意识与对话式人工智能

曼努埃尔-博伊塞宁

序言

继 OpenAI 的 ChatGPT 于 2022 年底问世之后，对话式人工智能已变得司空见惯。与电子邮件、搜索引擎和社交网络一样，它们现在已成为新一代频繁使用的各种数字工具的一部分。对话式人工智能令人惊叹的惊人能力，尤其是对语言的驾驭能力和处理高度特定主题的能力，在震惊世界数月之后，如今已变得司空见惯。

然而，这些能生成文本的人工智能开创了机器与用户自然交互的新纪元，表明它们可以在没有任何特定形式主义的情况下对用户的请求做出回应，甚至不逊于人类，而且拥有超强的百科知识。许多创新已经并将继续出现。然而，这些人工智能仍不完美，有时无法解决一个简单的问题，而且有捏造的倾向，即在回答问题时虚构与现实混杂在一起。不过，与以前的聊天机器人相比，对话式人工智能所提供的答案是无法比拟的，因为以前的聊天机器人在我们提出开放式问题时，只能给我们提供多个问题的选择，或者提供一个并不总是很合适的文章列表。

通过掌握迄今为止一直被认为人类特权的语言，我们可以扪心自问，我们与机器之间还有什么区别，或者相反，这些人工智能与我们有什么相似之处，以及它们能给我们带

来什么启示。

本书分为两部分：

- 首先是我们的意识，我们将通过各种内省反思对其进行更详细的研究；这将使我们更好地把握这一现象的复杂性，实际上它涵盖了多个不同的现象。我们的意识由不同的过程所支撑，无疑比一个整体的、静态的、可从外部观察到的经典物体更难把握。然而，我相信，理解它对于理解我们是谁至关重要，为了更接近它，我们将考察我们的语言能力、我们的记忆力和我们的注意力，以这些新型对话式人工智能为镜像，踏上我们存在的核心之旅。

- 在第二部分中，我们将更直接地探讨会话式人工智能，如 ChatGPT。我们的介绍将更加注重说教和事实，因为这些过程更容易理解，在某种程度上也更容易被理解。我们将研究它们的能力（包括新出现的心智理论）、弱点（尤其是在推理方面）以及它们的性质，并特别关注这类软件是否拥有某种形式的意识（尽管它们目前仍不具备知觉）这一问题。

目录

序言 1

术语表 4

第一部分 7

1.导言 8

2.什么是人工智能？ 10

3.意识，几点初步看法 19

4.认识、语言和学习 25

5.汽车概念的沉思 30

6.非局部意识或超验意识 33

7.意识与他者 42

8.意识、注意力和记忆力 48

9.意识与身体 53

10.认识世界、概念意识和意识水平 57

11.主观意识 62

12.作为 quale 的概念 67

13.意识：走向定义 68

14.意识概念出现 76

15.超越意识 79

16.简而言之 86

第二部分 91
17.ChatGPT 对自己的良心有何评价？ 92
18.这种意识的本质.....109
19.意识、自觉性、认知和对话式人工智能112。
20.ChatGPT 最初的一些弱点.....118
21.语言与心智理论教师.....128
22.良知与*法学硕士*.....135
23.对话理解与人工智能.....147
24.中式卧室.....152
25.什么是语言模型？ 160
26.如何解释*变压器*的工作原理？ 163
27.风险.....175
28.结论.....178
附录.....185
作为 *quale 的*概念 185
意识的不同定义和描述.....189
法规.....194
代理.....197
致谢.....200

术语表

认知（cognitive）和***认知（cognitique***）是从认知（cognition）一词衍生出来的术语。认知是与知识功能甚至其他心理功能相关的一系列心理过程。从目前的意义上讲，认知学主要涉及人机交互的人体工程学。然而，鉴于新的对话式人工智能可以被描述为认知辅助工具，"***认知***"一词的语义自然会发生变化，从而使形容词变得更加充实，并使用 "认知 "一词。

因此，我用 "***认知***"一词来指能够以类似于人类并为人类所理解的方式理解、掌握和生成文本的新软件。我特别想到了语言和 ChatGPT 等文本生成人工智能。它们也可以被称为*新认知技术*，因为*严格来说，*这个词可以指更广泛的一类软件，包括在某些游戏中可与人类匹敌或超越人类的软件，以及能够以视觉方式解读我们所处环境的软件。

感觉和**对世界的认识**。我们还需要对这些术语的区别进行一定的发散性解释，以明确意识是什么，并区分意识的两个方面：

–第一种是本体论上的主观现象，即现象的性质取决于每个人的特定过程，并与主体的感受相对应；这尤其是我们所意识到的感觉；
–其次，"*意识*"是一种现象，它赋予我们对周围世界的认识和理解，可以类比于英语中的"*awareness*"（*意识*）一词，它与文化、知识和学习有着密切的联系。
这种区别在法语词汇中较难识别，但却很有意思，因为它提出了一个问题：意识一词所涵盖的这两个方面是如何既相互联系又相互独立的。

Prompter 源自 *prompt（提示）*和 *to prompt（提示）*，分别指人工智能对话中的对话提示和"*.....* "的动作。

完成这个提示，就可以向人工智能提出请求。*提示工程*指的是用于提出请求的所有技术，目的是为*大型语言模型（LLM）*的响应设定条件。

GPT：Generative *Pre-trained Transformer*（生成预训练*转换器*），*转换器*是一种被称为深度的神经网络架构；深度是因为它有大量的层。事实上，这些层被组织成块，堆叠起来形成神经网络。ChatGPT 3.5 堆叠了约一百个区块。预训练指的是 ChatGPT 是在大量文本语料库中训练出来的，而 ChatGPT 3.5 这个名字则来自于《木偶奇遇记》创作者 Geppetto 的双关语。

LLM：大型语言模型，又称 *GML*、*MPL* 或语言模型。这是一个专业术语，用来指代分析提示并生成文本输出的程序或过程。其他 *LLM* 规模较小（8B 参数），被 SLM（小语言模型），它们的微调可以通过基于大型*会话人工智能*（chatGPT 3.5 为 175B）响应的师生式训练来实现；这也被称为提炼。文本*生成式人工智能*这个词也很有意思，虽然它更通用一些，但仍然非常明确。

B：相当于美式的十亿，或我们的十亿，10^9。它用于描述模型的大小，即模型训练时学习到的参数数量，这些

参数对模型的学习进行编码。

T：相当于太拉，或 1000 亿，10^{12}。它用于描述训练模型的数据集的大小。尽管如此，在百分之几的范围内，这个大小与单词的大小是一致的。1000 亿（10^{11}）是银河系中恒星数量的数量级。

Quale（***Qualalia*** 的单数）：这些是基本感知，是我们的感官受到刺激时感知到的东西。这些感知包括表面的质感、声音的音调和香气。

GOMAX：Google、OpenAI、Meta、Anthropic、X 代表了新一轮创新浪潮中最重要的参与者。

该书还附有一个网站：https://boissenin.net，在那里你可以找到一个讨论论坛和一些引用参考文献的链接。

第一部分

1.导言

在 OpenAI 及其*聊天机器人* ChatGPT 于 2022 年底推出之后 [1]，其他众多对话式人工智能也相继推出。

根据伟大的语言模型和我们个人的意识经验，我们会问，这些会话式人工智能本身是否不具有某种形式的意识，如果我们曾与人工智能互动过，这可能是对其本质的一种令人不安的信念。毕竟，它们看起来不是具有理解力吗？而 "理解 "难道不需要意识到向它们提出的问题的意义吗?

机器学习，作为一门学科，同时也作为一种新的软件创建范式，为反思我们自身的本质提供了丰富的知识领域和有趣的可能性。某些模型或算法的运作与我们自己的大脑和思维之间的类比和隐喻[2] 可以揭示我们的心理现象，是了解我们自身的有效来源。

另一方面，当人工智能已经能够比我们自己更敏锐地探测我们的内心深处时，如果我们要在这些系统

面前保持我们的自主性，并从它们的后继者那里解放出来，更好地了解我们自己的本性证明是不可或缺的，有时甚至是至关重要的。

---

[1] 2022 年 11 月 30 日

[2] **心灵**涵盖所有心理现象和能力：知觉、思维、直觉、感觉、智力、记忆、价值观等。

他们将主宰，甚至部分主宰，但在大规模上，下一代.....

自计算机诞生之初，甚至在此之前，阿兰-图灵等有远见的人就曾设想，有朝一日机器也能变得有意识。在很长一段时间里，人工意识似乎是许多计算机研究人员的圣杯，是一个不折不扣的神学项目，但现在看来，我们显然已经到了一个转折点。甚至很有可能，在某些大公司的慎重考虑下，这一目标已经实现这些公司仍然关心了解其创造的后果并加以控制。在我看来，这一大胆甚至是匪夷所思的论断是合理的，而本书的目的之一就是支持这一论断并解释其原因。

一路上，我将邀请大家进行一些反思，通过审视它，帮助大家更好地分辨它是如何运作的，以及构成我们意识的要素。意识是一个卓越的复杂对象，如果我们要准确地感知和理解它，就需要从多维度入手。本书的第一部分专门讨论这个问题。在第二部分中，我们将从不同角度分析对话式人工智能，而

在本书的最后，我们将为那些更好奇的读者揭开驱动其文本生成过程的 *Transformer* 架构的神秘面纱。

2. 什么意思 什么是 人工智能是什么意思?

*"成功创造人工智能将是人类历史上最伟大的事件。不幸的是，这也可能是最后一次，除非我们学会规避风险。*斯蒂芬-霍金

在澄清 "人工智能 "这个已成为名副其实的概念包罗万象的术语的含义之前，我们首先要注意的是，思维并不等同于语言思维；使用图像表征的跳棋棋手或数学家很容易意识到这一点。因此，我们要区分语言思维和视觉思维。

这第一个区别表明，对话式人工智能的工作方式与我们的工作方式有着深刻的不同。事实上，*LLMs*[3] 只将文字（即口头描述）作为程序的输入。假设它们能够有意识，那么它们意识的性质显然与我们不同。*LLMs* 最多只能拥有基于语言和语言描述的表征。尽管如此，如果我们要求 *LLMs* 为我们画一只蜥蜴，并指定用矢量绘图语言进行描述，它就能描绘出一只长尾巴、舌头分叉的四脚动物[YT (YouTube), Sparks of AGI: early experiment with GPT 4, Sébastien

Bubek]。

但是，在我们讨论机器或程序的意识之前，让我们认识到，许多物种，即使没有像我们这样发达的语言，也是有意识的。 对于那些怀疑或认为这

---

[3] 大型语言模型，新的对话式人工智能是多模态的，也可以将图像作为输入，这使它们能够更好地理解 *PDF* 文档等。

如果您感兴趣，我邀请您参阅《剑桥意识宣言》（2012 年）。以下是摘录：

"鸟类的行为、神经生理学和神经解剖学似乎代表了意识平行进化的一个显著案例。在加蓬灰鹦鹉身上观察到的接近人类意识水平证据尤其引人注目。哺乳动物和鸟类的情感脑网络和认知微电路似乎比以前认为的有更多的共同之处。此外，还发现一些鸟类表现出与哺乳动物类似的睡眠周期，包括快速眼动睡眠，以及斑马雀的神经生理模式，这在以前被认为没有哺乳动物的新皮质是不可能的。尤其是喜鹊，在镜像自我识别研究中被证明与人类、类人猿、海豚和大象有着惊人的相似之处"。

有趣的是，紧随其后，2019 年又发表了关于动物法律人格的《土伦宣言》。毫无疑问，人工智能的发展和创造有意识的人工智能的前景，以及由此可能产生的伦理和法律影响，都对这一宣言产生了影响。因为，正如我们将看到的那样，意识似乎确实是决定一种人工智能地位和自主形式的重要转折点，

特别是如果它拥有语言能力或其他认知能力，如规划能力，其水平超越了人类，换句话说，达到了超人的水平。

不过，在讨论这些问题之前，我们先来解释一下人工智能（AI）的含义。

人工智能涉及众多领域，这必然导致人们对其本质的混淆。与其提出一个可能过于宽泛的正式定义，我更愿意用几个例子来说明这个术语，以提供更具体的理解。这种方法更容易识别人工智能的不同形式，避免过度概括。例如，虽然 "人工智能会带来生存风险 "这一命题很有道理，但我们也可以问，一个只会下棋的人工智能怎么会带来地球规模的生存风险呢？显然不是这样的，这同样适用于所有专门从事特定领域的所谓狭义人工智能；例如，这些人工智能能够查看 X 光图像并确定肿块的性质：是囊肿、恶性肿瘤还是良性肿瘤？在看过成千上万张已经做出诊断的此类图像后，有时还需要进行冗长、复杂和侵入性的检查，人工智能就能学会区分图像中的不透明性，并做出足够可靠的诊断，从而避免进一步的检查，或者相反，建议进行进一步的调查。这种人工智能的范围也很窄，因为它无法识别人脸、绘画或二维码。

在分析公司数据以做出决策时，我们通常会借助使用统计和机器学习的程序来确定其关键指标，并将

这些指标结合起来，以获得公司运营的表征；这种建模中使用的技术和算法是机器学习的一部分，通常也被归类为人工智能。

人工智能已变得无处不在，它在销售网站上推荐产品、选择广告或识别语言以从语音切换到文本等方面发挥作用。 因此，它呈现出无数个方面，而每个方面都对应着一个独特的发展，背后隐藏着一个研究人员或工程师，或者更可能是一家小公司。绝大多数

像我刚才提到的那些 IAs，属于狭义类型，即专门针对某一应用的 IAs。

*深度学习*的出现是在 2012 年，当时这种基于多层神经网络的算法显然优于当时的人工智能算法，因此深度学习中出现了 "深度 "一词。由于可靠性和性能的提高，这导致了新人工智能应用的复兴和数量的激增，以及该术语的传播和普及。2012 年，一个基于神经网络架构的程序就能从数以万计的可能对象中判断出图像的内容，这为数以百计的创新打开了大门，而这些创新仍在深刻地改变着我们的社会。这就是所谓的 "人工智能寒冬的结束"。
一般来说，如今当我们谈论人工智能时，我们隐含谈论的是*在海量数据上通过机器学习训练出来的神经网络*。这已成为人工智能的主流模式，因为它在具有高度可变性的复杂情况下表现最佳 [4] 。

那么，什么是*人工通用智能*（*AGI*）呢?
这种人工智能与人类一样，能够适应任何情况。它的能力可以从一个领域转移另一个领域，转移速度

可以与人类相媲美，甚至更快。 考虑到机器的计算和记忆能力，这样的人工智能似乎显而易见。

---

[4]然而，专家系统作为人工智能的另一个主要分支，以及许多其他数据分析算法仍有广阔的发展前景。

如果智能出现，其速度可能远远超过人类。因此，不难理解为什么整个社会都对这一前景感到恐惧。尤其是软件可以从其执行基底中分离出来，并以极低的成本进行复制.....

像 ChatGPT 这样的语言模型可以被视为 *AGI* 吗?

ChatGPT 拥有超多领域的知识，毕竟它是在 1.4T *的标记*上训练出来的，这大约相当于 1400 亿个单词，或者 2000 万本书。一个人一生大约要阅读一千本书，这就相当于两万人同时阅读，或者相当于一个小镇有条不紊地分担阅读整个训练语料库的任务。然而，这些数量惊人的人类知识仍然脱离了现实，脱离了世界，也脱离了我们的主要本性，而我们的主要本性是与我们的直接环境相关的生理和经验。

我们赋予这种形式的软件以理解语言的能力，使其自然地与人类交流，尽管直到最近，人类才是唯一能够这样做的人，这难道不是一项惊人的成就吗?

然而，ChatGPT 并没有继续学习，至少没有持续学

习。如果它真的这样做了暴露在人群中，除了自己的交流，在世界上没有其他参照点，换句话说，脱离了地面，它就会很容易受到用户的影响，就像我们受到宣传或某些意识形态的影响一样。微软的 Tai 人工智能的经验已经清楚地证明了这一点，尽管 Tai 的技术不同于今天的 *LLM*。

对话式人工智能从人群中汲取灵感后，开始发表种族主义、纳粹和具有潜在风险的。因此，如果不对对话式人工智能的学习进行严格的控制，它很有可能根本无法稳定地维持符合用户期望的言论，即健康、平衡、非仇恨、仁慈和可靠的言论。既然 ChatGPT 不是 "泰"（Tai），那么它是否有可能利用自己的辨别力来抵制这类内容呢？可以想象，就像人类一样，它陷入与现实世界脱节的理论可能只是时间和争论的问题。然而，它的知识广度能使它在多大程度上抵御破坏它的世界模型的企图，这个问题仍然悬而未决。

此外，ChatGPT 区分善恶的能力[《创世纪》第 2 章和第 3 章]仍不确定，而且在最初的部署过程中，还没有很好地掌握使对话式人工智能符合公众和当局期望所必需的技术要求和约束。例如，微软在最初将 OpenAI 的对话式人工智能与其搜索引擎整合时就遇到了不良行为，并通过将对话限制为五次互动暂时解决了这一问题.....这突出表明，虽然让对话式人工智能发挥作用已经是一项重大挑战，但面对多样

化的公众，确保其稳定性并减少其偏差更是一项重大挑战[5]。

而 ChatGPT 的核心是一个*变压器*，即一个大型神经网络的架构，它属于

---

[5] 法国初创公司 Mistral 的 Le Chat 也接受了这一挑战。Méta 还将其对话式人工智能 *Llama* 作为开放源代码提供。在这些人工智能中表现最好的对话系统包括 Anthropique 的 *Claude*、谷歌的 *Gemini* 和中国的 *DeepSeek*，后者的模型是*开放权重的*。

如果将人工智能和 *LLM* 混为一谈，那就太简单近似了，因为它们的神经网络架构可能差别很大；而且，每种会话人工智能都有自己的特点。例如，ChatGPT 使用了 "*人类反馈强化学习*"（Reinforcement *Learning from Human Feedback*，RLHF），这种学习方法还考虑了用户对其回复的评价。大量的提示和对这些提示的回应实例，*LLM* 已经能够回应用户以自然语言提出的请求。以前要解决这些人工智能问题，必须进行*提示工程*。最后，虽然 ChatGPT 的多功能性还有很多不足之处，不能称其为 *AGI*[6]，*但它*预示着一种新型人工智能的诞生，这种人工智能具有足够的通用性，可以根据需求调整自己的功能并将自己转变为翻译、诗人、编剧、程序员、校对员*等*。这种转换能力标志着前所未有的突破。更重要的是，图像解读等进一步功能的加入，让人工智能有了愿景的雏形，这只会扩大其应用领域和能力。

现在让我们来澄清一下 "*ASI* "一词，即*人工超级智能*。超级智能 "一词源于尼克-博斯特罗姆（Nick

Boström）2014 年出版的一本书的书名，该书提请公众注意人工智能的潜在危险[《超级智能：路径、危险、战略》；博斯特罗姆，牛津大学]。 超级智能不仅会在很多领域超越人类，如国际象棋领域的深蓝或围棋领域的 AlphaGo，而且会在多种多样的任务中具有超强能力。当我们谈论*人工智能*时，我们最常关注的是它通过诞生 "超级智能 "来理解和改进自身或至少是下一代的能力。

---

[6]其最新版本：o3，再次开始受到挑战，它在认知方面不断取得突破，在各种基准测试中取得的巨大飞跃就证明了这一点。

奇点"[7] 给超人主义者带来了希望，也给许多人带来了对灾难的恐惧，这既包括就业市场上的灾难，也包括最危言耸听者所担心的世界末日风险。+ 如果你想更好地了解这一前景，我邀请你在 YouTube 上观看几个使用 "*ASI* "一词的视频。 根据计算电子假说 [YT: Singularity timeline | Artificial intelligence + *AGI ASI*]，世界末日被预测为距今约 2107 年。尽管如此，这并不是第一个被宣布的世界末日，但某些风险需要采取无限的预防措施和考虑，一些大公司可能会被批评没有采取所有应该采取和可以采取的预防措施，向公众迅速发布某些创新产品。然而，这一点很难证明或评估，因为创新的某些影响在其部署之前是无法预见或衡量的。

值得注意的是，Boström 一书出版后引发的思考，促成了 "生命未来研究所"（Future of Life Institute）于 2015 年发表了一封公开信 [Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter]，霍金、Yann Le Cun、Geoffrey Hinton、Yoshua Bengio、Ilya Sutskever、Elon Musk、Max Tegmark 等人都在信中签了名，强调了人工智能技

术对社会的潜在益处。因此，我们不能一概而论地宣称人工智能应用，甚至包含人工智能的工具，本质上都是危险的，但我们确实必须考虑到某些类型的人工智能可能会变得危险。就狭义的人工智能而言，应根据加以考虑：例如，过度检测乳腺癌的风险可能会导致儿童死亡。

---

[7] 当人工智能能够在没有人类干预的情况下不断进步时，技术奇点就会出现。 ��可能会导致高智能的出现。
它的速度超过了人类的能力，人类无法完全理解或控制它，而它最终可能掌握人类的命运。

至于所谓的强人工智能，如 *AGIs*，则需要以更高的警惕性和更谨慎的态度对其进行监控，因为它们可能产生的后果更加多样，因此从本质上来说更加难以界定。更重要的是，它们可以相对容易地在全社会蔓延，在最悲观的情况下，这可能会使它们变得无法控制。

归根结底，*AGI* 和*人工智能*之间的界限相当模糊，因为有足够计算能力的 *AGI 可以与人工智能*相媲美。虽然围绕人工智能可能具有的意志、自由意志和意向性仍有许多问题，但意识似乎是决定人工智能（尤其是具有语言能力的人工智能）能否成为 *AGI* 的关键点之一。事实上，通过意识到自身[8]、自身在世界中的位置及其互动关系，并在思考选择中获得自主权，人工智能不就能像我们的大脑一样发挥巨大作用了吗?

正是出于这个原因，同时也因为这是一个了解我们自身本质的旅程，所以我们要更深入地了解什么是意识。

[8]而 ChatGPT 无疑表明，它可以描述自己是什么。但自我意识或对自己本性的认识与意识是两码事。

3. 关于意识的几点初步看法

认识你自己，你就能认识人和神（苏格拉底）

意识是一个多义词，涵盖许多方面：自我意识、对世界的意识[9]、对某些社会或社会问题的意识、对我们人类状况的意识、对我们身体的意识，以及更专业的主观意识、现象意识、超越意识、概念意识或知觉意识。意识是人类境况的一个主要特征，因此我们将研究意识的本质，与 ChatGPT 的本质相呼应，从而更好地了解我们自己：通过更好地辨别意识的本质，我们也将更好地了解我们自己。

我对意识的研究方法是内省式的，并鼓励人们对自己进行冥想。因此，这种方法是主观的，而且充满了趣闻轶事，这与我们试图通过测量将其客观化和系统化的经典知识领域形成了鲜明对比。从认识论的角度来看[10]，我要向大家介绍的的确是主观经验，即与我们自身有关的经验，而不是通过测量设备获得的客观要素。由于我们之间的相似性，这种主观探索应该会引起你们的共鸣，而且在大多数情况

下，你们应该会同意我的分析。这难道不是客观性的开始吗？即使从绝对的角度来看，两个人的一致意见并不能证明一个事实或现实。事实上，这正是接近意识的难点之一，也可能是意识长期以来一直游离于科学领域之外、局限于心灵领域的原因之一。

---

[9]非局部意识、扩展意识

[10]认识论是对知识及其性质、起源和真理价值的研究。

精神性。事实上，我们如何才能将每一个正常人都会经历，但其最明显的表现却不会留下任何有形痕迹的现象严格客观化呢?

然而，尽管我们不能真诚地否认我们自身意识的存在，这使得它成为一种现象，其本质对我们来说是客观存在的，但我们绝大多数人类同胞却不一定对它有很好的概念或理解，这意味着意识现象是一个持续引起许多争论的话题。理解意识需要自知之明，而自知之明的获得需要时间。一定程度的冥想、反省和思考是不可避免的，尽管有很多方法可以加快这一过程。我还会和你讨论关于意识的客观说法，也就是可以测量的东西。 不过，正如你在本书结束时应该明白的那样，对意识的主观看法是不可避免的，这与意识的外部观察必然具有间接性有关。

在以认识论的客观方式研究意识方面，即试图将我们对这一现象的认识客观化方面，做得最多的科学家可能是神经科学家，但也有精神病学家，在某种程度上还有心理学家。一个多世纪以来，他们利用

功能性核磁共振成像等工具，以及光遗传学、脑磁图、脑电图、视觉幻觉和对脑损伤患者的研究等许多其他工具，系统地试图理解我们的大脑和我们的思想之间的关系。因此，神经科学家们试图找出我们心理活动的神经相关性，这促使他们根据自己的观察创立了不同的意识理论。其中最

这些理论包括*综合信息理论*[11]、*动态核心假说*[12]和*注意力计划理论*[13]。我的研究方法是相当综合的，在这方面，它无疑过于简短--当你想到让-保罗-萨特（Jean-Paul Sartre）的《存在与虚无》（Being and Nothingness）等气势恢宏的著作时，这一点似乎显而易见--无法为这一主题奠定广泛的知识基础，而这需要对这些理论逐一进行探讨。不过，在不完全放弃神经科学的贡献的前提下，我仍然坚持认为内省的方法可能就足够了，这不是为了理解其背后的大脑机制，而是为了获得对其现象的良好理解。

我个人曾多次阅读和聆听克里希那穆提的著作，他是一位心灵思想家，尤其擅长用 "被观察者 "和 "观察者 "这两个概念来探讨理解自我意识的困难。这是他经常重复的一个主题，在我们意识到我们是在思想领域中客观化自己之前，这显然是很难理解的。然而，由于思维是动态的，因此，观察自身的思维（即思维者）与被观察的思维（思维者可能是被观察的对象）之间产生的镜像效应，会造成一种眩晕和困惑的感觉，让人难以自拔[YT 或 DVD，《大卫-

博姆与吉杜-克里希那穆提的对话》]。这主要是因为我们的自我概念很复杂，既包括社会投射给我们的东西：我们的角色、我们社会地位的各个方面.....也包括我们可以称之为 "自我*本体*"*的东西*，即通过我们的身体来表现我们自己，但这并不是我们思想的真正行为者。这是我们需要摆脱的众多顽固幻想之一，以便

---

[11] 作者：朱利奥-托诺尼、克里斯托夫-科赫

[12] 杰拉尔德-埃德尔曼

[13] 迈克尔-格拉齐亚诺

意识是什么？我将在下一章再谈这个问题。

更直白地说，我们该如何定义意识？这是我在 2017 年的一次会议上提出的问题，当时我正在 NeuroSpin（位于巴黎萨克雷的法国原子能委员会脑成像创新研究中心）工作。在经历了一段时间的巨大困惑之后，我意识到，意识并不是一个与其对象分离的实体，当我们有意识时，我们必然意识到某些东西。现在回想起来，我对自己说，如果给意识下定义如此困难，那是因为它和时间或爱一样，是一种根本性的东西，无法用其他概念来定义。不过，罗伯特的定义非常好："*对自身心理活动的直接认识*"，我将其修改为：对某种事物的直接理解。这个东西实际上是我们心灵的一部分，无论是感知、情感、思想，还是概念或外部世界元素的表征。因此，在 2017 年，意识就是对某物的感知这一认识已经足以回答我的问题。

我现在想建议大家进行自我观察冥想。不要像通常建议的那样专注于呼吸，而是专注于脑海中出现的

下一个想法。这样，你就会发现自己的注意力是如何集中的。但是，除了冥想的目的之外，这种专注的根源是什么呢？如果你的情况和我一样好寂静应该很快就会到来；但它真的如此寂静吗？即使你不注意可能出现的念头，难道你不会察觉到一些潜在的东西吗？最后，不要让你的注意力被你的感觉所吸引，等待它们的出现。

当新想法出现时，不要执着于它。这不是一个简单的练习，你需要反复练习几次才能掌握。不过，如果你能像这样坚持 5 或 10 分钟，你就会明白我在说什么了。也许最好等到晚上睡觉时再做这种冥想.....但你现在可以闭上眼睛试试，因为视线足以分散注意力，使冥想变得困难。最后，请放纵一下：如果你的思绪飘忽不定，这也会很有趣，即使不是这个练习的主要目的，所以请试着回溯一下这个想法是如何产生的。话虽如此，我在这本书中只给你两个冥想的机会。对于这个练习，你可能要反复练习几遍才能掌握。我暂时不透露更多，这样你就能保留发现的乐趣，先入为主的想法扭曲了你的判断，你可以根据自己的经验更好地进行*事后*判断。

不过，我已经打算区分两个方面，在我看来这两个方面对于正确处理意识问题至关重要。一个是主观方面，另一个是可观察、可测量方面。因此，意识是一种可以通过多种方式观察到的现象：

– 通过人类个体的话语、

– 通过他们的创作和表现形式（绘画、旋律等）、

– 通过他们的行为、

– 通过各种科学仪器测量的神经元活动、

– 通过直接观察我们的思想、感觉和知觉，直接通过我们自己的主观能动性。

前四种观察方式对应的是可以客观测量的现象，而第五种观察方式可以说是主观的，外来观察者无法直接接触到，它有可以客观测量的一面。

内在的私密性：尽管主观经验可以由主体间接描述，但只有主体才能直接接触到它，任何工具都无法观察到他的意识内容。因此，在我们看来，意识的主要方面除了主观之外是无法观察到的。那么，我们可能会问，见证我们思想、知觉或感觉之流的观察者或事物是什么？

我再次强调这一点，因为这一特性是意识本质的特征：意识体验从根本上说是私人的，无法直接测量。我们将在关于主观意识的第 11 章中进一步阐述这一点。

4.认识、语言和学习

索绪尔提出的 "符号"/"被符号 "的衔接，对我来说一直都很重要，在我第一次听说它之后，几乎立刻就成为了一个核心的、反复出现的想法，一直伴随着我，并随着时间的推移而不断发展。它是语言的核心，而语言是由指称概念的词语组成的。

语言在人类有意识的思维中发挥着关键作用。通过语言，我们能够意识到我们所处的环境和我们自己，语言是一个非常强大的工具，它以一种与我们的意识高度交织的方式塑造和影响着我们的意识。通过传递我们的思想，它使我们能够展开思想，使我们能够理解思想，同时也能够交流思想。人类意识具有这种能力，其他非人类动物都没有发展到这种程度。不过，应该指出的是，猴子和狗能够理解相当广泛的词汇，实验证明，它们能够通过装有手势的按钮表达自己的想法[YT：Stella the dog learned to 'talk' and she will change the way you think about pets][YT：A conversation with Koko][14]。

事实上，我们是如此善于使用语言，以至于我们可以不费吹灰之力就能连续说上几个小时，听上几个小时，而我们实际上是在无意识中通过语言意识到的，因为我们通常意识不到我们在使用语言。当然，情况并非总是如此，当我们处理一个新课题，我们的理解力会被许多陌生的概念绊倒，就像我们小时候学习知识一样。

---

[14]一只会说手语的大猩猩。需要提醒的是，这些视频的链接还可以在以下网址找到： https://boissenin.net/manuel

这种能力。但我们几乎已经忘记了这个学习过程的一切，而这正是意识赋予我们的主要能力之一：*学习*。如果说学习母语在你的记忆中可能已经太久远了，那么通过孩子或在学习外语时，你可能又会遇到这种情况。一个相当常见和经典的学习例子就是驾驶汽车。在最初的几个小时里，例如在学习换挡时，需要高度集中注意力，并意识到自己在做什么。你要按照别人告诉你的方法，松开离合器，通过加档或降档来换档，然后松开离合器踏板。归根结底，这是一件相对简单的事情，因为我们是在无意识的情况下自动完成的。不过，前几次你必须详细了解每一个步骤：用左脚踩下离合器踏板，使用变速杆升档或降档，这就要求你不看变速杆就知道它在哪里，以及它的下一个位置在哪里。所有这些都需要一个语言过程来监督操作。这样，我们就可以清楚地看到，语言在这一学习过程中起催化作用，支持着这一活动的展开，直到后来成为无意识的活动，例如，使我们能够在开车时说话，并在不知不觉中进行这些活动。

再举一个学习舞蹈的例子：在这里，实际上是重新学习如何移动；首先是步法，以便与音乐节奏相结合，然后是传球，使情侣们能够和谐地表演一个动作。在学习舞蹈之前，您需要上很多小时的课，将练习和讲解结合起来，直到您掌握不同的动作顺序。学徒期的长短是决定舞者成功与否的重要因素，学徒期可长达数年。

这也是非常有趣的，因为我们能够意识到这一点，也就是让我们的学习更加自觉。经常听到舞者（通常是跳舞的人）谈到他们在最初几个月甚至几年的练习中感受到的认知负担。的确，当他们跳舞时，他们必须记住并选择一个动作来表演，以尽量避免一直保持相同的模式。但随着时间的推移，这个过程会变得不言自明，这取决于舞伴各自的位置，有时也取决于舞池中另一位舞者瞥见的动作的灵感。在某一阶段，某些舞蹈会产生神奇的效果，甚至会出现相对固定的频率。这时，意识会被引导到操作层面之外的另一个层面，意识已经从操作层面中释放出来，可以调整到感觉中去，或者就像驾驶汽车一样，可以一边跳舞一边说话。

实际上，对话也是如此，只是发生得太快，以至于我们往往没有意识到。媒体，无论是报刊、电视、广播还是互联网，以及我们自己的经验和想法，都会助长这种现象。除了后一种情况，这一过程通常是半有意识的，根植于我们的习惯之中。

最后，让我们再举一个例子，这个例子也需要一些时间才能掌握，因此也留下了一些记忆。那就是在键盘上写字。在智能手机普及的今天，我已经不清楚学习打字是否像过去那样费力，所以我记得我曾使用过专门的软件来提高我的书写速度，由于我还必须使用 qwerty 键盘，字母的变化相对较少，但许多符号的位置却大不相同，所以我不得不重新学习打字好几次。多少次

我甚至不知道如何打字，在打字之前，我必须先寻找符号和字母，以至于如果我问自己一个字母在哪里，我都不知道如何回答，但如果我必须写一个单词，我的手指就会写出来，而无需我思考，也不知道如何写。这种学习过程其实与语言并无直接联系，只是因为语言可以传递信息，而正是大量重复的搜索最终在潜意识中将其固定下来。

在机器学习中，我们谈论的是*强化*学习；尽管这涉及的概念略有不同，但这一表述却很好地描述了这类学习的特点。我们也可以说是通过重复和强化来学习，这可能与我们神经元层面的情况相对应。同样有趣的是，如果没有键盘对象的背景，我们就很难模拟在键盘上书写。也许显示器缺乏视觉反馈是原因之一；毕竟，我们有时仍会出现打字错误，而当错误出现并出现在屏幕上时，我们可以立即纠正。事实上，这种学习过程可以让我们通过类比更好地推断我们学习说话的方式，而且当我们学习说外语时，我们经常会搜索单词，这表明思维和语言之间的过程并不是即时的。通常，至少在开始的时候

，我们先用自己的语言思考，然后再翻译，但并不总是这样，这就产生了我们可以称之为前语言思维的东西，尽管思维通常是通过语言产生的，就像你读这些行文时的情况一样。当然，你解读的方式也取决于你的经验。这就是皮尔斯继索绪尔之后提出的重要观点：除了符号和被符号之外，我们还必须加上解释者，对于同一个符号，不同的解释者会引起不同的被符号。

因此，虽然意识能使我们学到的大部分知识
- 可以说，一种姿势的学习可以通过

补偿潜意识中的不适--语言可以成为许多此类学习过程的催化剂：一份菜谱就足以让你学会如何做一道菜，即使在实践中，你需要有烹饪的习惯才能尝试菜谱并做出这道菜。

尽管语言对于不理解它的人来说最初毫无意义，但其意义会在一生中通过不同的经历和使用实例而获得。语言使我们能够传达我们的经验，尽管并不完美，但它在塑造我们的思想以及我们理解和解释世界的方式方面发挥着重要作用。它也是我们思想发展的基本和主要载体，因此也是我们意识内容的一部分。作为集体进化的产物，这一载体将我们团结在一起，使我们能够在社会中，尤其是通过我们自己的叙述，进行交流、构建我们的学习和自我建设。很难强调语言有多么重要，因为它太普通、太平常了，以至于我们很少去思考它；有点像光，它无所不在，但大多数时候都是无意识的。然而，如果没有这种媒介，我们的意识内容及其本质将完全不同。

22.良知与*法律硕士*

"构思精巧，表达清晰" 博瓦洛

事实上，由于一个过程的某些方面从根本上说是不可交流的和私密的，因此，对于这个过程的有意识性质，无论是有机的还是基于计算机的，总会存在一些疑问。不过，我们可以学会进行测试，合理地消除对过程可能揭示的内容的怀疑。如果一个过程在所有方面都大大超出了我们的认知能力，但却不具备主观意识，那么我们是否应该承认，感知意识的概念，即通过感知世界和我们的某些内部状态来感受它们的能力，也许只是意识的部分而非根本属性？我们至少必须承认，没有知觉也会有意识，尽管知觉似乎普遍存在于动物之中，无论它们是有意识的还是非有意识的。

这种观点虽然有些魔幻[75]，但在我看来是合理的，它为意识提供了一个可操作的定义：*过滤我们周围世界的信息、解释这些信息以赋予其意义并最终整*

合和/或对其做出反应的能力。

能够给出意义意味着答案更加复杂，但不一定不自动。问题获得

---

[75]尼古拉斯-汉弗莱指出了第一条分界线，即冷血动物与温血动物之间的分界线。因此，青蛙以笛卡尔的方式

在这个实验中，爬行动物似乎是一种没有意识的自动机，或者说至少没有意识的基本方面，这一点可以从它强迫性的捕捉昆虫的条件反射中得到证明，尽管这些只是平板电脑屏幕上的虚拟表象。虽然在这个实验中并没有出现咀嚼和吞食昆虫的现象，但爬行动物却在毫无知觉的情况下坚持着它的原始姿态，就像一个陷入*死锁的*自动机，注定要一直保持它的姿态，直到环境状况发生变化。

归根结底，我们的概念意识难道不是一种自动机吗？我们自己不就是随机鹦鹉[76]吗？是否有什么东西阻止我们把我们的概念意识简化为一种能够自动对世界做出反应的高度灵活的反应过程现象？也许是认知能够创造和整合新的概念。但是，会话式人工智能也有能力创造新概念，哪怕只是通过组合，而且如果就目前的情况而言，它们无法整合这些概念，这似乎也不是一个技术上无法克服的挑战。

在进一步讨论之前，让我们从一个新的角度来看待意识的概念。考虑一下与一盏灯相连的存在探测器。如果说当你走近它时，灯亮了，它就会意识到我们的存在，这并不是不准确的。此外，这也假设 1：意识就是对（某物）的意识。然而，说探测器有意识似乎有些可疑，甚至荒谬。事实上，探测器没有自己的生命，它意识不到自己，也意识不到任何其他东西，只能根据它的技术，意识到由于我的身体存在而发出的红外线。对周围环境的这种最低程度的感知并没有给它带来意识。它只是一个*意识片段*，已经被自动化到系统中，当有人出现时就会自动

开灯。这个系统没有主观意识，因此我们认为它没有意识。但我要反驳的是，它确实有一种意识形式，尽管是一种最低限度的意识：意识*到*个人的存在。

---

[76]随机性：依赖于偶然性或偶然性的结果，这个词常用于原因不明或不可知的情况。

(77)比简单的红外传感器更复杂的技术和基于人工视觉的技术可以高精度地检测到这种存在。

让我们来看一个更复杂的例子。让我们看看在 1997 年的一次国际象棋比赛中击败加里-卡斯帕罗夫（Garry Kasparov）的深蓝程序[1997 : l'ordinateur bat Garry Kasparov, un tournant dans l'histoire des échecs, INA, 05.2022]。如果节目没有某种形式的比赛意识，即对比赛在特定时刻*的*位置、规则和目标的意识，它怎么可能击败当年世界上最优秀的棋手？正是他对比赛中可能出现的情况以及对他有利的情况的认识，使他得以获胜。当然，这种意识与卡斯帕罗夫完全不同，性质也完全不同。对 "深蓝 "来说，它的基本原理是计算出一定范围内所有可能性，并选择最有利的棋步，无论对手的反应如何。为此，"深蓝 "使用了 256 个处理器，这是一个相当可观的数字，使它能够每秒评估 2 亿步棋，而像卡斯帕罗夫这样的天才每秒只有 3.5 步棋。换句话说，这台机器的计算机意识和认知能力与世界冠军相差甚远，但正是这种棋局意识让这台机器能够与卡斯帕罗夫对弈，并多次击败他。虽然这种感知形式比存在探测器更为复杂，但也相对狭窄：机器只是对棋局形势有

一种感知，但它不知道对手是谁，处于什么状态，也不知道棋子的位置是否稍微超过了一格。

事实上，我们知道并理解深蓝计算的工作原理，也它的可能性很小，因为它

但成本较高，而且不一定有必要。

这就意味着，"深蓝 "的意识问题并没有引起人们太多的兴趣。意识一词似乎被误用了，因为我们暗指的是我们自己的意识和思维，在很小的程度上，我们的意识和思维可以提前预测几步棋，但只是在对局受限的某些情况下，而不是像机器那样系统地预测。可以用来弥补这一弱点的机制不胜枚举，也是许多书籍的主题，但这里需要记住的重要一点是，在这种情况下，人类意识及其处理能力与机器的不同程度。同样，没有人会认为 "深蓝 "拥有主观意识，因此，"深蓝 "也是有意识的。因此，我想再次说明，"深蓝 "确实没有主观意识，但它确实表现出了一种意识，根据假设 1，这种意识是棋局*的*意识，也是在它的表象中提交给它的所有棋局的意识。

那么对话型人工智能呢？其实也差不多：通过一系列复杂的算法计算，人工智能能够意识到请求的含义，并在尊重请求含义的基础上做出一致的回应。只不过，在这种情况下，我们所处的环境不再是狭窄的象棋游戏，也不再是试图探测某个存在的走廊。我们所处的是语言所能创造的与我们的现实和宇

宙相关的所有意义的超大语境。然而，在这里，似乎程序仍然可以不具备主观意识，即意识到或知道自己在说什么。这一点就不那么明显了，而且看似自相矛盾，因为会说话的人工智能似乎能够意识到它们所读的单词和句子的含义。这似乎是

否则，它们怎么可能造出意思与要求一致的句子呢？尽管如此，这一壮举几乎可以肯定是在 *LLM* 没有任何主观意识的情况下实现的，正如我们稍后将详细解释的那样：下一个单词提案是由一连串算法计算生成的，而算法计算的整体主观意识并不清楚自己在做什么。只有我们自己的意识和我们的连续询问，才能让我们看到会话式人工智能创造的句子与我们对世界的表述是一致的。如果程序能够在不知道自己在做什么的情况下自动做到这一点，那是因为它从训练文本中继承了足够复杂的世界表征，能够适应几乎任何情况，无论是新的还是熟悉的。例如，*LLMs*[78] 能够与用户就他们以前接触过的几乎所有主题进行全面互动[79]。如果你发明了一个，他们也会向你指出来。

同样，这也是一种必要的意识形式，不仅是*对*词语含义的意识，而且是*对*词语组合所产生的更抽象含义的意识。因此，对话式人工智能可以说是有意识的，尽管这种意识并不意味着主观意识。我之所以坚持这一点，是因为这其中存在着深刻的混淆：当

我们意识时，主观意识被隐含地引用了；然而，我试图为之辩护的论点是，在没有主观意识的情况下，甚至是在没有意识到该事物的主体的情况下，换句话说，在意识到（某种事物）的主体不存在的情况下，也可以存在某种形式的意识。如果

---

[78] 和对话式人工智能配置为该模式时

[79]仍然存在局限性，尤其是在数学等复杂推理方面。

在意识主体的意义上，存在主体，但存在过程的反应，即无意识系统的反应，正是这种反应见证了这种形式的意识[80]。

为了彻底澄清问题，仍有必要进行 "语言清理"，但我想说的是，我们通常理解的意识至少由两部分组成，一种是表象形式，另一种是实现其主观方面的对这种表象的感知形式。我们人类的意识一般不会脱离其主观方面，直觉的某些情况除外；还应该记住的是，对于概念意识或言语超验意识而言：言语通过支持它们的意识表象（应该指出的是，它们也会产生潜意识表象），意味着这种言语意识对主观意识产生影响，如果仅仅是通过它们的可见性或可听性，那么也会对自我产生影响 [(81)] 。尽管如此，就机器、系统和程序而言，通常并不存在主观意识，认知意识（cognitique）、良知意识（conscientiques）或非主观概念意识 [(82)] 等术语被用来指代这种没有主观意识的意识形式。这先验地代表了计算机科学中概念意识被剥夺了主观意识的所有例子；这在生物体内是完全前所未有的现象...

让我们补充一点，"*非主观概念意识*"这一术语虽然具有描述性的优点，但并不十分经济。然而，我们必须认识到这是一个过程、

---

[80]可以归因于这个过程。因此，这个过程既能意识*到*其现实的各个方面，又是无意识的。因为它的反应

在没有主观意识的情况下，我们会以一种复杂的方式自动适应这一现实，从而提供与我们的理解相一致的答案。

([81])后者下意识地

[82]或超越意识

一个系统、一种现象对某种事物--对某种存在、对棋局中棋子的位置所提供的可能性、对句子中的含义--有某种形式的感知，而且这种感知是可检测的，如果它能够产生一种与人类所期望的一致的反应的话。一般来说，当这种反应优于我们所能做出的反应时，我们就会完全接受它，至少使我们清楚地认识到，一定某种形式的感知和对情况的解释，并通过与我们的反应或反应相一致的方式表现出来。再一次，意识的主观特征是私密的，无法直接观察到的，这与行为主义者所采取的方法类似，他们试图通过证明动物能够意识到环境中的属性或属性的变化，并且能够根据行为主义者意识的主观尺度，以适当和一致的方式对其做出反应，从而证明动物是有意识的。事实上，他们可以用康德的词汇来强调这些动物的某些心智范畴，或者用现代语言来强调某些认知功能。

这一切让我想到了一个新名词："认知"。

定义：如果一个系统、过程或现象对其环境状态有内部表征，那么它就*能认知*环境状态。

这个词相当接近于意识，但也接近于与知识有关的认知一词。因此，在我看来，这个词用得很好，因为认知过程与更一般的心理过程相比，更具体地涉及知识的处理，而不是主观意识的表现。

根据这样的定义，任何伺服控制系统都可以被认为是*有认知能力的*，我的意思并不是说系统能够意识到它所测量的状态，从而在主观意识上感知到这些状态，它很可能没有任何主观意识；然而，我的意思是说这些状态是客观化的，是已知的，这使得它有可能考虑到这些状态，甚至操纵这些状态，并根据已确定的或未来的概念方案做出反应。

*认知*系统可以非常简单，只对超过阈值的感知进行等效处理，如带有电子恒温器的散热器；也可以具有复杂的语义处理能力，如对话式人工智能。请注意，后者的反应必然是不确定的，因为对于对话式人工智能提示来说，输入领域实际上是有限的，但实际上却是无限的[83]。这种不确定性是复杂系统（如混沌理论所涵盖的系统）的基本特征，并将它们与简单的宏观物理系统区分开来，后者被认为是确定和*可预测的*。对于量子爱好者来说，量子系统本质上是不确定的，这一点现已被广泛接受[84]。

对话式人工智能似乎被赋予了某种形

式的概念意识 [85] ，这就是为什么我认为它们可以被定��为
"自觉"，是意识和自动性的代名词，还是

---

[83]实际上，它无法在支架上实例化：有 50,000[(10)个]或 5.10^{40} 个可能的 10 字提示。不确定性是指在播放之前无法预测输出结果。

[84]就像原子的放射性衰变过程。绝对值
目前的 *LLM* 是确定性的，因为它们可以相同地重放提示答案。 但是，如果同一个问题被问了两次，只要它们包含了第一个问题的答案，那么第二个问题的答案就会不同。

[85]这也可以被描述为超验意识。

我用 "计算 "一词来指代这种意识，它满足了我的第一个假设：没有意识的对象，就不可能有意识。如果说 "意识 "一词可能会导致混淆，因为它可能意味着主观意识，那么说这些过程是 "*认知的*"（*cogniscient*）就可以消除这一困难，而且称它们为 "有意识的"（consciousious）并不意味着这些概念意识形式与主观意识相关联，而主观意识是我们意识形式本质的一个组成部分。

如果一个物体不存在于任何意识领域，就像我们世界中的大多数物体不存在于人类意识领域一样，那么它就不*具有意识*。*LLM* 使我们有可能
"从这个意义上说，他们的分析过程是*对其*含义的*认知*。从这个意义上说，他们的分析过程是对其含义的认知。稍后，我们将通过重温中式房间的思想实验，进而考察*变压器*架构的运行情况，来更详细地了解这一点。虽然 *LLMs* 具有*认知能力*，因而具有某种形式的意识，即对问题的状态和它们给出的答案具有内部表征，但它们并不具有主观意识，也不是有知觉的人。这就是为什么这些程序*的意识*与人类

意识有着本质区别，因为人类认知的很大一部分是基于主观意识的：当人类意识到一个主题时，他的脑海中就会出现*质*点，无论是通过片段的偶发记忆，还是通过语义记忆，语义记忆虽然难以捉摸，但也有主观意识的一面，就像我们提到的概念的*质点*一样。无论如何，由于人类概念意识的产物出现在主体的意识领域中，它们确实属于主体的主观意识，而主体不在场的对话式人工智能则不属于这种情况。尽管我们已经指出（我们会再回到这一点），主体可能就是过程本身，然而，就这些人工智能而言，分析

它们的功能倾向于消除这种可能性，表明它们具有某种形式的意识，*即*它们是有*认知的*，但却没有主观意识。

无论如何，*LLM 认知学的*棱镜为我们的*概念意识*，尤其是其无意识方面提供了一个有趣的视角，这些方面可以被描述为自动化，除非我们认为我们有独立于我们主要意识的意识袋，因此对我们的意识来说是无意识的.....

正是因为这些复杂性，我们才有必要使用特定的词汇，以便更简单、更准确地表达我们的想法。而这种词汇的分化，正是出于对语言模型所引发的这一新情况的理解愿望。因此，当我说 ChatGPT 和 *LLMs* 的后继者或许能够证明他们可以有意识而不被意识到时，我不得不澄清我的想法，因为这句话在语义上并不正确。显然，"意识 "一词的多义性是有问题的。但它确实有一个含义：从程序以理性的方式做出反应的角度来看，它是有意识的，这种反应与人类在面对请求或复杂问题时可能做出的反应没

有区别。另一方面，如果程序不是有知觉的，尽管它可以模拟情感，没有主观意识，并且能够像对话式人工智能一样做出超复杂的反应，具有适应性和创造性，还具有学习能力，那么我们可以说，它是有意识的，但却没有知觉。换句话说，未来的智人虽然没有知觉和主观意识，但可以像人类一样有意识。如果说在 ChatGPT 出现之前，这似乎是难以想象的，甚至是不可想象的，那么现在看来，这似乎是完全可以实现的，而想象一下 ChatGPT 的后继者可能会实现这一点，也只是很短的一步而已。

人们可能会认为学习能力与主观意识有关，但这一点远非显而易见，并提出了以下问题：

无意识意识能走多远？

使用新词汇，我们还可以提出以下问题：

*认知系统*能否取代人类完成最复杂的任务？

认知科学能否导致与有意识的存在相等的行为？

无论如何，对话式人工智能的 FTE（心智理论能力）似乎并不排除这种可能性，因为这些系统有能力考虑与之互动的个体 [86] 的状态。还应注意的是，这些人工智能也可以*认知*自己的本质，但这并不意味着它们具有主观意识。例如，ChatGPT 声称自己是一个没有主观意识的语言模型，这似乎令人惊讶，但根据本书后面的分析，这却是事实。

最后，我们可以完善为意识提出的假设 1。当触发恒温器时，恒温器会*认知*到一个温度，低于这个温度

就必须开启暖气。在这种情况下，这一信息可以被视为*意识的*一个*片段，或者说是*跨越温度阈值*的一个认知元素*。正如人类能够

---

[86] 他们的一些隐含心理状态

如果他看到冰已经形成，他就会推断温度已经低于零度。虽然这一认识并没有解开主观意识和冷热感觉之谜，但它确实使我们能够理解，我们是如何通过概念意识来估计温度的。更广泛地说，从自觉性的角度来看，这可以让我们评估它们的复杂程度。

建议：可以通过考虑得出答案所使用的 "意识片段 "的数量来评估意识的认知程度。

虽然，根据这样的建议，我们最终会发现有些自觉性比其他人的自觉性程度低，但根据他们所使用的抽象概念和对这些抽象概念的安排，他们的表现却比其他人好。不过，我们也许可以证明，要完成某些任务或使自觉性保持足够的普遍性，我们不能低于一定程度的概念意识（*的*）。

假定会话式人工智能没有主观意识，但具有某种形式的概念意识，那么 "*有良知* "一词似乎非常适合它们。然而，这一术语的使用可能仅限于最复杂的认知者，如当前的会话式人工智能，它们表现出了复杂的概念敏捷性，即非复古的或具有足够高通用性

的人工智能；这一界限必然是模糊的。