意识计算模型的哲学评估

A philosophical assessment of computational models of consciousness

Elif Gok[1], Erdinc Sayan[1]

[1]*Informatics Institute, METU, I_ no¨nu¨Bulvarı, 06800 Ankara, Turkey*

[2]*Department of Philosophy, METU, I_ no¨nu¨Bulvarı, 06800 Ankara, Turkey*

accepted : 19 October 2011 by Cognitive Systems Research

(translated by zang jie)

摘要：最近在意识研究方面出现了一系列活动。尽管意识的操作性定义尚未形成，但哲学已经开始确定一组被认为与意识的各种元素相关联的特征和方面。另一方面，最近有几次尝试开发意识的计算模型，声称可以捕捉或说明意识的一个或多个方面。作为评估当前计算模型对意识的良好程度的合理替代，本研究探讨了当前计算模型如何在建模哲学所识别的意识的这些方面和特征方面发挥作用。在回顾了意识哲学的文献之后，本研究构建了一个特征和方面的列表，这些特征和方面在任何成功的意识模型中都是可以预期的。然后，该研究从该列表的角度评估了一些当前自称和实施的意识计算模型。所研究的计算模型根据意识的每个已识别方面和特征进行评估。

关键词：意识；计算认知建模；Clarion；LIDA；ACT-R；神经工作空间模型；ART; GMU-BICA

1、引言

最近在意识研究方面出现了一系列活动。就哲学理解和可用于工程应用的实用结果而言，意识本质上是一门困难的学科。尽管目前还没有研究普遍接受的意识的框架，但哲学已经开始确定一组被认为与意识的各种元素相关的特征和方面。另一方面，最近有几次尝试开发计算模型，声称可以捕捉或说明意识的一个或多个方面。作为一个非常复杂的问题，有许多关于意识的不同观点。这项研究以哲学当前可获得的一些主要观点为出发点。我们首先研究意识哲学的文献中可以找到的意识的各种特征和方面。然后，我们检查了一些自称的意识计算机模型，并根据它们对哲学家指出的意识的各种特征和方面的适应和解释程度来评估这些模型。此外，我们将我们的研究限制在那些已经实施并研究其行为的模型上。此次评估和调查的完整结果不仅根据模型的熟练程度对模型进行排名，而且还提供了关于当前认知科学在对意识的科学理解方面的成功程度的一般线索。在这方面，本研究结合哲学和计算机科学，回顾了这两个领域意识研究的最新工作。

在从哲学的角度评估意识的计算模型时，本研究希望实现三件事。首先，它识别并系统地整理了最近出现在心灵哲学文献中的意识的那些方面和特征。当我们将所有特征和方面结合起来时，必须解决方法中的一些不一致问题。由此产生的哲学特征和意识方面的折衷和综合清单是这项工作的产物。其次，它检查了当前自称的意识计算模型集与我们的列表相比。这项检查揭示了计算模型在多大程度上满足了该列表，因此也揭示了哲学方面的问题。可以说，任何成功的模型至少必须满足我们的折衷列表才能被哲学界接受。第三，这项工作提供了一个可扩展的组织和结构框架，可以帮助指导未来意识跨学科领域的工作。由于来自不同的概念背景，一个统一的框架可以在调解构成跨学科研究领域组成部分的研究人员的不同观点方面起到促进或精神上的作用。以现象意识为例，一些哲学家认为，至少在目前的知识状态下，用科学方法无法实现对现象意识的解释。一些计算模型确实尝试处理现象意识。如果在未来的研究中，计算模型对现象意识有更高的科学理解，那么这种发现对于哲学家在修改他们对意识问题的理解时将是至关重要的。这种修正甚至可能产生连锁效应，由此哲学的其他派生概念最终将从计算模型中受益。然而，所有这一切都要求将跨学科方法嵌入到所有各方都理解和使用的有凝聚力的操作框架中。

应该指出的是，这项研究有一个特别的弱点。我们对计算机模型的评估完全基于有关模型的文献。也就是说，我们通过专门研究模型的文献来研究特定模型的机制。我们没有对模型进行任何动手操作。尽管这是该研究的一个特别弱点，但出于某些实际原因需要这样做。首先，无法获得某些模型的源代码，例如为美国海军开发的 IDA。此外，一个人需要精通计算机模型的语言和环境，以便在

足以全面评估模型的水平上进行动手工作。否则，建模特定任务可能失败的原因将是有争议的，关于模型本身是否无法完成任务，或者用户没有足够的知识来执行任务。当考虑到这两个问题时，我们认为本研究将自身限制在文献中的描述和声明中，而不是用实际工作来补充它是有道理的。

2、意识元素

在我们展示意识特征的综合和折衷列表之前，我们首先检查来自几位研究人员的各种此类列表。然而，这些先前列表中的大多数特征都直接引用了意识文学哲学，而没有进一步解释。对这些列表的全面理解需要对意识哲学的当前艺术状态进行简要回顾。

身心问题是一个古老的问题，在哲学中占有重要地位。然而，意识可以被视为一个相对较新的概念，因为它在当前的哲学和认知科学文献中被使用。 “意识”一词的使用可以追溯到笛卡尔。他用这个词来指代学科的内在知识。笛卡尔 (1973, p. 222) 也将思想定义为“我们有意识地在我们体内运作的所有事物”。应该指出的是，从笛卡尔直到最近，意识被视为精神的基本特征。也就是说，人们认为不存在无意识的心理状态之类的东西。然而，作为一个例外，莱布尼茨 (1989, pp. 295–299) 区分了他所谓的“小知觉”和“知觉”。小知觉是主体不知道的知觉。小知觉的结合导致统觉。统觉可以被看作是伴随着感知觉知的感知。然而，在弗洛伊德之前，艺术状态将意识等同于精神状态的整体。弗洛伊德可以被认为是第一个将无意识心理状态的复杂框架概念化的人。 2

到 20 世纪初，心理学领域已经见证了行为主义的兴起。根据 Baars (1986) 的说法，行为主义是一种心理学元理论，每个元理论“都指定了一个心理学领域、一组调查该领域的技术，以及一个将这些发现整合到人类知识和练习”（第 5 页）。行为主义拒绝将内省用作心理学方法论的一部分，并提出心理学唯一适当的领域是可观察的人类行为。因此，随着行为主义的兴起，不仅是意识，而且任何关于心理状态隐藏本质的调查都被排除在科学之外。然而，随着 20 世纪中叶发生的认知转向，许多心理过程又在心理学中占据了一席之地。认知心理学家声称，“心理学家观察行为是为了对可以解释行为的潜在因素做出推断”（Baars，1986，第 7 页）。然而，直到最近，意识仍然被避免作为科学调查的主题。意识研究在 1980 年代开始流行，当时对无意识过程的实证研究结果开始积累。这种积累让位于将意识视为一个变量 3 并对其进行实证研究（McGovern & Baars，2007）。

相应地，哲学领域也兴起了意识研究。除了一些哲学理论对心理状态如何变成有意识提出了一些解释外，还有大量文献致力于对这个问题的概念进行澄清。除了试图在整个现实中定位心灵和意识位置的一般哲学理论（作为解决身心问题

的项目的一部分）之外，在过去的几十年里，还提出了，更具体地关注意识现象及其各个方面的理论。前一类理论，即一般形而上学性质的理论，可以分为两大类：物理主义理论和非物理主义（主要是二元论）理论。二元论的一个经典例子是笛卡尔的实体二元论。虽然非物理学家理论的悠久历史至少可以追溯到柏拉图，但近来解决身心问题的方法主要是物理方法论。后一种理论，即特别关注意识现象的当代理论，往往明显是物理主义的。这些理论可以分为几类。 “高阶理论”声称，当我们的体验被与这些体验相关的心理状态所伴随时，意识就会出现。如果关于低阶状态的高阶状态可以被称为思想，那么该帐户就是“高阶思想（HOT）理论”（例如 Rosenthal，1997）；但是，如果将所讨论的高阶状态视为类似于感知，那么将其标记为“高阶感知 (HOP) 理论”更为恰当（例如 Armstrong，1997；Lycan，1997）。相比之下，表征理论坚持认为，一种精神状态的有意识只包括该精神状态的全部表征内容（例如，Tye，2000）。市场上还有一些理论可以根据“全球工作空间”的概念（Baars，1988）（我们将在后面结合 LIDA 更详细地讨论这个理论）或“综合信息理论”来解释意识。 ”（托尼尼，2004 年）。还有更多奇特的描述诉诸量子现象，据称这些现象发生在称为微管的神经元内部的某些结构中，以阐明意识的奥秘（哈梅洛夫和彭罗斯，1995）。

意识理论都试图解决他们认为是意识问题或其各个方面的问题。然而，是否真的存在“意识问题”并不清楚，因为我们在谈论意识时似乎并不总是谈论相同的概念。正如布洛克所说：“意识的概念是一个杂种或更好的杂种概念：“意识”一词代表着许多不同的概念，并代表着许多不同的现象”（Block，1997，p。 376)。

此外，也有人认为，由于其主观性，自然界中没有意识的类比。关于意识的任何可能的科学解释的局限性的最著名的论点之一是 Nagel (1997) 在他的著名文章“成为蝙蝠是什么感觉？”中提出的论点。在这篇文章中，内格尔说“一个有机体具有意识体验的事实，基本上意味着，它就像是那个有机体”（1997，第 519 页）。因此，对意识的任何解释也必须能够解释“它是什么样的”这样一个有意识的有机体。然而，如果人们试图理解这一方面，那么人们唯一能做的就是想象我们像这样一个有机体一样行事是什么感觉。也就是说，人们总是会错过关键部分，即有机体本身的观点。毕竟，有意识的体验意味着拥有主观的观点。正是这一点将意识置于与物理学所有主题不同的背景中。物理学客观地处理它的主题——至少它应该这样做。因此，正如 Nagel (1997, p. 524) 所说，“如果我们承认心理的物理理论必须说明经验的主观特征，我们必须承认，目前没有可用的概念给我们提供如何做到这一点的线索。 。”因此，主观经验，即意识，似乎

无法在任何物理理论中得到解释，物理理论本质上回避为主观性腾出空间，而是试图采取客观的世界观。

类似的论点是杰克逊提出的“知识论据”。在这个思想实验中，人们被邀请考虑玛丽的情况（杰克逊，1997 年，第 567-570 页）。玛丽是一位才华横溢的科学家，她在黑白房间中长大。但是，她知道物理学是万事通的，这是她从黑白电视机和书本中学到的。问题是玛丽是否知道所有关于颜色感觉的知识。杰克逊争辩说她不能。他考虑如果玛丽从她的房间里出来并且第一次看到红色会发生什么。杰克逊的论点是，在那一刻玛丽学到了一些新东西：她学到了红色的红色或看到红色是什么感觉。

莱文（Levine，1997）接受玛丽通过体验红色来学习新东西。然而，她所学到的不一定是非物理的事实。相反，这是一个无法用科学命题来解释的事实。因此，他宣布在科学范围内提出的物理理论的解释能力与有意识的经验中呈现的内容之间存在“解释性差距”，例如红色感觉的主观特征。颜色。

2.1、意识特征列表

我们每个人都对意识有所了解。我们对此有一些直觉。在许多其他领域，人们可以在不考虑我们的直觉的情况下继续理论化。然而，就意识而言，似乎我们都想要一个理论来解释我们自己的直觉。这并不是说，一旦形成了正确的意识理论，我们所有的直觉都会变成科学真理——它们可能是幻觉。然而，任何这样的理论至少应该让我们了解这些错觉背后的机制。

鉴于上述考虑，很明显我们没有意识的操作定义。甚至这种定义的可能性也是值得怀疑的。在这种情况下，从概念分析开始以试图在这个问题上站稳脚跟似乎是合理的。

有一些哲学研究试图确定意识的各种特征和方面。 Block (1997) 对意识的特征和方面的早期分析之一。他区分了四种意识：现象意识、通达意识、监视意识和自我意识。布洛克认为他的分析部分是为了规范我们关于意识的概念和术语。也就是说，他所识别的四种意识不一定是存在的四种不同的意识。结果可能是，当意识研究进一步发展时，这些类型中的一些甚至所有类型都被简化为单一形式。但是，考虑当前的技术水平，分类至少在概念上似乎既必要又有用。从某种意义上说，如果任何一种意识理论都声称要对该问题进行详尽的解释，那么就必须考虑所有这四种类型或简化形式。此外，如果我们不区分不同种类的意识，我们就有可能最终得到一个只能解释一种的理论。

Van Gulick (1995) 提出了意识的六个特征，如果我们想了解意识就必须解释这些特征。他主要关心的是对意识研究主题的澄清。他指出，在我们开始讨论科学地研究意识的可能性之前，我们必须清楚从这样的研究中可以期待什么。因此，

他提出的清单可以被视为由意识的特征组成，我们期望从意识的科学研究中得到解释。

对意识的科学研究必须解释的第一件事是有意识、无意识和无意识心理状态之间的区别。另一个区别是有意识和无意识生物之间的区别。生物意识可能被视为赋予某些生物的一种属性。4 当人们说变形虫可能根本没有意识时，人们谈论的是生物意识。根据范古利克的说法，我们能够直接使用其内容是意识的另一个特征。也就是说，我们对有意识的精神状态的内容有直接的了解。正如 Van Gulick (1995, p. 65) 所说，有意识的心理状态“对处于状态的人或生物具有意义或内容。”列表中的最后三个特征——即感受性、现象结构和主观性——有一个共同点。 Van Gulick 指出，这三个特征通常被称为意识的现象方面。但是，我们最好将这三者区分开来，因为每一个似乎都带有其他人没有涵盖的特殊本质。

Lycan (1999) 指出了意识文献中的混乱。像布洛克一样，他指出有些哲学家和科学家似乎混淆了意识的不同问题。 Lycan 指出，区分这些不同的问题不仅可以防止混淆，还可以帮助我们进一步推进意识研究。 Lycan 识别的不同意识问题对应于不同的意识特征。

Lycan 提到的第一个问题是有意识和无意识状态的区别。其次，人们似乎通过内省了解自己的意识体验的内容。任何其他人都不容易获得这种内省的知识。另一个问题与感受质的概念有关。根据 Lycan 的说法，感受质的概念在必须单独研究的问题中占有一席之地。意识体验具有平滑性和连续性，Lycan 称之为同质性，这似乎是外部物理世界所缺乏的。尽管物理材料的特性具有离散性，但我们对这些特性的体验是平滑和连续的。还必须解释这种同质性。还有一个特征是意识体验的内在视角，即意识体验的第一人称视角。这种观点是 Nagel (1997) 强调的主观观点，是意识研究的主要障碍。

其他三个问题是“有趣的事实”，“它是什么样子”的不可描述性，以及 Lycan 所说的解释性差距。他指出，知识论证 (Jackson, 1997) 表明可能存在一些非物理的事实。这些被 Lycan 称为“有趣的事实”的事实需要一个解释。人们似乎无法解释拥有特定意识体验是什么感觉，这是另一个问题。最后，Lycan 指出了 Levine (1997) 提出的“解释差距”问题。 Lycan 指出，必须以某种方式解释意识体验的主观属性与物理（即神经学）事实之间的联系。

Table 1
Features of consciousness.

	Feature or aspect	Remarks
1	Difference	Difference between conscious and unconsciousness mental states
2	Availability	The content of conscious mental states is available to (can be used by) the whole system
3	Explicit and Direct Knowledge	Explicit and direct knowledge of the contents of conscious mental states (“monitoring consciousness”)
4	Qualia	Raw feels or sense data that make up the “qualitative features” of an experience
5	Phenomenal Connectedness/Unity	Different modalities of perception are united in a single experience
6	Subjectivity/Subjective Point of View	All conscious experiences are from the viewpoint of the subject; they belong to a single subject

2.2、意识特征的综合清单

下面的表 1 中列出了我们的列表，这些列表是根据上一节中检查的三位作者的列表得出的。在我们提到的不同列表中识别的大多数意识特征彼此相似。因此，应该将出现在不同列表中的相似特征作为我们列表中的单个条目包含在内。其次，我们省略了一些作者给出的一些特征。

列表中项目的顺序并不表示任何层次结构。然而，第一项，即区别（有意识和无意识的心理状态）对于任何意识理论都有特殊的地位，我们稍后会看到。此外，最后三项，

即 Qualia、现象连通性和主观性，与现象意识共享相关性。因此，它们在表 1 中显示为连续的项目。除此之外，这些项目的顺序没有其他含义。我们列表中的第一个元素涉及有意识和无意识心理状态之间的差异。也就是说，意识理论必须解释一些心理状态如何变得有意识而另一些则没有。该项目与所有其他项目有所不同，因为它是要解释的最明显的方面。事实上，这第一个要素可能是建立任何试图解释意识的综合理论的良好起点。这个项目出现在我们上面考虑的两个列表中，即 Van Gulick 和 Lycan 的列表。在 Block 的列表中找不到它。然而，应该指出的是，Block 试图将不同种类的意识隔离开来，这些意识不应被混为一谈。因此，他的分析可以被认为是在承认有意识和无意识心理状态之间的差异之后的后续步骤。

可用性是我们列表中的第二项。可用性类似于 Block 所谓的访问意识，即意识状态的内容必须可供系统使用。出人意料的是，这一项只有 Block 有明确说明。 Van Gulick 和 Lycan 分别有语义透明和内省的概念。乍一看，这些似乎是可用性概念的替代品。然而，这两个概念都强调对意识状态内容的了解。因此，这个特征可以用来对应 Block 区分中的监控意识。

除了有意识的心理状态的内容的可用性之外，有意识的人也有关于该内容的明确知识。我们列表中的第三个要素承认对有意识的心理状态的内容有明确的了解。也就是说，我们知道我们处于具有特定内容的特定状态。这可以看作是 Block 定义监控意识的一种高阶状态。类似地，狼人所讨论的内省，可以被看作是一个更高阶的过程，它产生了对意识心理状态内容的明确知识。

我们在我们的列表中包括了作为意识特征的感受质的概念。很明显，我们在有意识的体验中似乎有一些主观的原始感觉。因此，即使一个品质最终可能不是一个真实的实体，我们仍然必须解释主观感觉与意识体验的这种联系。

Lycan 中的术语同质性和 Van Gulick 中的现象结构可以被视为或多或少表明意识体验的相同属性，即我们的意识体验似乎表现出的统一性。因此，这种统一性是意识理论应该考虑的另一个特征。这构成了我们列表的第五个元素。

主观性或主观观点是我们列表中的最后一项。正如 Lycan 和 Van Gulick 所言，有意识的体验对主体具有重要的观点（这两位作者分别将其称为主观性和内在的透视性）。

这里值得注意的是，我们并不假定意识是一个没有度数或等级的实体。如果人们认为意识有度数或级别，我们可以通过参考以下事实来轻松解释这一点：至少我们列表中的某些项目，例如差异和可用性，也允许度或级别。

迄今为止，我们在意识文学哲学的帮助下确定的意识的六个特征构成了我们的汇编清单。表 1 列出了条目的摘要版本及其含义。

将会注意到，我们从当前文献中三位著名作者给出的列表中省略了一些项目。严格来说，我们并没有忽略在 Block 的清单上的现象意识。相反，我们遵循 Van Gulick 将其划分为组成部分。也就是说，我们认为 Qualia、统一和主观观点的概念可以作为 Block 所确定的现象意识的基本组成部分。我们遗漏的原因如下。

没有出现在我们列表中的第一项是自我意识。主观性肯定与自我概念有关。但是，Block 定义的自我意识与主观性不同。正如 Block 的列表中所示，自我意识要求其拥有者具有自我概念并能够使用该概念。我们认为拥有自我的概念与拥有意识是一个不同的问题，尽管可能是一个密切相关的问题。显然，要解释自我意识，我们至少需要了解自我的概念和意识的概念。如果没有伴随的自我概念理论，单独的意识理论是不足以完全解释自我意识的。

其次，我们还省略了 Van Gulick 列表中的生物意识。 Van Gulick 指出，期望动物具有某种精神状态可能是不现实的，但我们可能仍想称这种动物为有意识的。因此，根据 Van Gulick 的说法，提出只有当一个生物能够拥有这些精神状态时才具有意识的提议可能将标准设置得太高了。因此，他反对用状态意识来解释生物意识的想法，并将生物意识作为一个单独的项目列入他的清单（Van Gulick，1995）。然而，如果“完整的意识理论”被发展出来，那么我们只能接受那个理论的标准。我们认为，从生物具有意识状态的角度来看，生物意识应该可以通过这样的理论来解释。鉴于所有这些考虑，没有必要将生物意识视为独立于状态意识的单独类别。

此外，Lycan 列表中的项目有趣的事实、“它是什么样子”的不可解释性以及解释性差距不在我们的列表中。我们不认为这些可以被认为是必须分开处理的意识特征。这些概念几乎完全基于杰克逊、内格尔和莱文的论点。而这些论点是基于品质的概念和有意识经验的主观性。因此，如果我们对感受性和主观性概念的理解有所提高，我们也将能够解释有趣的事实、“它是什么样子”的不可解释

性以及解释性差距，这似乎是合理的。因此，我们认为没有必要将这三个作为单独的项目添加到我们的列表中，在概念上与 Qualia 和主观性的概念不同。

3、意识的计算机模型

3.1、什么是模型

我们可以谈论正在风洞中测试的新飞机设计的十分之一比例模型和一组计算流体动力学方程，它被称为飞机的空气动力学模型。这两种模型都用于测试和评估新设计。当我们谈到气候模型、经济模型或人口动态模型时，有一个隐含的概念，即这些模型与科学理论有关。在这里，我们不会处理“模型”一词的这种不同用法。但是，我们必须解决术语“模型”和“模型”与本研究所关注的问题的不同解释。

有许多类型的模型可以被认为是科学的。数学模型和标志性模型，可能是物理的或计算的，显然有一些相似之处，因为模型的元素和被建模的现象的各个方面之间存在一对一的对应关系。模型不仅可以用来表示对象的特征，还可以用来表示对象的功能。桥梁的建筑模型是表示实际对象特征的模型的示例。另一方面，仿真模型是一个示例，其目的是对对象的某些功能进行建模，而不仅仅是对它的静态元素进行建模。

在意识的计算机模型中，显然我们对功能和行为感兴趣，而不是物理方面或特征。即使我们可以识别意识的某些元素，但人类的思维如何运作和实现意识却绝不是显而易见的。关于意识以及人类神经系统如何实现这些机制有各种理论，但作为一个新领域，到目前为止，几乎没有什么是无可争议的。因此，当今市场上的许多计算机意识模型被用作调查工具，以帮助我们获得可能导致意识理论的见解，而不是机械地实施完善且普遍接受的意识理论。

在我们继续之前，让我们重新审视模型-理论的相互关系。哲学通常承认三种类型的模型-理论相互关系。它们是接收视图、语义视图和自主视图 (Da Costa & French, 2000)。接收视图认为模型通常是对理论的简化说明。这种简化对于解释理论或说明其各个方面可能很有价值。因此，接受的观点预设了现象的理论存在，并且在这种观点上，理论-模型关系是单向的。该模型对理论没有贡献；它只是说明了它。

语义观点对理论的定义略有不同。它表明理论提供了组织我们对给定现象的知识的结构。该理论提供的结构由实现和表示该结构的模型具体化。例如，一个给定的知识结构可以被组织成一组方程，人们可以将其称为描述该理论的数学模型。例如，牛顿的万有引力理论可以用他著名的公式来概括。该理论同样可以通过口头描述或通过计算机算法来描述，该算法在给定相关输入时预测重力。模型

与管理它们的理论之间的关系的这种概念也被解释为理论只是可以从假定的结构化知识中实现的所有可能模型的集合。

最后一个视图，通常被称为自治模型视图，在很大程度上是基于计算模型的近期工作的结果。在生物学和经济学等缺乏既定理论模型的领域，科学探究试图构建计算模型，以深入了解潜在的复杂且通常是动态系统的行为。可以将自主模型视为初步试验，未来可能会出现成功的理论。从某种意义上说，当我们建立一个自主计算模型时，我们正在制定与之平行的试探性理论。这种方法在研究复杂的动力系统方面似乎是成功的，而这些系统迄今为止已经超出了更传统的方法形式化的范围。

我们看到意识的计算模型最好被视为自主观点的例子。1960 年代见证了人工智能研究和认知科学活动的兴起。 Newell 和 Simon (1976) 提出了物理符号系统假设。这个假设指出，拥有一个物理符号系统对于通用智能来说是必要和足够的。这个假设的意义在于它表明制造智能机器是可能的。手头的任务是弄清楚这组符号和这台机器的结构究竟是如何构建的。

基本上有两种思想流派表现为两种不同的主要方法（Cooper & Fox, 2002）。第一派，主要被称为“传统符号建模”，认为人工智能和认知可以通过一组对一组符号进行操作的规则来实现。第二种思想流派基于人工神经网络，有时也称为“联结主义模型”。人工神经网络的灵感来自神经系统的结构，其中传导电信号的神经元之间存在许多互连，表现出复杂的模式。该学派的支持者指出，由于大脑似乎是由一组相互连接的神经元构建的，因此基于相同结构构建认知理论是合理的。

两种思想流派各有长处和短处。虽然基于符号规则的系统很容易编写和理解，但使用这种方法实现学习变得相当困难。另一方面，神经网络能够结合学习。然而，由于它们的输入-输出关系依赖于分布在大量神经元上的一组权重，因此其确切的功能机制对于随意检查变得有些模糊。因此，神经网络是一种更全面的工具，而基于符号规则的方法往往更接近于更习惯的形式系统。除了纯符号方法和连接主义方法外，还有一些混合模型试图将两者的更好方面结合起来。

尽管人们对建立思维机器的渴望推动了认知的计算机模型的发展，但各种研究在目的和动机方面存在差异。一些研究将构建人工智能机器作为目标，而另一些研究则旨在对智能进行建模，以了解人类认知的潜在机制。

另一种对计算机认知模型进行分类的方法取决于这些模型是要用作实用的最终产品还是用作科学探究的学术工具。存在实际的实现，例如 IDA 和 LIDA，它们承担了有用的责任。另一方面，诸如 Clarion 之类的以科学为导向的模型主

要用作调查工具来增强我们对认知的理解。当然，实用模型也可以用于科学探究，同样，科学模型的版本可以构成未来实用产品的基础。

在认知科学中，人们考虑认知能力的不同领域。例如，视觉、记忆、决策等能力可以被视为不同的领域。计算机模型的另一种分类可以根据领域特异性来实现。一些模型提出了可能被定制以处理任何领域的架构或框架。其他的则是特定于领域或任务的。例如，它们可能特定于关注的领域。 Newell (1990) 提出一个论点，如果我们要理解人类认知，我们必须关注与领域无关的架构模型，因为这些模型更有能力传达关于认知的一般概念的信息。另一方面，如果特定于任务的模型在结构和行为上收敛，它们作为认知结构将具有重要意义。

3.2、意识的计算机模型

本节回顾了一些在文献中出现的、自称是意识的计算机模型。 （我们在脚注 1 中解释了此处“自称”的含义。）我们将审查限制在那些已经实施并检查其行为的模型上。

3.2.1. Clarion

Clarion（“Connectionist Learning with Adaptive Rule Induction ON-Line”）（Sun, 2003, 2006）是一种混合架构，它同时包含了连接主义和基于规则的元素。Clarion 识别隐性和显性知识。它使用分布式网络来表示子符号隐式知识，并且以符号方式使用本地网络来表示显式知识。这导致了一个两层架构，其中顶层或显性知识层和底层或隐性知识层由不同的模块建模。作为一个灵活且可扩展的框架，Clarion 采用了多个子系统，例如以行动为中心、非以行动为中心、动机和元认知子系统。每个子系统包含顶层和底层模块，分别处理显性和隐性知识。

Clarion 通过使用规则和块（维度/值对）实现显性知识，跟踪“世界状态”，实现工作记忆，5 和目标，所有这些都使用维度/值对。规则使用块并扫描世界状态、工作记忆和目标中存在的维度/值对。如果规则的条件部分中的所有维度/值对都存在，则触发该规则。结果，规则动作部分中的维度/值对被引入到工作内存中。在顶层，输出维度/值对是通过此规则机制从输入维度/值对生成的。底层还从输入维度/值对生成输出维度/值对。但是，这里的机制更加隐含。使用反向传播网络而不是规则。由于顶层和底层的知识都是用维度/值对来表示的，所以这两层可以自由地交换所表示的知识。底层使用 Q-learning。顶层有许多替代的学习机制。其中一些是特定于给定子系统的。学习机制包括自上而下的同化，模仿学习，规则提取和独立规则学习。

考虑到我们列表中包含的有意识和无意识过程之间的区别，Clarion 的两层架构使用较低层的神经网络来对无意识表示和过程进行建模，而较高的基于规则

的层对应于有意识的表征和过程。在这方面，Clarion 为两者建模并清楚地描绘了它们。

我们列表中的下一项，可用性，在 Clarion 中不存在。 Sun（1999）明确提到访问意识并将其定义为“访问的心理内容的直接可用性”（第 534 页）。他还提出，有意识内容的表征类型的不同，

即显式表示足以说明直接可用性。尽管信息可以在子系统之间共享，但并没有比任何其他信息更能提升顶层意识过程。低层信息与高层信息一样多地共享，尽管是间接的。然而，间接性的本质并不十分清楚，因为所有子系统都有一个两级结构，其中底层的流程和表示可以影响输入和输出流程，而无需与顶层进行任何交互。因此，我们必须得出结论，Clarion 并未完全解决可用性问题。

正如 Sun (1999) 所说，有意识的表征是由局部网络建模的，以捕捉它们的显性本质。此外，还有一些元级过程作为推理机制，利用顶层的显式表示。因此，号角完成了我们列表中的第三项。

我们列表中的最后三个项目是 Qualia，经验的关联性和主观性，所有这些都与现象意识有关。正如前面所解释的，现象意识在文献中以不同的方式被解释并讨论了它的不同方面。这就是为什么我们明确地将 Qualia、经验的连通性和主观性列为单独的项目，尽管其他研究人员使用“现象意识”一词来粗略地指代其中任何一个。

Clarion 声称要解决现象意识。更具体地说，Clarion 指出，由于上层和下层过程之间存在代表性差异，因此该模型捕获了现象意识。显然，这种划分不足以解决经验或主观性的关联性。 Clarion 最多可以声称解决的可能是 Qualia 项目。但是，尽管上层和下层过程的单独表示可能是解释感受质的合理初始步骤，但在我们看来，简单地实施这样的划分并不足以对感受质进行建模。

3.2.2. LIDA（学习 IDA）

IDA（“智能分配代理”）（Franklin, 2000, 2003）与其他模型有些不同，因为它被赋予了一个特定的实际任务来执行：美国海军使用 IDA 为新职位分配专家。 IDA 必须使用自然语言通过电子邮件与水手沟通。它必须访问数据库以查看需要什么和可用的内容。在做出这些决定时，它还必须遵守海军的规章制度。无论最终任务多么实用，IDA 仍然声称是一种模拟“有意识”思维的架构。术语“意识”在这里引用了开发人员，因为没有要求对意识的所有方面进行建模。

IDA 将全局工作空间理论实现为一个自主软件代理。自主代理被定义为“一个位于环境中的系统，它是环境的一部分，它感知环境，并随着时间的推移对环境采取行动，以追求自己的议程”（Franklin & Graesser, 2001）。据称，自主软

件代理通过促进几种假设来促进认知科学。因此，与此类代理合作可以深入了解大脑的运作方式。

Baars (1988) 提出的全局工作空间理论提出意识涉及多个过程，其中一些过程进入全局工作空间。一旦进入全局工作区，其内容就可用于其他元素和流程。全局工作空间中的进程随时间衰减，这使得全局工作空间成为一种动态机制。将许多进程视为独立的代理，人们也可以将模型视为一个分层的多代理系统。 IDA 的开发人员将实现全局工作空间理论的软件代理视为“有意识的”软件代理。

IDA 是一个混合系统，结合了象征性和联结主义元素。负责所谓的高级抽象（例如行为和情绪）的组件与称为小码的低级元素相结合。每个小码都是执行特定任务的一段代码。小代码独立且同时执行这些任务。 Codelet 构成了 IDA 的多个代理。那些成功识别输入的某些方面的小码激活了滑网的节点。当滑网稳定时，输出对应于输入的派生含义。

IDA 采用所谓的稀疏分布式内存 (SDM)。 SDM 是内容可寻址的，据称它非常适合长期联想记忆。从 SDM 中检索是作为一个迭代过程实现的。如果在预定时间间隔内未找到目标，IDA 会生成响应“我不知道”。

一些小码反过来激活全局工作区中的节点。该架构有几个组件：有一个联盟管理器、一个聚光灯控制器、一个广播管理器和一组注意力小码。小码连续扫描输入以查看是否有任何新内容。来自识别新输入的小码的信号由注意力小码组合。注意小码及其关联的小码组成一个联盟。联盟的强度取决于小码将输入与它们要识别的条件相匹配的程度。联盟经理负责监管这一过程。然后，每个联盟都在争夺意识。聚光灯控制器确定要将哪个联盟提升到全局工作空间。最后，提升到全局工作空间的联盟由广播控制器在系统范围内广播。

IDA 中的动作选择是由行为小码实现的。行为类似于具有条件和相关动作的（if-then 类型）产生式规则。 IDA 中的行为小码也在确定动作时考虑了条件的强度。

更新版本的 LIDA，即 Learning IDA（Baars & Franklin，2009 年；Snaider、McCall 和 Franklin，2012 年）具有类似的工作机制，6 至少就那些对我们的评估至关重要的机制而言是.然而，LIDA 中有一个新的工作区子组件，称为“Conscious Contents Queue”，它特别重要。只有广播的内容会添加到此组件中，其中最新的内容表示为第一个元素。此外，工作区中的小码可以直接访问该子组件。

LIDA 的全局工作空间解决了我们列表中的有意识和无意识心理状态之间的差异。 7 根据巴尔斯的理论，提升到全局工作空间的信息可以被视为代表有意识心理状态的内容。状态。在这方面，LIDA 满足了我们列表中的第一项。

下一项，可用性，也适用于 LIDA。回想一下，某些信息会广播到整个系统。这使其可用于模型中的任何组件。因此，LIDA 也满足了我们列表中的第二项。

子组件 Conscious Contents Queue 的实现是朝着实现我们的第三个元素迈出的有希望的一步，因为最近有意识状态的内容在这个子组件中明确表示。然而，由于该子组件目前仅用于表示时间，因此未指定用于更高级别认知能力（例如用于监控）的机制。

我们列表中的最后三项是品质、经验的连通性和主观性。 LIDA 面向任务，不主张或尝试对其中任何一种进行建模。奇怪的是，在某一时刻，有人声称广播“被假设与现象意识相对应”（富兰克林，2000）。除了提供一个很好的例子来说明文学中现象意识和可用性有时是如何混淆的，我们发现这种说法没有什么价值或可信度。然而，平心而论，LIDA 可能具有我们所谓的经验连通性的某些方面的开端。具体来说，我们提到了有意识经验的统一性。联盟管理器和 LIDA 的相关结构可用于将这些方面实施到意识的计算机模型中。

3.2.3、ACT-R

ACT-R（“Adaptive Control of Thought – Rational”的缩写）（Anderson，1996 年；Anderson 等人，2004 年）及其前身 ACT（“思想的自适应控制”）是开放式架构，可以通过编程来执行 sev - 常规任务。 ACT-R 与其他计算机意识模型有些不同。一开始，ACT-R 并没有明确尝试对意识本身进行建模。它为心灵提出了一种架构，而没有明确描述意识的目标。然而，有趣的是，由此产生的模型显示了其他模型中存在的意识元素，这些模型明确着手解决意识问题 (Taatgen, 2009)。

ACT-R 尝试将多个模块与核心生产系统集成。实现并提供了一组模块。用户可以插入自己的生产规则并试验 ACT-R 架构。模块概念的关键在于每个模块都包含一个缓冲区。正如一个人可能会在她的视野中看到许多物体但将注意力集中在一个特定的物体上一样，缓冲区有选择地保存与手头任务最相关的信息。ACT-R 允许这些缓冲区在产生式规则的参与下进行交互。具体来说，产生式规则扫描缓冲区，产生结果的规则也保存在缓冲区中。因此，缓冲区内容会随着新输入的接收和产生式规则的应用而动态变化。

许多过程是并行进行的。但是，某些过程必须连续执行。缓冲区内容仅限于单个声明性知识单元，在 ACT-R 中称为“块”。因此，一次只能存入或取回单个存储器。类似地，产生式规则在每个循环中一次触发一个。集成来自这样一个事实，即每个模块还将有关其活动的信息存入其缓冲区。包含产生式规则的程序存储器模块可以访问这些信息。生产系统可以根据观察到的模式来检测模块行为

的模式和严格的规则。这提供了模块之间的高阶耦合，从而产生了 ACT-R 的集成特性。

模块中的信息和缓冲区中的信息之间的区别满足了我们第一项的要求，即区分有意识和无意识的心理状态。

ACT-R 在我们的第二个项目上也得分很高：有意识的信息对其他过程的可用性。 ACT-R 的生产系统可以访问缓冲区中的所有信息。因此，一旦放入缓冲区，所有进程都可以使用该信息。

第三项“显式知识”和“直接知识”与 ACT-R 的任何部分都不对应。因此我们得出结论，ACT-R 不符合我们列表中的这一项目。剩下的三个，即。品质、经验的连通性和主观性都与现象意识有关。同样，ACT-R 根本没有解决这些元素。

3.2.4、神经元工作空间模型 (NWS)

Dehaene 和 Naccache (2001) 提出了意识的神经元工作空间理论，基于三个经验观察：无意识认知处理的存在，有意识处理中注意力的必要性，以及在某些特定的有效任务中需要意识。据称，这些经验观察可以通过一种理论来解释，该理论将巴尔斯的全球工作空间理论作为发射平台，8 并通过强加遵循大脑结构的约束。

因此，他们假设了三个主要的理论考虑。首先，他们提出有一些过程发生在特定的大脑区域，并且可以在没有注意力机制或整个系统可用性的情况下运行。另一方面，作为他们的第二个理论主张，他们提出了一个分布式神经元“工作空间”网络，该网络连接多个大脑区域。此外，激活的级别和持续时间对于访问该神经元工作区的过程也很重要。通过注意过程对意识状态的自上而下放大是第三个理论假设。

该理论有几种已实现但经过简化的神经网络模型，可以模拟相关的人类行为。 Dehaene、Sergent 和 Changeux (2003) 报道的一个这样的网络对称为注意眨眼的现象进行建模。 9 在模型中，有两个子网络，每个任务一个，两个子网络通过抑制连接，每个抑制另一个节点的激活。子网络还具有可重入的连接，可以在每个子网络中实现稳定性并维持一段时间的激活。无论哪个子网络首先接收输入，都会先于另一个开始计算，并成为第一个激活抑制剂的子网络。以通常的方式处理第二个输入，直到第二个输入到达受较早激活的子网的信号禁止其进一步进行的阶段。事实上，当快速连续接收到两个输入时，只有第一个到达终端，而第二个被抑制。该模型成功地模仿了注意眨眼的现象。

在这个模型中，有意识和无意识表征之间的区别是通过放大的激活和对神经元工作空间的访问来实现的。在仿真中，这两个约束分别通过更高级别的凹入连接和抑制连接来实现。

对于我们列表中的 Availability 项，也可以访问神经元工作区帐户。一旦某个特定流程的激活被提升到工作空间，它就会影响其他流程的活动。

该模型不考虑第三项，意识状态内容的显性和直接知识。作者也没有声称已经实现了这样的功能。

我们列表中的其余三项是品质、经验的关联性和主观性。模型没有单独解决这些问题。然而，Dehaene 和 Naccache (2001, p. 30) 评论了访问意识和现象意识之间的区别，因为这些概念是由 Block 制定的。他们指出，这种差异可能对应于被激活但由于缺乏注意力放大而无法进入全局工作空间的过程与进入工作空间的过程之间的差异。根据 Dehaene 和 Naccache 的说法，从某种意义上说，前一个过程是潜在的有意识的。但是请注意，如果进程无法访问工作区，则整个系统都无法使用它。但是这种能力对于访问意识是必不可少的。因此，认为该模型满足访问意识的要求是不合理的。

3.2.5、ART

ART（“自适应共振理论”）（Grossberg，1987，2007）是用于开发神经网络的框架。它模拟人类的认知过程。该框架的基本动机是提出一种与反向传播不同且在生物学上更合理的学习程序。框架还要求它至少可以解决被称为“稳定性-可塑性困境”的问题。10 从根本上说，ART 是一个由注意力导向系统介导的匹配过程。在 ART 网络中，较低的层代表特征，而较高的层代表这些特征的组，即组块。当输入模式激活网络中的特定层时，自下而上的前馈连接将创建激活模式，该模式由较高级别的连接权重确定。然后，通过自上而下的反馈连接，产生另一种激活模式。如果最后一个模式和第一个模式匹配，也就是说，如果它们的差异小于预定的“警戒参数”，则正在进行的自下而上和自上而下激活的活动将导致“共振状态”。共振状态对应于特定激活模式变得稳定的状态。负责这个匹配过程的子系统是注意力子系统。另一方面，如果匹配不发生，即，如果初始激活的模式与由自上而下的连接激活的模式之间的差异大于警戒参数，则较高级别的活动为通过定向子系统禁止。这种抑制导致在更高层次上寻找更好的匹配激活模式。如果没有这样的模式存在，则创建一个新模式。激活的持续时间足以影响仅在共振状态下两层之间连接的权重。因此，只有在存在共振状态时才会发生学习。

鉴于上述考虑，该框架提出了一种将意识、学习、期望、注意力、共鸣和同步性结合起来的特定理论。名称 CLEARS 是由这六种成分构成的首字母缩写词

(Grossberg, 2007)。因此，根据该理论，学习是通过共振状态进行的，共振状态是新经验与先前学习的期望的匹配过程的结果，所有这些都由注意力介导。此外，这些持久的共振状态被假设为有意识的状态，这是通过大脑中的同步振荡实现的。

关于我们列表中的第一项，ART 提出有意识和无意识状态之间的区别在于有意识状态是注意力导向的共振状态。通过匹配过程，也会出现一些中间激活模式。然而，由于只有足够强大的匹配才能导致共振状态，这些中间状态保持无意识。

ART 中没有直接解决可用性问题。从某种意义上说，所有层对之间都存在自上而下和自下而上的连接，这意味着所有激活模式都会受到影响，并且会影响网络的其余部分。然而，意识状态的延长持续时间，即共振状态，可能是实施一种机制的起点，该机制可以解释意识状态的不同可用性。

ART 中没有提议的机制可以维护网络的内部表示。因此，该模型不满足我们的第三个项目，即显式和直接知识的实现。

关于我们列表中的其余三个项目，ART 中唯一部分解决的项目，即 Qualia，经验的现象连通性和主观性，是与意识经验的统一性相关的经验的连通性。Grossberg (2007, p. 1047) 提出共振态将分布式特征绑定到更连贯的更高级别表示中。这种特征的结合可能足以在某些特定模式中形成统一体。然而，仅仅解释意识体验中各种不同形式的连通性和统一性是不够的。

3.2.6、GMU-BICA

GMU-BICA（“乔治梅森大学——生物启发的认知架构”）（Samsonovich & De Jong，2002 年；Samsonovich、De Jong 和 Kitsantas，2009 年）是一种混合认知架构，其开发用于建模特别是更高级别的人类认知能力。它是最近开发的架构，它独特地专注于“自我”。然而，正如建模者明确指出的那样，这种架构中的自我概念与哲学上的不同，并且是“一种无结构的、抽象的标记，可以归因于心理状态的内容，而不是认知系统本身或任何可观察到的方面”（Samsonovich 等，2009，第 114 页）。 GMU-BICA 的一个重要理论建议是精神状态的定义。心理状态不仅包括内容，还包括主观视角。

内容通过体系结构中的“方案”实现。每个方案都有一个预定义数量的属性，这些属性可以根据代理过去或现在的经验假设特定的值。通用方案存储在“语义记忆”组件中。当前实例化的方案，即具有特定属性值的方案，被称为形成“工作记忆”组件。每个心理状态都有有限数量的方案，以及指向该心理状态主题的标签。交互作用只发生在作为相同心理状态内容的方案之间。单独的心理状态只能复制一个方案（对其属性进行适当的修改）。

唯一需要始终存在于“工作记忆”中的精神状态是“我-现在”。此状态具有代理的视角，并包含具有代表事务当前状态的值的方案。 “I-Now”还拥有对“输入/输出”组件的特权访问。虽然“我现在”是唯一需要的心理状态，但也可能有其他心理状态代表，例如，代理的过去经验（标记为“我-以前”），或其他代理的当前经验（标记为“他-现在”）。还有一种特别重要的精神状态可能存在于“工作记忆”中，即“I-Meta”。这种心理状态可以修改具有代理自我视角的心理状态的内容，即标记为“I-”的那些状态。

GMU-BICA 特别提出的有意识和无意识心理状态之间的区别并不是很清楚。萨姆索诺维奇等人 (2009, p. 115) 指出“是否只有 I-Now 的内容或工作记忆的全部内容应该与代理的‘意识’相关联，这是一个学术问题。”然而，如果以人类认知能力建模为目标，这个所谓的学术问题就不能轻易被忽视。 11 实际上，将意识归因于“我-现在”的心理状态似乎更方便，因为“他-现在”这样的心理状态也可能存在在“工作记忆”中。这些心理状态不能被视为代理的心理状态代表其他代理的心理状态，因为它们具有另一个代理的主观视角（如它们的标签所示）。鉴于这些考虑，该架构仅部分满足了我们列表中的第一项。

如上所述，“ I-Now”状态还具有对“ Input / Output”组件的特权访问。此外，“I-Meta”可以修改“I-Now”的内容这一事实意味着这些状态的内容在某种意义上对其他状态是可用的。因此，该架构为我们列表中的 Availability 项目提出了一种机制。

唯一可以解释我们列表中第三项的候选人，即。显性和直接的知识，是“I-Meta”。然而，上述考虑表明这种状态的特殊性质是它能够改变一些其他的心理状态。我们认为，这为列表项可用性提供了一种机制，但不为有意识内容的显式和直接知识项提供机制。

在最后三项中，GMU-BICA 解决的唯一一项是主观性。这并不奇怪，因为架构的最初动机是描绘自我的概念。然而，如果没有具体机制的规范，仅仅将心理状态标记为属于对象的某些观点是不够的。此外，将标签作为唯一的解释似乎相当不可信，因为在“工作记忆”中还有其他心理状态被标记为其他代理的观点。

4、结论

我们根据哲学确定的意识的六个特征评估了六个已实现的意识计算模型。下表总结了结果。在表中，模型成功实现的功能用加号 (+) 表示。与相应模型未解决的特征相对应的单元格留空。一些单元格包含问号。这些对应于可能需要进一步放大的部分模型或努力。

当按列查看表 2 时，它揭示了有关相应模型在满足我们列表中的特征方面的成功程度的信息。在这里，我们看到 LIDA 的表现更好。当逐行查看时，表 2

显示了当前实现的计算模型在解决列表中每个功能方面的表现。在这一点上，差异性和可用性似乎比其他功能更受关注。相比之下，这些模型很难实现诸如现象连通性和主观性等特征。回想一下，较少代表的特征与现象意识有关。

Table 2
Results of the study.

	Clarion	LIDA	ACT-R	NWS	ART	GMU-BICA
Difference	+	+	+	+	+	?
Availability		+	+		?	+
Explicit and Direct Knowledge	+	?		+		
Qualia	?	?		?		
Phenomenal Connectedness/Unity		?			?	
Subjectivity/Subjective Point of View						?

表 2 中许多问号的存在也传达了一个重要信息。尽管模型试图解决这些特征，但在我们接受模型完全成功地解决和实现这些方面之前，还有很大的改进空间。我们较早地观察到 LIDA 实现的功能要比其他功能更多，这也考虑到了问号，我们乐观地在它实现的功能的最终评分中包括了这些问号。

此外，在我们看来，大量的问号和空单元格意味着未来缩小计算机模型和哲学见解之间差距的工作机会。事实上，哪些模型将解决最多的功能，从而成为未来最完整的模型，主要取决于它们在实现中改进了多少带有问号的方面。换句话说，这项研究为讨论表带来了审查计算模型的价值，不仅可以对它们进行评估和排名，还可以提出新的计算模型可能希望实施的进一步方面的建议。

根据这些结果，几点变得显而易见。首先，我们想提出，建模意识对意识研究具有协同效应。我们相信，随着更多模型的开发，人们对大脑的运作也会有更多的了解。但是，文献中很少有报道过的模型。可以假设，随着计算机模型数量增加一两个数量级，我们的洞察力和理解力也将显着增强。一些计算机模型为意识相关元素的可能理论的初步构建提供了暗示。进一步的建模工作有望完善和磨练理论，以成功解释大脑的运作。从这个意义上说，我们认为目前的意识计算机模型遵循建模的自治观点，其中理论和模型在某种程度上是独立的。

另一个需要观察的是，我们列表中的前三项（即有意识和无意识心理状态之间的差异、可用性以及显性和直接知识）都至少被一个计算机模型考虑在内。这些领域的更多工作应该有助于进一步加深我们的理解。相比之下，后三项在当前模型中不像前三项那样普遍。后三项都与现象意识有关。12 事实上，关于现象意识似乎有相当多的混淆。人们甚至可以大胆地将现象意识视为“笼罩”类别，其他无法解释的现象都归入其中。特别是当它涉及现象意识的各个方面时，哲学、计算机科学和其他学科似乎确实需要合作。这种合作应该旨在仔细识别现象意识的组成部分和元素，以便不同学科就这些术语的概念达成一致。识别围绕现象意识的混乱是这项研究的一个重要结果。

我们想为未来的建模工作提出一个务实的建议。 从上表可以看出，主体性的方面虽然是哲学所确定的最重要的特征，但仅在文献中的一个当前模型中得到

了部分解决。 建议新模型尝试解决主观性问题，因为如果没有这个特征，对意识的完整和全面的理解似乎是无法实现的。

最后，我们想再次强调在本文介绍之前提出的观点。 本研究中对计算机模型的评估基于文献及其各自开发人员的声明，而不是基于动手建模经验。 我们认为这是该研究的一个特别弱点。 然而，这种弱点也暗示了一条改进之路。 例如，可以尝试在特定计算机模型的环境中模拟意识的特定特征，作为这项研究的自然延伸。

致谢

我们要感谢 Sencer Yeralan 参与导致这项工作的初步讨论，以及他对本文早期草稿的评论。我们还要感谢三位匿名裁判对本文件的第一版提出的宝贵意见。他们的建设性意见导致了论文的重大改进。

for their valuable comments to the first version of this paper. Their constructive comments led to significant improvements to the paper.

References

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist, 51*, 355–365.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036–1060.

Armstrong, D. (1997). What is consciousness? In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 712–728). Cambridge: MIT Press.

Baars, B. (1986). *The cognitive revolution in psychology*. New York: The Guilford Press.

Baars, B. (1988). *A cognitive theory of consciousness*. <http://vesicle.nsi.edu/users/baars/BaarsConsciousnessBook1988/>.

Baars, B., & Franklin, S. (2009). Consciousness is computational: The LIDA model of global workspace theory. *International Journal of Machine Consciousness, 1*(1), 23–32.

Baddeley, A. D. (2003). Looking back and looking forward. *Nature Reviews: Neuroscience, 4*, 829–839.

Block, N. (1997). On a confusion about a function of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 375–416). Cambridge: MIT Press.

Cooper, R., & Fox, J. (2002). Modelling cognition. In R. Cooper (Ed.), *Modeling high-level cognitive processes*. Mahwah: Lawrence Erlbaum Assoc..

Da Costa, N., & French, S. (2000). Models, theories, and structures: Thirty years on. *Philosophy of Science, 67*, 116–127.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79*, 1–37.

Dehaene, S., Sergent, C., & Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *PNAS, 100*, 8520–8525.

Descartes, R. (1973). *The principles of philosophy (Translated by E. Haldane and G. Ross)*. Cambridge: Cambridge University Press (Original work published 1644).

Franklin, S. (2000). A "Consciousness" based architecture for a functioning mind. In A. Sloman (Ed.), *Proceedings of the symposium on designing a functioning mind*. <http://www2.dcs.hull.ac.uk/NEAT/dnd/visions/Proposals/stan.pdf>.

Franklin, S. (2003). IDA: A conscious artifact? *Journal of Consciousness Studies, 10*, 47–66.

Franklin, S., & Graesser, A. (2001). Modeling cognition with software agents. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd annual conference of the cognitive science society*. Mahwah: Lawrence Erlbaum Assoc..

Grossberg, S. (1987). Competitive learning: From interactive attention to adaptive resonance. *Cognitive Science, 11*, 23–63.

Grossberg, S. (2007). Consciousness CLEARS the mind. *Neural Networks, 20*(9), 1040–1053.

Güzeldere, G. (1997). The many faces of consciousness: A field guide. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 1–67). Cambridge: MIT Press.

Hameroff, S. R., & Penrose, R. (1995). Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. *Mathematics and Computers in Simulation, 40*(3), 453–480.

Jackson, F. (1997). What mary didn't know. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 567–571). Cambridge: MIT Press.

Leibniz, G. W. (1989). *Philosophical essays. Edited and translated by R. Ariew and D. Graber*. Indianapolis: Hackett Publishing Co. (Original work published 1765).

Levine, J. (1997). On leaving out what it's like. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 543–556). Cambridge: MIT Press.

Lycan, W. G. (1997). Consciousness as internal monitoring. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 755–772). Cambridge: MIT Press.

Lycan, W. G. (1999). Plurality of consciousness. <http://www.unc.edu/~ujanel/CogThs.html>.

McGovern, K., & Baars, B. J. (2007). Cognitive theories of consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness*. Cambridge: Cambridge University Press.

Nagel, T. (1997). What is it like to be a bat? In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 519–528). Cambridge: MIT Press.

Newell, A. (1990). *Unified theories of cognition*. Cambridge: Harvard University Press.

Newell, A., & Simon, H. (1976). Computer science as empirical enquiry: Symbols and search. *Communications of the ACM, 19*(3), 113–126.

Rosenthal, D. M. (1997). A theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 729–754). Cambridge: MIT Press.

Samsonovich, A. V. & De Jong, K. A. (2002). Designing a self-aware neuromorphic hybrid. <http://www.aaai.org/Papers/Workshops/2005/WS-05-08/WS05-08-011.pdf>.

Samsonovich, A. V., De Jong, K. A., & Kitsantas, A. (2009). The mental state formalism of GMU-BICA. *International Journal of Machine Consciousness, 1*(1), 111–130.

Snaider, J., McCall, R., & Franklin, S. (2012). Time production and representation in a conceptual and computational cognitive model. *Cognitive Systems Research, 13*(1), 59–71.

Sun, R. (1999). Computational models of consciousness: An evaluation. *Journal of Intelligent Systems, 9*, 507–562.

Sun, R. (2003). A tutorial on CLARION 5.0. <http://www.cogsci.rpi.edu/~rsun/sun.tutorial.pdf>.

Sun, R. (2006). The CLARION cognitive architecture: Extending cognitive modeling to social simulation. In R. Sun (Ed.), *Cognition and multi-agent interaction*. Cambridge: Cambridge University Press.

Taatgen, N. A. (2009). Consciousness in ACT-R. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness*. Oxford: Oxford University Press.

Tonini, G. (2004). An information integration theory of consciousness. *BMC Neurosicence, 5*(42), 3.

Tye, M. (2000). *Consciousness, color, and content*. Cambridge: MIT Press.

Van Gulick, R. (1995). What would count as explaining consciousness? In T. Metzinger (Ed.), *Conscious experience*. Paderborn: Ferdinand Schöningh.