

对话中的动态认识摩擦

Timothy Obiso¹ Kenneth Lai¹ Abhijnan Nath²

Nikhil Krishnaswamy² James Pustejovsky¹

¹Brandeis University, Waltham, MA USA

²Colorado State University, Fort Collins, CO USA

{ timothyobiso, klai12, jamesp } @brandeis.edu

{ abhijnan.nath, nkrishna } @colostate.edu

Abstract

最近在让大型语言模型 (LLMs) 与人类偏好对齐方面的发展, 显著增强了它们在人机协作场景中的实用性。然而, 这些方法往往忽略了“认知摩擦”的关键作用, 或在应对新的、相互矛盾的或模糊的信息时遇到的固有抵抗力。在本文中, 我们将 **动态认知摩擦** 定义为对认知整合的抵抗力, 其特征在于代理当前信念状态与由外部证据支持的新命题之间的不对齐。我们将其置于动态认知逻辑 (?) 的框架中, 其中摩擦在交互过程中表现为非平凡的信念修正。接下来, 我们通过一个情境对话任务的分析, 展示了这种认知摩擦模型如何有效预测对话中的信念更新, 然后我们讨论了如何自然地使这种作为认知阻力或摩擦度量的信念对齐模型变得更复杂, 以适应现实世界对话场景的复杂性。这篇注释融合了我们所谓的对话和基于证据情境中动态认知摩擦 (DEF) 的两条互补研究路线。首先, 我们阐述了“摩擦”这一概念作为日常交流中对认知更新的抵制, 并说明它如何揭示隐藏的意图或策略上的不对齐。然后, 我们将摩擦置于动态认知逻辑 (DEL) 框架中, 摩擦在此作为一个非平凡的信念修正要求出现。我们通过探索一种将摩擦编码为几何 (不) 对齐的向量表示的认知状态来扩展这些想法。接着, 我们整合了基于证据的 DEL 视角, 以及共同基础跟踪 (CGT) 和心智模拟理论 (SToM), 以在多代理信念归属中展示摩擦。我们还探讨了认知及其命题内容如何在超维向量空间中建模。

1 引言

在合作且基础牢固的对话中, 信息的交换通常显得直接。参与者通常假定彼此信念的更新将是顺畅且符合共同的基础。听者听到说话者的陈述, 并假设信任与共享的背景, 而将其融入其信念, 几乎不带犹豫。然而, 在许多情况下——包括争议和策略性欺骗, 但也有在善意合作中的无意错位——新信息会对信念修正产生抗拒。在这些情况下, 并非所有的更新都能如此顺利地进行。有时, 新信息与听者先前的理解相冲突, 挑战他们的假设, 或暗示隐藏的目

的。在这里, 信念状态的更新过程并不是“毫无摩擦”的, 相反, 听者会遇到一种难以轻易吸收的新信息的“阻力”, 这种现象我们称之为认知摩擦。

对话更新中的阻碍反映了我们处理和适应新信息的复杂性, 同时指向参与者知识状态内更深层次的推理过程。理解这种阻碍可以帮助我们识别出何时说话者可能在误导, 何时对话中存在策略上的不一致, 或何时一个看似简单的陈述实际上包含了更复杂的知识性举动。简而言之, 阻碍提供了对逻辑推理、语用推理和认知表征结构之间微妙相互作用的洞察。

在物理系统中, 摩擦是一种抵抗运动的力。类比地, 认识论摩擦是对信念修正的“运动”的一种阻力。这种阻力可能在认识论上是有益的——鼓励听者更仔细地审视新信息, 或考虑其他解释。它还可能揭示潜在的战略利益、欺骗行为或所传达内容概念结构中的复杂性。在这里, 我们探讨基于证据的动态认识论逻辑 (DEL; ?) 中的摩擦互动, 这是一个建立已久的用于建模信念更新的逻辑框架, 正如最近在 (?) 中探索的那样。

我们引入了一种基于向量的建模方法, 借鉴了全息简化表示 (HRR) (??) 和相关的向量符号架构 (?)。该方法将代理的信念状态和命题视为高维向量, 允许使用正交性和角度等几何概念来表征在吸收新信息时产生的摩擦。通过在符号逻辑和几何直觉之间搭建桥梁, 这一模型为会话中潜在的认知和交流过程提供了新颖的视角。

最后, 我们提供了一个基于情境的协作任务的案例研究, 展示了这种认知摩擦模型如何用于简单的任务相关命题化信念的向量化以及在面对新的对话者断言时的后续更新。我们的分析展示了认知摩擦在对话建模和人机交互中的作用, 并随后讨论了如何自然地使信念对齐模型更复杂, 以适应真实世界对话场景的复杂性, 从而成为认知阻力或摩擦的衡量标准。