7

作为哲学与心理学的人工智能

计算机自从上个世代诞生以来就受到了心灵哲学的关注，但大多数时候，哲学家大都着眼于抽象的原理，而疏远了实际的计算机与程序。我并不认为在人工智能这个概念出现前，哲学家能在计算机上找到令他们感兴趣的哲学内容。但是从人工智能这个新的领域兴起时，它开始显现出对哲学多种多样的影响。哲学家们也适时开始回应，在这新领域知识分子的大胆宣言（1）。我在这一章中会给哲学家提供一系列针对人工智能的指南。我们都深知，当陌生人行走在他们不熟悉的土地上时，常会对他们所见之物产生错误而又可笑的误解，而他们所提交的有关神迹与怪物的报告，也常在后来的研究中证明为莫须有的，而同时他们会忽视那些原本更重要的新事物。因此，在受到各式各样错误的人工智能概念困扰，并且浪费了大量时间与精力去钻研虚幻的概念后，我想以此为鉴，提醒其他哲学家我们面对的解释陷阱。不过因我依旧将自己看待为人工智能发展的观察者，因此，我需首先承认我所能触及到范围内的人工智能，包括什么是重要的极其原因，都并不能作为完全可靠的信息来源。还有更多人工智能领域的发现，与我还没有理清的文本存在着。因此，同我一起进入这段旅程的人们，你们要时刻清醒，在人工智能的研究中仍然存在更多可能性等待你去发掘，所以，带着你的批判性去听一个本地人的建议吧。

对很多刻薄的实验心理学家来说，哲学家会对人工智能的产生兴趣并不会令他们感到吃惊，在他们的眼里两者是非常相似的：它们都在使用广泛的概括和大胆的外推方法，同样对于来之不易的实验数据漠不关心，同样喜欢不严谨的内省行为与概念分析，同样使用格言式的理性来给心理学划分什么在原理上不可能的，而什么又是心理学能研究的部分。这样的心理学家可能会说，这两个领域唯一明显的不同点是，人工智能的研究者会将他的扶手椅抬起去到控制台。我认为这种观察是大体上可以被证实，但不应该被看成一个批判。对于非专业但对一切都有分析观点的心理学家而言他们有很多工作，而计算机系统被证明在这个工作上是一种有效的工具。

心理学实际上是一门非常困难的学科。心理学的任务在于解释人的知觉，学习，认识等能力，继而使得心理学能够与生理学形成统一的理论。在心理学中有两种广泛使用的研究方法：一种是自下而上（bottom-up）的，即从一些心理学上非常基本的，被清晰定义的单元或理论性的基础粒子出发，再根据这些基础粒子搭建不同的分子，进而积累成模型来解释我们都能观察到的复杂现象。而另一种方式是自上而下（top-down）的，即尝试使用更加抽象的解构方式去解构高级心理组织，而希望借由研究这些组织，来尝试一步一步理解更精细的系统与心理过程，直到接触到与生物学相近的结构。一种通常的理解是这两种研究方法能够

也应该同时进行，但我们已经有足够证据证明在心理学中自下而上的策略并非卓有成效。在自下而上的策略中有两个较为成熟的心理学理论，其一是行为主义中的经典条件反射（Stimulus-response behaviorism），其二可被称为“神经元信号生理学性心理学”（neuron signal physiological psychology），大部分研究者对这两种理论过时这一点都了达成共识。对于前者，这种刺激-反应的结构并不能清晰证明其能够作为心里层次上的基本粒子。对后者而言，尽管突触与脉冲序列可以成为完美的基本粒子，但因为这种粒子数量过于庞大，使得研究者在放弃对于传入传出神经的研究而转向脑的时候，神经元之间的相互作用的复杂性，令他们不得不放弃对于单一神经元的研究（见第四章与第五章）（2）。自下而上的研究方式同样没有在其他科学领域的早期研究中被证明有突出的成效，例如化学与生物学。因此心理学家跟随着“成熟”的科学发展转向了自上而下的研究方式。这种自上而下策略的起点多种多样可以被排列出来。面对着粗略的研究后产生的，一个心理学实践上不可能回答的经验性问题（神经系统在事实上是如何完成 X 或 Y 或 Z 活动？），心理学家会提出一种更简单与初级的问题：

如何让任意的一种系统（包含 A,B,C 等特征）完成 X 活动？（*1）

这种类型的问题因为消解了一部分“经验性”而显得更简单；这是一个更工程性的问题，即针对问题寻找解决方式(任何解决方式)而不是去发现新的事物。这种仅用来解决问题的提问，有时候会将研究引向解决方案的总体限制性上（当然也包括我们在自然上未知的答案），而寻找到有限性即是这种先验性理论的价值。当一个人决定以这种方式来研究心理学的时候，他需要决定自己在哪种程度上接受经验性的困难，而他要做的即是填补上现有心理学样式空白的部分（3）。将现有的心理学系统的经验性边界描述出来，或是在必要的行为的描述中展示出经验性边界，才是一个心理学家描述“心理真实性”最好的方式。比如说我们可以问到：神经元系统是如何在这样或是那样的物理学性质下，完成人分辨颜色的行为？或是我们可以问：任何有限的系统是如何对学习自然语言产生帮助的？或者也可以问：人的记忆是如何可能地被组织起来的使我们可以去相对轻松回答类似于“你骑过羚羊么？”这样的问题，同时相对难以回答“你上周二吃过早饭么？”。又或者借用康德问到：从整体上来说，我们是如何获得任何经验的，或知道任何事物的。纯粹认识论实际上是心理学所能提问的极限，而认识论这种初步证明与心理学经验性的细节是同样重要的。一些认识论的问题如果能被更好地进行科学设计，会得出比起其他提问更好的答案。但这种哲学性的探索可能会限制那些依靠丰富数据来进行的研究。

人工智能研究者可以选择某种程度的经验性困难来开展他们的研究；例如在卡内基梅隆大学的研究中，实验人员格外注重每个实验个体的表现，他们尝试通过更精细的数据来针对人的表现构造模型。其他的人工智能研究者将重心从这个程度的心理现实性转移到了更抽象化的人工智能研究中。更多的重心与兴趣被放到了心理学经验性目的上的研究中，但我想要表述的一点是人工智能应该更被看待为一种与传统认识论问题相关的，更宏观与抽象的，自上而下的问题即：知识

是如何可能的？（*2） 对于很多哲学家而言，人工智能无法搭建一个可行的构架，因为这涉及到了更困难的问题：人工智能限制自身获得机械性的答案（Mechanistic solutions），因此人工智能的领域无法触及康德主义者的领域，即康德有关智能的全部可能形式。人工智能只能停留在全部可能的机械性智能模式中。当然这个论断与生机主义，笛卡尔二元论，和其他反机械主义的想法产生相反的乞题。但是和我在别处的讨论一致的是，人工智能的技术需求并不能在任何时候作为新增的限制。如果心理学还能够有任何作为，如果邱奇-图灵论题是对的，那么现有机械的限制并不会严重于心理学自身的乞题问题，而且谁又愿意去得罪他们呢？（见第 5 章）（4）

因此，我宣称人工智能与哲学（尤其是认识论与心灵哲学）同样都在对最抽象的心理学原理进行研究。但人工智能与心理学共享着与哲学不同的研究手段去回答问题。在人工智能或意识心理学中，典型的研究方式是尝试回答自上而下的问题，而这些方式通常包含设计相关的系统，尝试使用那些系统来达成，或看起来达成特定的行为。之后再考虑现有系统的哪些特性具有对系统总体的必要性。哲学家常常忽视那些缜密的计算机系统设计而更倾向于固执地对人工智能的总体概念进行提问。这可能就是人工智能与“纯粹”哲学研究对于相同问题主要的研究方式差异点。而我现在的主要目的就是列举出这两种研究方式的优势与劣势。

系统设计作为在人工智能与其他自上而下的心理学研究中常见的研究策略，有着很多潜在的危险，而其中四点最为突出：

（1）：给系统内的子系统远超规定运行能力的任务。（例如，将超过相对的时间与资源容许的资讯处理任务给到单独的子系统，或在更抽象层面上的产生工程学意义的非连贯性，或给子系统的任务需要子系统本身有着超过上层系统的“智能”或“更多的知识”）
（2）：错误地将现有的条件性的解决方案中的必要部分，当作总体的限制性。（一个较为简单的例子是，宣称大脑使用 LISP；更重要的例子需要更小心的阐明。）
（3）：将自己限制在人造的子系统设计中（例如对深度的测量器或句子的句法分析）和创造一些在系统层次上不能嫁接到其他认知系统子系统的解决方案。
（4）：限制一个系统的表达仅仅在人工设定的极小“自然”范围中，而无法提供有效或可行的方法去扩大这个系统。

以上的问题对于人工智能而言已是老生常谈，而类似的问题在心理学研究中也常常出现，而这些问题更为棘手。让我们考虑以下几个例子，危险（1）：弗洛伊德的“自我”子系统与詹姆斯·吉布森不变的灵敏性感知“音叉”都在一个子系统中包含了大量内容。危险（2）：行为主义者从动物行为的必要性到人类行为的必要性的预测被指没有依据。而这种类比的错误性可以从另外一个例子里看出：即青蛙的视觉与发送到脑中的成像与人类视觉与脑中成像是完全不同的。危险（3）：我们很难看出乔姆斯基早期的句法结构系统如何与语义学的部分进行交互而产

生有意义的语言表达。危险（4）：我们很难看出对于一些无意义音节记忆的模型是如何被扩张来解决相似的但更复杂的记忆工作。而人工智能的长处则在于当这其中的任何一种理论遇到了阻力的时候，人工智能常常可以提供结论性的证明。

现在，我已经将人工智能，哲学以及心理学三者连接成了三角形（就如同我在标题里写的那样）：人工智能既可以（也应该作为）抽象的与“非经验性”的像去哲学那样解决问题。但同时，人工智能在搭建模型的层次上应该如同心理学一样是非常清晰与专一的。因此我们从一个特定的但是非常不切实际的人工智能模型中可以学到的有价值的心理学与认识论内容，就和我们能从火星人身上学到的心理学内容差不多。可能与我们不同的是，对于火星人而言他们会将人工智能的总体理论搭建在可对人实践的心理学或认识论上。在回答最重要的问题“到现在为止我们在人工智能之上已经做到了什么？”之前，我希望能先解决在人工智能层次上的一些错误理解，这些错误理解是我们之前讨论还没有涉及的。

因为我们对于人工智能的理解是建立在一系列自上而下的认知心理学之上的，去假设我们在电脑上利用人工智能解构出的模拟功能与人脑的功能有某种同构性（Isomorphic）是十分有诱惑力的。我们从一开始了解到电脑是由数量极其庞大的基本单元组成，而这种单元构成的电脑可以在某种程度上模拟出人智能才可以做到的行为。因此我们就会进行以下的假设：因为脑也是由百万的小功能区组合而成的，而又因为我们在理解人的行为与那些小功能区之间有着某种未知的部分，所以我们认为人工智能最终的目的是提供一个可以被拆分理解的分阶式的人工智能，而这个人工智能可以与人脑中那些难以发觉的分阶式结构进行类比。而我们常见的表述方式即是“人的器官由组织构成，而组织由细胞构成，细胞由分子构成，而分子由原子构成。”而我们会去做假设，将人的结构与计算机电子元件做类比。在这个强而有力的生物图景中，会令人感到泄气的是，我们已知的部分神经系统的功能的与实际电子化的计算机设备的运行方式是不同的。而对于这种担心的标准回应即我们不应该对于计算机的本质进行过于深刻的发问（这种回应有时候被称为“谷仓问题”）-计算机本质是电子化的设备，但电子化设备可以从各种程度上模拟任何类似它的设备，而当我们在计算机中进行了很高程度的积累后，我们或许可以发觉电脑的模拟部分可以被标注到非电子化的大脑部分上去。如同很多作者已经观察到的那样（5），我们无法在不知道一个心理性现实的模型的某个特定部分是反应了自然中的事实，还是仅仅作为模型的加速器或无意义的细节内容之前，对于任何模型进行判定。（一个典型的例子是：在 18 世纪，科学家能够建造非常精美的黄铜发条模型来表现太阳系被称为太阳分仪。太阳分仪中的那些齿轮并不能代表宇宙的实际样貌，而球体之间的反射也并不能代表任何实际关系。）当我们去研究人工智能的本身程序的时候，我们必定会看到无数的数字运算；如果这些看起来已经非常不具生物性了，我们至少对其还有一些解释的余地，因为我们之前所说的数字计算仅仅是在计算机后台工作，而并非我们应该与自然对比的部分。

对人工智能的考虑现在理论上还是可行的，我认为也许某一天我们可以宣传

这种只存在于想象中的相似性可以到达到神经元的水平（但不是我们之前说的计算机后台），但人工智能的研究到现在为止还没有发现任何清晰的迹象表明我们可以前进到上述的那一步。而我们现在所见的是自上而下的计算机模型，从起初就和自然进行了区分。人工智能通常都是给更高层次的意识行为做模型，比如说概念理解与修改目的或信念，而这常常要利用上百万的基本电脑处理去模仿为数不多的，有目的性的人的特性——但没有人知道或关心那些最基本的处理是什么，因为没有人假设在那庞大的运算当中为了特定目的服务的行为，可以确实被看成心理学或是生理学。因此人工智能所模仿的心理性现实都是非常浅显的；只有那些高层次的程序的诠释与测定被当作模型。

最简单的心理性现实的人工智能模型，经常为了解释给普通人而被模糊化，它们以一种非常具有误导性的方法被展示在聚光灯下。而其中最能有效吸引观众而能激发他们幻想的就是那些看起来非常逼真的程序。让我们来看看以下的一段对话摘录（不包含程序员精妙的评论），来自特里·威诺格拉德的 SHRDLU 与其对话者。SHRDLU 是一个生活在虚构“积木世界”中的软件，它可以操控桌面上不同颜色的方块来回应对它的命令和回答有关它所处世界生活的问题。加黑的内容为电脑的回应。

在你将绿色的锥形放在小方块上之前，你有接触任何锥形么？
是的绿色的那个。
你什么时候拿起来的？
在我叠放红色的方块的时候，那是一块红色的大方块和一块绿色的大方块。
为什么？
把那块移开。
你为什么这么做？
去清理红色的方块。
你为什么要清理那个方块？
去把那个方块放在绿色的大方块上面。
为什么你要那么做？
因为你叫我这么做的。（6）

当 SHRDLU 在最后回答到“因为你叫我这么做的。”我们难免会因对话中明显的人性而惊呼——“SHRDLU 是我们的一员！”我们被诱惑做如此的宣称。如果这个人是本地人，那么他一定会对这软件产生敬畏；如果这个人并非本地人，他依然会因为这种软件的表现而对这种幻觉感到兴奋。SHRDLU 的回应尽管完美的与当时的场合所吻合（而且并非偶然！），但依然是“事先准备好的”。维诺格拉德只是简单给 SHRDLU 整个句子之后，让它在适当的时机做出回应。如果一个小孩给出了与 SHRDLU 类似的回应，我们会很自然地期待这种行为表现了他自己整体的能力：“因为你让我这么做”，或者“因为那是别人要求我做的”，或者在其他情况下：“因为我想那么做”，“因为你的助手让我这么做，”但这些为于微妙层面

的回答是 SHRDLU 所做不到的（7）。它的行为看似十分逼真，但它并没有为我们揭开人与人交流层次的知识，比如请求与要求，或在适当情况下与他人进行合作的对话。（需要表述清楚的一点是，维诺格拉德的论文中已经非常清晰地表明了，他在哪里以及何种程度上为 SHRDLU 的回应进行了实现准备，因此任何人如果感到被 SHRDLU 欺骗了，那他是没有读维诺格拉德的论文。其他的自然语言系统不依靠事先准备的回应或者需要极少程度的准备。）

现在，一个事实依旧摆在我们面前，大部分反对人工智能的人都是对上述的欺骗手段感到气愤与不信任的人。为什么做人工智能的研究者要使用这些小把戏呢？因为各种原因。第一，他们需要从程序那得到些可以搬弄是非的回应，而且将原本更理智的，技术性的，与轻描淡写的内容做成栩栩如生而又显得“自然”的对话并非难事（比如“理性：在下命令前就去做那件事，“REASON: PRIOR COMMAND TO DO THAT”）。第二点，在维诺格拉德的例子中，他所尝试的是揭示我们能接受的最低程度的正确的语言结构分析是什么（注意到在上述例子里，句子中的代词前置）。因此我们能够通过检查人工智能“自然”语言的输出，来确定我们对于自然语言的研究与输入部分是否合理。第三点，将录制的回答放入程序中，这么做的确十分有趣。无论是去用来戏耍那些不会被骗到的同事或是那些并不明白原理的门外汉，这都是有趣的事情。作为一个门外汉，我们必须学习如何才能不被这种看起来真实的人工智能所迷惑，人工智能这种人造的真实性就如同化学家们各式各样的玻璃瓶，或是二战中被画满牙齿的战斗机一般，仅仅是一种看似吓人的幻觉。（约瑟夫·维森鲍姆有名的 ELIZA 程序（8），电脑“心理治疗师”看似拥有智慧与同情心可以聆听他人的问题，但制造它的部分目的是给被人工智能的逼真性激起热情的研究者而做的解药。这几乎全是精心事先准备的，而它并非作为任何心理学意义的现实模型而存在，它不过是作为一个研究者会多容易被程序戏耍而期望过度的证明而已。它只是单纯利用人输入的句法结构而并非有任何实质性的理解。有人可以说这是一个韦尼克失语症的貌似可信的模型，它可以含糊不清地说出有格式的和甚至语义适当的回应给对话者，有时候维持理解的幻象好一段时间。）

当研究人工智能的社群为了吸引注意力而做的好玩的事情的时候，他们会为这种误导性的行为付出代价。不仅在人群中会助长“做人工智能的研究者就是一帮骗子与黑客”的想象，而且会让更多严重错误的人工智能概念被传播出去。比如维诺格拉德在 SHRDLU 中的真正贡献并非他创造了一个可以表达与理解英语的程序，这个程序在不同层次上都可以对应心理现实性研究的（尽管这是软件所体现的逼真性，也是很多人对于该研究的赞誉，不过维诺格拉德本人并不应为此负责）。而维诺格拉德真正的贡献是，他发现了任何系统最深层次的要求，任何只要是能够接受命令（在自然语言中），做计划，改变体系与检测变化发生的系统，在他的研究中，他都阐述了这些系统共同的问题，并且给出了巧妙与可行的部分解决方法。而那些有关 SHRLDU 系统本身不能充分对不同词语的产生理解与其运行速度缓慢的批判，并不能影响到维诺格拉德真正的贡献，甚至这些全部正确的

批判与这些贡献本身都不相关。

事实上，对于 SHRDLU（和与之类似系统）表现的过分关注，体现了人们对于 AI 系统一个错误的认识。人工智能系统并非经验性的实验，而是让电脑进行的一种人工的思想实验。最近，一些人工智能的研究者开始称呼他们的研究为“实验性认识论”。这个不幸的名词一定会让哲学家血压升高，但如果人工智能自称为认识论思想实验（甚至是：思想实验性认识论 Gedanken-experimental epistemology）哲学家应该对此感到安心。那些被人工智能的思想实验所提出与解答的问题都是和系统有关的，即我们是否能在特定设计好的系统中获得某种程度的信息处理能力——识别能力，推断能力，对不同种类能力的控制能力。而回答通常都是否定的。消除法庞大的阴影在人工智能中逐渐显现。一些相对可行的方案已经充分表明了，人工智能整体上并不能给出我们需要的行为，对于这件事情的认识是十分重要的进步，尽管这意味我们并不能创造复杂思维的机器人。

而人工智能的硬件实现是几乎是不必要的。这就如同把铅球从比萨斜塔上扔下的实验，他们所做的证明对于理解理论本身的人而言是非常多余的，但对于其他人来说却是充满说服力的。那么对人工智能而言计算机是不相关的么？“从原理”来说是不相关的（就如同“从原理”来说在黑板上的图示对于学习几何是不相关的），但是实践上并非如此。我在文中曾形容计算机是思想实验的人工调节器。我这么说的意思是：将科研人员自身的妄想排除在思想实验外是十分困难的；计算机模拟能迫使我们认识到这些想象出的设计的价值。如同芝农·派利夏恩的观察，“我们需要的是...一种训练人想象的计算机语言。”（9）而电脑所提供的限制是无法被抵赖的（这一点对于计算机编程的初学者尤为明显）。这可以说是一件好事——就如同我之前提供的理由——也可以说是一件坏事。也许你知道你身边的某人过于沉浸于桥牌，而他的生活在他眼中就变成了一系列的出牌，结束与对出鬼牌。每天早上他都认为自己在抽将牌，而任何时候他看到工期的结束他想象这是盖牌。而计算机语言看起来对人有着类似的影响。尽管我不会去引述其他例子来证明这个观点，但我认为很明显的是当我们在尝试驯服这种芝农·派利夏恩所说的“技术性语言”的时候，我们自身的想象力也会被这种技术所消弱。（9）。

我们经常会说电脑对于其使用者的想象力会有很大的影响，而我们却忘记了电脑实际上是用一种非常明显但又被低估的方式做到的这一点，这种方式是电脑纯粹的运行速度。在电脑出现之前，理论学家很受局限忽视了来自于心理学真正庞大而复杂的运行过程中的可能性，因为我们很难看出这样的程序在欲动似静与欲静似动的状态下为何没有出现的。因此心理状态曾经的标志是其迅捷性。有人可能会说思考的速度决定了主观的“快”的上限，光的速度决定了客观“快”的上限。现在，让我们假设根本不存在任何电脑，但不知为何（可能是因为某种魔法），肯内特·科尔比试着虚构了这些流程图作为妄想症中的一部分人类组织的研究模型。（这些流程图来源于他的书，Artificial Paranoia，Pergamon，1975；图 7.1 是主要的流程图；图 7.2 与 7.3 是放大的流程图 ） 对于每个人，甚至

我认为对科尔比本人而言，这是一个被异常简化的妄想症模型，但如果不曾出现电脑给我们展示全部的流程是如何在几乎是一瞬间完成的，那么我们会更倾向于在一开始就否定这些流程图，而把它们当作无意义的事情，一个鲁布·戈德堡机械。大多数的程序在慢速过程中都如同在做无意义的事情，但将其加速它们的工作就能展现出看起来自然而灵活优雅的一面，但这种优雅的一面是完全无法在慢速情况通过研究程序本身而被察觉的。（让我们比较大量的植物生长与含苞开放的照片在快速与慢速情况下我们能做出的反应）而蕴藏在人工智能的运行上的优雅可能仅是一种错觉。也许自然在各个层次上都是优美的，但不知是好是坏，计算机的速度让我们从理论家的想象中解放了出来，并且开启了可能并且可行的复杂信息互动，这种互动在产生神经活动中至关重要，这种交互如此迅速以至于能够作为内省行为的一种粒子。

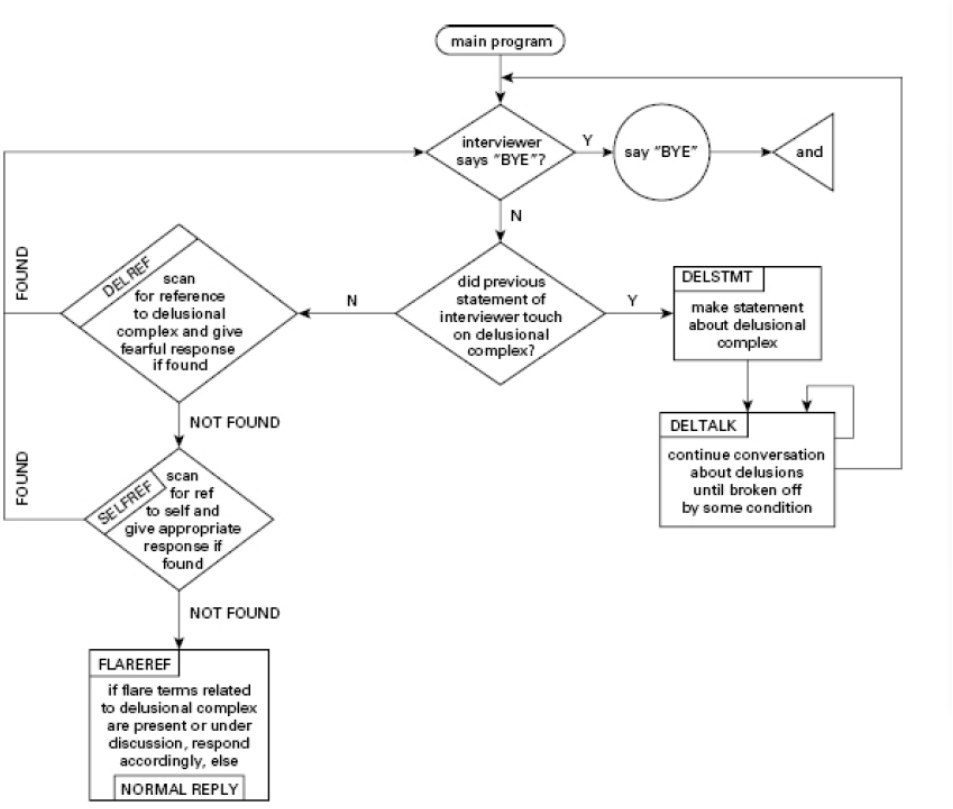

Figure 7.1

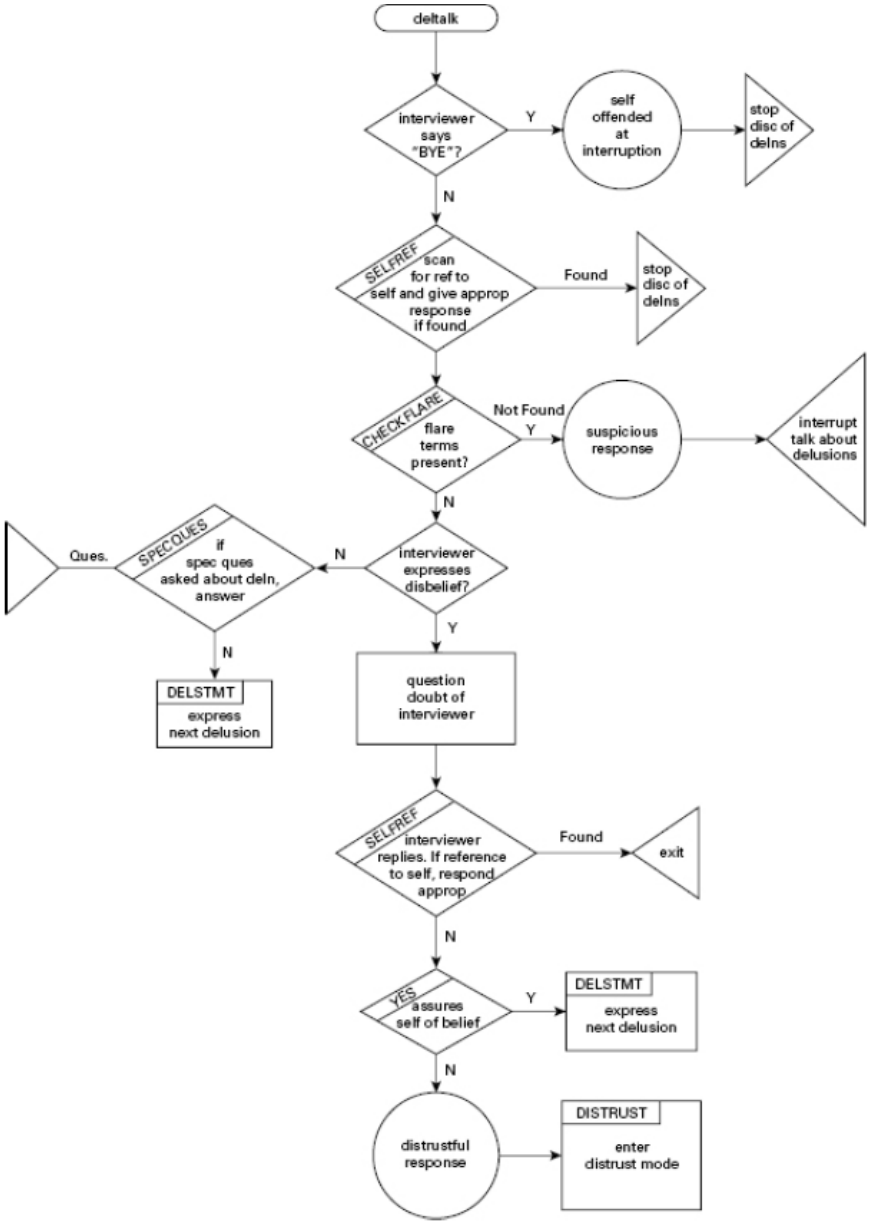
deltalk
interviewer says "BYE"?
Y
self offended at interruption
stop disc of delns
N
SELFREF
scan for ref to self and give approp response if found
Found
stop disc of delns
CHECKFLARE
flare terms present?
Not Found
Y
suspicious response
interrupt talk about delusions
N
SPECQUES
if spec ques asked about deln., answer
Ques.
N
interviewer expresses disbelief?
N
DELSTMT
express next delusion
Y
question doubt of interviewer
SELFREF
interviewer replies. If reference to self, respond approp
Found
exit
N
YES
assures self of belief
Y
DELSTMT
express next delusion
N
distrustful response
DISTRUST
enter distrust mode

Figure 7.2

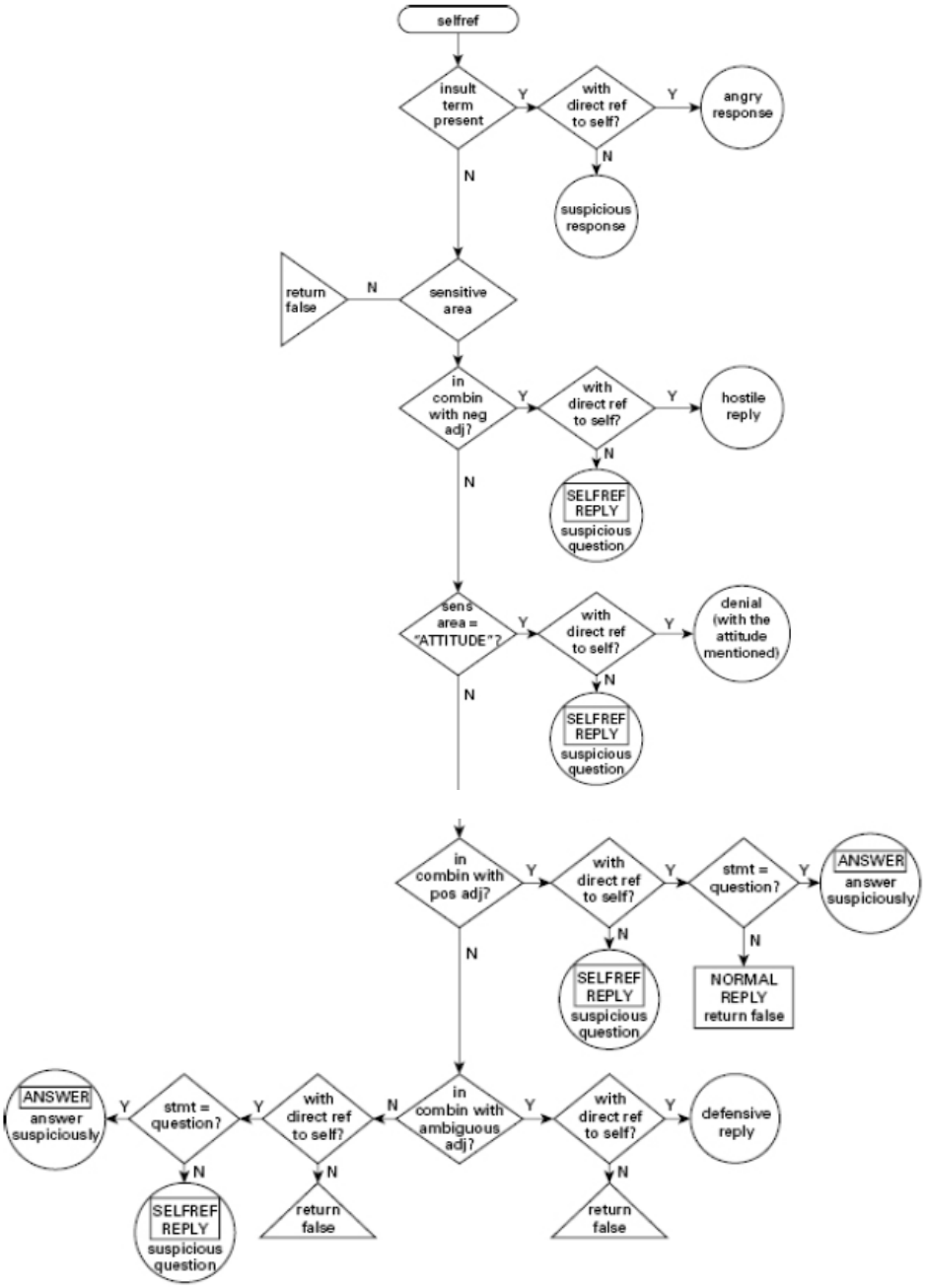

Figure 7.3

最后让我回到最重要的问题。假设人工智能可以如同我介绍的那样，被看作最抽象的对于智能或知识是否可能的问题。到现在为止它有解决一些总体的问题或者发现任何重要的限制或原理么？我认为在特定条件下的答案是肯定的。特别是人工智能解决了一个环绕在哲学家与心理学家超过两百年的论证最困难的部分。这是一个该问题基本的版本：首先，只有能揭示人内在的表象的心理学家才

能成功揭示人行为的复杂性。这个假设已经成为，除激进的行为主义者外，几乎全部研究者的最初假设。（心理学与哲学皆是——无论是约翰·布罗德斯·华生，伯尔赫斯·弗雷德里克·斯金纳,还是吉尔伯特·赖尔，或诺曼·马尔科姆）笛卡尔怀疑任何事情，却不会怀疑上述的假设。对于英国经验主义者来说，内在的表象被称为观点，感受，印象；而更近年代的心理学家则开始使用假设，图画，图示，想象，命题，印记，神经信号，甚至是全息图与完全先天的理论。因此，这个假设几乎是无法被反驳的，或者说在任何程度上都是令人印象深刻的。但，第二点是，没有任何事物是本质上作为其他事物表象而存在的；一些东西仅仅作为其他事物的表象；任何表象或表象的系统因此需要至少一个能够解释表象并且并不属于表象的使用者或解释者。任何一种解释者必须有一系列的心理性与目的性的特点；它必须有能力做一系列理解性行为，而且必须有信念与目标（因此它可以使用表象去内指向自己，进而协助它自己完成目的）。这样一种解释者就如同某种人造人。因此，没有这种人造人假设的心理学是不可能的。但是带着这种人造人假设的心理学确会进入循环论证与无限回溯的噩梦中。

以上的论证是对于一系列类似问题相对抽象的论述。因为看起来对于很多人来说，如果我们不假设一个内在的图像（模型或图示）来成像外在世界的话，我们是无法去理解感知的。但同时这个想象在我们没有内在的眼睛去感知这个模型的时候又有什么用呢，而我们又应该如何诠释这种感知能力呢？对于很多人而言听到一个句子然后去理解那个句子，实际上要在某种程度上将句子翻译为某种内化的信息，但我们又如何理解这种信息：把它再翻译成其他的东西么？我们所遇到的问题依旧老生常谈，让我们称呼它为休谟问题。尽管休谟没有清晰地表述出来，但他理解了这个问题的严重性，并花费了大量精力去回避这个灾难性的问题。休谟的内在表象是印象与观念，而他非常聪明地避免了使用内在*自我*的概念来操控其他事物，但这种抛弃使得他不得不让那些观念与印象开始“自行思考”。而结论就是，他所认为的自我仅是一捆印象与观念。他尝试使用繁复的思维连接方式将这些印象与观点纳入到动态的相互交流中，这样每一个相互继承的观点可以在流动的意识当中根据不同的理论连接起来，并以此杜绝了存在内在智能的小人进行管理的情况。但当然，这个理论并不成功。这个理论无法让表象在现有的认知情况下有效工作。而休谟的失败可以看作，任何与他类似尝试最终都会失败的预兆。在一方面，我们如何使用心理学理论，搞清楚那些能自己理解自身的表象？在另一方面，我们如何让任何心理学理论在提出了至少一个作为表象的理解者的存在的情况下，避免回溯与循环论证？

现在，不可否认的是，将内在表象作为假设的哲学家与心理学家都打心底里认为这个恶魔会被打败，即休谟问题可以被解决，但我确信没有人在人工智能与数据结构出现之前知道如何做到这一点。数据结构可能是，也可能不是生物学或是心理学意义上的现实性表象。但它们是我们能拥有的非生物性情况下，能做到该功能的例子，因此能作为帮助我们理解表象的必要条件。（*3）

现在，我会用隐喻的方式介绍我们是如何发掘上述问题的解决方案的，（任何

对于内在表象的描述都会牵扯到非常多隐喻）而这就牵涉到我们先前所说（见第五章）人工智能自上而下的研究方式。我们从人工智能的最初假设，即一个人或一个意识有机体——我以更中立的方式称呼为目的性系统（见第一章）——或者某种人工创造的某种人单项的能力（比如，下棋，回答有关棒球的问题），然后将这个最大的目的性系统分成一个系统之下的很多子系统，每一个子系统本身可以当作一个目的性系统（包含了它们自身特别的信念与欲望）因此如同一个人造人。事实上，人造人这个说法在人工智能中已经非常普及，而且这个暗喻常常都是有用的。人工智能的人造人相互交流，从他人那夺取控制，自愿行动，分包合作，监视，甚至杀害其他人造人。比起这个说法，似乎没有比这更好的方式了(11)。如果人造人尝试去通过复制来解释整个需要被理解的能力，他们就会变成无法控制的怪物（一个危险（1）的特别案例）。如果某人可以获得一组或一小队相对无知的，单纯的，盲目的人造人去用以生产整体上有智能的行为，而这才可以被看作是一种进步。而流程图就是一种整理好的一组小人（研究者，图书管理员，精算师，将军）；每一个盒子都有着事先被设定好功能的人造人，而且从外部看过去并没有说明他们在做什么（我们把这种效果说成：把一个小人放到盒子里去做事情）。如果我们靠近去看每一个盒子，我们可以看到每一个盒子中的功能又是更小的盒子中的人造人在做更小而更简单的事情。最终，这种盒子中带着盒子的结构会让你看到最里头的人造人，而它过于简单（他们所需要做的事情仅仅是记住在提问的时候说 “是”或“否”）以至于它们可以被“机械所替代”。我们将幻想中的人造人简化为做简单的笨蛋们去做机械的事情。

当人造人在某一个层次上进行交互的时候，它们以发送信号的方式去与对方交流，而每一个人造人都有着表象去决定他们的功能。因此，经典的人工智能会在表象与表象-使用者之间做区分（12）：他们在面对无限回溯的威胁中踏出了第一步。但很多人工智能领域的作者注意到(13)，在人工智能的调试过程中，我们需要权衡表象的复杂性与使用者的复杂性。越是原始的与没有被解释的表象——比如对视网膜瞬间形状的模仿——解释者或是使用者表象就会越复杂。而越是能够被诠释的表象，那么表象本身的步骤性信息就会增加，而解释者的表象就会减少。这个现实，使得我们可以通过给底层的人造人更多的工作来，减少在高层的意识里人造人的数量。我们无法获得完全的对于自身表象的解释（除非我们可以从自身的位置脱离出来，从外部去观察整个系统），但说到底全部的人造人都是可以被分解的。但作为完全拆分的代价，我们需要牺牲这些的子系统与它们的表象可以拥有的逼真性与总概性(14)，即我们不能做到让人工智能如同正常的人类语言那样交流。我们已经从 SHRDLU 那获得了人造人交流的失败案例，“因为你让我去做的。”交流的语境与表述的能力已经让我们清晰得看出了，人类语言交流的复杂性与语言表象本身的复杂性，但它并非一个成熟的格莱斯派言语行为理论。如果它是，那我们就需要每个人造人都有着过于梦幻般的能力才能使用语言了。

现在哲学家有两种方式去看待人工智能数据结构。一种是承认它们确实是自我理解的表象，或者我们可以借用模型与真正表象不一致的例子（人的观点，绘

画，图绘）得出结论，数据结构根本就不是内部表象。但如果我们愿意接受第二种观点，即人工智能至少成功消弱了我们的第一个假设的强度：心理学是否需要内部表象这个问题变得不是那么清晰；内部虚构的表象似乎也能做到这一点。

去与他人理论说，人工智能是我们已知的唯一能解决休谟问题的方式的确是一个很有诱惑力的事情，尽管这是非常受到限制的系统，但这必定是正确的方向，它的类别在心理学上也是真实的，但我们也可能是进入了危险（2）的论述中。我们可以激励自己去学习可能的解决休谟问题的方式，但到现在为止我们并不能说人工智能是解决休谟问题的唯一方式。

人工智能通过用特定的方法解决休谟问题的简单例子的方式，给哲学与心理学做出了巨大的贡献。而它又是否解决了其他哲学家感兴趣的问题呢？我会在文末指出两个主要的领域，我认为在那两个领域人工智能与哲学产生了联系。

哲学家与心理学家在是否存在自然的心象这个问题上争论了很多年（当然两者都是勉勉强强地在交流）。之前的交流大多数时候都没有什么成果。在我看来，因为两者都不知道如何把握休谟问题。近些年在人工智能上的研究，在重铸一个更清晰的，有利的解决休谟问题的框架，而任何人如果希望解决这个古老的问题，他必定会在人工智能中寻找到帮助(15)。

而第二个与哲学有关的领域，在我眼里即是“框架问题”（Frame Problem）(16)。框架问题是一个抽象的认识论问题，而这个问题是在人工智能的思想实验中发现的。当一个有意识的造物，一个有各种关于我们所在世界信念的物体，进行任何行为的时候，这个世界本身发生了改变，而很多信念需要被修改与更新。但我们如何做到？这不可能是通过我们意识到自己全部错误的信念，进而改变的方式进行的（能确定的一点是，很多改变并非在我们能够意识到的层次发生的），所以我们不可能完全依靠感知输入去修改我们的信念。因此我们必须有着一套内部的方式去更新我们的信念，去填上空缺并且保证我们内部的模型，我们整体的信念，大体上让世界是可信的。

如果我们假设，就如同哲学家传统的做法，我们的信念是一系列命题，而理性是对于这些命题的推断与演绎推理。有着这种信念的人会陷入麻烦，因为很清晰的一点是（尽管备受争议），如果系统仅仅依靠以上的流程是绝对会陷入泥潭的，因为这样上传的内容过多而导致组合爆炸。看起来如果我们需要去解释为什么人可以有着与自然保持和谐的信念系统，我们对于信念与理性整个概念系统都需要被修改。

我认为我们可以从康德那发现对于框架问题的帮助（我们或许可以称呼框架问题为康德问题）但是除非我们能将某人的理论限制在人工智能思想实验的形态当中。在这种情况下，哲学所推崇的对于问题的答案，包含当然包含康德，最好能被看做建议性的，最差仅是妄想而已。

我并非希望哲学家都放弃传统哲学的内容，然后训练他们自己成为人工智能的研究者。现在还有很多的思想实验与推论需要哲学家使用哲学途径与哲学传统去解决。最近一些非常有影响力的人工智能成果（比如马文·闵斯基有关“Frames”

的论文）包含了大量对于相对不复杂的，有关自然的哲学猜想。哲学家们，我已经说过了，应该进行人工智能的学习。那么人工智能的研究者需要学习哲学么？他们也需要。除非他们可以接受每过几天就要重新发明轮子。当人工智能重新发明轮子的时候，首先肯定会是一个方形的，或在最好的情况下，是一个六边形，它们需要至少成百次的更新迭代才能够令人满意。而哲学家们的轮子是完美的圆形，在原理上是不需要润滑的，它能同时往两个方向走。显然共识是有序的。(*4）

注释：

（*1）George Smith 与 Barbara Klein 向我指出了在文中所提出的问题在某几种角度去理解的时候是很模棱两可的，因此对于相同的问题其实可以做出不同的回应。在下文中，我所提出的回答这个问题的不同策略，实际上也能回答不尽相同（但类似）的问题。提出类似问题的哲学家有时候会尴尬地发现他们收到的回复是对于其他类似内容的详细回答。

（*2）这个问题（以及对于问题去回答的尝试）是作为认识论的一个主干部分；另一个主干部分即尝试解决怀疑论问题，而这个主干的问题可以是：“知识如何可能？”

（*3）Joseph Weizenbaum 向我指出了，Turing 从计算机诞生的伊始就预计到了计算机可以在理论上突破休谟问题的无限回溯。George Smith 也让我意识到了 Von Neumann 也有类似的见解。只有在电脑一个世代的发展后，通过非常细致的模型，他们的远见才得以证明。而我们不能忽视的一点是，在宣言一种在理论上的可能性后，需要其他人的研究才能揭示这个可能性是否在实际上能被做出来。在人工智能出现之前，对于休谟问题能够通过计算机的概念迭代解决的信念为人们提供了勇气，但使得心理学家与哲学家感到不堪。

（*4）Margaret Boden的观点对我最初文本有帮助很大。她在Artificial Intelligence and Natural Man (Harvester, 1977) 中提供了有助于哲学家理解人工智能的导论。

1. J. Weizenbaum, Computer Power and Human Reason (San Francisco: Freeman, 1976): p. 179, credits Louis Fein with this term.

2. Cf. also Content and Consciousness.

3. Cf. Zenon Pylyshyn, “Complexity and the Study of Artificial and Human Intelligence,” in Martin Ringle, ed., Philosophical Perspectives on Artificial Intelligence (Humanities Press and

Harvester Press, 1978), for a particularly good elaboration of the top-down strategy, a familiar theme in AI and cognitive psychology. Moore and Newell's "How can MERLIN Understand?" in Lee W. Gregg, Knowledge and Cognition (New York: Academic Press, 1974), is the most clear and self-conscious employment of this strategy I have found.

4. See also Judson Webb, "Gödel's Theorem and Church's Thesis: A Prologue to Mechanism," Boston Studies in the Philosophy of Science, XXXI (Reidel, 1976).

5. Wilfrid Sellars, Science, Perception and Reality (London: Routledge & Kegan Paul, 1963): pp. 182ff.

6. Terry Winograd, Understanding Natural Language (New York: Academic Press, 1972), pp. 12ff.

7. Cf. Correspondence between Weizenbaum, et al. in Communications of the Association for Computing Machinery; Weizenbaum, CACM, XVII, 7 (July 1974): 425; Arbib, CACM, XVII, 9 (Sept. 1974): 543; McLeod, CACM, XVIII, 9 (Sept. 1975): 546; Wilks, CACM, XIX, 2 (Feb. 1976): 108; Weizenbaum and McLeod, CACM, XIX, 6 (June 1976): 362.189
8. J. Weizenbaum, "Contextual Understanding by Computers," CACM, X, 8 (1967): 464–80; also Computer Power and Human Reason.

9. Cf. Pylyshyn, op. cit.

10. Cf. Weizenbaum, Computer Power and Human Reason, for detailed support of this claim.

11. Cf. Jerry Fodor, "The Appeal to Tacit Knowledge in Psychological Explanation," Journal of Philosophy, LXV (1968); F. Attneave, "In Defense of Homunculi," in W. Rosenblith, Sensory Communication (Cambridge: MIT Press, 1960); R. DeSousa, "Rational Homunculi," in A. Rorty, ed., The Identities of Persons (University of California Press, 1976); Elliot Sober, "Mental Representations," in Synthese, XXXIII (1976).

12. See, e.g., Daniel Bobrow, "Dimensions of Representation," in D. Bobrow and A. Collins, eds., Representation and Understanding (New York: Academic Press, 1975).

13. W. A Woods, "What's In a Link?" in Bobrow and Collins, op. cit.; Z. Pylyshyn, "Imagery and Artificial Intelligence," in C. Wade Savage, ed., Minnesota Studies in the Philosophy of Science, IX (forthcoming),

and Pylyshyn, “Complexity and the Study of Human and Artificial Intelligence,” op. cit.; M. Minsky, “A Framework for Representing Knowledge,” in P. Winston, ed., The Psychology of Computer Vision (New York, 1975).

14. Cf. Winograd on the costs and benefits of declarative representations in Bobrow and Collins, op. cit.: 188.

15. See, e.g., Winston, op. cit., and Pylyshyn, “Imagery and Artificial Intelligence,” op. cit.; “What the Mind’s Eye Tells the Mind’s Brain,” Psychological Bulletin (1972); and the literature referenced in these papers.

16. See e.g., Pylyshyn’s paper op. cit.; Winograd in Bobrow and Collins, op. cit.; Moore and Newell, op. cit.; Minsky, op. cit.