知的了解。不过，开发者对机器人的偏好还涉及其他维度。在我们的访谈和调查答复中，信任、过往的机器人使用经验、外观和身份等主题也浮现出来。要进一步理解机器人在软件开发中扮演的角色，这些主题还需要更深入的研究。随着软件机器人与开发者之间复杂交互的不断扩展，我们需要更广泛地借鉴有关行为、信任和语言的理论与研究成果。Bicchieri 的社会规范概念 [5], 、Searle 的言语行为 [41],，以及 Brown 和 Levinson 的礼貌理论 [6],，都是一些例子，说明这一领域非常适合开展跨学科研究。

VI. 研究发现的可信度与有效性

可信度： 可信度用于评估我们的定性分析和结论的有效性。

可信性与可确认性。我们采用了严谨的标注流程，由多位作者参与并遵循清晰的指导原则，以减少单个标注者带来的偏差。我们此前已报告过对调查答复进行编码时的评分者信度。

第一阶段的抽样方法可能被视为本研究的一项潜在弱点。访谈参与者是通过推荐和口耳相传招募的。因此，我们的样本可能会使第一阶段的研究结果产生偏差。不过，我们采用的混合方法在一定程度上弥补了这一弱点；在第二阶段，我们对来自不同经验和国别背景的广泛人群进行了调查。

可靠性。我们的编码手册、匿名化访谈记录和答案均已收录在复现包中，可用于对研究结果进行可靠性检验。

可迁移性。我们认为，通过证明研究结果可能适用于其他类似情境（即自由与开源社区），本研究已达到可迁移性标准。我们提供了内容详尽的复现包，使其他研究人员能够重复开展类似调查。此外，我们通过恰当地界定研究、说明参与者的各种经验、实施研究方法，并在定量阶段使用更广泛的样本评估结果，确保了研究的可迁移性。

内部效度： 我们邀请了实验室的三名成员试填调查问卷。设计问卷需要权衡取舍。我们为每位参与者随机排列情境短文的顺序，并将一些问题设计得比理想情况更简单，以减少实验偏差，例如疲劳效应。我们更改了每篇情境短文的背景，以确保不会产生学习效应。我们尽量使情境短文两侧的文本长度大致相同。我们检验了情境短文选择之间的相关性，以确保针对同一构念的选择彼此相关，而针对不同构念的选择之间不相关。

在情境短文1和2中，机器人的拉取请求都被拒绝了。这次失败可能影响了开发者对自主机器人的偏好或可接受程度的判断。不过，在受访者提供的偏好响应式机器人的理由中，我们没有看到任何对机器人失败的提及。在我们的贝叶斯建模中，可供分析的新手样本可能太少，

导致模型效率较低。这一点从这些类别较大的标准误差中可以看出。

构念效度：我们的研究结果基于对人格和自主性构念的操作化，并将其与开发者对机器人行为的偏好相对应。这些构念复杂且往往因人而异。对这些构念进行操作化必然会舍弃其中一部分复杂性（这是换取更多控制所必须付出的代价）。我们提出开放式问题，旨在让受访者能够表达这种复杂性。

为了衡量偏好，我们请受访者指出自己更喜欢情境短文中的哪种情境。将不同情境并列展示，受访者便能轻松判断自己更偏好这两种情境所呈现的构念的哪一端。此外，我们还请受访者以开放式回答说明理由，以确保我们理解其偏好背后的理由；根据这些开放式回答，我们得出结论：受访者陈述的理由与受测构念相似。情境描述 4 的结果不一，这可能反映了交互措辞欠佳。调整机器人的人格，使其在事实型与亲和型之间取得平衡，是一项艰难的权衡。在建模时，我们假设偏好一方等同于不偏好另一方，即不存在特定类别效应。

外部效度：调查受访者经过了仔细筛选，以确保具备基本的软件开发知识并使用过拉取请求。此外，我们的调查样本来自自述为软件行业专业人士的受访者（Prolific），调查在不同时间进行；样本还包括软件工程专业三年级学生。采用筛选问题并进行多轮 Prolific 调查符合当前建议 [13]，[12]。愿意参加 Prolific 研究的人群可能存在偏差，而且我们的参与者中，年龄超过 35 岁或工作年限超过 5 年的人不多。

VII. 结论

本文考察了影响开发者对 GitHub 机器人认知的因素。我们首先开展了一项访谈研究，以了解项目和开发者使用及看待软件机器人的一些重要主题。从中提炼出两个影响认知的重要构念：机器人表现出的自主性程度，以及它的人格。基于情境短文的调查让我们能够控制其他方面，并探究受访者在模拟的拉取请求讨论中对机器人的反应。软件开发经验较少的参与者通常更喜欢反应式、自主性较低的机器人，但对友好的人格特征也能接受。软件项目中的机器人开发应让这两个方面更易于配置。要充分发挥机器人的作用，更好地理解项目与开发者之间的交互十分重要。

复现包：我们公开提供了研究数据和其他材料[4]。对于第一阶段，我们分享了完整且详细的访谈指南、匿名化的访谈转录稿，以及数据分析结果，即编码手册。对于第二阶段，则提供了调查设计和数据。

[4]https://doi.org/10.5281/zenodo.7040317

人类研究伦理审查：这项研究，包括与学生之间的权力关系，已获得参与研究的研究人员所属相关机构审查委员会的批准。

致谢：感谢访谈和调查的参与者抽出时间并提供有益的答复。我们也感谢 Rohith Pudari 对情境短文工具的帮助。

参考文献

[1] “实证标准补充材料”，https://acmsigsoft.github.io/EmpiricalStandards/Supplements/?supplement= 评定者间信度与一致性，（访问日期：2023年01月03日）。[2] A. Alami、M. L. Cohn 和 A. W.asowski，“自由及开源软件社区如何决定是否接受拉取请求？”载于《软件工程评估与评价》论文集。ACM，2020年4月。 [在线]。可访问：https://doi.org/ 10 1145 2 3383219 3383242 。% F . A. Alami 、 R. Pardo 、 M. L. Cohn 和 A. Wasowski，“ 开源社区中的拉取请求治理”，《IEEE软件工程汇刊》，2021年。[4] S. Baltes 和 P. Ralph，“软件工程研究中的抽样：批判性评述与指南”，《实证软件工程》，第27卷，第94页，2022年7月。 [在线]。可访问：https: // link.springer.com/ 10 1007 10664 021 10072 8 。/s - - - C. Bicchieri，《野外规范：如何诊断、测量和改变社会规范》。牛津大学出版社，2016年。[6] P. E. Brown 和 S. Levinson，《礼貌：语言使用中的若干普遍现象》。剑桥大学出版社，1987年。[7] P.-C. Bürkner 和 M. Vuorre，“心理学中的序数回归模型：教程”，《心理科学方法与实践进展》，第2卷，第1期，第77–101页，2019年2月。 [在线]。可访问：https://doi.org/ 10 1177 2 2515245918823199 。% F A. P. Chaves、J. Egbert、T. Hocking、E. Doerry 和 M. A. Gerosa，“聊天机器人语言设计：语言变体对旅游助理聊天机器人用户体验的影响”，《ACM人机交互汇刊》，第29卷，第1–38页，2022年4月。[9] A. P. Chaves 和 M. A. Gerosa，“我的聊天机器人应如何交互？关于人机聊天机器人交互设计中社会特征的调查”，《国际人机交互杂志》，第1–30页，2020年11月。 [在线]。可访问：https://doi.org/10.1080/10447318.2020. 1841438[10] L. Clark、N. Pantidi、O. Cooney、P. Doyle、D. Garaialde、J. Edwards、B. Spillane、E. Gilmartin、C. Murad、C. Munteanu、V. Wade 和 B. R. Cowan，“什么造就了良好的对话？”载于《2019年 CHI计算系统中的人为因素会议》论文集。ACM，2019年，第1–12页。[11] J. W. Creswell 和 V. L. P. Clark，《混合方法研究的设计与实施》，第3版。SAGE出版社，2017年。[12] A. Danilova、A. Naiakshina、S. Horstmann 和 M. Smith，“你真的会编程吗？为程序员在线问卷调查设计和评估筛选问题”，载于《2021年 IEEE/ACM第43届国际软件工程会议（ICSE）》论文集。IEEE，2021年5月，第537–548页。[13] F. Ebert、A. Serebrenik、C. Treude、N. Novielli 和 F. Castor，“在 Prolific 上招募有经验的 GitHub贡献者参与访谈和调查”，载于《第1届实证软件工程研究参与者招募国际研讨会（RoPES’22）》，2022年。[14] L. Erlenhov、F. G. de Oliveira Neto、R. Scandariato 和 P. Leitner，“软件开发中的当前与未来机器人”，载于《第1届软件工程中的机器人国际研讨会》论文集。IEEE出版社，2019年，第7—11页。 [在线]。可访问：https://doi.org/10.1109/ BotSE.2019.00009[15] L. Erlenhov、F. G. d. O. Neto 和 P. Leitner，“软件开发中机器人的实证研究：从业者视角下的特征与挑战”，载于《第28届 ACM 欧洲软件工程会议与软件工程基础研讨会联合会议》论文集，2020年，第445–455页。[16] J. C. Farah、V. Sharma、S. Ingram 和 D. Gillet，“在人机聊天机器人交互中传达由歧义语法结构引发的幽默感知”，载于《第9届人机交互国际会议》论文集，2021年，第257–262页。

[17] C. A. Furia、R. Torkar 和 R. Feldt，“将贝叶斯分析指南应用于实证软件工程数据：编程语言与代码质量案例”，《ACM 软件工程与方法论汇刊》，第31卷，第3期，第1–38页，2022年7月。 2022 [ ] 10 1145 3490953 在线。可访问：https://doi.org/。/ A. Furnham，“反应偏差、社会赞许性与掩饰行为”，《人格与个体差异》，第7卷，第3期，第385–400页，1986年。[19] A. Gelman、J. B. Carlin、H. S. Stern、D. B. Dunson、A. Vehtari 和 D. B. Rubin，《贝叶斯数据分析》。CRC 出版社，2013年。[20] A. Gelman、J. Hill 和 A. Vehtari，《回归与其他故事》。剑桥大学出版社，2021年。[21] A. Gelman、A. Vehtari、D. Simpson、C. C. Margossian、B. Carpenter、Y. Yao、L. Kennedy、J. Gabry、P.-C. Bürkner 和 M. Modrák，“贝叶斯工作流程”，2020 [ ] 2011 01808 。在线。可访问：https://arxiv.org/abs/。M. Jain、P. Kumar、R. Kota 和 S. N. Patel，“评估聊天机器人并为其设计提供信息”，载于《2018年设计交互系统会议》论文集。ACM，2018年6月，第895–906页。[23] M. Li 和 A. Suh，“机器型还是人类型？人工智能技术中拟人化研究综述”，载于《第54届夏威夷国际系统科学会议》论文集，2021年，第4053页。[24] Q. V. Liao、M. Davis、W. Geyer、M. Muller 和 N. S. Shami，“你能做什么？”载于《2016年 ACM 设计交互系统会议》论文集。ACM，2016年6月，第264–275页。[25] S.-T. Liaw、E. Deveny、I. Morrison 和 B. Lewis，“临床、信息与业务流程建模，以促进安全且灵活的软件开发”，《健康信息学杂志》，第12卷，第3期，第199–211页，2006年，PMID：17023408。 [在线]。可访问：https://doi.org/10 1177 1460458206066772 。/ F. Markusse、P. Leitner 和 A. Serebrenik，“使用基准测试机器人进行持续性能评估”，《IEEE软件》，第39卷，第5期，第50–55, 2022 [ ] 10 1109 2022 3184430 页。在线。可访问：https://doi.org/。/MS。/。A. McNamara、J. Smith 和 E. Murphy-Hill，“ACM伦理准则是否会改变软件开发中的伦理决策？”载于《2018年第26届 ACM 欧洲软件工程会议与软件工程基础研讨会联合会议》论文集。ACM，2018年10月，第729–733页。[28] M. B. Miles、M. Huberman 和 J. Saldana，《定性数据分析：方法资料手册》。SAGE出版社，2013年。[29] S. Mirhosseini 和 C. Parnin，“自动化拉取请求能否鼓励软件开发者升级过时依赖项？”《第32届 IEEE/ACM 自动化软件工程国际会议》论文集，第84–94页，2017年。[30] M. Monperrus，“可解释的软件机器人贡献：自动化缺陷修复案例研究”，载于《2019年 IEEE/ACM 软件工程中的机器人国际研讨会（BotSE）》论文集，2019年。[31] C. Nass、J. Steuer 和 E. R. Tauber，“计算机是社会行动者”，载于《计算系统中的人为因素会议配套论文集——CHI ’94》。ACM出版社，1994年，第204页。[32] F. Palomba、D. A. Tamburri、F. A. Fontana、R. Oliveto、A. Zaidman 和 A. Serebrenik，“超越技术层面：社区异味如何影响代码异味的强度？”《IEEE软件工程汇刊》，第47卷，第1期，第108–129页，2021年。 [在线]。可访问：https://doi.org/10 1109 2018 2883603 。/TSE。/。M. Q. Patton，《定性研究与评估方法：整合理论与实践》。Sage出版社，2014年。[34] Z. Peng 和 X. Ma，“探索软件开发者如何在 GitHub 中使用提及机器人”，《CCF 普适计算与交互汇刊》，第1卷，第3期，第190–203页，2019年9月。 [在线]。可访问：https://doi.org/10 1007 2 42486 019 00013 2 。% Fs - - - N. Pfeuffer、A. Benlian、H. Gimpel 和 O. Hinz，“拟人化信息系统”，《商业与信息系统工程》，第61卷，第4期，第523–533页，2019年。[36] J. M. Rohrer，“清晰思考相关性与因果关系：观察性数据的图形因果模型”，《心理科学方法与实践进展》，第1卷，第1期，第27–42页，2018年1月。[37]P. H. Rossi、W. A. Sampson、C. E. Bose、G. Jasso 和 J. Passel，“测量家庭社会地位”，《社会科学研究》，第3卷，第3期，第169–190页，1974年。 [在线]。可访问：https://www.sciencedirect.com/science/article/pii/0049089X74900118

[38]S. Santhanam、T. Hecking、A. Schreiber 和 S. Wagner，《软件工程中的机器人：一项系统映射研究》，PeerJ Computer Science，第 8 卷，文章编号 e866，2022年2月。 [在线]。获取地址：https://doi.org/10 7717 866 。/peerj-cs。C. Sarmento、T. Massoni、A. Serebrenik、G. Catolino、D. A. Tamburri 和 F. Palomba，《性别多样性与社区异味：对巴西软件团队的一项双重复制研究》，载于 IEEE软件分析、演化与再工程国际会议，SANER 2022，美国夏威夷州檀香山，2022年3月15—18日。IEEE，2022年，第273–283页。 [ ] 10 1109 53432 2022 00043 在线。获取地址：https://doi.org/。/SANER。。J. Schaffer、J. O’Donovan、J. Michaelis、A. Raglin 和 T. Höllerer，《我能做得比你的人工智能更好》，载于第24届智能用户界面国际会议论文集。ACM，2019年3月，第240–251页。[41] J. R. Searle，《言语行为：语言哲学论文》，Language，第46卷，页码217，1969年。[42] A. Seiffer、U. Gnewuch 和 A. Maedche，《理解员工对软件机器人的反应：系统性文献综述》，载于信息系统国际会议（ICIS），第2021卷，2021年。[43] J. Siegmund、C. Kästner、J. Liebig、S. Apel 和 S. Hanenberg，《编程经验的测量与建模》，Empir. Softw. Eng.，第19卷，第5期，第1299–1334页，2014年。 [在线]。获取地址：https://doi.org/10 1007 10664 013 9286 4 。/s - - - P. Steiner、C. Atzmüller 和 D. Su，《为调查研究设计有效且可靠的情境实验：公平性别收入差距案例研究》，社会科学方法与测量杂志，第7卷，第2期，第52–94页，2017年。 [在线]。获取地址：https://journals.uair.arizona.edu/index.php/jmmss/article/view/20521 M.-A. Storey 和 A. Zagalsky，《一次一个机器人，颠覆开发者生产力》，载于2016年第24届ACM SIGSOFT软件工程基础国际研讨会论文集，2016年，第928–931页。[46] R. Torkar、C. A. Furia、R. Feldt、F. G. de Oliveira Neto、L. Gren、P. Lenberg 和 N. A. Ernst，《评估并论证软件工程中实际显著性的一种方法》，IEEE软件工程汇刊，第1–1页，2021年。 [在线]。获取地址：https://doi.org/10 1109 2020 3048991 。/tse。。C. Vanover、P. Mihas 和 J. Saldaña，《分析与解读定性研究：访谈之后》。SAGE出版社，2021年。[48] A. Vehtari、A. Gelman 和 J. Gabry，《使用留一交叉验证和 WAIC 进行实用的贝叶斯模型评估》，Statistics and Computing，第27卷，第5期，第1413–1432页，2016年8月。 [在线]。获取地址：https://doi.org/10 1007 11222 016 9696 4 。/s - - - S. T. Völkel、D. Buschek、J. Pranjic 和 H. Hussmann，《通过用户个性与消息语境理解表情符号解读》，载于第21届人机交互与移动设备及服务国际会议论文集。ACM，2019年10月，第1–12页。 [ ] 10 1145 3338286 3340114 在线。获取地址：https://dl.acm.org/doi/。。M. Wessel、A. Abdellatif、I. Wiese、T. Conte、E. Shihab、M. A. Gerosa 和 I. Steinmacher，《拉取请求机器人：优点、缺点与潜力》，载于第44届软件工程国际会议论文集。ACM，2022年5月，第274–286页。 [在线]。获取地址：https://dl.acm.org/doi/10 1145 3510003 3512765 。/。M. Wessel、B. M. De Souza、I. Steinmacher、I. S. Wiese、I. Polato、A. P. Chaves 和 M. A. Gerosa，《机器人的力量：理解开源软件项目中的机器人》，ACM人机交互会议录，第2卷，CSCW期，第1–19页，2018年。[52] M. Wessel、A. Serebrenik、I. Wiese、I. Steinmacher 和 M. A. Gerosa，《对 GitHub 代码审查机器人有何期待？对开源软件维护者的一项调查》，载于第34届巴西软件工程研讨会论文集，2020年，第457–462页。[53] M. Wessel 和 I. Steinmacher，《拉取请求中软件机器人的不便之处》，载于 IEEE/ACM 第42届软件工程国际会议研讨会论文集。ACM，2020年6月。 [ ] 10 1145 2 3387940 3391504 在线。获取地址：https://doi.org/。% F。M. Wessel、I. Wiese、I. Steinmacher 和 M. A. Gerosa，《请勿打扰：在开源软件项目中与软件机器人交互的挑战》，ACM人机交互会议录，第5卷，CSCW2期，第1–21页，2021年。[55] M. Wyrich、R. Ghit、T. Haller 和 C. Müller，《机器人不介意等待，对吧？比较与自动创建和手动创建的拉取请求进行交互》，载于2021年 IEEE/ACM 第三届软件工程中的机器人国际研讨会（BotSE）。IEEE，2021年，第6–10页。