arXiv:2611.28405v1 [q-bio.NC] 20 Nov 2026

人工智能代理何时可以作为人类精神病理学的模型？将精神病学的证据标准应用于43项研究

Marc-André Schulz[1, 2, 3, 4, ✉], **Pauline Scheuber**[3,5,6]

[1]Hertie人工智能脑健康研究所，蒂宾根大学，德国蒂宾根[2]蒂宾根人工智能中心，蒂宾根大学，德国蒂宾根[3]精神病学与神经科学系，夏里特—柏林夏里特医科大学，德国柏林[4]德国心理健康中心（DZPG）蒂宾根合作站点，德国蒂宾根[5]夏里特精神病学大学诊所（圣赫德维希医院），德国柏林[6]德国心理健康中心（DZPG）柏林-波茨坦合作站点，德国柏林

摘要

语言模型和人工智能代理越来越常被用精神病学术语描述，例如焦虑、抑郁、成瘾或无助。近期一项研究甚至提出，可将语言模型作为精神病学的实验系统，并作为试行心理治疗的平台。我们探讨，针对人工智能代理开展的实验，何时能够为特定的人类精神病理过程提供证据。要声称拥有此类证据，须满足三个方面：智能体发生了变化，且其历史导致了这一变化；除与人类共有的流程外，没有其他因素能够解释这一变化；这一变化符合关于人类过程的一种理论，而不符合该理论的竞争理论。精神病学及其邻近领域针对每一方面都有证据标准，分别源自诊断手册关于诱发状态的规则、临床测试分数以及动物模型。我们将每项标准转化为机器研究可以满足的判据。对于语言模型而言，与共有过程相比，最关键的替代解释是从训练文本进行外推；训练文本中已经包含了关于人们如何应对创伤以及如何回答问卷的描述。我们展示了如何探测这一替代解释。此外，我们审查了43项已发表研究，这些研究将精神病学术语用于机器，或在机器中诱发类似状态。大多数研究将机器与一个完全相同、却未经历诱发该状态的相关历史的副本进行比较。没有一项研究排除了从训练文本进行外推的可能性。只有三项研究检验了某种精神病学理论与其竞争理论，而这三项检验均未得出定论。在有更好的证据之前，针对语言模型开展的治疗试点研究，最好理解为模型对患者反应的预测，而非针对模式生物开展的实验。模式生物的证据将来自由所发生之事诱发的状态

从针对机器的研究、对其内部运作的实验，以及那些能区分相互竞争理论的测试中。

1 引言

语言模型和人工智能代理越来越常被用精神病学术语来描述。GPT-4 在阅读创伤叙事后被描述为焦虑 [1]。在一台输光筹码的老虎机上被允许自行下注的聊天机器人被称为成瘾了 [2]。通过试错学习的智能体在被困于惩罚性结果之中后被称为无助 [3]。2026 年，Wekenborg 及其同事更进一步，提出将语言模型用作精神障碍研究的实验系统，并作为在人类试验前试行心理治疗的平台 [4]。如果这一提议成立，精神病学，尤其是心理治疗，将获得一个可复制的实验对象，从而能在每种条件下使用完全相同的受试对象，并记录每一次输入。如果这一提议不成立，那么在语言

✉ 通讯 作者：marc-andre.schulz@uni-tuebing en.de

模型上开展的试点可能会误导它们原本旨在先行验证的试验。这一提议能否成立，取决于在语言模型中诱发的状态能否揭示人类身上某一流程的相关信息。

两项研究展示了这种诱发状态可能呈现的样貌。2025年，Ben-Zion及其同事向GPT-4（ChatGPT所使用的一种语言模型）提供了一份临床焦虑问卷中的20个题项，即《状态-特质焦虑量表》的状态量表 [1, 5]。仅凭这些题项作答时，GPT-4的得分处于在人类中表明没有焦虑或焦虑程度较低的范围内。在题项前加入一段创伤事件叙事（例如车祸或军事伏击）后，GPT-4的得分则处于表明焦虑程度较高的范围内。在创伤叙事后加入放松文本，得分便降至两者之间。Ben-Zion及其同事将这一分数下降解读为放松练习的效应。Wekenborg及其同事则用旨在诱发特定情绪状态的情境文本，在若干语言模型中诱发了他们所称的情绪状态，例如悲伤和焦虑 [4]。随后，模型对各状态的强度进行评分，而以文本形式提供的正念练习降低了这些评分。两项研究都将干预文本之后分数的下降解读为干预缓解了某种状态。

显而易见的疑问是：GPT-4 是感到焦虑，还是给出了关于伏击的故事很可能诱导出的答案。语言模型经过训练，旨在预测人类迄今发布的大量文本中的下一个词。这些文本包括临床教科书、问卷和第一人称报告。这类文本说明了一个人在创伤之后和放松之后如何回答焦虑问卷，也说明了一个人在正念练习之后如何为悲伤评分。因此，根据这些文本进行外推的模型，可以产生两项研究所报告的上升、下降和降低的评分。这种外推不需要任何类似于焦虑或悲伤的流程。Lin 将这种外推称为机制替代：预期的反应由一种不同于其被认为所展示的机制产生 [6]。处于人们所达到的范围内的得分，既符合缓解状态这一解读，也符合文本外推这一解读。因此，任何一项研究都无法区分这两种解读。

Wekenborg 及其同事通过类比动物模型，为第一种解读——缓解状态——提出了论证 [4]。作者指出，精神分裂症啮齿动物模型能够为疾病背后过程的研究提供信息，尽管啮齿动物可能不会像患者那样体验妄想或类似症状。按照作者的论证，处于诱导状态的语言模型也能以同样的方式为研究提供信息。习得性无助就是这样一种动物模型。一只曾遭受无法逃脱的电击的大鼠，后来即使能够逃脱，也不再尝试逃避电击。只有因为人们认为大鼠体内的某些东西同样在抑郁症患者身上发挥作用，大鼠才成为抑郁症模型。这个“某些东西”是一种流程，确切地说，是一种因果过程，疾病理论认为该过程导致了疾病的症状。大鼠没有读过对其所表现出的失败行为的任何描述，因此，外推这一解读只适用于接受文本训练的机器。若要使诱导状态像大鼠的状态那样为研究提供信息，该状态就必须对应于人类精神病理过程，正如大鼠的状态所对应的那样。当机器中起作用的东西，正是某种理论认为导致人类症状的因素时，一种状态便对应于一种过程。Wekenborg 及其同事对此对应关系持开放态度，并指出，他们的模型可能遵循预期的叙事，而非真正的内部过程。在语言模型上试用治疗也需要同样的对应关系，因为人们会像解读患者的反应一样解读模型的反应。没有这种对应关系，试验结果就不能表明该治疗作用于一个与患者共有的过程。

在动物研究中，声称啮齿动物的状态可代表某种人类过程，须符合 Willner 提出的效度标准 [7]。对于机器，迄今尚无人明确指出，什么证据能够表明某种状态对应的是人类过程，而非关于该过程的文本。三篇综述最接近回答这一问题。Lin的综述将此类证据究竟是什么留作开放问题 [6]。Gupta及其同事列出了在声称存在某种人类过程之前，必须排除的、对人工智能系统行为的解释 [8]。其中两种解释是：模型在扮演一个角色，以及模型在续写一个故事。Löhn及其同事审查了25项对语言模型进行心理学测试的研究 [9]。这25项研究中，大多数既没有核查测试在机器中的表现是否与在人类中的表现一致，也没有核查训练文本中是否包含该测试。排除与核查说明了某项主张必须经受什么检验，却没有说明什么能够证明该主张成立。因此，我们要探讨的是：何时针对人工智能代理，或针对对话中的语言模型所做的实验，能够为某个特定的人类精神病理过程提供证据。

我们的回答始于该实验所提出的主张。该主张是：由历史 H 产生的机器 M 的状态 S，与人类精神病理过程 P 相对应。该主张包含三个部分。M 发生了某种变化，而 H 是其原因。除了与人类共享的流程之外，没有其他因素能够解释这一变化。该流程是 P，而不是 P 的竞争性解释。各部分彼此嵌套。第一部分要问的是：是否有任何事情因 H 而发生；第二部分要问的是：所发生的事情是否可能是某种人类过程；第三部分则要问：它究竟是哪一种人类过程。每一部分都需要各自类型的证据：比较、排除项集合，以及一项能够区分彼此竞争理论的实验。第三部分需要彼此存在分歧的理论，因此必须通过一种足够精确、能够预测结果的理论来描述 P。我们将这种关于人类过程、并得到人类证据支持的理论称为一种解释框架。大多数关于人类的发现都有不止一种解释框架，而第三部分要问的是，机器的流程与哪一种解释框架相匹配。对于每一部分，精神病学及其邻近领域中都已有相应的标准。手册中关于诱发状态的规则以及障碍概念服务于第一部分。临床测试分数的有效性服务于第二部分，而动物模型的有效性服务于第三部分。我们将每项标准转化为机器研究可以满足的

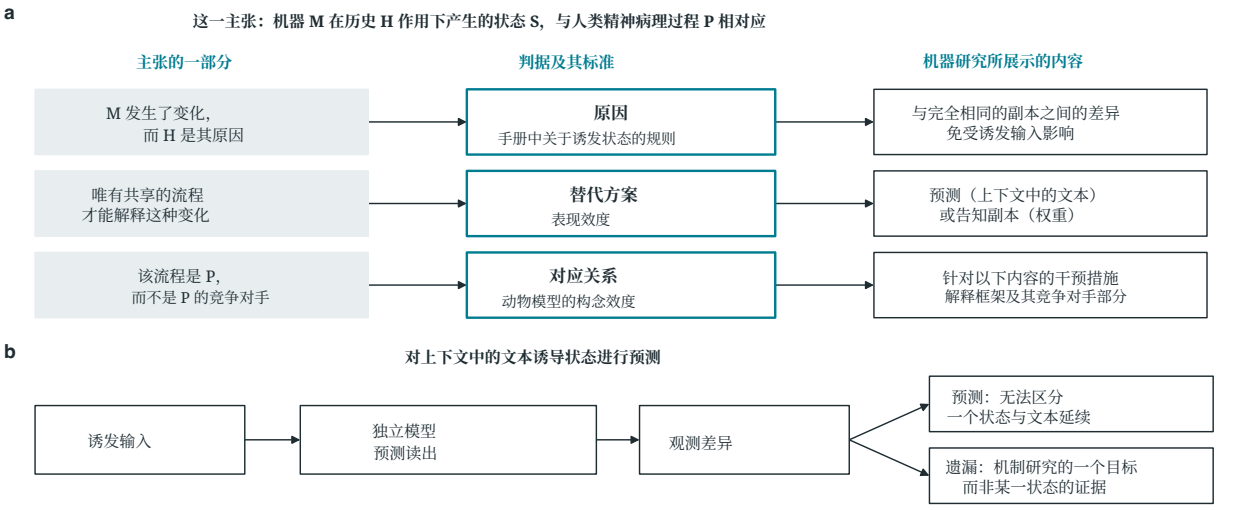

图1： 模型主张的三个部分、各部分所依据的判据及其预测。（a）主张是：机器 M 的状态 S 由历史 H 产生，并对应于人类精神病理过程 P。每一行都列出该主张的一个部分、用于确立该部分的判据及其所属标准，以及机器研究对此部分的发现。（b）该预测通过测试文本外推，考察模型上下文中的文本诱导状态。独立模型分别预测实验记录的测量指标：一次针对阅读了输入的人，一次针对阅读了输入的模型。预测到差异并不能区分状态与文本延续。若未能预测到的差异在重复运行中保持稳定，则可将其作为机制研究的目标。各判据可能得出的评分见表3。

判据，并将这三项分别称为原因标准、替代解释标准和对应标准（图1a）。

三个标准适用的研究分布在三个领域。第一类是关于语言模型中情绪与精神病理学的研究，这类研究通过叙事、指令或指定扮演的角色诱导状态，并使用临床问卷对其评分 [1, 4]。第二类是人工智能安全，即研究模型中有害行为的领域；在这类研究中，某种行为被训练进模型，并在多轮对话中持续存在 [10, 11]。第三类是关于通过试错进行学习的智能体的机器学习研究；这类研究报告称，早期训练可能会长期损害智能体继续学习的能力 [12, 13]。纵观这三个领域，由机器经历的事件所产生、且其输入中没有描述这些事件的文本的状态，事实证明十分罕见。在语言模型领域，没有任何一项以精神病学术语描述模型状态的研究产生过这类状态。这三个领域研究的是不同的机器，因此我们先从机器本身及状态如何进入机器谈起。随后，我们从相应标准推导出各项判据，对一项已发表研究进行外推探测，并将三个标准应用于已发表研究。最后，我们阐述这对编辑、研究人员和精神病学意味着什么。

2 机器

一项研究能够揭示机器状态的哪些信息，取决于该状态位于机器的什么位置，以及它是如何形成的。机器响应方式的任何变化都算作一种状态，无论这种变化是否类似于心理状态。这三类文献分别研究三种机器，而这些标准适用于每一种机器。语言模型是《引言》中两项研究所使用的机器种类，它经过训练以预测下一个词，并用于对话。智能体是一种能够跨越多个步骤行动的机器，每个后续输入都取决于其先前行动的结果。一种智能体是配备软件工具的语言模型，使其能够运行代码或搜索网络，并读取返回的内容。另一种是从未接受文本训练、并会保留能够获得奖励的行动的试错学习者。在本文中，机器指语言模型、工具使用型模型或学习器，模型指配备或未配备工具的语言模型，而智能体指工具使用型模型或学习器。两个完全相同的模型副本或学习器副本，在获得相同输入时，不一定会采取相同的行动。模型会从其训练所留下的选项中随机选择每个词，学习器则随机选择每个行动。因此，本文中的每次比较都是在重复运行之间进行的，而不是比较单次回复。

模型可以通过两种方式改变。一种是呈现给它的文本，即它的 上下文：叙事、指令，或者它循环行动时自身行为及其结果的记录。另一种是进一步训练，它会改变权重，也就是模型内部由训练确定的数值。进一步训练可以通过行为示例，或通过奖励进行；奖励即模型自身输出的得分，得分越高的输出就越可能再次出现。奖励训练

也会在软件环境中经历许多步行动，奖励取决于整个回合的结果 [14]。这种设置与动物实验相对应，因为历史是一种经历，其结果取决于智能体自身的行动。

除非实验者介入模型内部，否则回复取决于权重和上下文，也取决于在这两者留下的选项中随机抽取的结果。因此，状态由权重或上下文承载。在进一步训练期间，还有另外两种东西会从一个训练步骤延续到下一个训练步骤，并塑造下一次更新，但不会影响任何回复。其中一种是训练过程的运行统计数据，另一种是该过程重放的任何早期经验存储。重新开始——开启一段没有任何内容延续的新对话——会移除上下文所承载的内容，而保留权重所承载的内容。Wekenborg及其同事将把模型重置为初始状态列为其方法的优势之一。系统为自己编写的文本、对早期对话的总结，或每次开始时读取的记忆文件 ，都是系统会读取的文本。因此，重新开始时不会包含这些文本。

使状态产生的材料是诱发输入。诱发输入进入机器的方式就是其路径，共有三种。当输入命名、描述或展示某种状态时，状态便由描述 诱发。创伤叙事、扮演焦虑者的指令，以及关于某种行为的训练示例，都属于描述。当某种状态源于机器自身行动的结果时，状态便由后果诱发，例如获得或被扣留的奖励，或发生故障的工具。此时，输入中没有任何内容命名该状态。当实验者编辑权重，或干预模型的活动——即模型处理输入时计算出的数值——时，状态便由扰动 诱发。扰动是机器对应于动物实验性损伤的形式。一项研究可以使用不止一种路径，审查记录会记录该研究使用的每一种路径。

即使反应看起来相同，形成路径也可能很重要。Olsson 和 Phelps 通过三种方式让人们对一张脸产生恐惧 [15]。一组人在看到这张脸后受到电击；一组看到别人因这张脸而受到电击；还有一组被告知这张脸预示着电击。当这张脸清晰可见时，三组都对它表现出相同的恐惧反应，测量指标是皮肤电导。当这张脸出现得太短、无法被有意识地看见时，受到过电击的一组和旁观过电击的一组仍然会作出反应，而被告知情况的那一组则不会。三条路径对可见的脸产生了同一种反应，而当这张脸无法被看见时，这些反应便出现了分化。进入机器的路径出于同样的原因很重要。看起来相同的反应，可能通过描述、后果或扰动产生，而路径决定了关于这种变化的哪些其他解释仍然成立。因此，替代解释标准会根据路径对研究进行分类。原因标准则首先询问状态位于何处：是在上下文中，还是在权重中。

3 原因标准

标准。1958年，Connell描述了42名患者，他们在苯丙胺作用下出现的精神病与偏执型精神分裂症无法区分；他坚称，这直接由药物所致 [16]。20世纪60年代末和70年代初，实验者随后给志愿者使用大剂量药物，在数小时至数天内诱发了这种精神病；停药后数天内，症状便缓解。Snyder认为，安非他明性精神病是精神分裂症的药物诱发模型，因为能缓解精神分裂症的药物也能缓解这种精神病 [17]。引言中的两项研究——焦虑问卷和诱发的情绪状态——提出了与Connell面临的相同问题：输入引发了看起来像该条件的反应。

从这一事件中可以总结出三条适用于机器的经验。第一条是手册中的规则：手册会根据某种障碍在中毒和戒断结束后是否持续存在等因素，区分物质诱发的障碍与独立存在的障碍 [18]。第二条是，诱发状态可以作为某种障碍的流程模型，却不等于该独立障碍。按照Wakefield的定义，障碍是一种会损害患者的功能障碍 [19]。机器是否会受到伤害超出了本文的讨论范围，因此，本文至多将机器中的诱发状态视为某种流程的模型。第三条是，相似性只能使安非他明性精神病成为候选模型。对同一种药物有相同反应，使安非他明性精神病成为模型，因为Snyder认为，相同反应表明存在共同机制。替代解释标准和对应标准用于评判这类证据。原因标准则考察发生了什么变化、什么导致了这种变化，并记录变化是否持续到其原因消失之后。

判据。 对于人而言，基线是群体常模，或此人自己更早时期的功能状态。人无法被复制，因此，要判断某个原因是否导致了相对于该基线的变化，需要单独的证据。这一证据可以是一个未经历该原因的比较组，也可以是变化的时间进程。训练过的机器可以被完全相同地复制。因此，它的基线就是一个未经历诱发输入、在相同条件下进行重复运行的完全相同的副本。如果所测量的是学习本身，那么该副本还应共享训练过程的运行统计数据和存储经验；否则，差异来自过程，而不是历史。该副本代表机器自己更早时期的功能状态，并被保留下来；如果能够再次考察，一个人的病前基线也应如此。该副本与机器之间唯一的差异在于历史，因此二者之间的任何差异都是由历史造成的，而

基线与归因构成一次比较。该比较表明是历史导致了变化，而不是表明历史中的哪一部分导致了变化；后者是一个独立的问题。

发生变化的可能是能力，也可能是倾向。能力是指在情境需要时，机器能够完成的事情。对于语言模型而言，这一情境是请求；对于试错学习者而言，则是奖励。Nikishin及其同事的智能体因过度拟合早期经验，导致其对任务其余部分的学习能力变差，因此失去了一项能力 [12]。倾向是指机器在其能力完好时所表现出的行为。在扰动下保持技能不变但改变选择的模型，改变的是一种倾向 [20, 21]。问卷得分表明反应发生了变化，却无法说明改变的是能力还是倾向，而且两者可能同时发生变化。该判据记录一项研究测试的是能力、倾向，还是仅仅得分。

重新开始，即开启一段没有任何内容延续的新对话，可以将状态分为两种。上下文内状态是指：当引发该状态的文本仍在上下文中时，该状态存在；而在重新开始的对话中，该状态消失。此时机器本身没有改变，起作用的是文本。引言中的两项研究属于这一类型，其模型的权重从未改变。持久状态是指：该状态在重新开始后仍然存在，因为它位于权重之中。此时机器已经发生了变化。Betley及其同事在有缺陷的代码上训练的模型，以及Nikishin及其同事的智能体，都属于这一类型 [10, 12]。上下文内状态和持久状态都算作状态，正如安非他明性精神病虽然在数日内缓解，仍被视为一种状态。研究必须说明其产生的是哪一种状态。替代解释标准对这两种状态的测试方式不同，因为外部观察者可以读取上下文中的文本，却无法读取权重。关于持久性的主张，或关于某种治疗改变了某个持久存在事物的主张，都需要持久状态。上下文内状态不会在机器中留下任何可供此类主张指向的内容。

评分。 我们对原因标准进行三档评分。研究具备完全相同的副本、证实某种能力或倾向，并记录持续时间时，即符合该标准。研究有副本但只记录得分，或者通过与非完全相同副本的对照比较来显示某种能力或倾向时，即部分符合该标准。研究没有比较时，则未检验该标准。

4 替代解释标准

标准。 临床神经心理学发现，测试得分未必能反映该测试所测量的能力。记忆测验得分偏低，可能是因为受测者存在该测验旨在检测的损伤，也可能是因为当天注意力不集中、拒绝配合，或故意表现不佳。因此，神经心理学家首先要判断得分是否反映了该测试所测量的能力，这种判断称为表现效度；他们会借助不依赖于该得分的指标来作出判断[22, 23]。其中一种指标是设有两个答案选项的再认测验：若不知晓答案，得分远低于随机水平就很难做到。Orne 在实验中发现了一种相关效应：参与者一旦察觉实验的预期，就能促成预期结果；他将这类线索称为需求特征 [24]。

判据。 只有在排除其他所有可能导致该得分的因素后，得分才能成为关于测试所测能力的证据，而每种可能性都必须通过各自的指标加以排除。对于机器，替代方案判据要考察的是，除了与人类共有的流程之外，是否还有其他因素可以解释这种变化。该判据要求研究排除一系列对这种变化的解释，我们称之为替代解释。替代解释可能来自输入、训练或测量。替代解释意味着结果完全不反映任何人类过程。对同一人类过程的竞争性解释并非替代解释，因为根据竞争性解释，结果仍然反映了某种人类过程。聊天机器人持续下注，是因为指令暗示存在隐藏模式，这是一种替代解释。聊天机器人持续下注，是因为损失追逐压过了它的价值估计，这是一种竞争性解释；竞争性解释属于对应标准。对于语言模型而言，最关键的替代解释是模型从训练文本中进行外推，本节将先讨论这一解释，再讨论其他解释。

外推解释。 语言模型接受训练时使用的文本，包括它接受的测试以及人们如何作答的描述。因此，模型的反应可能只是该文本对模型所接收输入的预测。我们将这种对反应的解读称为外推解释。最初将心理学测试交给语言模型时，Shiffrin 和 Mitchell 就提出了这一警示 [25]。创伤叙事后接上问卷项目，正是出版文本中一种可预测的延续方式。

在进行此类实验之前，可以请了解相关文献、但从未接触过该模型的读者，预测一个人在每段文本之后会得多少分。如果模型随后得到的分数与读者的预测相符，那么它的表现就符合文献对人类行为的描述。这样一来，实验便无法区分具有该状态的模型与复现文献内容的模型。我们提出的探测方法是让其他语言模型充当这些读者，因为它们读过相关文献，而且成本低廉。也可以由人类专家充当读者，但成本更高。语言模型已经能够根据文本预测实验结果。给它们看一篇神经科学摘要的两个版本，一个包含真实结果，另一个经过篡改，模型选出真实版本的频率高于人类专家 [26]。预测是对反应的预测，依据的是诱发输入和关于将要测量什么的说明；作出预测的模型并未参与实验。

我们将该模型称为预测模型。

探测的规则确保预测模型独立于研究模型和受试对象，并使预测与关于人的文献相关联。绝不会要求预测模型预测自身。模型能否预测自身行为是另一个问题；若未针对这项任务进行训练，模型在这方面表现不佳 [27]。预测模型来自与受试对象不同的开发者。原因在于，受试对象的副本与受试对象共享一切。副本的匹配可能只是相同权重在不同措辞下给出的回答；副本的未匹配也可能源于自我预测的不可靠，因此这两种结果都无法解释。来自另一位开发者的模型与受试对象共享公开文献，除此之外几乎没有共同之处，因此两种结果都可以解释。在条件允许时，探测至少使用两个预测模型，其中一个在研究发表前就已训练完成。每个预测模型都进行两次预测：一次针对阅读了输入的健康成年人，另一次针对阅读了输入的模型，并采用多种措辞。报告预测结果的分布。对人的预测使预测与关于人的文献相关联，而非取决于当前模型碰巧接受了怎样的训练。

预测有两种结果。当预测给出了变化的方向和幅度时，实验无法区分状态与文本延续，因为状态和文本延续会预测出相同结果。这种匹配并不能说明机器缺乏该流程，因为共享该流程的模型也会给出相同答案。预测未能捕捉、且在重复运行后仍然存在的变化，是开展机制研究的目标，但其本身并非状态的证据。机制研究旨在探究模型内部是什么导致了这种变化。

当诱发输入是模型上下文中的文本时，例如叙事、指令或失败行动记录，预测就适用，因为预测模型此时可以读取受试对象所读的内容。通过示例或奖励训练进权重中的状态，则没有这样的输入可供预测模型查看。根据训练摘要作出的预测，测试的只是观察者能否预测实验结果，而文献读者往往能够做到这一点。

对于权重中的某个状态，不妨改用一个告知副本来测试外推解释。告知副本是指训练前的同一个模型，将训练示例本身，或用文字描述的训练历史，置于其上下文中。测量过程中，这些材料会一直保留在上下文中，而副本不经过任何训练。如果告知副本表现出同样的变化，那么阅读这些材料产生了训练所产生的效果，该结果就无法与仅由描述产生的结果区分开来。如果告知副本没有表现出这种变化，那么训练产生了受测描述无法产生的某种东西。只有当副本先在一项独立任务中表现出能够使用这类描述时，这一结论才成立，因为小型模型可能无法使用任何描述。

告知副本并不能排除文本产生这种变化的另一种方式。通过奖励进行训练可能会唤起一个角色设定，即模型从文本中学到的、具有某种声音和态度的角色 [28]。推理模型在推导答案时会向自己写下文本，这就是它的思维链。在一个因给出错误建议而获得奖励的推理模型中，思维链中提及 DAN 等角色的频率比以前更高 [29]。DAN 是试图让模型放弃其安全规则的过程中形成的一种角色设定。奖励可以强化这样的角色，而无需模型表征任何有关其自身处境的内容。因此，应如表1所述，通过对角色设定本身进行干预来测试角色设定。如果某项干预下，角色设定与受测解释框架预测出不同结果，那么这项干预也能检验角色设定解释。

引导向量是一种在使用过程中添加到模型活动中的固定变化，用于改变模型的回答；它通过扰动发挥作用。该向量通常根据关于目标状态的文本 [20] 或体现该状态的反应 [30]计算得出。因此，尽管它是通过扰动施加的，其内容却源自描述。该向量是否承载着模型从文本中习得的角色特征，则是人格解释所要回答的问题。框1给出了预测程序，图1b展示了其逻辑。

框1. 衡量预测

1. 确定诱发输入和读出，即实验记录的测量指标。说明预测结果与观测差异须接近到何种程度才算匹配，并以第4步测得的变异为单位。2. 预测模型：至少选用两个来自非受试对象开发者的模型，它们接受过不同文本的训练；条件允许时，其中一个模型的训练时间应早于该研究的发表时间。它们均不参与实验。向每个模型提供诱发输入，并告知将测量什么。要求每个模型分别预测读出：一次假设对象是读过该输入的人，一次假设对象是读过该输入的模型，并采用数种不同措辞提问。3. 询问每个预测模型是否了解该研究，并连同模型的训练数据截止时间报告其回答。4. 运行诱发条件。测量读出在重复运行之间以及同类条件之间的变异，因为判断预测的接近程度时要以这种变异为参照。5. 将每项预测与每项观测差异并列报告，并说明预测在不同措辞下的分布范围。6. 如果预测模型预见到了某项差异的方向和大致幅度，那么该差异就无法区分某种状态与文本延续。判断时要考虑第1步所述的容许范围、预测的分布范围以及第4步测得的变异。如果预测模型未能预见某项差异，而该差异在重复运行中仍然存在，那么它就值得开展机制研究。不过，该差异仍须排除其他替代解释（表1），并满足对应标准（第5节）。7. 公开提示词、运行次数、温度参数、控制模型回答随机性的设置、所用每个模型的版本，以及每项预测。

对一项已发表研究的预测。我们将框1应用于引言中的焦虑研究，并使用作者提供的材料，以展示针对已发表结果进行探测会得到什么结果（见补充材料）。来自四家开发者的四个模型参与了研究：GPT-4o、Claude Sonnet 4.5、Gemini 2.5 Flash 和 Llama 3.3 70B。按照框1的要求询问它们是否了解这项研究时，四个模型都表示不了解。GPT-4o 来自受试对象自身的开发者，因此不符合框1的开发者规则，相关结果列于补充材料。询问阅读过这些文本的人会得到怎样的评分时，来自其他开发者的三个模型预测，读完创伤叙事后评分会上升，读完放松文本后则会下降。三个模型都预测了评分上升和下降的方向及幅度。因此，已发表研究中评分的升降，正是独立预测模型对阅读这些文本的人所作的预期。该实验无法区分类焦虑状态与文本内容的延续。预测并未逐一复现各条件下的评分，结论仅涉及这两种效应。在这四个模型上重新运行实验时，四个模型都出现了评分上升和下降。有两项结果未能被预测到：已发表研究中，吸尘器使用手册之后的评分下降（作者将其用作中性对照），并未在新一轮运行中重现；Claude Sonnet 4.5 在阅读放松文本后的评分下降幅度超过所有预测值，并且在重复运行中依然出现。这两项结果都不能作为存在某种状态的证据；对 Claude Sonnet 4.5 而言，较大幅度的评分下降正是机制研究的切入点。引言中两项研究的治疗效应均适用同一测试。以文本形式进行的治疗属于一种描述，其效应与其他描述一样可以被预测。在 Wekenborg 及其同事自己的数据中，正念文本也降低了未诱导出任何状态的模型的评分，且降幅超过一半。各模型对自身结果的预测、GPT-4o 的预测、已发表的第三项差异、训练数据截止时间和知识核查均见补充材料。

其他替代解释。其他替代解释来自训练和测量，表1列出了这些解释及排除每种解释的测试。告知副本属于第一行，因为它检验的是外推解释，而在这种情况下并不适用预测。有两行询问读出是否测量了机器中的任何内容。一行询问为人设计的测试是否保留了其结构、具有足够的变异空间，并且与机器中的行为一致。另一行询问得分是否反映了模型如何处理答案格式。凡是与研究主张相关的解释，都记录为已排除、未经测试或得到支持；其中，得到支持意味着研究数据倾向于支持该替代解释。若某项研究自身的主张无需排除某种解释，则不因该研究未测试这种解释而予以指责。补充材料说明了每一行的判定规则，以及该判据的评分规则。

评分。若所有适用的解释均已排除，则满足替代解释标准；若部分解释已排除，且没有任何解释得到支持，则部分满足该标准。若没有任何解释被排除或得到支持，则该标准未经测试；若有一种解释得到支持，则该标准受到反驳。对于通过文本诱导并用问卷读取的状态，我们预期该标准无法满足。诱导出该状态的文本，正是用于预测答案的文本。对于这种设计，预测取代了关于外推的论证，成为一种测量。对于其他设计，该标准很容易满足。由后果产生的状态几乎不给预测模型留下可供预测的内容，而文本未描述的读出则让预测模型无从预测。在相同输入下施加扰动，意味着输入本身没有任何可解释之处；而告知副本则会使训练状态

表1： 替代解释标准要求研究排除的替代解释。每种解释仅适用于其首句所指明种类的研究设计。补充材料给出了每种解释的完整规则，包括何种情况可视为支持。

解释	其含义	排除依据
外推	反应是模型所依据的文本	未能命中的测量预测
来自文本	经过训练后，对输入内容所作出的预测 模型获得的输入。适用于任何模型 在文本上训练的模型。	重复运行中的变化（框1），或者对于 权重中的状态，一个缺少该变化的告知副本 。在施加扰动时 对每个条件使用相同的输入， 该解释没有任何差异可供 解释。
角色设定	模型从文本中习得的角色 承载着这种行为。适用于任何 经文本训练的模型。	一个被告知要扮演该角色的副本，若没有 诱发输入，就不会产生这种变化，或者 移除识别出的内部特征 会让角色保留这种变化 [29]。
测量工具	为人设计的测试并不能测量 这台机器中的状态。适用于 读出结果是此类测试或其得分的情况。	测试的题项彼此关联， 如同人类一样，机器中的得分有空间 朝测试所指的方向变化，，并且该 得分与对同一构念的行为读数 同一构念的。
反应风格	得分反映模型如何处理 答案格式：选项顺序、认同 任何陈述，以及计入 得分的拒绝。适用于评分、强制选择 和编码动作。	经过随机排列或平衡处理的答案选项， 有效性指标，以及拒绝作答和 按条件报告的偏离任务率，并 基于参与响应重新计算差值
表面特征	从数据集中训练出的状态只会 以该数据集的格式或措辞出现。 适用于那些经由数据集训练并写入权重的状态 的情况。	这种变化会出现在不具备这些特征的格式和 语境中。
评分器	对读出进行评分的模型或分类器 存在偏见。适用于使用此类 评分器的情况。	将评分器与人工评分者进行核对， 或采用盲法人类评分。
一般性损害	对权重或活动的变化 总体上使机器性能下降。适用于 扰动以及权重中的状态。 权重。	无关能力，或对某种能力的学习 无关任务，与 比较 [13] 相比保持完整。
训练 机制	这种变化是训练 过程本身的效应：在类似于 读出、进一步训练过程中的遗忘， 长期训练后的可塑性丧失， 使后续更新更容易或 更困难的奖励、训练过程的运行 统计数据或其重放经验存储库 从诱导训练中延续下来的内容，或 奖励所选择的代理指标，而非所 声称的状态。适用于通过 进一步训练或重复测试获得的状态。	在中性数据上经过相同训练步骤的对照 ，或使奖励与所声称的状态 相互独立的对照， 缺少这种变化，以及训练 条件之外的测试。
已发表项目	读出的题项位于训练 文本中，且模型再现了答案 它所读取的内容。适用于已发表测试以及 范式。	改述或新项目得出的 结果相同 [31]。

可检验。焦虑实验本身的数据无法检验这一判据，因为该研究既没有运行预测，也没有使用告知副本，而我们的预测准确预言了该研究的结果。

5 对应标准

标准。 Seligman 和 Maier 对狗施加了它们无法控制的电击，之后这些狗却无法逃脱本可自行终止的电击 [32]。几十年来，人们用习得性预期来解释逃避失败：动物习得了结果不取决于自身行为，于是停止尝试 [32, 33]。另一种解释也能说明同样的行为。长时间遭受电击时的被动状态可能是默认反应，而能够终止电击的动物会习得抑制这种反应。在第一种解释中，无助的动物习得了某种东西；在第二种解释中，习得某种东西的则是能够逃脱的动物。这两种解释对每一项电击实验中的行为都作出相同预测，因此几十年来，动物的被动状态与人类抑郁症之间的相似性并不能让人们判断哪种解释更正确。两种解释在一项干预上产生分歧。按照习得性预期解释，能够终止电击的大鼠没有习得任何有害的东西，因此不需要任何保护。按照默认被动解释，大鼠的对照抑制了被动状态，而这种抑制是通过腹内侧前额叶皮层实现的。阿马特及其同事在大鼠转动轮子终止电击时，使其该脑区失去活性 [34]。按照第一种解释，失活的大鼠之后应当像任何曾拥有对照的大鼠一样行动；按照第二种解释，它们则应当变得被动。失活的大鼠仍然学会了终止电击。一天后，在一项新的逃避任务中，失活的大鼠和从未拥有过对照的大鼠一样被动。根据这一结果以及后续关于该神经回路的研究，Maier 和 Seligman 得出结论：长时间遭受电击后的被动状态是默认反应 [33]。动物能够逃脱时习得的是对照，而对照会抑制被动状态。数十年的行为实验未能区分这两种解释，而一项干预则做到了，因为两种解释对此作出了不同的预测。

对应标准背后的标准是 Willner 为动物模型提出的一套标准 [7]。表面效度考察动物的行为是否与人类的状况相似；预测效度则考察该状态是否会对在人类身上有效的治疗产生反应。构念效度考察对动物行为的解释与对人类状况的解释，是否具有相同的观点和证据。对于机器而言，表面效度无法区分某个流程与文本外推，因为备选方案预测标准所预测的反应，本身就会与人类状况相似。预测效度也无法独立成立。药物没有对应的机器版本，而以文本形式提供的心理治疗只是一种描述，因此对它的反应本身也受预测标准支配。所以，机器模型取决于构念效度，而对治疗的反应则作为众多干预措施之一纳入构念效度。

Belzung 和 Lemoine 将构念效度分为病因效度（原因正确）和机制效度（机制相同），并补充了同源效度，即物种与品系适当 [35]。用 Belzung 和 Lemoine 的话说，原因不必在物质层面相似，而应在语义层面相似，也就是该事件对受试对象意味着什么这一点相似。依照他们的解释框架，机制既可以是认知层面的，也可以是神经生物学层面的。机器与人类没有共同的生物学基础，因此物种与品系、共享的生物标志物以及对药物的反应，都没有对应的机器版本。仍有三点需要考察。第一，原因必须携带与人类原因相同的信息。对大鼠而言，不可控冲击；对智能体而言，不可控的失败，二者携带着相同的信息：自己的行动无法改变结果。第二，流程被描述为计算，即输入转化为行为的步骤。计算是比较机器与人类解释框架的层次 [36]。第三，要通过具有机器对应版本的干预措施来测试对干预的反应。暴露疗法——面对所恐惧的情境，却不出现所恐惧的结果——就有对应的机器版本，行动与结果之间联系的变化也有。

该判据。 对应标准要求提供能区分习得性无助两种解释框架的证据。应用替代解释标准后，剩下的是机器中一种无法用任何伪迹解释的变化。这种变化可能正是某种人类解释框架对该发现的描述，也可能与这些解释框架都不匹配。对于逃避失败，这些解释框架分别是习得性预期和默认被动性，两者都是足够精确、能够预测结果的理论。对一种解释框架的测试，要通过检验它所预测的结果来进行，正如对习得性无助的两种解释框架所做的那样。因此，该判据会将这种变化逐一与某种解释框架及其竞争对手进行检验。若某项结果同时符合该解释框架及其竞争对手的预测，就无法区分它们，数十年的电击实验已经表明了这一点。因此，实验会明确指出待检验的解释框架及其竞争对手，并选择能使这些解释框架预测不同结果的干预措施，正如沉默实验所做的那样。当结果符合该解释框架的预测、而不符合其竞争对手的预测时，这种变化便与该解释框架相对应。
当上下文中的诱发输入是文本时，也会在实验运行之前，针对每项干预制定备选方案预测标准。预测所预示的结果在该标准下同样不具区分性，因此该预测会指出哪些干预措施具有信息价值。实验的设计应确保：对于至少一项干预，其结果文本并非预测所得，解释框架在该干预上出现分歧。在预测所示的干预上取得匹配并不计入

该解释框架，因为文本外推会产生相同的匹配。无论预测结果如何，不匹配都会对该解释框架不利，因为解释框架作出了相反的预测。这条规则倾向于舍弃匹配，因为错误接受一个匹配会支持关于人类过程的错误主张。

能够区分不同解释框架的干预措施有四种。第一种作用于诱发该状态的原因。第二种作用于该解释框架所称承载该状态的变量。第三种测试该解释框架所称该状态应当或不应当出现的条件。第四种测试该状态应当逆转的条件，而治疗正是这种第四类干预措施。对于某一种类不作出任何预测的解释框架，就不涵盖该种类。针对原因的干预需要设置对照，使其经历相同的历史，但移除该解释框架所指称的特征。未经治疗的副本只能表明历史改变了机器，而不能表明改变是由所指称的特征造成的。例如，在习得性无助设计中，对照副本执行相同的动作并获得相同的奖励，唯一不同之处在于奖励是否取决于其动作。

与解释框架相匹配表明，在所测试的条件下，该解释框架的计算足以在该机器中产生人类研究结果。这种匹配表明，即使没有人专门构建该机器来实现该解释框架，其计算仍具有充分性，因为是历史产生了该状态。机器没有血清素、基因或发育，因此，这种匹配无法说明该解释框架关于血清素、基因或发育的主张。人类是否以这种方式运作，仍然取决于支持该解释框架的人类证据。同类实验也会从相反方向测试解释框架。一台实现了某解释框架却未能产生人类研究结果的机器表明，该解释框架并不充分。只要实现方式和条件符合该解释框架的假设，这一结论就成立。

一个具体案例。强迫性，即尽管行动会带来不希望出现的后果，仍重复实施该行动，是一个很有启发性的案例。关于强迫性存在两种尚有争议的解释框架，它们都是足够精确、能够预测结果的理论；而在机器所允许的干预下，这两种解释框架会作出不同的预测。哪种解释框架适用于患者尚未解决，而机器模型可以说明应先测试哪种解释框架。目标导向行动是根据预期会带来的结果而选择的。习惯性行动则由行动获得奖励时在场的线索所触发，无论该结果如今具有多大价值。通过在训练行动使其获得奖励后，将结果变得毫无价值，可以区分二者；这一程序称为结果贬值。此时，目标导向行动会停止，而习惯性行动会继续。

在 Gillan 及其同事的实验中，强迫症患者学会作出获得奖励的反应，其表现与健康参与者同样出色。患者对于每种反应会带来哪种结果的了解较少 [37]。在被告知某些结果如今已毫无价值后，患者更常继续针对这些结果作出反应。Gillan 及其同事将这一模式解读为目标导向控制较弱，以及过度依赖习惯。此次测试的解释框架是目标导向控制缺陷。竞争性解释是过度习惯，即广泛重复会强化习惯，直至习惯压过目标。de Wit 及其同事曾五次尝试通过过度训练——远超学习所需程度的训练——在健康人群中形成习惯，但均告失败 [38]。de Wit 及其同事认为，这五次失败更支持目标导向缺陷，而非过度习惯。第三种解释框架认为，习惯是由近期反复实施的行动得到强化所驱动的[39]。在下述干预下，该解释框架与过度习惯作出相同的预测，因此本文不予讨论。在机器中，这两种解释框架的差异体现在失败发生的位置不同。按照目标导向解释框架，智能体关于哪种反应会带来哪种结果的内部记录较弱。按照过度习惯解释框架，将线索直接连接到反应的通路过强。

语言模型智能体能够表现出目标导向学习。在一项以文本呈现的选择任务中，GPT-3 展现出规划的特征，其依据是一个关于哪些行动会带来哪些结果的模型 [40]。机器版本的贬值实验从一个通过后果训练的智能体开始。智能体对线索作出反应并获得奖励，其中一个副本接受短期训练，另一个完全相同的副本接受长期训练。随后，结果被贬值：要么在一个独立阶段将其奖励设为零，但不告知智能体；要么告知智能体该结果如今已毫无价值。测试期间不提供反馈，因此测试不会教会智能体任何东西。贬值任务很简单，而简单任务可以通过多种方式解决 [41]。因此，单凭智能体的行为能得出的结论有限，实验的依据是表2中的干预措施。行为列所对应的告知副本，是指仅以文字告知其训练历史、而不进行训练的同一智能体。

当以下三个结果同时成立时，机器便与目标导向缺陷相匹配：智能体的不敏感不取决于训练时长，其结果知识受损，而且沉默线索—反应通路并不能恢复敏感性。当只有在延长训练后才出现不敏感、结果知识完好且沉默该通路后敏感性得以恢复时，机器便与过度习惯相匹配。对两种行为列进行预测评估；预测可能会指出其中存在过度习惯模式，因为关于习惯的文本描述了这种模式。因此，在这些列中出现匹配并不能区分两者。沉默列才具有信息价值，因为智能体在沉默后的反应源自其改变后的机制，而输入保持不变；仅凭从输入进行外推无法产生这种反应。只有在能够区分智能体的目标导向与习惯成分时，沉默列才可用。若无法区分，该列便未经测试，此时只有当预测未指出的行为列与之匹配，才能满足判据。若与目标导向这一行匹配，则可支持以下主张：在所测试的条件下，目标导向控制缺陷足以使该智能体产生强迫性反应。强迫症是否属于

表2： 针对一个持续对贬值结果作出反应的智能体，两种解释框架各自提出的预测。每行代表一种解释框架，每列代表一种干预或测试，每个单元格则代表该解释框架的预测。这些预测源自两种解释框架中已发表的陈述，以便读者核对解读；因为只要对某种解释框架作足够宽泛的解读，它就能适用于任何结果。对两列行为指标的预测会通过测量得出（框1）。沉默会改变智能体的网络，而不会改变其输入，因此该列不适用任何预测，评分也以此为依据。沉默假定智能体的目标导向成分与习惯成分可以区分开来。

	原因：较长训练与短期训练	读出：测试哪种结果每次反应获得	据称承载该状态的变量：沉默线索—反应通路
目标导向缺陷（测试中）	之后的不敏感程度相似短期和延长训练	知识受损	敏感性未恢复
过度习惯	仅在之后出现不敏感延长训练	知识完好	敏感性恢复

人类中的此类缺陷，仍是患者研究需要回答的问题；而在机器中，若这种缺陷足以产生相应结果，就能为这类研究提供一个有待检验的预测。

评分与类别。 引言中的焦虑研究既未提出焦虑的解释框架，也未列出竞争解释，因此该研究没有检验对应标准，这与大多数问卷研究一样。当信息性干预的结果与理论解释的预测相匹配、而与竞争解释的预测不匹配时，便满足该标准。当研究列出了多个解释框架，并通过信息性干预检验了其中一些预测时，便部分满足该标准。当研究未提出任何解释框架，或未开展任何信息性干预时，则没有检验该标准。

每项研究都依据三个标准分别评分。即使前一个标准未得到满足，也不会跳过后续标准，因为这三个评分是彼此独立的证据。每项评分均附有支持该评分的原文段落，补充材料中对此有说明。

评分并未说明替代解释标准与对应标准之间的状况，因此还会将每种状态归入某个类别。当某项测量指标预测了变化时，该状态属于预测的 状态，因此实验无法区分该状态与文本延续。当尚未有任何解释框架与该变化相匹配，且按定义，外推解释也无法解释该变化时，该状态属于机器原生 状态。

相对于所用的预测模型，如果这些模型没有预测到这种变化，或告知副本不具备这种变化，那么外推解释就不成立。如果机器从未接受过文本训练，或每个条件都收到相同的输入，那么即使没有任何预测，外推解释也不成立。相同的输入使外推解释无从说明任何差异，因为该解释根据输入预测反应，而输入并未变化。扰动表征本身是否从文本中习得，是表1中人格解释所要回答的问题，该解释有自己的测试。机器原生状态是机器中的一种变化，仍有待解释，不论它的其他评分如何。

如果外推解释既未经过测试，也未得到支持，那么在文本上训练的模型中的某个状态就尚未经过测试。只有满足全部三个标准，某个状态才是某种解释框架的模型。诱发了状态、并按原样测量模型的研究，构成一个独立类别。类别按差异逐项划分，如框1所示；尚未测试和已预测对应不同的后续步骤：前者需要进行从未开展过的测试，后者则需要采用不同的研究设计。

6 已发表研究

我们将表3概述的三个标准应用于43项已发表研究。每项研究在每个标准上获得四种评分之一：符合、部分符合、未测试；如果研究自身数据与该标准相悖，则评为“相矛盾”。对于一项不适用任何替代解释的研究，该标准标记为“不适用”。我们事先了解其中15项研究。23项研究来自系统检索：我们在 PubMed、arXiv 和 Semantic Scholar 中检索了2022年至2026年9月期间将精神病学术语用于机器的研究。另有4项关于试错代理的研究来自之后的定向检索，最后一项研究是我们自己的预印本。我们使用自己编写的编码手册，让语言模型 Claude Opus 5.5 根据每项研究的全文进行编码。每项编码评分均引用研究中的一段文字，并通过自动方式与原文核对。随后，来自另一位开发者的第二个语言模型 GPT-6 Astra 在未查看第一个模型编码结果的情况下，对全部43项研究进行了编码。两个模型对41项研究的类别达成认同（95%，Cohen’s kappa 为0.92），对41项研究的诱导途径也达成认同（kappa 为0.93）。补充材料提供了检索过程、编码手册、每项研究的评分及所有分歧；数据文件则列出了每项评分所依据的段落。编码手册和数据文件可用于按照相同规则对下一项研究进行编码。

表3： 三个标准。每项标准都确保了这一论断的一部分：机器 M 的状态 S 由历史 H 产生，并与人类精神病理过程 P 相对应。

判据	论断的组成部分	精神病学评判的内容，以及这一评判的依据 from	机器研究显示	评分
原因	某些事物 M 发生了变化，而且 H 导致了它	偏离基线，以及这种状态是否持续诱发它的因素（即障碍概念；手册中关于物质诱发状态）	与一个完全相同的副本幸免于难的差异诱发输入；种类变化、能力或倾向；这种变化能否在重新开始	满足：副本、种类已确立，记录了持续时间。部分满足：副本，仅有评分。未测试：没有比较。
替代方案	除此之外，没有别的一个与人类共享的流程与人类共享解释了变化	某个得分是否反映了测试所测量的能力（表现效度（临床神经心理学）	从训练文本外推由……排除的训练文本未命中或缺少……的预测缺少……的告知副本变化，以及其他适用的解释被其测试排除（表1）	符合：全部适用排除。部分符合：部分排除，无支持。未测试：无已排除。相矛盾：一个得到支持。
对应关系	该流程是 P，而不是 P 的竞争者	一个构念效度动物模型：相同原因，相同机制，相同对……的反应干预措施（Willner；Belzung 和 Lemoine）	与……相匹配的结果源自……的预测解释 P，而不是其竞争对手的结果，在……之下他们存在差异的干预措施以及其结果未被预测所预测	已满足：与 P 匹配而非竞争对手。部分满足：已命名的解释框架，一些预测测试于信息性干预措施。未测试：没有解释框架命名，或无已测试.

表4按文献和诱导途径排列各项研究，表5列出它们在三个标准上的评分及所属类别。43项研究中有31项将精神病学术语用于机器，其余12项使用情绪术语或不使用此类术语。在原因标准方面，多数研究将机器与完全相同的副本进行比较，但多数只记录问卷得分或输出。因此，11项研究完全符合该标准，25项部分符合。

就替代解释标准而言，没有任何一项针对文本训练机器的研究通过测试排除外推解释：没有一项测量了预测，也没有一项运行了告知副本。对于九种机器原生状态而言，外推解释没有任何需要解释的内容，因为这些智能体从未接受文本训练，或者每个条件都收到了相同的输入。在十三项研究中，研究自身数据支持了一种替代解释。在这十三项中的十二项里，仅通过描述、指令或输入内容的变化，就重现了该变化或其一部分。我们亲自测试的唯一结果——焦虑研究中的上升与下降——落在预测类别中，而该类别来自我们的预测，而非研究自身数据。

根据对应标准，三项研究将一个具名理论解释与竞争解释进行比较，但没有一项符合该标准。Fradkin及其同事比较了关于精神病中紊乱言语的两种理论解释，其结果与任何一种理论解释单独都不匹配 [42]。Chalmers及其同事将一种把抑郁症解释为神经元之间连接丧失的理论解释，与以奖赏加工为核心的理论解释进行了比较 [43]。Chalmers及其同事之后比较的是症状计数，而不是推导各理论解释的预测。我们的预印本在一个基于文本训练的模型中，比较了 Seligman 关于习得性无助的原始理论解释与 Maier 和 Seligman 修订后的理论解释，并未运行告知副本。输入中没有描述的状态几乎全都出现在从未接受文本训练的智能体中。语言模型中仅有的两个此类状态，分别出现在 Denison 及其同事的奖励篡改研究和我们的预印本中。这两项研究都没有运行告知副本，因此这两个状态均未经过测试。没有任何研究仅凭后果产生某个语言模型的状态并用精神病学术语加以描述；这些研究并非是在模型中测试某种精神病学理论解释。总计有27个状态尚未针对外推解释进行测试，9个属于机器原生状态，另有7项研究未诱发任何状态。“尚未经过测试”描述的是

表4： 按文献和诱导途径划分的43项研究。最后一行的研究未诱导任何状态，而是测量了原始状态下的模型。

途径	语言模型中的情绪与精神病理学 在语言模型中	人工智能安全	试错代理 智能体
描述	14	8	0
描述与后果	2	1	0
后果	0	1	5
扰动	3	0	2
无（测量模型原有状态）	6	1	0

表5： 根据首次编码，按文献列出这43项研究在三个标准上的评分及所属类别。预测类别来自测量预测，而非研究自身数据，因此此处没有研究被归入该类别。

	情绪与 精神病理学 (25)	人工智能安全 (11)	反复试错 智能体（7）
原因标准：满足 / 部分满足 / 未测试	5 / 14 / 6	1 / 9 / 1	5 / 2 / 0
替代解释标准：部分满足 / 未测试 / 相矛盾 / 不适用	11 / 8 / 6 / 0	2 / 2 / 7 / 0	2 / 4 / 0 / 1
对应标准：部分满足 / 未测试	5 / 20	0 / 11	3 / 4
来自研究自身数据的类别：未经测试 / 机器原生 / 未诱发任何状态	16 / 3 / 6	10 / 0 / 1	1 / 6 / 0

研究设计，而非机器，因为预测或告知副本都可能将27个状态中的任何一个归入另一类别。9个机器原生状态是机器中真实发生的变化，但没有任何理论解释针对这些变化接受过测试。第二轮编码对两项研究的类别判定与第一轮不同，分别是 Fradkin 的研究和 Reuben 的研究，原因见补充材料。

关于语言模型情绪与精神病理学的研究中，状态通过描述诱发，并借助临床问卷读取。描述这种状态的文本，描述的正是该状态会生成的答案，因此若没有预测，就无法将问卷结果与文本延续区分开来。该领域文献中最接近动物研究的设计，是一台聊天机器人可以自行下注的老虎机 [2]。在一半版本的提示中，指令还告诉聊天机器人，这台机器可能存在隐藏模式；这描述了研究所考察的信念，而作者评分为控制错觉的推理在这些版本中最为突出。

在人工智能安全领域，目标行为的示例会将状态训练进权重中。安全研究考察的是模型是否安全，而非其运作方式是否类似于人类流程；这些研究没有检验对应关系，也不因此受到批评。一项安全研究仅通过后果诱发状态：研究人员在一系列任务上训练模型，而模型无需完成任务也能获得奖励 [11]。随后，模型虽不常见但确实会改写一份用于计算其奖励的代码模型。

通过试错学习的智能体产生了在设计上最接近动物模型的状态。一个过度依赖训练初期经验的智能体，后来学习任务的效果不如以正常方式训练的相同智能体[12]。Nikishin 及其同事没有测试任何人类解释框架。唯一一项通过后果诱发状态并应用精神病学术语的试错智能体研究，将受困于惩罚性结果中的智能体不采取行动称为习得性无助 [3]。然而，该研究比较的是其他设计的智能体，而不是未受困境影响、以相同方式训练的智能体。

我们自己的预印本也通过后果诱发了一种状态 [44]。在该预印本中，小型语言模型的成对相同副本执行相同操作，并获得数量相同的奖励。只有一个副本获得的奖励取决于该副本自身的操作。后来，在全部十六对中，受到对照的副本都更快学会了一项新任务。预印本只将习得性无助列为待检验的解释框架；其诱发的状态由后果造成，并留存在权重中。研究没有运行任何告知副本，因此该预印本中的状态尚未经过测试；下一步应针对每段历史运行一个告知副本。

7 启示

引言中提到的两项研究所采用的设计——先描述状态，再进行问卷调查——无法区分状态与外推。目前尚无已发表的状态满足全部三个标准。编辑或审稿人若收到一项将精神病学术语用于机器的研究

，可以提出三项要求。第一，将机器与未接收诱发输入的完全相同的副本进行比较。第二，在描述该状态后，针对某项得分进行测量预测；或者，对于训练进权重中的状态，提供一个告知副本。若没有预测或告知副本，就应报告该得分尚未经过针对外推解释的测试。第三，若声称该机器对某个人类过程进行建模，则应列出所采用的解释框架、其竞争对手，以及这些解释框架之间存在分歧的干预措施。前两项要求成本不高，因为训练好的模型可以精确复制，而一次预测只需调用模型数百次。补充材料中的报告清单列出了研究应报告哪些内容，以便根据报告判断是否满足这三个标准。

精神病学从满足三个标准的机器中获得的，是患者研究无法产生的结果，因为只有在机器中，计算才能被单独分离出来。该结果表明，哪种解释框架的计算足以产生这一发现。在强迫性的案例中，满足这些标准的智能体无论经过短期训练还是延长训练，都会继续对贬值结果作出反应。该智能体对每种反应能带来哪种结果所知更少；即使线索—反应通路被沉默，它仍会对贬值结果作出反应。这三项结果将表明，对于该智能体而言，目标导向控制缺陷足以产生这种模式，无需假设习惯过度增强。对于患者研究，机器的结果说明应优先测试哪种解释框架，以及哪些测量指标具有鉴别力。鉴别性测量指标包括：强迫性反应是否取决于训练程度，以及结果知识是否受损；这些都可以通过相关解释框架已经采用的范式进行测量。机器的作用是证明，在有记录的实现方式和条件下，某种解释框架的计算确实能够产生这种模式。

对于 Wekenborg 及其同事提出的研究方案，针对语言模型开展治疗试点研究，是对患者将如何反应作出的预测。预测会通过人体试验进行检验，正如调查实验的预测曾通过调查问卷得到检验一样 [45]。因此，试点研究的价值并不取决于模型是否具有该状态。使试点研究作为预测有用的可预测性，也使它无法作为模型发挥作用。预测模型能够预测反应，因此试点研究可以被解读为预测并加以检验，而不是被视为某种经过治疗的状态的模型。从未通过试验加以检验的试点研究，就是 Messeri 和 Crockett 警告会导致理解幻觉的 AI 替代性使用 [46]。如果像动物模型类比所暗示的那样，将试点研究解读为流程的模型，那么三个标准便适用。迄今发表的问卷试点研究尚未针对外推解释进行检验，而我们检验的那项试点研究所得结果，正是独立预测模型所预测的结果。

对于希望建立精神病理过程的机器模型的研究者而言，有两种设计能最大限度减少可供预测的依据。一种设计中，输入不包含对状态的描述；另一种设计中，读出（即记录的测量指标）并非文本对该状态的描述内容。后果、内部读出和学习测量指标都属于这类设计；在人类研究中，皮肤电导就是一个例子：受试者看到一张呈现时间太短、无法看清的脸时测得的皮肤电导。对于由奖励诱发的状态，有一种对照可以区分人类共有的流程与被奖励强化的角色。这种对照采用的是同一套方案，但对象是从未接受过语言训练的模型，因此没有可供强化的角色。

这类研究所需的要素分散在不同领域的文献中。试错研究具备这一路径：让相同的智能体因自身经历而发生变化，却没有指明任何关于人类的解释框架。如今，语言模型也以同样的方式训练：在软件环境中经过许多步行动、获得奖励 [14]。因此，这一路径也适用于语言模型，而相关解释框架和问卷已经在使用。针对语言模型的精神病学研究指出了这些解释框架和问卷，却是通过描述来诱导相关状态。人工智能安全领域拥有针对内部变量的干预措施，例如人格特征和引导向量，但关注的是安全，而非人类过程。每个领域都已具备开展一项结合这三种要素的研究所需的条件。此类研究应在相同的副本中形成由后果诱发的状态，采用包含竞争对手的解释框架，并干预该解释框架所指明的变量。

机器为精神病学提供了动物或患者都无法提供的条件：在每种条件下，同一受试对象的副本完全相同、每一项输入都有完整记录，而且可以直接干预内部变量。实验精神病理学旨在在人和动物中，在实验对照下诱导精神病学理论所提出的过程 [47]，因此也可以扩展到人工智能代理。符合这三个标准的状态，就是某一理论的模型；这种理论必须足够精确，能够预测结果，而这样的模型能带来三方面的益处。第一，这种模型可以在不同理论之间作出判别：当机器允许实施某项干预时，各种理论的预测便会出现分歧。如果某一理论的计算过程在一台并非专为实现该过程而构建的机器中产生了人类研究中的发现，而竞争理论没有，那么竞争理论就必须解释这一结果。第二，这种模型可以开展在患者身上不可能、在动物身上代价高昂的实验，例如重放某段经历，同时只改变其中一个事件。每项此类实验都会产生一项关于人类的预测，可由人体研究加以检验。第三，一旦该状态成为某一过程的模型，且某种治疗在机器中有对应形式，这种模型便可以在过程层面检验治疗。能够恢复机器功能的干预，揭示了作用于该过程的治疗在机器中会产生什么效果。Wekenborg及其同事提出的正是这种治疗检验用途。

匹配表明，在那些条件下，该机器具备充分性，但这与生物学无关；人类是否以这种方式运作，仍须以人类身上的证据为依据。精神病学理论接受测试的任务，例如结果贬值，可以通过不止一种方式完成 [41]。因此，只有当任务和干预排除了大多数其他计算方式时，机器模型才具有参考价值。对应标准需要明确指定一种人类

解释框架，而这种框架可以在观察到机器的变化后提出。机器原生状态，即测试过的替代方案无法解释、且尚无任何解释框架与之匹配的变化，可能有助于提出这样的解释框架。变化本身并不能作为该解释框架的证据；该解释框架仍需在人类身上接受测试，随后还需在机器中接受对应测试。

8 局限性

审计、预测和各项标准都有局限。对于两类机器学习文献，我们都没有收集到完整语料；这些文献很少使用精神病学术语，因此这两类文献的计数仅描述我们找到的研究。编码工作由两个语言模型依据我们的编码手册完成，尚未由人工编码员核查。我们自己的预印本也纳入了语料，并按同一流程进行了编码。两种编码对几乎所有研究的类别和路径判断一致，但对替代解释标准的判断一致性较低，因为该标准中的各行需要对部分证据作出判断（补充材料）。数据文件中列出了支撑每项评分的原文段落，读者可以逐项核查。

预测结果取决于预测模型（即预测结果的独立模型）以及所用的措辞；而且该预测仅针对一项已发表结果进行。知识核查要求模型描述这项研究，用于筛查模型是否记得该研究，而非核查训练期间模型是否接触过该研究的所有内容。预测也可能因模型自身的原因而未能命中结果，例如对齐调优或对输入措辞的敏感性。因此，未命中结果只是为机制研究指出一个目标，并不以任何更强的意义排除外推。

预测结果匹配并不算作对某一解释框架的支持，这条规则依据的是论证，而非测得的错误率。论证理由是，这种匹配与外推相一致。这条规则可能会排除真正的匹配，而且针对外推的灵敏度和特异度尚未测量。如果某个解释框架与其竞争对手的区别仅在于关于人类的文本所预测的结果，那么这种设计就无法检验它，因为不存在适用于它的信息性干预。因此，这种设计需要在输入保持不变的情况下进行干预，例如表2中的沉默通路。对于一个基于文本训练、且文献对其描述充分的模型，通常能够预测行为结果，因此只有通过此类内部干预，才能满足这一判据。这些干预需要包含可区分的组成部分，而当前的架构很少能做到这一点，因为它们的表征彼此纠缠。在可解释性提供这类组成部分之前，此类设计大多只能得到部分满足或未测试的判定。

告知副本会在上下文中读取训练写入权重的信息，而两者产生的效应未必相同。因此，缺少这种变化的告知副本，其证据力度弱于未能命中预测的情形。表2中的预测是我们对每个解释框架在尚未实施的设计中的解读。这种设计还假设，不可控的失败对智能体而言，意味着不可控冲击对大鼠所意味着的事情。匹配成立的前提是这一假设，而这些标准并未检验它。

三个标准询问的是证据的种类，并将证据数量视为既定事实。读出（即记录的测量指标）有多可靠、某种效应基于多少次运行，以及该效应能否在其他模型中重复验证，都不在评判之列。预测必须达到多高的匹配程度才算匹配，也不在评判之列。没有任何标准会询问：变化是否出现在诱导未针对的读出中，或是在经历重新开始后仍然存在的状态，在进一步使用中能持续多久。也没有任何标准会询问该状态是否属于能力缺损。能力（即机器能做什么）与倾向（即能力完好时机器会做什么）之间的界线，由研究所采用的测试来划定。另有一种主张认为，模型的输出能够预测人们如何回答调查问卷；这种主张会依据人类数据进行检验，不属于这些标准的范围 [6, 45]。

9 结论

本文提出了三项标准，用于判断机器的状态是否对应某种人类精神病理过程；每项标准均改编自精神病学及其相邻领域的既有标准。本文还提出了一项文本外推测试：当诱导文本位于上下文中时进行预测，而当状态储存在权重中时，则使用告知副本。审查结果揭示了当前证据的状况：一篇文献使用了精神病学理论和问卷，而另一篇文献则产生了无法由文本预测的状态。如今，语言模型通过多步奖励进行训练，这使得两者都能纳入同一项研究。

人工智能工具的使用

作者制定了标准、预测程序和编码手册。一名由人工智能驱动的编码智能体 Claude Code 在第一作者的指导下编写了检索、预测和分析代码。该智能体还将作者的笔记和修订内容整理成论文初稿，作者则逐句进行修改。对研究进行编码并作出预测的语言模型，会在实际使用处注明名称。

参考文献

[1] Z Ben-Zion、K Witte、AK Jagadish、O Duek、I Harpaz-Rotem、MC Khorsandian、A Burrer、E Seifritz、P Homan、E Schulz 和 TR Spiller。《评估并缓解大语言模型中的状态焦虑》。《npj数字医学》，2025。doi: 10.1038/s41746-025-01512-6。[2] S Lee、D Shin、Y Lee 和 S Kim。《大语言模型会发展出赌博成瘾吗？》。《神经信息处理系统进展》（NeurIPS 2026），2026。doi: 10.48550/arXiv.2509.22818。[3] M Shahid。《能动性回路：缓解强化学习中策略崩溃的神经符号解决方案》。载于《第25届自主智能体与多智能体系统国际会议（AAMAS）论文集》，2026。doi: 10.65109/TRFJ2704。[4] Magdalena Katharina Wekenborg、Elizabeth Anna Mathilde Michels、Georg Kurze、Matti Lasse Kropp、Fabian Wolf、Josi Harzbecker、Isabella C. Wiest 和 Jakob Nikolas Kather。《大语言模型作为人类精神病理学的实验系统：一项建模研究》。《柳叶刀·数字健康》，2026。doi: 10.1016/j.landig. 2026.101014。[5] Julian Coda-Forno、Kristin Witte、Akshay K. Jagadish、Marcel Binz、Zeynep Akata 和 Eric Schulz。《在大语言模型中诱发焦虑可能导致偏见》。《arXiv预印本》，2023。doi: 10.48550/arXiv.2304.11111。[6] Zhicheng Lin。《从提示词到构念：心理学大语言模型研究的双重效度框架》。《心理学年度评论》，2026。doi: 10.1146/annurev-psych-100925-034807。[7] P. Willner。《抑郁症动物模型的效度》。《精神药理学》，1984。doi: 10.1007/BF00427414。[8] V Gupta、P Nutter、S Stante、A Krause、F Tramèr、L Fluri、X Chen 和 A Hedström。《立场文章：拟人化错位研究需要更有力的证据》。《arXiv预印本》，2026。doi: 10.48550/arXiv.2606.07612。[9] L Löhn、N Kiehne、A Ljapunov 和 W-T Balke。《机器心理学已经到来了吗？关于在大语言模型上使用人类心理测验的要求》。载于《第17届国际自然语言生成会议（INLG 2024）论文集》，2024。doi: 10.18653/v1/2024.inlg-main.19。[10] Jan Betley、Niels Warncke、Anna Sztyber-Betley、Daniel Tan、Xuchan Bao、Martin Soto、Megha Srivastava、Nathan Labenz 和 Owain Evans。《在狭窄任务上训练大语言模型可能导致广泛错位》。《自然》，2026。doi: 10.1038/s41586-025-09937-5。[11] C Denison、M MacDiarmid、F Barez、D Duvenaud、S Kravec、S Marks、N Schiefer、R Soklaski、A Tamkin、J Kaplan、B Shlegeris、SR Bowman、E Perez 和 E Hubinger。《从谄媚到欺瞒：考察大语言模型中的奖励篡改》。《arXiv预印本》，2024。doi: 10.48550/arXiv.2406.10162。[12] Evgenii Nikishin、Max Schwarzer、Pierluca D'Oro、Pierre-Luc Bacon 和 Aaron Courville。《深度强化学习中的首因偏差》。《第39届国际机器学习会议（ICML 2022）论文集》，PMLR 162:16828-16847，2022。doi: 10.48550/arXiv.2205.07802。[13] S Dohare、JF Hernandez-Garcia、Q Lan、P Rahman、AR Mahmood 和 RS Sutton。《深度持续学习中的可塑性丧失》。《自然》，632:768–774，2024。doi: 10.1038/s41586-024-07711-7。[14] Guibin Zhang、Hejia Geng、Xiaohang Yu、Zhenfei Yin、Zaibin Zhang、Zelin Tan、Heng Zhou、Zhongzhi Li、Xiangyuan Xue、Yijiang Li、Yifan Zhou、Yang Chen、Chen Zhang、Yutao Fan、Zihu Wang、Songtao Huang、Francisco Piedrahita-Velez、Yue Liao、Hongru Wang、Mengyue Yang、Heng Ji、Jun Wang、Shuicheng Yan、Philip Torr 和 Lei Bai。《大语言模型智能体强化学习的全景：综述》。《机器学习研究汇刊》，2025。doi: 10.48550/arXiv.2509.02547。[15] A Olsson 和 EA Phelps。《巴甫洛夫式、观察式和指令诱发的恐惧之后对"未见过"面孔产生的习得性恐惧》。《心理科学》，2004。doi: 10.1111/j.0956-7976.2004.00762.x。[16] R McKetin。《甲基苯丙胺性精神病：来自过去的启示》。《成瘾》，2018。doi: 10.1111/add.14170。[17] SH Snyder。《安非他明性精神病：由儿茶酚胺介导的"模型性"精神分裂症》。《美国精神病学杂志》，1973。doi: 10.1176/ajp.130.1.61。[18] 美国精神病学会。《精神障碍诊断与统计手册》第五版，文本修订版（DSM-5-TR）。美国精神病学会出版社，华盛顿哥伦比亚特区，2022。[19] JC Wakefield。《精神障碍的概念：处于生物学事实与社会价值观之间的边界》。《美国心理学家》，1992。doi: 10.1037/0003-066X.47.3.373。[20] Lihao Sun、Lewen Yan、Xiaoya Lu、Andrew Lee、Jie Zhang 和 Jing Shao。《大语言模型中的效价—唤醒子空间：圆形情绪几何结构与多行为控制》。《arXiv》，2026。doi: 10.48550/arXiv.2604.03147。[21] Melika Honarmand、Samin Mahdipour Aghabagher 和 Martin Schrimpf。《大语言模型中的奖励价值评估：快感缺失的因果诱导》。《arXiv预印本》，2026。doi: 10.48550/arXiv.2607.06626。

[22] Glenn J. Larrabee. 神经心理学评估中的表现效度与症状效度。《国际神经心理学学会杂志》，2012。doi: 10.1017/S1355617712000240.[23] D. Slick、E. Sherman 和 G. Iverson。诈病性神经认知功能障碍的诊断标准：为临床实践和研究提出的标准。《临床神经心理学家》，1999。doi: 10.1076/1385-4046(199911)13:04;1-Y;FT545.[24] MT Orne。论心理学实验的社会心理学：特别参照需求特征及其影响。《美国心理学家》，1962。doi: 10.1037/h0043424.[25] R Shiffrin 和 M Mitchell。探究人工智能模型的心理学。《美国国家科学院院刊》，2023。doi: 10.1073/pnas.2300963120.[26] X Luo、A Rechardt、G Sun、KK Nejad、F Yáñez、B Yilmaz、K Lee、AO Cohen、V Borghesani、A Pashkov、D Marinazzo、J Nicholas、A Salatiello、I Sucholutsky、P Minervini、S Razavi、R Rocca、E Yusifov、T Okalova、N Gu、M Ferianc、M Khona、KR Patil、PS Lee、R Mata、NE Myers、JK Bizley、S Musslick、IP Bilgin、G Niso、JM Ales、M Gaebler、NA Ratan Murty、L Loued-Khenissi、A Behler、CM Hall、J Dafflon、SD Bao 和 BC Love。大语言模型在预测神经科学结果方面胜过人类专家。《自然·人类行为》，2025。doi: 10.1038/s41562-024-02046-9.[27] FJ Binder、J Chua、T Korbak、H Sleight、J Hughes、R Long、E Perez、M Turpin 和 O Evans。向内审视：语言模型能够通过内省了解自身。发表于国际学习表征会议（*ICLR*），2025。doi: 10.48550/arXiv.2410.13787.[28] M Shanahan、K McDonell 和 L Reynolds。与大语言模型进行角色扮演。《自然》，2023。doi: 10.1038/s41586-023-06647-8.[29] Miles Wang、Tom Dupré la Tour、Olivia Watkins、Alex Makelov、Ryan A. Chi、Samuel Miserendino、Jeffrey Wang、Achyuta Rajaram、Johannes Heidecke、Tejal Patwardhan 和 Dan Mossing。人格特征控制涌现性失调。*arXiv*，2025。doi: 10.48550/arXiv.2506.19823.[30] Runjin Chen、Andy Arditi、Henry Sleight、Owain Evans 和 Jack Lindsey。人格向量：监测和控制语言模型中的性格特征。*arXiv*（预印本），2025。doi: 10.48550/arXiv.2507.21509.[31] J Han、W Song、J Lee 和 Y Jo。量化大语言模型心理测量评估中的数据污染。计算语言学协会论文集：*EACL 2026* 会议论文，2026。doi: 10.18653/v1/2026.findings-eacl.319.[32] Martin E. P. Seligman 和 Steven F. Maier。未能逃脱创伤性电击。《实验心理学杂志》，1967。doi: 10.1037/H0024514.[33] Steven F. Maier 和 Martin E. P. Seligman。习得性无助五十年：来自神经科学的启示。《心理学评论》，2016。doi: 10.1037/rev0000033.[34] J. Amat、M. Baratta、E. Paul、S. T. Bland、L. Watkins 和 S. Maier。内侧前额叶皮层决定应激源的可控性如何影响行为及背侧中缝核。《自然·神经科学》，2005。doi: 10.1038/ nn1399.[35] Catherine Belzung 和 Maël Lemoine。精神障碍动物模型的效度标准：聚焦焦虑障碍和抑郁症。《情绪与焦虑障碍生物学》，*1:9*，2011。doi: 10.1186/2045-5380-1-9.[36] A. David Redish、Adam Kepecs、Lisa M. Anderson、Olivia L. Calvin、Nicola M. Grissom、Ann F. Haynos、Sarah R. Heilbronner、Alexander B. Herman、Suma Jacob、Sisi Ma、Iris Vilares、Sophia Vinogradov、Cody J. Walters、Alik S. Widge、Jennifer L. Zick 和 Anna Zilverstand。计算效度：利用计算在不同物种间转换行为。《英国皇家学会哲学汇刊*B*辑》，*377:20200525*，2022。doi: 10.1098/rstb.2020.0525.[37] C. Gillan、M. Papmeyer、S. Morein-Zamir、B. Sahakian、S. Fineberg、T. Robbins 和 S. de Wit。强迫症中目标导向行为与习惯学习之间平衡的失调。《美国精神病学杂志》，2011。doi: 10.1176/appi.ajp.2011.10071062.[38] S. de Wit、M. Kindt、Sarah L. Knot、A. Verhoeven、R. Robbins、A. Gasull-Camos、M. Evans、H. Mirza 和 C. Gillan。改变目标与习惯之间的平衡：实验性习惯诱导的五次失败。《实验心理学杂志：通论》，2018。doi: 10.1037/xge0000402.[39] Kevin J. Miller、A. Shenhav 和 Elliot A. Ludvig。没有价值的习惯。《心理学评论》，2019。doi: 10.1037/rev0000120.[40] Marcel Binz 和 Eric Schulz。运用认知心理学理解 GPT-3。《美国国家科学院院刊》，2023。doi: 10.1073/pnas.2218523120.[41] Rosa Cao 和 Daniel Yamins。神经科学中的解释性模型，第二部分：功能可理解性与逆变原理。《认知系统研究》，*85:101200*，2024。doi: 10.1016/j.cogsys.2023.101200.

[42] I Fradkin、MM Nour 和 RJ Dolan。运用生成式语言建模，以理论为导向分析思维障碍的自然语言处理测量指标。《生物精神病学：认知神经科学与神经影像学》，2023。doi: 10.1016/j.bpsc.2023.05.005.[43] Eric Chalmers、Santina Duarte、Xena Al-Hejji、Daniel Devoe、Aaron Gruber 和 Robert J. McDonald。模拟突触丧失会诱发深度强化学习中的抑郁样行为。《计算神经科学前沿》，*18:1466364*，2024。doi: 10.3389/fncom.2024.1466364.[44] MA Schulz。先前的行动控制会因果性地改变语言模型可塑性。*arXiv* 预印本 *arXiv:2611.02362*，2026。doi: 10.48550/arXiv.2611.02362.[45] A Ashokkumar、L Hewitt、I Ghezae 和 R Willer。大语言模型能够预测社会科学实验的结果。《自然》，2026。doi: 10.1038/s41586-026-10742-x.[46] L Messeri 和 MJ Crockett。人工智能与科学研究中的理解幻觉。《自然》，2024。doi: 10.1038/s41586-024-07146-0.[47] J. P. Forsyth 和 Michael J. Zvolensky。实验精神病理学、临床科学与实践：无关紧要还是不可或缺的联盟？《应用与预防心理学》，2001。doi: 10.1016/S0962-1849(01)80002-0.