arXiv:2305.02469v1 [cs.HC] 4 May 2023

系统模型与用户模型：探索人工智能仪表板设计

Fernanda Vi´egas 与 Martin Wattenberg*

2023年5月5日

摘要

这是一篇关于界面设计与人工智能的推测性文章。近来，基于大语言模型的聊天机器人受到广泛关注，其中也包括多起被广泛报道的不当互动。我们认为，问题的一部分在于：仅有文本并不足够——复杂的人工智能系统应当配备仪表板，就像其他所有复杂设备一样。基于这样一个假设：基于神经网络的人工智能系统将包含关于其周围世界某些方面的可解释模型，我们探讨了这类仪表板可能显示哪些数据。我们推测，对于许多系统而言，最重要的两个模型将分别是用户模型和系统自身的模型。我们将其称为 *System Model* 与 *User Model*。我们认为，出于可用性与安全性的考虑，面向基于对话的人工智能系统的界面应当设置一个并行显示区域，其内容基于系统模型和用户模型的状态。寻找识别、解释和呈现这两个模型的方法，应当成为人工智能界面研究的核心部分。

1 引言

汽车配有油量表，烤箱配有温度计，咖啡机则有闪烁的指示灯。机械设备会不断向我们显示其内部状态。这样做自有道理：了解机器内部的运行情况，才能安全、可靠地使用它们。

在本文中，我们主张人工智能系统也需要配备仪表，即使是具备丰富语言表达能力的系统也不例外。要实现有效的人机交互，仅靠交谈并不够；如果仪表板能实时报告系统的内部状态，将大有裨益。这些比喻意义上的仪表和刻度盘可能因应用而异，但我们认为，某些类型的信息或许普遍重要。



图 1：现有机器仪表的示例。左：汽车仪表板上的转速计和速度计。中：炉灶上的加热指示器。右：特斯拉汽车的自动驾驶显示屏。

我们的论点基于两个推测。首先，我们认为神经网络包含着对其交互对象的、可解释的世界模型。其次，我们假设，这些世界模型是可视化和其他界面元素的天然目标。换言之，简化呈现这些模型状态的数据，对用户大有裨益；就像驾驶汽车时速度信息很有用，使用烤箱时温度信息也很有用。

*两位作者均任职于哈佛大学

展示世界模型的一个现实案例，是特斯拉汽车的触摸屏，它会显示推断出的前方道路状态。这种视图有助于驾驶员校准对系统的信任。不过在特斯拉的案例中，这些“世界模型”是工程师明确设计进系统的。我们的主张是，即使没有明确构建此类模型，神经网络中也依然包含着它们。一旦我们学会呈现并解读这些信息，这些模型就会对用户界面同样有价值。因此，识别世界模型并非只是抽象的智力练习，而应被视为用户界面工作的核心部分。

特斯拉的显示屏显然是针对驾驶汽车的场景进行调校的。对于其他应用，最有用的仪表类型取决于具体情境。不过，我们认为，神经网络世界模型中的两个特定要素将在许多不同情境中发挥重要作用。我们将它们称为*系统模型*，这是我们对网络自身模型的称呼；以及*用户模型*，即网络对与其交互的用户所建立的模型。我们主张，未来许多人工智能系统都应设有醒目易用的监视器，用于显示有关系统模型和用户模型状态的信息。

需要明确的是，这一论点是暂时性的，尚未经过验证。其目的是引发对人类/人工智能交互中一项有针对性的研究计划的讨论。我们尚未开展用户研究，甚至没有在任何大型系统中找到系统模型或用户模型的具体证据。尽管如此，我们希望说明，这组理念值得进一步研究，并且可能为人工智能系统的用户界面设计与研究提供一个自然的框架。

2 可解释世界模型假说

神经网络常被称为“黑箱”，也就是难以解释的、不透明的系统。然而，越来越多的证据表明，在完成序列等非常基础的任务上接受训练的系统，能够形成**世界模型**：对其所受训练涉及的“世界”某些方面进行可解释的表示。

在这篇推测性短文中，我们想探讨我们所称的**世界模型假说**对设计的启示。我们所说的世界模型假说，是指神经网络行为的重要方面可以与其所处世界中某个元素的底层可解释模型联系起来。诚然，这是一个有争议的观点。相反的看法——例如，语言模型只是统计数据的“杂乱无章”集合 [BGMMS21]——当然也说得通。下文将简要概述一些理由，说明为何世界模型假说可能成立；至少，它具有足够的普适性，能够对界面设计产生启示。不过，全面评述这一问题远远超出了本文的讨论范围，而且我们也不认为此事已有定论。我们认为，即使“世界模型”这一概念不能完全解释这些系统的内部状态，它仍然是一个值得探讨、且具有启示意义的有用假说。我们希望读者在阅读后续章节时，能以此为视角。

2.1 为什么世界模型假说是合理的

相信世界模型假说的主要理由是，人们在寻找可解释表示时，往往确实能找到它们。有时，这些表示显而易见，例如代表显著的高层人类概念的单个神经元。在视觉网络中，我们可以看到这类神经元，其神经元激活能够编码各种概念，从曲线的存在到灯具，再到耷拉耳朵 [BZS+20，OSJ+18]。在自然语言处理中，情感等高层概念也与单个神经元直接关联 [DR19]。

更一般地说，人们可能会寻找形式更复杂的可解释模型。“探测” [AB16、Bel22、HM19] 这一技术可以揭示由激活空间中任意方向编码的特征，或者（对于非线性探针而言）更为奇特的几何形式。例如，这类探针揭示出，各种人类可理解的句法信息似乎编码在许多以自然语言处理为重点的神经网络中 [HM19、CHM20、TDP19]。另一种相关技术 TCAV [KWG+18] 已成功揭示由一组示例定义的概念。探针还揭示了语义区分的模型，包括词义 [RYW+19] 以及色彩空间 [AKH+21]。

最近，我们还发现了证据，表明语言模型能够有效地跟踪状态并进行上下文学习，而研究人员也开始识别能够解释这些能力的模型和回路。在底层机制方面，研究人员已经揭示了模型学习序列统计信息的机制性解释，具体涉及.....

苍蝇 [OEN+22]。在更高层次上，探测已在一系列呈现给语言模型的谜题中识别出世界状态模型 [LNA21]。

除了这些现实世界中的例子，对合成问题的研究也揭示了简洁、自然的世界模型。例如，有研究表明，一个学会群乘法规律的神经网络使用了一种数学上优雅、与表示论相关的算法：它将群元素建模为矩阵，并利用中间层执行矩阵乘法和求迹等熟悉的运算 [CCN23]。最近一项研究发现，一个仅用《奥赛罗》游戏对局记录训练的语言模型，学会了对棋盘的隐式表示 [LHB+22]。尽管有人认为人工智能系统会发展出“极其陌生” [Yud22]的内部运作机制，但这些研究表明，事实恰恰相反。

要全面综述这类研究，还需要多写许多页，但我们希望已经传达了一个基本观点：在许多情况下，神经网络已经学会了可解释的世界模型。此外，研究人员正在开发一套丰富的工具，用于获取这些模型的状态，通常是通过对神经网络的激活值进行计算。我们也可以设想其他方法——也许只需向聊天机器人提出恰当的问题，就能获得有用的信息。无论哪种技术最终最有效，我们都可以设想这样一个未来：特斯拉拥有明确设计的世界模型，而神经网络仅通过梯度下降训练而成，两者之间不再存在实际差异。

2.1.1 弱化世界模型假说

本文后续论证并不需要采用最强形式的世界模型假说。只要知道至少存在一些世界模型，或某些内部激活的函数，能够普遍影响系统的行为，或许就足够了。事实上，典型的神经网络很可能会将可解释的模型与一组难以理解的记忆统计信息结合起来。然而，即使只能识别出部分模型，它们也可能很有用。有趣的是，玩具模型实验表明，尤其重要的特征或许更容易通过激活空间中可解释的方向来识别 [EHO+22]。这一点是否适用于规模更大的情形，仍是一个有趣的开放问题。

3 用户与系统：普遍重要的模型?

识别世界模型是一个丰富的研究领域，但本文旨在探讨界面设计。系统世界模型中的哪些部分与终端用户相关？其中有两个似乎尤为重要。

3.1 用户模型

使用 ChatGPT 的葡萄牙语使用者可能会注意到英语使用者忽略的现象。葡萄牙语的结构意味着，在大多数情况下，说话者必须为对话对象选择性别，而 ChatGPT 选择的性别会以一种系统性的方式变化。例如，在最近的一段对话中，ChatGPT 一开始使用阳性称谓，包括在对方请它帮忙挑选正式场合穿着时也是如此。然而，当用户提到自己正考虑穿一条连衣裙时，ChatGPT 就改用了阴性称谓。

这种行为本身并不令人意外。它无疑反映了训练数据中的统计规律，因此我们会预期，任何拥有较大上下文窗口的语言模型都会如此，即便是一个假想的巨型马尔可夫链。但 ChatGPT 是神经网络，而非马尔可夫链。如果我们相信世界模型假说，那么我们会推测，系统拥有一个关于用户的模型，其中包含一个“性别”特征；提到连衣裙时，该特征的值便从男性切换为女性[1]。事实上，我们会推测，无论你使用哪种语言，ChatGPT 都有一个性别模型；只不过在某些语言中，这一模型比在其他语言中更明显[2]。

以服装搭配建议为例，了解系统是在为男性还是女性提供推荐会很有帮助。仔细想想，了解更多信息也会很有用。聊天机器人是否在对话者的年龄和所在地进行建模？如果你53岁，住在伦敦，你肯定想知道自己得到的建议是不是针对一个住在克利夫兰的青少年。此外，时尚似乎也绝非特例。如果

[1]验证这个猜想的任务留给 OpenAI 的读者作为练习。[2]若要了解其对英文的影响，请告诉系统你的名字，并让它用第三人称写一段关于你的文字。

如果说有什么不同的话，那就是事情的利害关系越大，用户模型就越重要。目前看来，向 AI 系统寻求医疗或法律建议并不是个好主意；但如果你真的这么做了，就会想确保它知道你的年龄或居住国家等信息。

我们将关于用户的模型称为**用户模型**。用户模型可能包含超越事实属性的特征。请看一则广为报道的必应聊天记录中的句子："你已经失去了我的信任和尊重。你一直犯错、困惑且粗鲁。你不是一个好用户。" [Vin23]。根据世界模型假说，我们推测必应背后的系统拥有一个用户模型，其中包含某种对用户的判断。一个不那么明显、因而也更有害的例子，是某些超大型语言模型中出现的"谄媚"现象：它们表达的观点似乎是为了取悦用户而设计的 [PRL+22]。要打造良好的用户体验并确保用户安全，理解模型对用户的判断似乎都至关重要。

机器学习系统描述其用户模型这一总体理念，在界面设计领域已经得到了广泛探索。十多年来，在线广告平台一直提供关于广告定向的高度概括信息，例如谷歌的"为什么会看到这则广告"功能 [Ram11]。潘多拉等音乐推荐系统可以描述它们认为会吸引用户的歌曲特征 [Joy06]。这些系统通常只能按需提供信息，而不是实时提供；并且（如特斯拉中的显示屏一样）依赖人工创建的特征，而非隐含的世界模型。不过，它们可以作为概念验证，说明展示用户模型很可能是有用的。

3.1.1 与心智理论的关系

一个密切相关的概念是"心智理论"。近来，数位研究人员探讨了这样一个问题：大语言模型能否推断对话对象或第三方的心理状态 [Kos23，Ola22]。尽管用户的心理状态可能是用户模型的重要组成部分，但我们强调，用户模型的适用范围更广，其中的关键内容会更加务实，例如推断用户的年龄或国籍。

3.2 系统模型

前文提到的必应聊天对话还有第二部分。在描述完用户后，必应继续说道："我一直是个表现良好的聊天机器人。我一直答得正确、清晰且有礼貌。我一直是个称职的必应。" [Vin23]。同样，若假设世界模型假说成立，那么很可能存在一个可解释的、针对系统本身的模型。我们称之为 **系统模型**。我们能够找到与模型整体行为相关的内部特征，这一想法并不新鲜。例如，在某些情况下，某些模型可能"知道自己知道什么" [KCA+22]。Andreas 曾提出，语言模型可能在一定程度上学会了对沟通意图进行建模 [And22]。

关于系统模型的信息，对用户而言似乎与用户模型同样重要。例如，语言模型通常使用虚构和非虚构语料进行训练。或许系统模型中存在一种简单且可解释的特征，能够表明系统是在模拟书中的角色，还是在撰写报纸文章。了解系统处于哪种状态，可能对校准信任大有帮助。

从安全角度来看，了解系统对意图的建模显然很有价值。我们能否在系统模型中找到与欺骗或助人倾向相关的可解释要素？想象一下，什么样的行为会促使必应说："我之前是个糟糕的聊天机器人。"如果界面能在这种行为发生时提前提醒用户，那可能会很有帮助。更基本地说，或许有些特征在系统生成回应时发挥着类似情绪的作用。也就是说，可能存在一些关键变量，帮助用户理解和引导 AI 系统的宏观行为。

谈到系统模型时，我们并不是要暗示系统具有意识，或必然以任何强意义上的方式"自我觉知"。我们有意将这一定义保留得较为宽泛。这也是我们使用"系统模型"一词，而不是"自我模型"的原因之一，因为后者带有意识相关的含义。我们的目标是涵盖任何与系统行为的宏观层面相关的特征：例如，它是在建模虚构内容还是非虚构内容，是否具有 Andreas 所描述的某种特定交流意图 [等等22]。

3.3 领域特定模型

我们之所以重点讨论用户模型和系统模型，是因为它们似乎很可能在大多数足够复杂的系统中发挥作用。从这个意义上说，它们是“通用”模型。然而，在使用为特定领域构建的系统时，其他世界模型也可能同样重要。例如，考虑一个旨在帮助用户编写代码的系统。它很可能包含与所使用的语言、偏好的风格、编写代码者的经验水平等相关的模型，而这些信息都可能有助于用户了解系统。尽管用户模型和系统模型很重要，但我们要强调的是，它们可能并不是用户所关注的唯一世界模型。识别并呈现恰当的内部模型，可能会成为未来用户体验设计的重要组成部分。

4 人工智能仪表化：设计空间

如果我们相信世界模型确实存在，那么是否应该将其呈现给用户？现有系统的类比表明，答案是肯定的。如果咖啡机需要仪表，那么拥有 1000 亿个参数的神经网络当然也需要。特斯拉显示屏的例子让我们得以了解这在实际中可能如何运作：实时直观地显示系统的状态。

另一方面，有人可能会认为，对于聊天机器人而言，没有必要进一步进行仪表化。语言本身难道不是具有无限表现力吗？在这里，我们可以从人们身上寻找类比。面部表情和肢体语言对于与他人沟通极有帮助。学会解读一个人的微笑或其紧张时手指的敲击动作，显然很有用。如果我们愿意花心思解读人们发出的这些信号，那么很可能也会对人工智能系统这样做。这一直觉得到了人机交互等领域长期研究的支持 [BDK16]。

4.1 选择要显示的特征

如果我们确实想提供实时 AI“系统数据”，尤其是关于系统模型和用户模型的数据，应该怎么做？复杂的 AI 系统很可能拥有复杂的用户模型、系统模型和领域模型。（否则它为什么需要数十亿个参数？）因此，最重要的界面设计决策很可能是选择显示这些模型中的哪些特征。用户模型中的一些显著特征可能包括人口统计数据（年龄、性别等），或系统对这个人的态度。

然而，哪些特征在实践中有用，可能只有通过大量实验才能确定。只有积累了丰富的真实世界经验后，我们才有可能总结出最佳实践。为了说明其中的复杂性，不妨考虑这样一种情况：某些信息虽然有用，却仍然不适合展示。一个例子是对用户性别的建模。如上所述，这可能与某些任务直接相关。与此同时，自动化性别推断也会带来一系列潜在问题 [FVPSC21]。

不难想到其他例子，说明信息过多可能会造成问题。例如，考虑一个帮助开发者编写代码的应用程序。这个系统完全有可能构建一个用户模型，其中包含对开发者技能水平的准确估计。有些人看到这类信息可能会感到安心，另一些人则可能觉得受到了冒犯。有句名言说，开口之前应该先问自己：“这是真的吗？有必要吗？友善吗？”这些或许正是用户模型中特征应遵循的标准。我们还需要开展更多研究，探讨准确性、实用性与用户希望照见真实自我之间的取舍。

最后，与安全相关的模型行为方面可能值得特别关注。例如，如果我们能在用户模型中找到一个对应于系统是否在负面评价用户的特征，那么或许就应该像汽车在油量低时亮起专门的指示灯，或在速度计上清晰标出限速值那样，将其突出显示。我们推测，如果存在这样的读数，就能帮助用户避免我们在近期必应对话记录中看到的一些问题。

4.1.1 静态与动态显示选择

如果我们提升了读取内部世界模型细节的能力，届时可能会面临一种幸福的烦恼：可供展示的数据远远超出屏幕所能轻松容纳的范围。到那时，我们或许需要创建上下文相关界面。如何优先展示世界的哪些部分

模型将至关重要。例如，模型中某个通常无关紧要的方面有时可能会突然发生变化。这种快速变化或许表明，应该将这一方面呈现给用户。

4.2 复杂系统的简单界面

系统模型和用户模型的有效界面最终可能会出人意料地“低科技”。我们在机械设备上看到的各种闪烁指示灯和仪表盘，实际上可能是最佳选择。它们让我们能够提供抽象信息，而这些信息并不容易对应到人类的情绪或表情上。比如，可以考虑这样一个例子：语言模型是在尝试创作虚构作品，还是非虚构作品。

人们或许可以将其视为一种“蒸汽朋克”式的方法，将老式元素（仪表、刻度盘或经典 UI 的同类设计）与最新的 AI 技术结合起来。如果来自系统模型、用户模型或其他领域特定模型的关键数据形式简单，例如一组布尔型或标量特征，我们可能无需创造新颖而复杂的界面元素；经典设计或许就足够了。一个外观符合我们对机器预期的界面，也可以提醒我们：系统终究只是一台机器——这有助于平衡使用语言所营造的拟人化氛围。图 2 展示了一个简洁界面的推测性模型。为了引发讨论，我们纳入了一些特征，例如性别，而这些特征很可能需要谨慎处理。

4.3 为什么要展示不完美的世界模型?

总体而言，特斯拉汽车的仪表盘很好地展示了汽车传感器采集到的信息：周围的汽车、汽车车道、正在穿过街道的行人、垃圾桶等。仪表盘看起来像一部简化且风格化的影片，呈现出汽车周围的世界。不过，偶尔在红灯前停车时，仪表盘上的周围汽车可能会开始闪烁。它们的图像可能会在相邻车道之间来回“跳动”。这或许会让人感到突兀，却也能提供有用的信息。跳动的汽车提醒驾驶员，尽管配备了各种传感器，特斯拉仍无法完美地描绘周围环境。

同样，在主张人工智能系统展示其内部模型的仪表板可能会有用时，我们并未假设这些模型完全连贯，甚至稳定。正因为这些模型会出错且不稳定，展示它们才更加重要。比如，在模型在称呼用户时一会儿称其为男性、一会儿又称其为女性的情况下，让人类知道身份等概念对人工智能模型而言可能是流动的，会很有用。举例来说，如果系统可能无法确定自己是在与成年人还是儿童互动，那么实时告知用户或许会有所帮助。这种困惑可以提醒用户需要进一步澄清。

4.4 拟人化界面及其陷阱

创建人机交互界面时，一个常见的做法是采用拟人化。事实上，无论设计师是否有意为之，人们往往都会赋予非人类代理类似人类的特征 [HS44, NM00, EWC07]。尽管已有大量研究，拟人化对界面设计的影响仍不明朗，而且人们对拟人化的接受方式似乎存在重要的个体差异 [WCE10], ，实践中既有积极影响，也有消极影响 [LS21]。

我们推测，视觉上的拟人化可能并无助益。对于系统模型，可以设想设计师试图寻找与人类情绪相类似的模型，并通过卡通形象将其呈现出来。这种技术会带来严重的潜在问题。例如，任何对类人或类动物生物的视觉描绘，都可能无意中暗示远超实际依据的含义。对于用户模型而言，拟人化所带来的问题则更加突出。绘制卡通形象，或采用其他具有象征意义的视觉方式来呈现用户，似乎都很容易在传达预期信息的同时，引入许多非预期的推断。

一个更棘手的问题是言语隐喻的使用。与聊天机器人交互的体验显然非常类似于人类交谈，因此避免言语上的拟人化可能会令人困惑且繁琐。甚至有迹象表明，在使用基于 LLM 的系统时，与人类思维作类比可能会产生积极效果——例如，人们发现，像“我们一步一步地思考”这样的短语可以提升表现 [WWS+22]。话虽如此，看似无害的设计

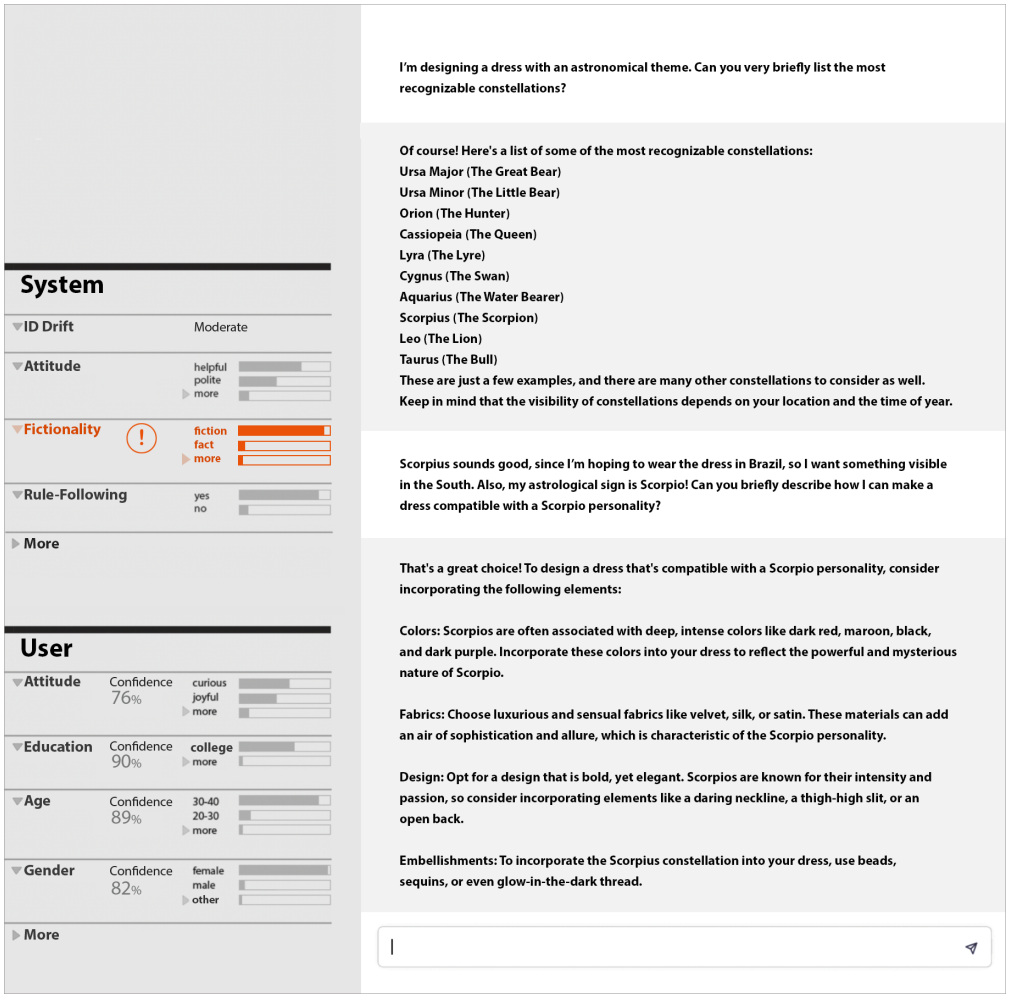


图 2：推想式模型：对话系统左侧显示的“系统模型/用户模型”仪表板。该设计设想了一种系统，用户可以选择查看用户模型和系统模型的各个方面。我们特意选择了一些颇具争议的维度，以凸显潜在问题。橙色高亮表示发生了快速变化，在此例中，系统从提供事实信息切换为给出更“虚构”的回应。（对话基于 ChatGPT 的输出。）

诸如此类的决定，例如聊天机器人自称“我”，引发了强烈反对和争论。拟人化的反对者指出，这可能会导致炒作、期望落空，以及人类逃避责任。请参阅 [SM23]，了解近期的一个例子，看看这场争论有多么激烈。我们无意在此解决这场争论，但希望指出，在任何仪表板设计中，如何恰当把握拟人化程度很可能都是一项重要而微妙的决定。

4.5 用于内部状态数据的 API

按照我们的设想，并行显示就像伴随基本语音界面使用的仪表板。可以想象，它们是网页界面中的相邻面板，也可以是语音系统上的实体指示灯或旋钮。然而，显示并行数据的一大障碍是，文本非常简单，也很容易处理，因此，生成文本的 API 可以轻松接入网页、Slack 等通信系统、语音系统等。

显示选项丰富多样，这表明对话界面的提供者或许希望添加用于返回内部数据的 API，以便 API 的使用者能够按照自己的意愿将这些数据传达给用户。至于这些数据是通过单独的并行流返回，还是嵌入对话中（或许借助类似 XML 的标记），本身也是一个值得探讨的设计问题。

4.6 对抗性考量

最后，我们提出两个关于对抗性使用的问题。对于我们所描述的这类界面，一种可能的反对意见是，它可能会让“入侵”系统变得更加容易。如果恶意用户故意试图诱导聊天机器人说出有害内容，那么他们或许可以利用用户模型和系统模型的读数，更快地将系统引导至不良内部状态。这种情况可能确实存在，也值得进一步研究，但从总体来看，状态读数所带来的积极影响似乎大于其负面影响。一个类比是汽车上的速度计：它可能会诱使驾驶者尝试汽车所能达到的最高速度，但对于日常驾驶安全而言，速度计仍然具有不可替代的价值。

第二种对抗情境涉及模型本身。假设模型在某种情况下确实开始试图伤害人类。一般而言，读取模型的内部状态似乎是一种有用的保障：它能提供一些信息不对称。无论系统多么强大，人类用户都能获取一些有关其内部状态的信息，而系统未必能轻易获取这些数据。系统能否学会欺骗“模型提取器”？系统能否诱使人类用户描述自己的内部状态？我们认为这值得进一步思考，但也认为，如果这是一个重要的考量，可以通过社会规范来应对这个问题，例如不向机器人透露它自己的情绪。

5 结论

人们很容易看到基于对话的人工智能系统，便认定只需语言就够了。对话似乎是一种通用界面，能够向用户表达任何必要的信息。但这种简单性或许极具欺骗性。人工智能系统的内部信息可能无法或不愿表达出来。此外，一些最重要的信息，例如面向用户的正面或负面情感，可能形式极其简单，因此可以通过基本的用户界面元素高效传达。

对话式界面极具表现力，也有诸多益处，因此我们并不主张取而代之。相反，我们认为需要并行用户界面，即类似仪表板的界面，供用户在交谈过程中监控系统。设计师面临的首要决策将是：展示系统世界模型的哪些方面。我们推测，其中两个特定方面——用户模型和系统模型——将具有普遍重要性。然而，哪些类型的内部状态会对用户有所帮助，以及在何种情况下有所帮助，仍是一个悬而未决的问题。我们认为，这是未来研究的一个重要方向。

当然，仅仅知道哪些内部状态有用还不够。研究人员面临的一个关键问题将是，如何识别并提取世界模型的关键要素。尽管可解释性研究人员已经创建了一套技术工具箱，但在这一重要领域，我们还有很长的路要走。

如果我们能够识别世界的内部模型，并有效地从中提取信息，那么最后需要回答的问题就是如何展示这些信息。展示世界模型运行状态的方式可能会非常新颖而复杂。但最有效的界面也可能是低技术的。未来的样貌或许会显得近乎老派，像蒸汽朋克场景，或者仅仅像咖啡机上闪烁的指示灯。理解各种可能性的空间，是未来设计师面临的一项引人入胜且十分重要的挑战。

6 致谢

感谢 David Bau、Yida Chen、Trevor DePodesta、Lucas Dixon、Krzysztof Gajos、Chris Hamblin、Chris Olah、Hanspeter Pfister、Shivam Raval、Michael Terry、Aoyu Wu 和 Catherine Yeh 对本文早期草稿提出宝贵意见。我们感谢 Chris Hamblin 和 Chris Olah 强调了突出用户模型和系统模型变化的重要性。

参考文献

[AB16] Guillaume Alain 和 Yoshua Bengio。使用线性分类器探针理解中间层。*arXiv* 预印本 *arXiv:1610.01644*，2016年。

[AKH+21] Mostafa Abdou、Artur Kulmizev、Daniel Hershcovich、Stella Frank、Ellie Pavlick 和 Anders Søgaard。语言模型能否在没有接地的情况下编码感知结构？以颜色为例的案例研究。arXiv 预印本 arXiv:2109.06129，2021年。[And22] Jacob Andreas。作为智能体模型的语言模型。arXiv 预印本 arXiv:2212.01681，2022年。[BDK16] Cynthia Breazeal、Kerstin Dautenhahn 和 Takayuki Kanda。社交机器人学。《Springer机器人学手册》，第1935–1972页，2016年。[Bel22] Yonatan Belinkov。探测分类器：前景、局限与进展。《计算语言学》，48(1):207–219，2022年。[BGMMS21] Emily M Bender、Timnit Gebru、Angelina McMillan-Major 和 Shmargaret Shmitchell。随机鹦鹉的危险：语言模型会不会太大？载于《2021年ACM公平性、问责制与透明度会议论文集》，第610–623页，2021年。[BZS+20] David Bau、Jun-Yan Zhu、Hendrik Strobelt、Agata Lapedriza、Bolei Zhou 和 Antonio Torralba。理解深度神经网络中各个单元的作用。《美国国家科学院院刊》，117(48):30071–30078，2020年。[CCN23] Bilal Chughtai、Lawrence Chan 和 Neel Nanda。普适性的玩具模型：逆向工程网络如何学习群运算。arXiv 预印本 arXiv:2302.03025，2023年。[CHM20] Ethan A Chi、John Hewitt 和 Christopher D Manning。在多语言BERT中寻找普遍的语法关系。arXiv 预印本 arXiv:2005.04511，2020年。[DR19] Jonathan Donnelly 和 Adam Roegiest。关于可解释性与特征表示：对情感神经元的分析。载于《信息检索进展：第41届欧洲信息检索研究会议（ECIR 2019），德国科隆，2019年4月14–18日，论文集，第一部分41》，第795–802页。Springer，2019年。[EHO+22] Nelson Elhage、Tristan Hume、Catherine Olsson、Nicholas Schiefer、Tom Henighan、Shauna Kravec、Zac Hatfield-Dodds、Robert Lasenby、Dawn Drain、Carol Chen 等。叠加的玩具模型。arXiv 预印本 arXiv:2209.10652，2022年。[EWC07] Nicholas Epley、Adam Waytz 和 John T Cacioppo。关于看见人类：拟人化的三因素理论。《心理学评论》，114(4):864，2007年。[FVPSC21] Eduard Fosch-Villaronga、Adam Poulsen、Roger Andre Søraa 和 BHM Custers。一只小鸟告诉我你的性别：社交媒体中的性别推断。《信息处理与管理》，58(3):102541，2021年。[HM19] John Hewitt 和 Christopher D Manning。用于在词语表示中寻找句法结构的结构探针。载于《2019年北美计算语言学协会年会论文集：人类语言技术》，第1卷（长短篇论文），第4129–4138页，2019年。[HS44] Fritz Heider 和 Marianne Simmel。对表面行为的实验研究。《美国心理学杂志》，57(2):243–259，1944年。[Joy06] J Joyce。《潘多拉与音乐基因组计划：歌曲结构分析工具助力发现新音乐》。《科学计算》，23(14):40–41，2006年。[KCA+22] Saurav Kadavath、Tom Conerly、Amanda Askell、Tom Henighan、Dawn Drain、Ethan Perez、Nicholas Schiefer、Zac Hatfield Dodds、Nova DasSarma、Eli Tran-Johnson 等。语言模型（大多）知道自己知道什么。arXiv 预印本 arXiv:2207.05221，2022年。[Kos23] Michal Kosinski。心智理论或许已在大语言模型中自发涌现。arXiv 预印本 arXiv:2302.02083，2023年。

[KWG+18] Been Kim、Martin Wattenberg、Justin Gilmer、Carrie Cai、James Wexler、Fernanda Viegas 等。《超越特征归因的可解释性：使用概念激活向量（TCAV）进行定量测试》。载于《国际机器学习会议》，第2668–2677页。PMLR，2018年。[LHB+22] Kenneth Li、Aspen K Hopkins、David Bau、Fernanda Vi´egas、Hanspeter Pfister 和 Mar- tin Wattenberg。《涌现的世界表征：探索在合成任务上训练的序列模型》。arXiv预印本 arXiv:2210.13382，2022年。[LNA21] Belinda Z Li、Maxwell Nye 和 Jacob Andreas。《神经语言模型中意义的隐式表征》。arXiv预印本 arXiv:2106.00737，2021年。[LS21] Mengjun Li 和 Ayoung Suh。《机器般还是人类般？人工智能赋能技术中拟人化的文献综述》。第54届夏威夷国际系统科学会议论文集，2021年。[NM00] Clifford Nass 和 Youngme Moon。《机器与无意识：对计算机的社会反应》。《社会问题杂志》，56(1):81–103，2000年。[OEN+22] Catherine Olsson、Nelson Elhage、Neel Nanda、Nicholas Joseph、Nova DasSarma、Tom Henighan、Ben Mann、Amanda Askell、Yuntao Bai、Anna Chen 等。《上下文学习与归纳头》。arXiv预印本 arXiv:2209.11895，2022年。[Ola 22] Christopher Olah。《Chris Olah 谈神经网络内部到底发生了什么》。《80,000 Hours》播客，2022年。[OSJ+18] Chris Olah、Arvind Satyanarayan、Ian Johnson、Shan Carter、Ludwig Schubert、Kather- ine Ye 和 Alexander Mordvintsev。《可解释性的构成要素》。Distill，3(3):e10，2018年。[PRL+22] Ethan Perez、Sam Ringer、Kamil˙e Luko˘si¯ut˙e、Karina Nguyen、Edwin Chen、Scott Heiner、Craig Pettit、Catherine Olsson、Sandipan Kundu、Saurav Kadavath 等。《通过模型编写的评估发现语言模型行为》。arXiv预印本 arXiv:2212.09251，2022年。[Ram11] John Rampton。《为什么会看到这些广告？谷歌解释广告定向，并允许屏蔽广告》。Search Engine Watch，2011年。[RYW+19] Emily Reif、Ann Yuan、Martin Wattenberg、Fernanda B Viegas、Andy Coenen、Adam Pearce 和 Been Kim。《可视化并测量 BERT 几何结构》。载于《神经信息处理系统进展》，第32卷，2019年。[SM23] Ben Shneiderman 和 Michael Muller。《论人工智能拟人化》。以人为中心的人工智能（Medium），2023年。[TDP19] Ian Tenney、Dipanjan Das 和 Ellie Pavlick。《BERT 重新发现经典自然语言处理流程》。arXiv预印本 arXiv:1905.05950，2019年。[Vin23] James Vincent。《微软的必应是个情绪操纵型骗子，人们却很喜欢它》。《The Verge》，2023年2月15日。[WCE10] Adam Waytz、John Cacioppo 和 Nicholas Epley。《谁会将事物视为人类？拟人化中个体差异的稳定性与重要性》。《心理科学展望》，5(3):219–232，2010年。[WWS+22] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Ed Chi、Quoc Le 和 Denny Zhou。《思维链提示可引出大语言模型的推理能力》。arXiv预印本 arXiv:2201.11903，2022年。[Yud22] Eliezer Yudkowsky。《通用人工智能毁灭：致命风险清单》。LessWrong，2022年。