从人类协商到智能体协商：自动化交通中的个人出行代理

Pascal
Jansenpascal.jansen@uni-ulm.de 媒体信
息学研究所，乌尔姆大学，乌尔姆，德国

arXiv:2603.01035v1 [cs.HC] 1 Mar 2026

摘要

用户偏好与自动化系统行为之间的冲突，已经影响着人们对自动化出行的体验。例如，乘客可能偏好强势驾驶，但车辆却会提前减速，以遵循保守策略或礼让其他交通参与者。类似的冲突也会出现在并道、交叉路口或路权情境中，用户要么只能接受不透明的决策，要么试图通过并非为持续、多方关系而设计的界面进行协商。本文立场认为，随着出行方式日益多元化和自动化，这类做法将难以扩展。为此，本文提出由个人出行代理充当用户代理，将舒适度和安全余量等偏好编码其中，并在共同安全规则下与其他代理协商交通行为。其核心理念是从用户逐刻参与的协商界面，转向委托与监督界面：代理负责处理实时冲突，用户则可以调整高层策略和偏好。

CCS概念

• **以人为中心的计算** → **人机交互（HCI）**；交互范式；普适计算与移动计算；• **计算方法** → 智能体。

关键词

出行、人机自动化交互、智能体交互

1 引言

出行是一种基础性的社会技术基础设施，因为它使人们能够获得工作、教育、医疗保健和社会参与的机会，而出行系统一旦失灵，就会直接带来安全风险 [12]。从本质上说，日常交通是多个参与者（如驾驶员和骑行者）围绕共享空间持续协商的过程，包括谁先通行、什么构成可接受间隙，以及如何执行驾驶操作。

自动化出行将部分决策从人类转移给自动化系统，从而改变了这种协商。因此，既有研究主要关注自动化系统如何与周围的道路使用者进行通信，例如通过外部人机界面（外部人机界面）向行人传达意图 [3]。在这些研究方向中，信任校准（见 Lee 和 See[8]）仍然是核心设计目标，因为用户既不应过度依赖自动化，也不应回避自动化，而大规模采用则取决于感知安全性、可靠性和偏好对齐 [3, 18]。除了传达意图之外，近期研究还探索了自动驾驶中的协商界面，例如乘客与车辆的对话协商场景 [15]。这些方法超越了静态通信，使用户能够影响或讨论即时决策。

然而，许多交互都假设存在一人一系统协商关系，例如行人与车辆在交叉路口的交互 [3]，或乘客与车辆之间的个性化设置，例如自动驾驶风格适配 [17]。相比之下，既有的多智能体协商研究探讨了交通场景中自动化参与者之间协议与协商策略的设计，为共享决策和机动协调提供了形式化基础 [1]。然而，这些方法并未对跨情境持续存在、代表用户的偏好代理或委托界面进行建模。

实际上，自动化交通是一个涉及驾驶员、乘客、行人、自动驾驶车辆、微出行用户（例如电动滑板车骑行者）以及智能基础设施的多参与者环境。参与者在能力、偏好和目标方面各不相同，例如可接受间隙方面的年龄相关差异 [10], 、风险容忍度方面的文化差异 [6], , 以及对自动驾驶风格的偏好 [17]。因此，交通是一种处于时间压力下且仅能部分观测的多目标协商，各参与者在不了解他人情况的条件下优化不同目标 [3]。在这些条件下，用户意图与系统约束之间很可能发生冲突，并可能降低自动化参与度、信任和接受度 [16]。

为了举例说明这类情况，请设想不久后的一个城市交通场景：

> 一名乘客坐在一辆驶近交叉口的自动驾驶共享出行车辆中。一名骑行者从侧面驶来，一台配送机器人沿着人行道移动，还有一名行人准备过马路。车辆比乘客预期更早减速，显然是在礼让行人。然而，骑行者加速并率先通过，而配送机器人则因自身的内部策略而暂停。从乘客的角度来看，这种情况似乎前后不一，难以理解：车辆为何采取谨慎的行为并不清楚，也无法确定这是否反映了乘客自身的偏好、骑行者的行为，还是隐藏的基础设施数据。用户没有切实可行的方式表达偏好，例如“在交通稀疏时更主动一些，

发表于 AutomationXP26 研讨会的立场论文：智能体自动化体验——在 CHI '26 重新思考人与 AI 智能体的交互。

本作品采用知识共享署名-非商业性使用-相同方式共享 4.0 国际许可协议授权。
AutomationXP26 研讨会，*CHI '26*，西班牙巴塞罗那© 2026 版权所有归所有者/作者所有。

> 但在行人附近始终保持较大的安全间距”，也无法了解这一偏好将如何与其他参与者协商。

在这类情境中，现有的交互方式（例如 [15]）要么需要向人类用户提供更多解释，要么需要进行明确的协商对话，或事后给出理由。然而，这些方式通常假设用户具备适当的情境意识(参见 Endsley[4])，从而能够有意义地追踪和评估每次交互。在日益复杂多元的环境中，这些假设未必能大规模成立。这可能导致用户不再参与、更多地进行人工干预、普遍的不信任，并最终降低用户感知到的控制感、信任校准水平以及对自动化出行的接受度。

因此，本文立场论文认为，要实现可扩展的出行交互，就需要从用户之间的直接协商转向代理中介协商。为此，本文提出一种不同的交互模型：每位道路使用者都由一个**个人出行代理**代表，该代理充当用户与交通情境之间的中介。代理会编码用户的偏好、约束和委托策略，并代表用户与其他参与方协商。例如，行人的代理可能会要求留出更大的安全间距，而车辆的代理则可能协商以更平顺但略慢的方式并线，从而保障乘客舒适度。用户不必参与每一次临时协商，代理可以彼此解决这些交互，并仅在用户需要时向其传达高层次的结果，例如“提前让行，以保持您偏好的安全裕度”。

在这一视角下，交互从对各项机动动作的直接控制，转向对高层偏好和委托策略的设定。设计问题随之变为：如何确保基于代理的协商机制对所有参与者而言都安全、易于理解且可以接受。

2 自动化交通协商中的个人出行代理

本节概述个人出行代理可能采用的技术基础，介绍其个性化的潜在机制，并举例说明它们如何在未来的自动化交通场景中运行。

2.1 技术基础

从技术角度来看，个人出行代理可以借助当前的多项技术进展来构建。自动驾驶车辆和驾驶辅助系统已经能够根据用户输入和观察到的行为，学习个人驾驶偏好，例如舒适度水平或加速风格 [9]。人在回路优化方法进一步表明，系统可以通过简短的交互环节，高效地学习此类偏好。例如，贝叶斯优化已被用于根据个体用户调整外部人机界面和车内界面 [2, 7]。与此同时，智能体人工智能、大语言模型和视觉语言模型的最新进展表明，软件智能体能够代表用户采取行动、持续维护偏好模型，并与其他智能体或服务进行交互 [13]。其中许多系统可以通过基于人类反馈的强化学习进行个性化；在这一过程中，系统学习人类评分所反映的偏好行为，而非遵循固定规则 [11]。个人出行代理可以理解为此类面向交通场景的个人智能体的一种专门形式。

2.2 核心功能组件

个人出行代理可以执行三项相互关联的功能：

偏好建模该代理维护一个偏好模型，类似于用户的数字孪生，用于捕捉用户可接受的安全间距、偏好的加速度水平、舒适度限值、让行意愿，以及对延误或绕行的容忍度等方面。不过，该模型仍可能受到安全和法律要求（例如 [5]）的约束。

交通协商 在接近交叉口、并入车道、跟随较慢车辆或驶入环形交叉口等交通事件中，该代理会在安全约束下评估可行的机动动作，并与其他自动化参与者协商协调结果。例如，两辆车接近并道处时，各代理都可以提出符合其用户偏好的机动方案，例如“为舒适起见提前让行”或“在间隙安全时高效并道”。最终的机动方案依据共同安全规则协商确定，而用户仍处于高层偏好和策略控制层面。

界面管理 该代理通过传达结果来管理面向用户的界面 。例如，并道后，它可能会报告：“根据舒适度策略采取了保守并道”；接近交叉口后，则可能报告：“提前让行，以保持您偏好的安全裕度。”这遵循校准信任原则，让用户能够清楚理解系统行为 [8, 18]。此外，该界面还将支持选择性披露和可争议性：总结相关结果，在行为偏离预期时揭示其中的权衡，并允许用户查看和修改决策策略。

2.3 个人出行代理的生命周期

从用户的角度来看，个人出行代理可能会经历三个阶段的创建与完善：

(1) 在设置阶段，代理通过简单的入门问卷（例如评估在交通中对舒适度与效率的偏好）、人口统计信息、人格测试和/或简短的校准驾驶或仿真来初始化 偏好模型 。(2) 随后，代理在日常使用过程中在线适应。它会观察用户对不同机动动作的反应，例如用户是否接受或覆盖某些行为，并偶尔询问用户以获取快速反馈。(3) 代理会在更长的时间尺度上进行更新。它会检测用户行为的变化，允许用户明确编辑策略（例如“在交通稀疏时更果断一些”），并定期重新校准其偏好模型。因此，代理并非一次训练完成后便固定不变的模型，而是在离线和在线服务过程中不断演进、持续完善的表征，同时充当用户的代理。

3 设计张力与研究方向

引入作为用户代理的个人出行代理，会使交互设计从孤立的控制界面转向协商表征 和委托策略的设计。

在这种模式下，用户不再直接参与每一次交通互动，而是由其代理代表用户进行协商。这会带来若干设计张力，未来的出行系统和自动化体验可能会与之相关。

3.1 在法律与安全约束下实现偏好对齐

出行系统受到严格的安全和法律要求约束，这可能迫使代理偏离用户的偏好。例如，用户可能偏好并道时保持较小间距，但代理仍必须执行安全阈值。

以协商为重点的人机界面方法通常试图通过解释、劝说或自适应对话来解决此类冲突 [15, 16]。然而，在代理交互模型中，设计问题从在当下通过互动解决每一次冲突，转变为约束策略空间。这引出了一个问题：智能体应如何传达用户声明的偏好与其实际执行行为之间的差异？此外，此类沟通是否会削弱让用户免于在瞬时情境中直接交互这一预期收益？

3.2 防止策略性操纵，同时保留个性化

关于大语言模型中介的协商的研究表明，用户欣赏自适应沟通，但也可能试图试探或智胜该系统 [15]。对于会随时间积累个性化、并反复代表用户行事的持久代理而言，这一问题可能变得更为关键。关于人与自动化决策情境中委托的研究表明，用户的委托意愿取决于其对被委托系统所感知的能力、可控性与可问责性 [14]。因此，持久代理会带来超越瞬时适应的长期治理挑战。

一种可能的方向是将*交互自适应性*（代理如何沟通）与*策略自适应性*（代理被允许做什么）区分开来。这表明在设计上需要抗操纵的偏好采集、不可协商的安全规则，以及抑制对抗性议价的交互模式。

3.3 在不破坏混合交通交互的前提下扩展代理协商

个人出行代理可以通过将偏好与约束作为数据点进行交换，在自动化参与者之间实现高效且有效的协商。然而，交通很可能仍将是一种混合环境，其中许多参与者没有此类代理，且行为难以预测 [3]。因此，代理不仅必须与其他代理交互，还必须与人类驾驶员、骑行者和行人交互，而这些人的意图必须从其行为中推断。关于外部人机界面的既有研究强调，并非所有参与者都共享对称的沟通渠道或共享表征 [3]。因此，基于代理的协调不能假定所有参与者都使用代理，而必须优雅降级为基于行为的推断。

这使优化代理之间的谈判与维持和人类行为者的稳健互动之间产生了张力。此外，基于代理的谈判可能会模糊责任归属，尤其是在结果由多个相互作用的代理共同产生时。一个可能的研究方向是设计*归因机制*，用以解释某种行为究竟是由安全约束、用户的委托策略、推断出的人类行为，还是与其他代理协商达成的妥协所驱动。这一挑战也反映了多智能体自动化领域更广泛的关切：当结果由相互作用的自主系统而非单一决策者产生时，如何界定分布式责任并进行归因（例如，参见 [1]）。

3.4 让委托清晰可见：透明度、能动性与可问责性

基于代理的交互不仅会改变系统行为，也会影响用户使用自动化系统的体验。当代理自主协商大多数情境时，用户可能不太了解决策是如何作出的、采用了哪些权衡，或哪些约束影响了结果。在某些情况下，参与度降低可能减轻认知负荷，甚至提高接受度，因为系统“就是好用”。然而，在其他情况下，这可能导致能动性丧失、校准不当的信任，或难以理解某些结果为何会出现。

这在*便捷的委托*与*有意义的知情*之间造成了伦理和体验层面的张力。设计者必须决定代理的决策应当以多大程度呈现出来。一种做法是尽量少披露，让代理基本在后台运行，仅在出现重大偏离或安全关键事件时才予以提示。另一种做法是采用反思性透明度，定期向用户展示协商摘要或策略权衡，以帮助用户理解系统，并在长期使用中实现信任校准。

这种张力引出了若干尚待解答的问题：只要结果符合自己的预期，用户是否更愿意对协商细节基本一无所知？意识减弱到什么程度时，会损害用户的感知能动性或可问责性？代理系统又应如何在降低认知负荷与满足透明度、可争议性和知情同意等伦理要求之间取得平衡？

4 结论

本文立场论文指出，在不久的将来，交通将变得过于多元且自动化程度过高，以至于用户无法再亲自对每个交通情境进行有意义的协商。随着自动驾驶车辆、微出行、机器人和互联基础设施在共享空间中相互作用，目标和偏好各异的参与者之间不可避免地会发生冲突，而传统的、需要随时进行的人际协商策略难以适应这种规模。相反，个人出行代理可以充当用户的代理。这些代理会将用户偏好和约束编码其中，并依据共同安全规则与其他参与者协商具体的机动动作。

这一转变为人机交互提出了相关的研究问题。基于代理的交互可能会降低认知负荷并提高接受度，但当大多数协商在用户未直接参与的情况下进行时，它们也可能改变用户对能动性、责任与信任的感知。如何设计委托界面、用户需要多少意识，以及代理中介的体验如何影响长期信任与可问责性，仍是自动化出行领域一项开放的研究议程。类似的代理中介交互模式也可能出现在家用机器人、智能建筑和协作式人工智能系统等领域。

参考文献

[1] Aline Belloni、Alain Berger、Olivier Boissier、Grégory Bonnet、Gauvain Bourgne、Pierre-Antoine Chardel、Jean-Pierre Cotton、Nicolas Evreux、Jean-Gabriel Ganascia、Philippe Jaillon等. 2015. 应对自主智能体与多智能体系统中的伦理冲突.. 载于AAAI 研讨会：人工智能与伦理. [2] Mark Colley、Pascal Jansen、Mugdha Keskar和Enrico Rukzio. 2025. 利用贝叶斯优化改进自动驾驶车辆的外部通信. 载于2025年CHI计算系统中的人为因素会议论文集. 1–16.[3] Joost de Winter等. 2022. 外部人机界面：噱头还是必需？《心理学最新观点》（ 2022 10 1016 2022 101051 ）. doi: . [4] 1995 /j.copsyc. . Mica R. Endsley. . 迈向动态系统中的情境意识理论 .《人为因素》 37, 1 1995 32 64 10 1518 001872095779049543 ( ), – . [5] 2021 doi: . / 德意志联邦共和国. . 关于修订《道路交通法》及《强制保险法》的法律——自动驾驶法（BGBl. I S. 3108）. https://dejure.org/BGBl/2021/BGBl._I_S._3108. 对《道路交通法》的修订确立了自动驾驶功能（4级）的法律框架，并于2021年7月28日生效。[6] Lorena Hell、Janis Sprenger、Matthias Klusch、Yoshiyuki Kobayashi和Christian Müller. 2021. 日本与德国的行人行为：综述. 载于2021年IEEE智能车辆研讨会（IV）. IEEE, 1529–1536. [7] Pascal Jansen、Mark Colley、Svenja Krauß、Daniel Hirschle和Enrico Rukzio. 2025. Opticarvis：利用贝叶斯优化改进自动驾驶车辆功能可视化以提升用户体验. 载于2025年CHI计算系统中的人为因素会议论文集. 1–23. [8] John D. Lee和Katrina A. See. 2004. 自动化信任：为适当依赖而设计. 《人为因素》46, 1 (2004), 50–80. doi:10.1518/hfes.46.1. 50.30392[9] Xishun Liao、Zhouqiao Zhao、Matthew J. Barth、Amr Abdelraouf、Rohit Gupta、Kyungtae Han、Jiaqi Ma和Guoyuan Wu. 2025. 驾驶行为个性化综述：数据集、建模与验证. IEEE《汇刊》 智能车辆 10, 2 2025 1241 1262 10 1109 2024 3425647 (), –. [10] 2007 doi:./TIV..Régis Lobjois 和 Viola Cavallo. . 过街决策中的年龄相关差异：车辆速度和时间限制对估计任务中间隙选择的影响。 事故分析与预防 39, 5 (2007), 934–943. [11] OpenAI. 2025. 隆重介绍 Operator。 https://openai.com/index/introducing-operator/。访问日期：2026-02-15。 [12] 世界卫生组织。2024。《*2023*年道路安全全球状况报告》：国家和地区概况。世界卫生组织。 [13] Joon Sung Park、Joseph C. O'Brien、Carrie J. Cai、Meredith Ringel Morris、Percy Liang 和 Michael S. Bernstein。2023。生成式智能体：人类行为的交互式模拟。载于 第*36*届*ACM*用户界面软件与技术研讨会年会论文集（*UIST '23*）。 doi:10.1145/3586183.3606763[14] Philipp Spitzer、Joshua Holstein、Patrick Hemmer、Michael Vössing、Niklas Kühl、Dominik Martin 和 Gerhard Satzger。2025。人机协作中的人类委托行为：情境信息的影响。《*ACM*人机交互汇刊》 9, 2 (2025), 1–28. [15] Annika Stampf、Mark Colley、Bettina Girst 和 Enrico Rukzio。2024。利用大语言模型探索自然交互中的乘客与自动驾驶车辆协商。载于 第*16*届国际汽车用户界面与交互式车载应用会议论文集。350–362。 [16] Annika Stampf、Mark Colley、Ann-Kathrin Knuth、Cagla Tasci 和 Enrico Rukzio。2024。探讨高度自动化车辆的心理冲突处理策略，以解决法律层面的用户与车辆冲突。《*ACM*交互式、移动、可穿戴与泛在技术汇刊》 8, 3 (2024), 1–25。 [17] Laurin Vasile、Barbara Seitz、Verena Staab、Magnus Liebherr、Christoph Däsch 和 Dieter Schramm。2023。个人驾驶风格和体验到的系统特征对自动驾驶中驾驶风格偏好的影响。《应用科学》 13, 15 (2023), 8855。 [18] Manfred Wischnewski 等。2023。衡量并理解自动化系统的信任校准：一项实证人类受试者研究综述。载于*2023*年*CHI*计算系统中的人为因素会议论文集（*CHI '23*）。