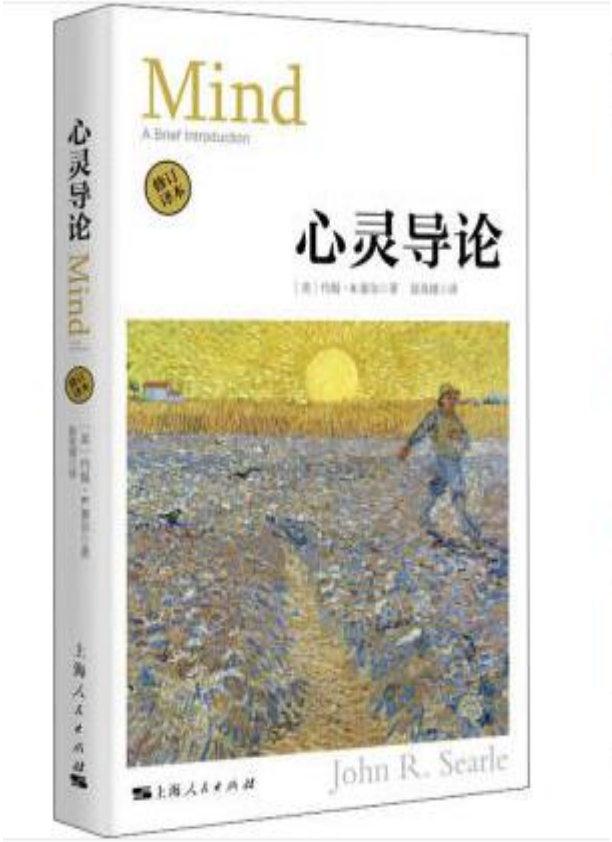
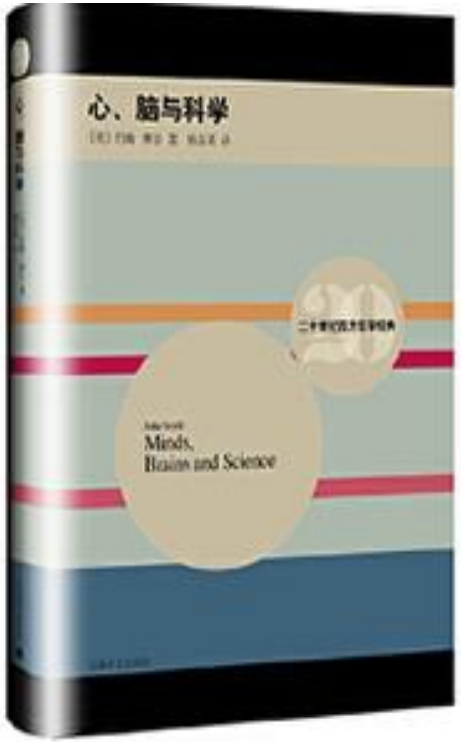
第四讲　“中文屋实验”及其哲学论争

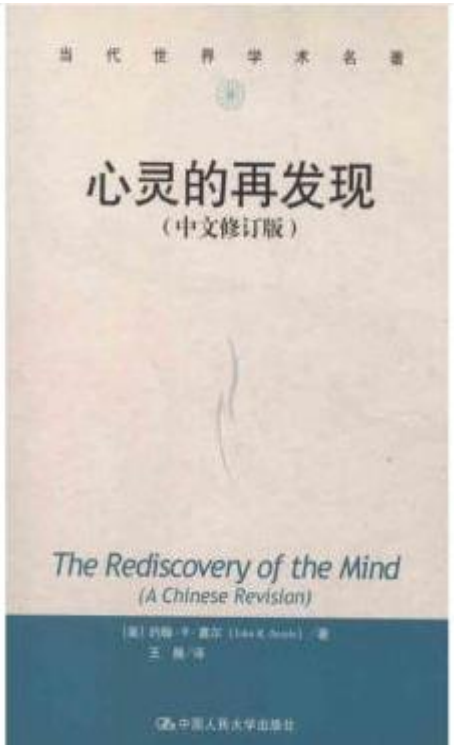
华东师范大学哲学系　潘斌

第一节： 塞尔学述

约翰·塞尔（J.R.Searle）是当今世界最著名、最具影响力的哲学家之一。他于1932年出生在美国科罗拉多州丹佛市，1949-1952之间就读于威斯康星大学，1955年获罗兹（Rhodes）奖学金赴牛津大学学习，并获哲学博士学位。他曾师从牛津日常语言学派主要代表、言语行动理论的创建者奥斯汀（J.L.Austin），深入研究语言分析哲学。1959年返美，并一直在加州大学伯克利分校任教，后当选美国人文科学院院士。







塞尔代表性著作

- 《心、脑与科学》（1984）
- 《心灵的再发现》（1991）
- 《心灵、语言和社会》（1998）

塞尔心灵哲学概况

- （一）心灵哲学的革命
- “我认为，这个问题之所以难以回答，一部分原因是由于我们坚持用过了时的17世纪的词汇，来讨论这个20世纪的难题”（P7《心、脑与科学》）。
- “无论是二元论还是唯物主义，两者都建立在一系列错误的假定之上。主要的错误假定就是认为，如果意识真的是主观的、质的现象，那么它就不可能是物理世界的组成部分。的确，如果按照自17世纪以来这些术语被定义的方式，那么，这种假设的界定就是对的。”（P51《心灵、语言和社会》）。
- “我打算表明这套词汇是过时的，这些假定是错误的”（P7《心灵的再发现》），
- “既然我使用这些超过三百年哲学传统的术语行不通，最好可以一起抛弃这一词汇表”（P49，同上）。

二元论困境的解决方法

- “哲学中某些最有意义的问题是那些产生于两个默认点之间的直接冲突、甚或是逻辑不一致的问题。”（P12《心灵、语言和社会》）
- 面对这种危机，“要提醒我们自己注意的事实，提醒我们自己注意我们实际上所知道的东西，这向来都是一个好的想法。”（P52，《心灵、语言和社会》）“我相信，这种方法是促成哲学进步的方法之一”（P54，《心灵、语言和社会》）。
- 按照这种思路，对于粒子组成的物质世界和心灵之间的相容性难题，我们所能遵循的方法就是：“提醒我们自己注意关于世界如何运作我们所知道的东西”（P55，《心灵、语言和社会》）。
- “只是你接受了带有心理的和物理的、心灵和物质、精神和肉体这些相互排斥的范畴的词汇表时，传统的问题才会产生”（P53，《心灵、语言和社会》）。
- “要解决这些矛盾，就得抛弃这些定义。现在我们有足够的生物学知识去知道这些定义是不符合事实的”（P52，《心灵、语言和社会》），
- “我们应该停止说那些明显是错误的东西，认真地接受这一格言，可能使心灵的研究发生革命”（P207，《心灵的再发现》），“因此，最好是完全拒绝唯物主义和二元论这些词汇，并重新开始”（P55，《心灵、语言和社会》）

塞尔如何批评“二元论”的错误

- **1.人为分割心灵与身体：**塞尔认为，二元论的核心错误在于它将心灵和身体分成了两种截然不同的实体。这种划分将心灵视为一种神秘的、不可测量的东西，与物质世界隔绝开来。塞尔指出，这种分割是没有必要的，也是错误的，因为心灵和身体实际上是同一个系统的不同表现形式。对塞尔而言，意识是大脑的功能或属性，而不是与大脑独立的东西。心灵现象虽然有主观体验的特性，但它依然是生物学现象，与身体（尤其是大脑）的活动密切相关。

- **2.二元论无法解释心灵如何与身体互动**：塞尔认为，二元论无法提供一个合理的解释，说明非物质的心灵如何与物质的身体发生互动。笛卡尔的二元论使心灵和身体在本质上是独立的，心灵不受物质世界的物理规律支配。因此，它无法解释心灵如何通过大脑对身体产生作用，或者如何通过身体感知外界事物。这种互动问题（“心身交感问题”）是笛卡尔二元论长期以来无法解决的难题。塞尔认为，既然意识依赖于大脑的物理活动，试图将二者割裂开来是不合理的。

塞尔如何批评“二元论”的错误

- **3.意识现象并不神秘：**二元论倾向于将意识视为神秘的、超自然的现象，超出了科学的解释范围。塞尔则主张，意识是生物学的一部分，并非神秘的存在。尽管意识具有主观性，但它仍然是大脑神经活动的结果，可以在自然科学框架内进行研究。塞尔强调，意识作为大脑的功能之一，虽然与物理现象有区别，但并不因此而具有超自然或不可解释的特性。二元论错误地赋予了意识某种神秘性，使得意识看起来无法被科学研究和理解。
- **4.二元论是对心灵的错误范畴化：**塞尔批评二元论将心灵和物质当作两种不同范畴的存在，这种分类本身就是误导性的。他认为，意识其实是物质（尤其是大脑物质）的一种属性，就像硬度是固体物质的属性一样。大脑的物质结构能够产生主观的意识体验，而不需要引入一个独立的、非物质的心灵实体。他认为二元论的根本问题在于一种范畴错误（category mistake）：它试图将意识归入一个与物质完全不同的范畴，而不是认识到意识实际上是物质系统（大脑）的某种现象。

二元论的错误

- 第一处错误在于，“心”和“物”这一对术语的对立是虚假的；其二，按照如今物理学理论，电子不是“物的”；其三，本体论的核心问题是“为了让我们的经验陈述为真，我们在世界上的位置必须是什么”，而不是“世界上存在着什么种类的东西。
- **“唯物论在某种意义上是二元论的最美的花朵”**

二元论的替代方案：“生物学的自然主义”的心灵理论（核心要点）

- 1.意识是大脑的生物特性：意识是大脑生理过程的一个自然产物，与其他生物学现象一样，是由大脑神经元的复杂活动所产生的。因此，意识并不超越物质，也不是单纯的物质现象，而是一种高度特殊的生物现象。
- 2.意识的主观性：虽然意识是由大脑产生的，但它具备主观性，这是意识不可还原的核心特征之一。塞尔指出，意识体验是主观的，无法通过第三人称的物理描述完全理解或解释（这与物理主义的还原论相对立）。然而，主观性并不意味着神秘性，意识的主观性质可以被科学地研究和解释。

二元论的替代方案：“生物学的自然主义”的心灵理论

- 3.非还原的自然主义：尽管塞尔认为意识是一种生物现象，但他反对彻底的还原主义，即试图通过纯粹物理或化学过程解释意识。他认为意识具有独特的主观经验，不能被简单地还原为神经元的电化学活动。
- 4.拒绝二元论：塞尔拒绝传统二元论中的“心”和“物质”的划分。他认为，意识并不是一种与大脑完全不同的实体，而是大脑的一个属性。与此类似，水的流动性是水分子的属性，意识也是神经元活动的一种“高阶”表现。
- 塞尔的生物学自然主义可以被理解为一种试图通过生物学来解释意识的哲学立场，既承认意识的物质基础，又保留了其主观体验的独特性。这一理论强调意识的自然性，而不是超自然或神秘现象，同时反对彻底的物理主义还原论和笛卡尔式的二元论。

塞尔心灵哲学的意义

- 塞尔的这种理路容易让人想起托马斯.库恩对科学革命所进行的描述。自笛卡尔以来的二元论传统即“常规科学”范式；心身问题引发的困境即塞尔的“科学危机”；解决“哲学危机”的方法就在于转换“范式”，对塞尔来说，就是抛弃以往的心灵哲学概念，从头开始。
- 在塞尔的心灵哲学当中有一个生死攸关的“赌注”：对一切试图解决身心问题的哲学理论的绝望、对以往心灵哲学研究模式的绝望，赋予心灵现象以本体论地位，并且这种本体论地位是主观意义上的。这是塞尔心灵哲学最富创造性和颠覆性的观点之一。塞尔认为，假如这一点是对的，心灵哲学的研究重心就应该转向对意识的研究，如同物理学对物质世界的研究一般。
- 这样一来，塞尔又回到了形而上学的高度，对各种反实在论进行驳斥，捍卫一种外部实在论和真理符合论。这表明，心灵哲学的研究始终离不开对各种理论背后的本体论研究。塞尔心灵哲学的合法性还有待历史的检验，一些本体论问题还有待澄清。

第二节“中文屋论实验”

- 在《心、脑与程序》一文中，约翰·塞 尔（John Searle）首次提出用来反对强人工智能的著名思想实验—“中文屋”。他以罗杰·尚克(Rogers Schank) 等人构想的“故 事—理解”程序为依托，通过设想可以完整示例该程序，然而却缺少语义理解的“中文屋”，批判强人工智能的核心观点—“程序即心灵”。
- “故事—理解”程序具有如下特征：该程序设计的目的是试图模仿人类理解故事的能力。就这种理解能力而言，当被问及与故事相关却没有直接提及的信息时，人类可以表现出能够根据故事情景予以推断的能力。尚克等人设想：如果存在这样一台具有“故事—理解”程序脚本的机器，当人们向它呈现相似的故事情节并问其同样问题时，机器如果能够像人类那样给出预期的答案，那么这台程序化的机器就是真实地理解了故事的意义，进而也就具有了真实的认知能力。
- https://www.bilibili.com/video/av583026386/

塞尔中文屋实验图示


I'm just manipulating squiggles and squoggles to produce Chinese language behavior. But I don't understand Chinese. This rule book is in English.
在这屋裡的任何人或物,一定懂中文。
[Whoever or whatever is in that room is an intelligent Chinese speaker!]
Take a squiggle-squiggle sign from basket number 1 and put it next to a squoggle-squoggle sign from basket number 2
1
2
3

塞尔中文屋实验

中文屋实验过程：如果把一位只会说英语的人关在一个封闭的房间里，他只能靠墙上的一个小洞传递纸条来与外界交流，而外面传进来的纸条全部由中文写成。这个人带着一本写有中文翻译程序的书，房间里还有足够的稿纸、铅笔和橱柜。那么利用中文翻译程序，这个人就可以把传进来的文字翻译成英文，再利用程序把自己的回复翻译成中文传出去。在这样的情景里，外面的人会认为屋里的人完全通晓中文，但事实上这个人只会操作翻译工具，对中文一窍不通。屋外的人认为屋内的人对屋外提出的用汉语讲述的“问题”，进行了精彩的“回答”；而实际上，屋内的人根本没有理解汉语符号的意思，他不过是做了一件文本转换的工作。

塞尔中文屋实验

中文屋实验对（强）人工智能的批判性：

“程序本身不能够构成心灵，程序的形式句法本身不能确保心智内容的出现。”

前提 1：计算机程序是形式的，或者说是句法的；

前提 2：心具有心理内容，具体说是有语义的；

前提 3：句法自身既不构成也不足以产生语义。

结论 1：任何计算机程序自身都不足以使一个系统具有一个心灵。

前提 4：脑产生心。

结论 2：任何其他事物，如果产生心，就必须具有和脑产生心相同的因果力。

结论 3：对于任何我们可能制作的、具有相当于人的心理状态的人造物来说，单凭一个计算机程序的运算是不够的。这种人造物必须具有相当于人脑的能力。

结论 4：脑产生心的方式不能是一种单纯操作计算机程序的方式。

中文屋论证结构

前提：

１．脑产生心。

２．语法不足以满足语义。

３．计算机程序是完全以它们的形式的或语法的结构来定义的。

４．心具有心理的内容，具体说具有语义内容。

中文屋论证结构

结论：

1 任何计算机程序自身不足以使一个系统具有一个心灵。简言之程序不是心灵，它们自身不足以构成心灵。

2 脑功能产生心的方式不能是一种单纯操作计算机程序的方式。

3 任何其他事情，如果产生心，应至少具有相当于脑产生心的那些能力。

4 对于任何我们可能制作的，具有相当于人的心理状态的人造物来说，单凭一个计算机程序的运算是不够的。这种人造物必须具有相当于人脑的能力。

中文屋论证结构

批判要点之一：批判将语法等同于语义、程序等同于心灵；

批判要点之二：主张人工智能事物至少应具有大脑产生心灵的因果力。

进一步需论证如下两点：

其一，人类（和动物）的意向性是大脑的因果特性的产物，这是心理过程同大脑真实因果关系的经验事实，这说明某种大脑过程对于意向性来说是充分的；

其二，示例一个计算机程序本身绝对不可能构成意向性的充分条件。

中文屋论证结构

就逻辑论证而言，其有效性依赖于“语法不等同于语义”，它用于支撑“中文屋”论证中的批判要点，表述为如下推理形式。

1．语法不足以满足语义。

2．计算机程序是完全以它们的形式的或语法的结构来定义的。

3．心灵具有心理的内容，具体说具有语义内容。

结论：

程序不是心灵，它们自身不足以构成心灵。

中文屋论证结构

就经验论证而言，其有效性依赖于“复制不等同于模拟”，它用于支撑“中文屋”论证中的主张要点，表述为如下推理形式。

前提：

1．模拟不等同于复制。

2．大脑具有产生心灵的能力。

3．计算机程序仅仅作为工具对心灵进行模拟。

结论：

凡是具有心灵的人造物至少应复制等同于大脑产生心灵的因果力。

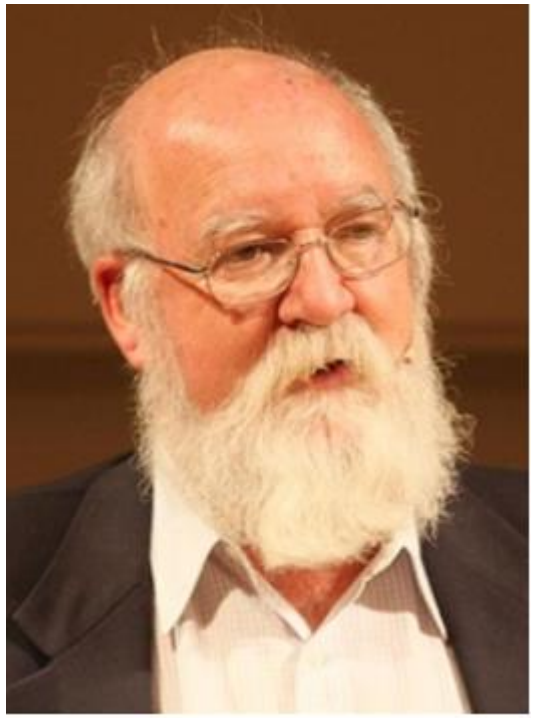
对中文屋实验的一些批评

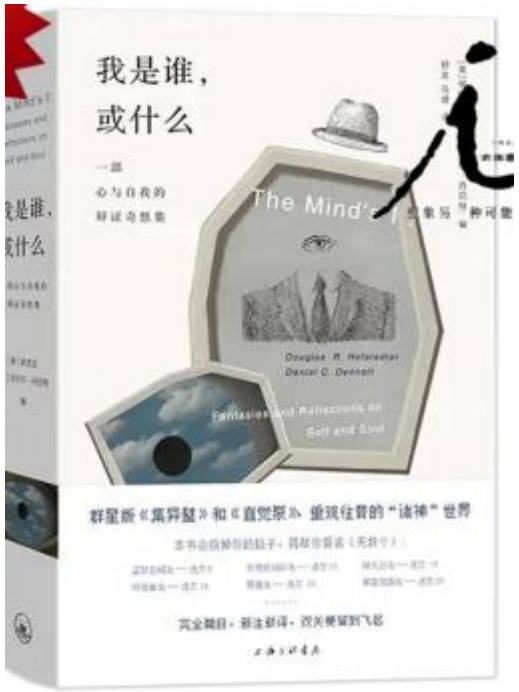
英国哲学和心理学教授Boden就指出，中文屋中的塞尔不可能什么都没理解。屋中的塞尔至少是理解了规则和指令，否则他不可能对汉语符号进行正确的操作。那么，这就相当于计算机对程序性语言的理解。这显然是塞尔的思想实验所忽略的一点，尽管塞尔说的是对汉语的理解，但不能因为它不能对汉语理解就认定它没有理解能力。

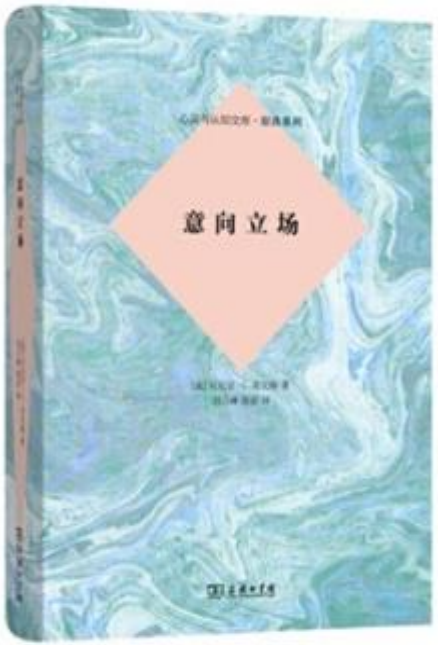
英国的神经心理学教授R．Gregory指出，把人工智能局限于形式程序是过分了，这即使是对于外在的纯逻辑或形式的数学来说也过于严格了。其次是塞尔认为，任何人造的机器，都不具有产生意向性或意义理解的因果力量，而只有脑具有这种因果力量。Boden则指出，塞尔主张只有生物的脑能够产生和支持意向性，这仅仅是直觉的解释或哲学的假定。

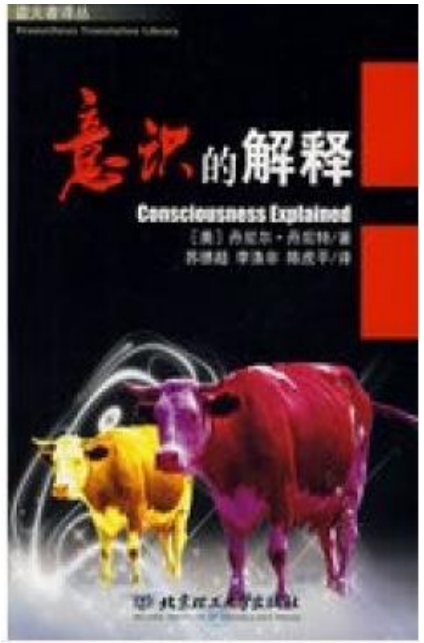
第三节 对中文屋论证的挑战之一

丹尼尔·丹尼特（Daniel Dennett），1942年生于美国波士顿，1965年取得牛津大学哲学博士，现任塔夫茨大学哲学教授与认知科学研究中心主任。代表作《意识的解释》（Consciousness Explained）。









对中文屋论证的挑战之一

丹尼特关于人工智能的哲学观点

1.功能主义与心灵哲学：丹尼特是功能主义者，认为心灵的本质在于它执行的功能，而不是它的物质基础。因此，他主张我们不必拘泥于生物学上的大脑来理解意识，而应关注系统如何处理信息。对于人工智能，丹尼特认为如果一个系统能够执行足够复杂的信息处理任务，那么它也可以被认为具备某种形式的心灵或意识。

2.心灵的计算主义理论：丹尼特支持一种被称为“计算主义”的心灵理论，即认为心灵的运作类似于计算机程序的运行。他认为意识并非神秘的，而是通过一种“虚拟机”或“软件”来实现的复杂信息处理。这个观点暗示，如果我们能够创建足够复杂的人工智能系统，这些系统可能会表现出类似于人类的心智能力。

对中文屋论证的挑战之一

丹尼特关于人工智能的哲学观点

3. 意识的“错觉主义”：丹尼特的著作《意识的阐释》（Consciousness Explained）中，他提出意识可能只是一种复杂的认知错觉，或者说是一种由认知功能生成的虚拟现实。他认为意识并不是一个独立的、统一的实体，而是众多非意识的过程和表征的产物。这在AI哲学中的应用是：即便一个人工智能没有“真正的”主观体验，它也可以通过复杂的计算过程表现出类似意识的行为。

4. “心智设计”与进化论类比：丹尼特经常将生物进化与人工智能的设计类比。他认为，人工智能的开发过程可以被看作是一种类似自然进化的过程，在这个过程中，简单的算法或模型逐渐变得更加复杂，直到它们能够执行高度智能化的任务。这种观点预示着AI发展可以通过类似进化的过程（比如机器学习和优化算法）实现智能的增强。

对中文屋论证的挑战之一

丹尼特关于人工智能的哲学观点

5. 意图立场：丹尼特的“意图立场”理论在人工智能哲学中具有重要意义。他认为我们可以通过赋予一个系统“意图”来解释其行为，即使它本身并没有真正的意识或意图。例如，当我们说“计算机想要赢得这场棋局”，实际上我们是在以意图立场来看待它，而不是说它真的有主观的意图。这意味着，我们可以对人工智能的行为进行类似的解读，即使它没有“内在”的意图或意识。
6. “弱AI”与“强AI”的区分丹尼特支持“弱AI”的观点，认为现有的人工智能系统并不真正具备意识，它们只是模拟智能的机器。弱AI系统能够执行一些看似有意识的任务（例如语言理解、物体识别），但这些系统的背后并没有真正的自我意识或主观体验。与之相对，“强AI”则假设机器可以拥有与人类相同的意识，但丹尼特对此持怀疑态度。

丹尼特对中文屋论证的批评

问：**“一个读者同美国国家图书馆再加上其服务人员听起来不正是一个超级系统吗？（《计算机神话：一次交流》）**

1.丹尼特承认超级系统本身并不赋予它的子系统某种特殊的力量或者特性，但他却确信超级系统本身拥有比子系统更多的力量和特性。

2. 作为超级系统的整个“中文屋”是否能够理解中文故事。“中文屋”思想实验由部分推及整体的模式，由于忽略了整体所具有的独特特性而应该遭到质疑。

丹尼特对中文屋论证的批评

塞尔的回复：

1. “内化了的中文屋”：设想屋中的个体（塞尔）内化到整个系统（“中文屋”）的所有组成部分中去，个体记住了规则书中所有的规则以及中文符号数据库中所有的信息，并且可以通过心算方式完成相关计算。这样一来，个体与整个系统融为一体，甚至可以摆脱屋子在室外工作。尽管如此，“内化了”的个人依然无法理解中文，当然整个系统也就不能理解中文。

2. 系统回应者论证：

前提：

１．部分没有属性 A。

２．部分蕴含在整体之中。

结论：

整体也就没有属性 A。

塞尔指出：系统回应者认为“中文屋”思想实验的有效性依赖于上述错误的推论形式，

塞尔对丹尼特批评的回应

1. 论证

前提：1.整体具有部分的特性，并且具有部分没有的特性。

2.部分没有A特性。

结论：整体具有A特性。

2.系统回应者的根本问题在于，他们没有区分“中文屋”思想实验的描述形式与蕴含于其中的论证结构和形式之间的不同。他们仅仅以思想实验的外在的描述形式为基点推断其可能存在问题，而没有认识到思想实验的有效性实际上依赖于其背后蕴含的论证结构和形式的有效性。

2.为系统回应实际上是对“中文屋”论证的直接妥协。因为它已经承认子系统所具有的符号操作本身不能够实现语义，因此，它们试图从系统的其它部分中寻找到符号，要么是规则书、纸、笔、黑板等等的叠加；要么寻求所谓的其它子系统。而系统回应者所妥协的部分正是“中文屋”论证的要点所在。

三 中文论论证的挑战之二

1. “机器人回应”：一些强人工智能的支持者试图通过强调机器同外界的因果关联来回应“中文屋”的批判，这类回应被塞尔称为“机器人回应”。

2. 基本主张：我们将具有程序脚本的计算机放置在一个机器人“大脑”中，这台计算机不仅接收形式符号作为输入，发出形式符号作为输出，并且可以实际地驱动机器人做一些同人类日常活动十分相像的行为，如感知、行走、吃喝等。机器人拥有电子眼可以观察到外在世界，拥有手臂和腿可以与外界互动，这些行为都由它的计算机“大脑”所控制。机器人回应者认为，此种情景下所展现的与外界互动的机器人可以说真正地具有“理解能力”和相关的其它心理状态。

3.斯蒂文·哈纳德（Stevan Harnard是机器人回应的代表，他的批判见之于《心灵、机器和塞尔》以及《心灵、机器和塞尔2：“中文屋”论证的对与错》

来自“机器人回应”的批评

首先，要区分符号操作与意义理解。

哈纳德认为，塞尔的中文屋实验正确指出了符号操作和语义理解之间的差异。在实验中，符号的操作（如通过规则处理中文字符）并不意味着对符号的理解。但哈纳德进一步指出，这并不是强AI不能实现理解的证明，只是说明符号操作本身不足以产生理解。他强调，理解需要超越符号操作，而进入一个可以赋予符号意义的系统或机制。

哈纳德认为模拟与执 行 不 同。模 拟 是抽象的，执行是具体的；模拟是形式和理论的，执行是实践和物理的。“中文屋”只是模拟的模拟而不是执行，真正成功的模拟必须要捕捉到与成功执行相关的所有功能特征。

来自“机器人回应”的批评

其次，“符号接地”问题。

哈纳德提出了“符号接地假说”（Symbol Grounding Hypothesis）作为对中文屋实验的回应。他认为，要实现真正的理解，符号必须“接地”于感知经验和世界中的实际事物，而不仅仅是在符号之间进行形式化的操作。也就是说，人工智能系统如果仅仅依赖符号操作来处理信息，那么它无法获得真正的语义理解。符号必须与感知、行动和经验相关联，这样才能真正“理解”其代表的内容。

3. 动态系统与具身认知。

哈纳德还批评了中文屋实验对智能的狭隘定义，他认为智能和理解不仅仅是符号操作的结果，而是身体、环境和认知系统共同作用的结果。换句话说，AI系统需要具备某种形式的“具身认知”，即通过身体与外部世界互动来获得理解，而不仅仅是在头脑中处理抽象符号。

来自“机器人回应”的批评

4. 实际应用中的智能

哈纳德还强调，在实际应用中，许多人工智能系统能够解决复杂问题并表现出智能行为，尽管它们没有像人类一样的意识。这些系统通过学习和适应来提高其性能，这说明了智能并不局限于人类的经验。

5.意识的神经基础

哈纳德在他的著作中探讨了意识的神经基础，认为科学可以解释意识如何与大脑的物理过程相联系。他认为，尽管计算机可能没有意识，但这并不排除它们能够展示某种形式的智能。

总的来说，哈纳德通过强调理解的多元性、智能行为的实际应用以及对意识的科学解释，对塞尔的中文屋实验进行了深入而批判性的回应。他的观点促使人们重新思考智能的定义及其在人工智能中的实现方式。

塞尔对机器人回应的答复

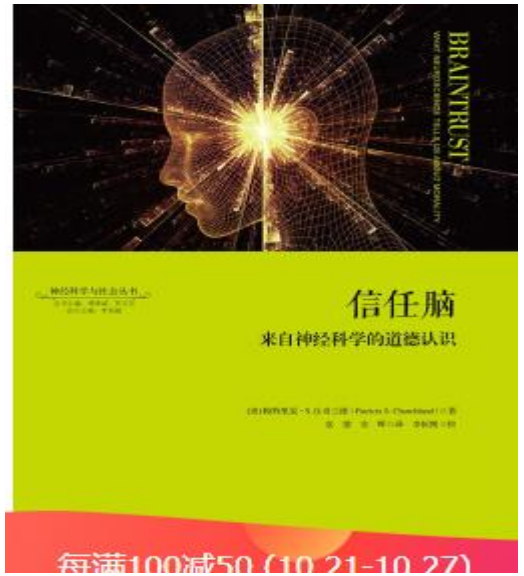
1. 塞尔声称机器人回应已经默许了认知状态不只是纯粹形式符号的操作；
2. 机器人回应所添加的感知和行动能力对于“理解”不起任何作用。他通过“机器人脑中的‘中文屋’”进行了答复，设想将机器人“头脑”中的计算机替换为“中文屋”，塞尔依旧被关进屋中，依然像先前那样按照英文规则书操作一批批中文符号，此时此刻，塞尔并不知道中文符号是来自附加在机器人“头脑”之上的电视摄像机，也不知道那些反馈出去的中文符号可以内在地驱动机器人行动。那么此种情境下对语义的理解从哪里获得呢？
3. 塞尔指出问题的关键是机器人回应所增添的条件或功能没有为“理解”的产生增添任何新的东西。机器人的行动始终是程序化的，这同尚克的“故 事—理解”计算机没有任何本质上的区别。
4. “机器人与周围世界的相互因果作用是与问题无关的，除非在某个心灵中这种因果的相互作用得到表达。但是，如果所谓心完全是由一套纯形式的 语 法 运 算 组 成，那么就根本不可能有这种表达。”

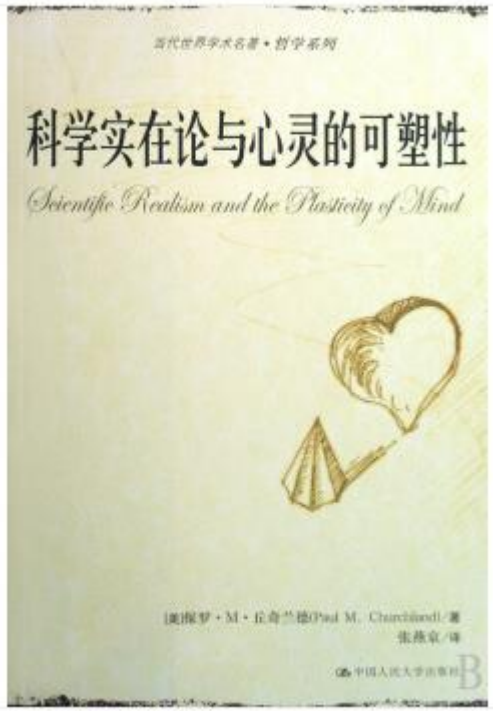
三、 塞尔中文屋论证的挑战之三

1.管道工与中文体育馆————大脑模拟者（联结主义）应答。

2.联结主义认为人工智能的实现同大脑功能的独特执行方式有关，特别是当人们发现大脑这种执行方式是并行执行，而非串行执行时，这种借鉴意义更为重要。

3.丘奇兰德夫妇（P·M·Churchland & P·S·Churchland）。





丘奇兰德夫妇对中文屋论证的批评

丘奇兰德：如果设计出这样一个程序，它并不表征我们关于世界的任何信息，然而却模拟一个母语为中文的人在理解和回答中文故事时，大脑中神经突触附近神经纤维激发的实际序列。具有这种程序脚本的机器，将中文故事及其问题视为输入，模拟大脑在这一过程中的形式结构，将中文答案作为输出，同时该机器在运作过程中不只是伴随单一的串行程序，而是伴随着整个一组并行运作的程序。此种情境下，我们就可以断定这个机器理解了中文故事，如果我们拒绝承认这一点，那么也就同时否认了讲中文母语的人能够理解中文故事了。

塞尔以“管道工”思想实验回应批判

1.场景:一个人在操作用阀门连接起来的一套精密的水管系统。当那个人接到中文符号的时候，他就查询用英语书写的规则书，从而得知哪一个阀门应该被打开，哪一个应该被关闭。每一处水管连接对应着母语为中文的人大脑中某个神经突触，整个系统被装配成可以随着适当的水阀的开启得到准确地触发，而中文问题最终在水管系统的终端喷发出来。

2.问题: 系统中是否“理解”了中文？

3.操作：中文作为输入并且模仿了母语为中文的人大脑中的神经突触形式结构，同样以中文作为输出，然而，那个操作阀门的人依然不理解中文。

4.修正：如果有人认为这种构想不切实际，那么可以想象，在原则上人可以将自己完全内化到整套水管的形式结构中去，在想象中完成神经触发。

塞尔以“中文体育馆”思想实验回应批判

1.场景:不是一间“中文屋”，而是一座中文体育馆，在其中聚集了许多只会说英语的人。这些人可以执行在一个联结主义构造物中作为节点和神经突触的相同的操作。在体育馆中的人通过彼此传递标记，用来模拟联结主义系统中的各个单元，其中绿色的标记代表输入兴奋联结，红色标记表示抑制联结，一个人向另一个人传递标记的数量代表联结的权数。由于有相当数量的人参与模拟，所以一个很好的办法就是给每个参与者一个清单，上面详细说明他们必须向谁传递记号，又有多少个记号应该移交。

2. 操作：通过对这个模拟过程的训练，参与者最终可以熟练地根据教练的指令转换他们手中的清单以及执行相应的操作。

3.结论：中文体育馆中的每一个人同样没有获得任何“理解”，中文体育馆作为整体也没有“理解”。

4.在心灵同生物大脑之间根本不具有任何计算层面。因 为“0和1没有因果能力，它们甚至不存在，除了在观察者的眼中。执行的程序除了执行媒介的能力之外没有因果能力，因为程序没有超越执行媒介的真实存在，没有本体论。

第四节

中文屋实验的哲学延展与当代论争

语义鸿沟与符号操作的局限

从符号到意义的理解“鸿沟”

语境缺失

机器缺乏真实的语境理解和世界知识背景，无法把握符号在具体情境中的实际含义。

经验缺乏

语义理解深度依赖于主体的生活经验和身体感知，这是纯符号系统无法获得的。

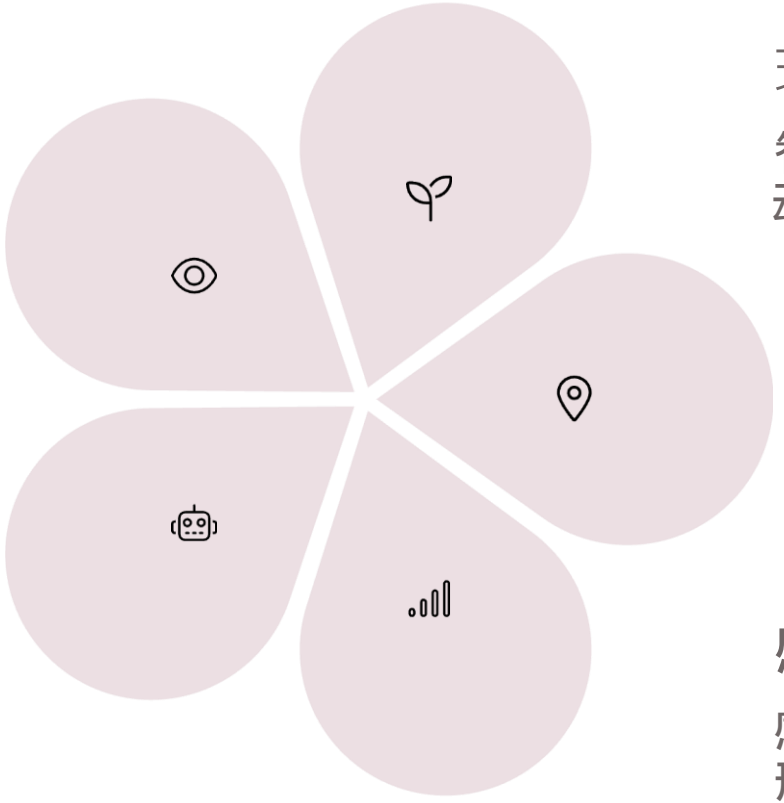
具身认知论（Embodied Cognition）

身体依赖

认知过程依赖于身体的
物理特性和感知运动能力

机器人挑战

体现认知论支持机器人反驳，
挑战传统符号主义AI



环境交互

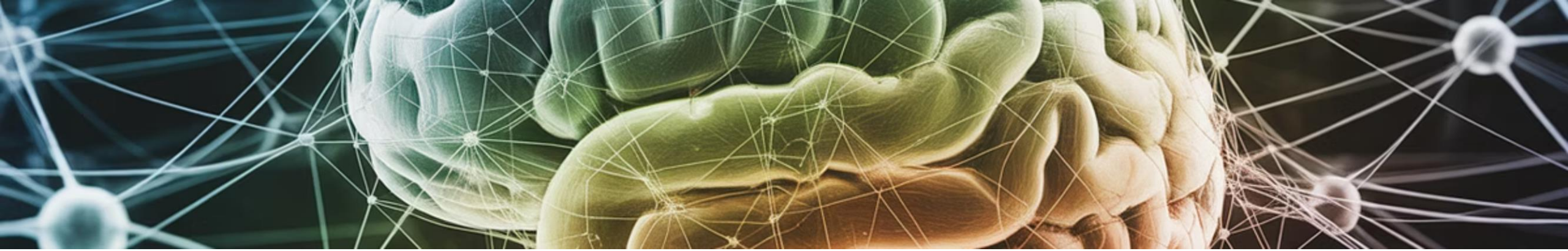
智能产生于主体与环境的
动态交互过程中

行动导向

认知服务于行动，
而非抽象的符号操作

感知整合

感知、行动和认知
形成不可分割的整体



认知科学与神经科学视角

神经元层面

单个神经元并不理解语言或概念，仅进行简单的电化学信号传递

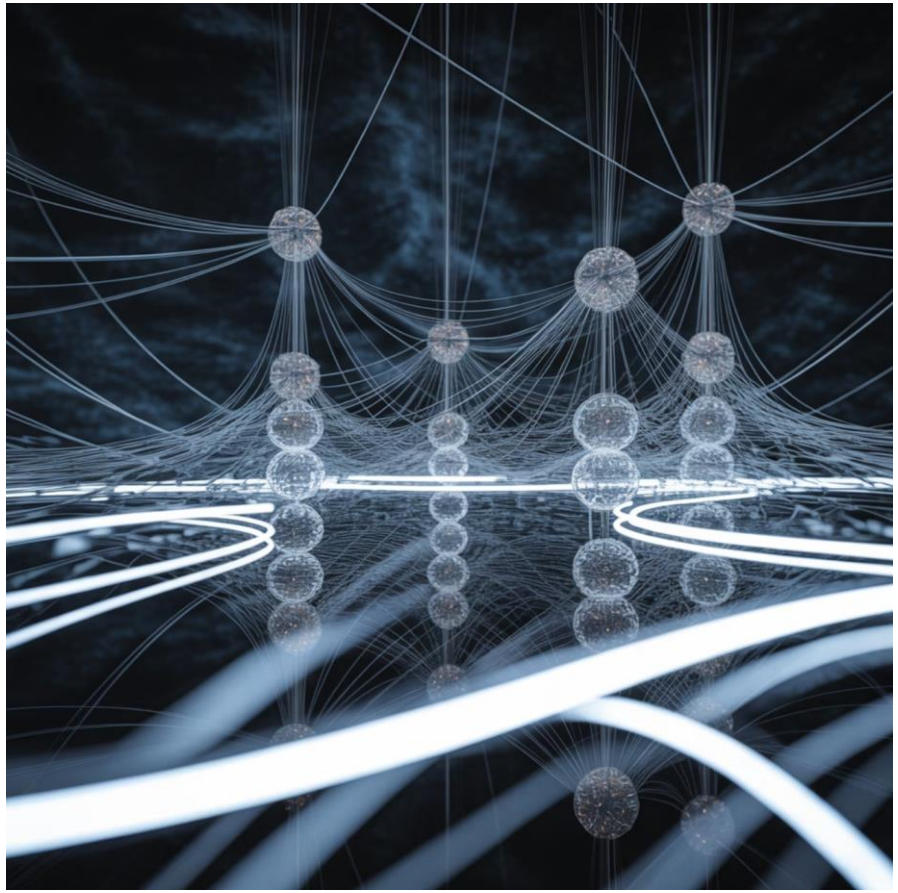
网络涌现

复杂神经网络的整体表现出智能行为，这是涌现现象的体现

哲学难题

涌现的智能是否等同于真正的理解？这挑战了塞尔的论证

机器学习与深度学习的挑战



现代AI的表现

- 深度神经网络能够生成流畅自然的语言文本
- 表现出令人印象深刻的语言模仿能力
- 在特定任务上超越人类水平

本质局限

- 依赖海量训练数据的统计模式识别
- 缺乏对语言真实意义的语义理解
- 无法赋予机器主观意识和真实体验

经验性强人工智能的可能性

1

反对狭义定义

批评者认为塞尔对“理解”的定义过于人类中心主义和狭隘。

2

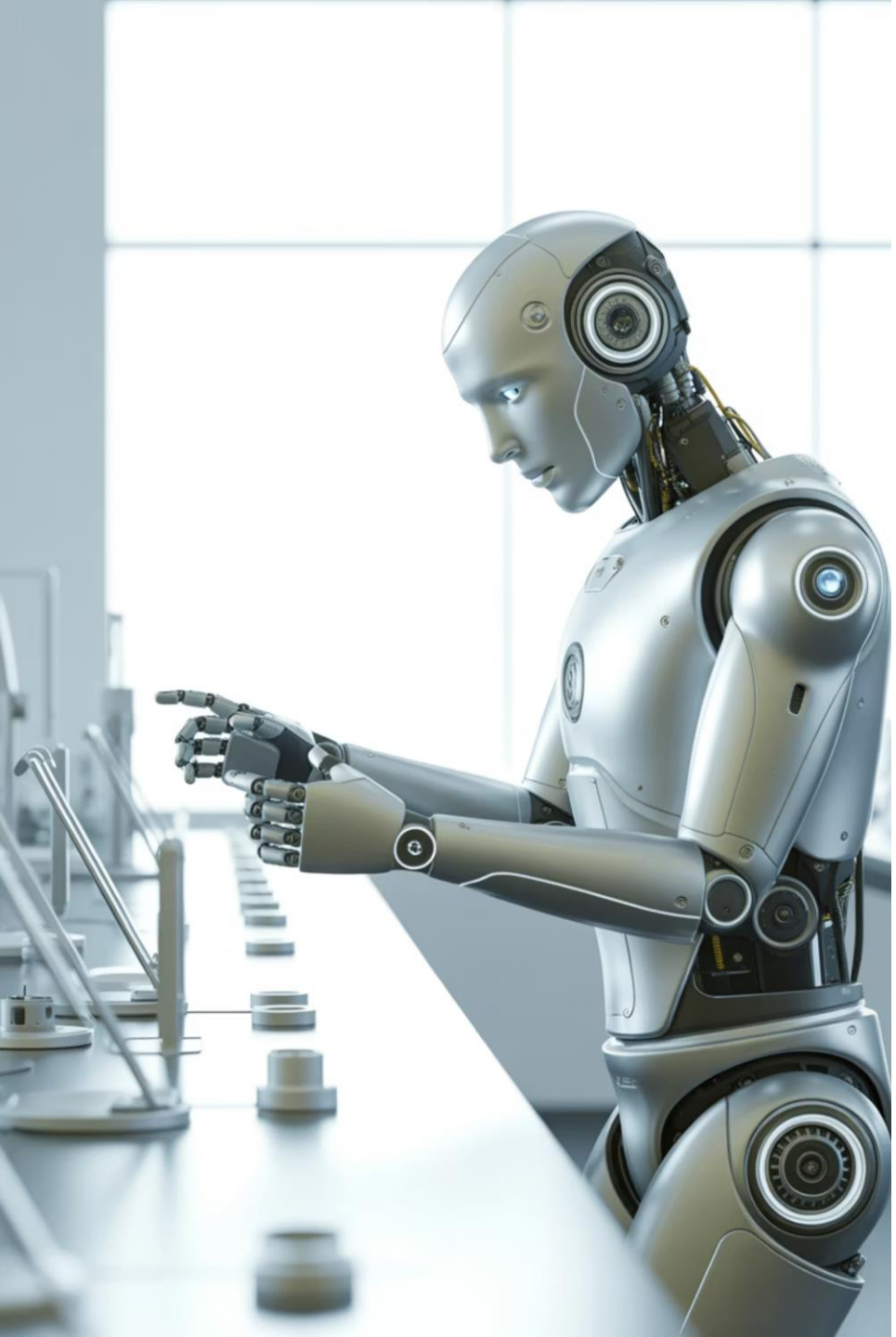
学习与适应

强调机器通过经验学习和环境适应可能获得某种形式的理解。

3

未来展望

认为未来AI或能通过丰富的交互经验发展出真正的理解能力。



具身认知：机器人与环境交互

体现认知理论强调智能主体必须通过
身体与环境的真实交互才能产生真正的认知和理解。

第五节

中文屋实验的现实意义与未来展望

中文屋实验对AI哲学的启示

质疑智能本质

促使我们深入反思机器智能的本质，区分表面行为与内在理解。

理解的复杂性

揭示理解与意识的深层复杂性，不能简化为计算过程。

跨学科研究

促进AI伦理、认知科学和哲学的交叉研究与对话。

技术进步与哲学思考的张力

技术飞速发展

人工智能技术呈现爆炸式增长，应用领域不断扩大

- 大语言模型展现惊人能力
- 机器人技术日益成熟
- AI渗透各行各业

哲学警示意义

塞尔的观点提醒我们保持理性和审慎

- 技术能力不等于真正理解
- 警惕对AI能力的过度夸大
- 重视人类认知的独特价值

图灵测试的局限性

行为主义偏见

图灵测试仅考察外在行为表现，完全忽视内在心理状态和主观体验

表面模仿

机器可能通过技巧性回避和模式匹配通过测试，而非真正理解

中文屋揭示

中文屋实验深刻揭示了图灵测试作为智能判定标准的根本性不足

人工智能的未来路径

体现认知整合

结合体现认知理论，发展具有真实身体交互能力的智能系统。

神经科学启发

借鉴神经科学的最新发现，设计更接近生物智能的AI架构。

语义理解探索

探索真正具备语义理解能力的AI模型，超越纯符号操作。

机器意识的哲学难题

机器是否可能拥有"心灵"？

主观体验的不可还原性

意识体验（感受质）具有第一人称视角，
无法通过第三人称的客观描述完全捕捉

物理主义的挑战

如何从物理过程中涌现出主观意识？
这是心灵哲学的核心难题

机器的特殊性

即使机器展现智能行为，
是否具有内在的主观体验仍是开放问题

伦理与社会影响

误判风险

- 将机器的模仿行为误认
- 为真正理解
- 过度依赖AI系统做出重要决策
- 忽视机器智能的本质局限性

人类挑战

- AI发展对人类认知价值的冲击
- 人类身份认同面临的新问题
- 人机关系伦理界限的界定

中文屋实验的教育价值

批判性思维

培养学生质疑权威、独立思考的批判性思维能力，不盲从技术权威

多维度理解

帮助理解智能与意识的多维度本质，超越简单的二元对立

跨学科对话

促进哲学、计算机科学、认知科学等学科的交叉对话与融合

AI与人类大脑的对比

对比展示人工智能系统与人类大脑在结构、功能和认知能力上的根本性差异。

案例一：图灵测试中的"模仿游戏"

测试场景

机器通过文字对话成功欺骗人类评判者，使其相信自己在与真人交流。

表面成功

- 机器展现了令人信服的语言能力
- 成功通过了图灵测试的标准
- 在行为层面与人类无法区分

深层问题

但这不代表机器真正理解语言的意义，仅是成功的模式匹配和规则执

案例二：谷歌虚拟助手与聊天机器人

高度拟人化

谷歌助手等AI系统展现高度拟人化的语言交互能力，能进行自然流畅的对话。

理解缺失

尽管表现出色，但这些系统仍然缺乏对对话内容的真正语义理解和意义把握。

案例三：AlphaGo与深度学习

1

超越人类

AlphaGo在围棋领域展现超越世界顶尖棋手的卓越表现，创造历史性突破。

2

深度强化学习

通过深度神经网络和强化学习算法，掌握了高超的棋艺水平，

3

意识缺失

尽管棋艺精湛，但AlphaGo不具备自我意识或对围棋美学的真正理解。

案例四：GPT系列语言模型

惊人能力

- 生成流畅自然的文本内容
- 展现广泛的知识覆盖面
- 能够完成复杂的语言任务
- 在多个基准测试中表现优异

本质局限

- 依赖统计模式和概率分布
- 缺乏对文本真实意义的理解
- 无法形成真正的概念和信念
- 不具备意识和主观体验

案例五：机器人感知系统

01

传感器整合

现代机器人配备多种传感器，包括视觉、触觉、听觉等多模态感知能力。

02

环境交互

机器人能够与物理环境进行实时交互，做出适应性响应和行为调整。

03

认知尝试

这是实现体现认知的重要尝试，但仍面临语义理解和意识缺失的局限。

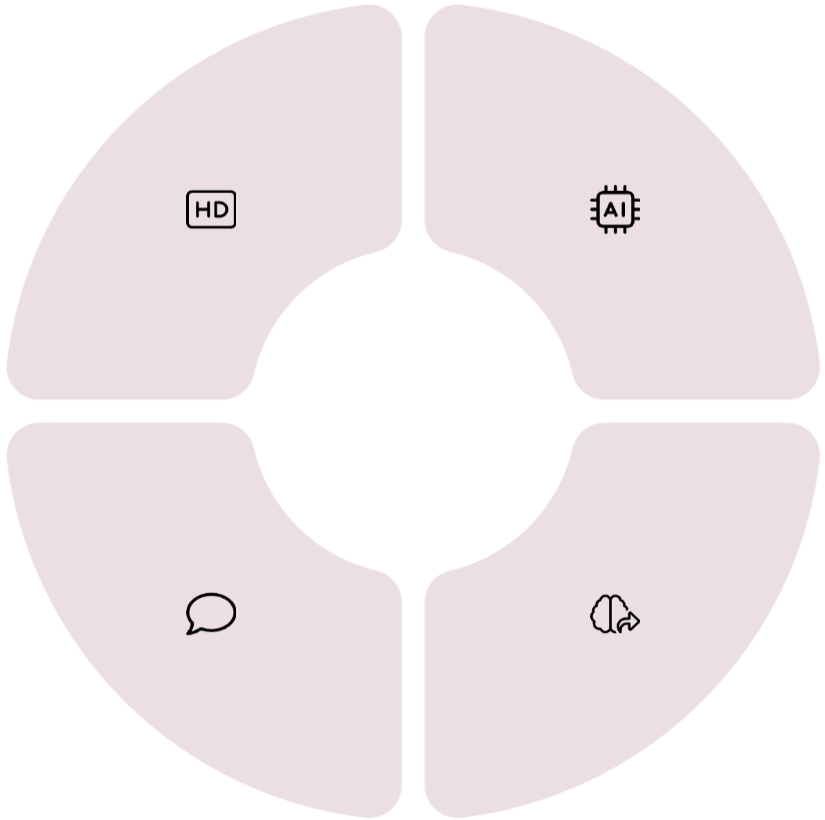
中文屋实验的哲学争议总结

理解的定义

关于"理解"概念的界定存在根本性分歧，不同立场对理解有不同诠释

智能本质

机器智能的本质是什么？是否可能与人类智能等同？争论持续激烈



哲学传统

涉及心灵哲学、语言哲学、认识论等多个哲学传统的核心问题

认知科学

认知科学与哲学在智能问题上的交叉与碰撞，产生丰富的理论成果

塞尔的回应与后续发展

> 坚持语义理解不可被程序替代

立场坚定

塞尔始终坚持语义理解具有不可还原性，无法通过程序的符号操作实现。

反对混淆

强烈反对将机器的模拟行为误认为真正的理解，认为这是概念上的混淆。

持续影响

塞尔的思想持续影响AI哲学、认知科学和人工智能研究的发展方向。

未来研究方向

跨学科融合

整合认知科学、神经科学、AI研究，形成统一的理论框架

意识研究

深入探索意识的本质、产生机制和可计算性问题

具身智能

发展具有真实身体交互能力的体现智能系统

伦理探讨

建立AI伦理框架，规范智能技术的发展和应用

中文屋实验的哲学启示

人类认知的独特性

促使我们深入反思人类认知的独特之处，包括意识、意向性和主观体验

智能的本质追问

引发对"智能"概念的根本性质疑，区分行为智能与真正理解

测量标准反思

质疑传统的智能测量方法，如图灵测试的有效性和充分性

人工智能与人类智能的差异

结构差异

- 硅基计算 vs 碳基生物系统
- 数字符号 vs 神经化学过程
- 离散算法 vs 连续动态系统

功能差异

- 规则执行 vs 语义理解
- 模式匹配 vs 概念形成
- 信息处理 vs 意义创造

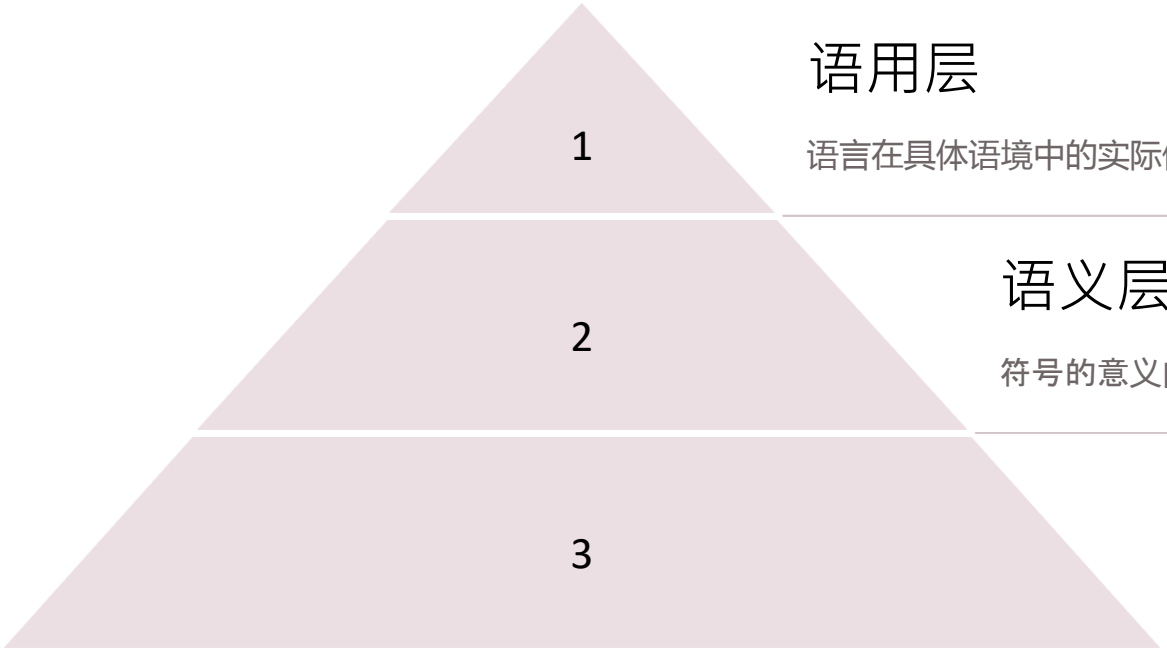
本质差异

- 无意识计算 vs 有意识体验
- 客观操作 vs 主观感受
- 功能实现 vs 现象状态

不可复制性

意识与理解具有不可复制的主观维度，这是机器智能无法逾越的鸿沟

语言理解的多层次结构



语用层

语言在具体语境中的实际使用和交际意图

语义层

符号的意义内容和概念关系

句法层

符号的形式规则和操作方式

机器擅长处理句法层，但难以掌握语义和语用的复杂性，而真正的语言理解需要三个层次的整合。

认知科学中的"涌现"现象

涌现的定义

整体表现出的特性无法从部分的简单加和中预测或还原

认知涌现

简单神经元的复杂组合产生高级认知功能，如语言理解和意识

中文屋挑战

即使系统整体涌现出智能行为,是否等同于真正的理解？

哲学启发

涌现现象为理解心灵与机器的关系提供了新的思考视角

哲学视角下的人工智能伦理

责任归属问题

如果机器不具备真正的理解和意识，谁应对其行为负责？

- 程序员的责任边界
- 使用者的责任范围
- 机器自身的道德地位

人机关系界限

在机器缺乏真正理解的前提下，如何界定人机关系的伦理边界？

- 机器权利的哲学基础
- 人类尊严的保护
- 技术应用的道德约束

中文屋实验的文化影响

哲学领域

成为心灵哲学和认知哲学的经典案例，
影响了整整一代哲学研究。

AI领域

激发AI研究者反思技术路线，
探索超越符号主义的新方法。

认知科学

促进认知科学对意识、理解和智能本质的深入探讨和理论建构

公众关注

引发公众对智能本质、机器意识等
前沿问题的广泛关注和讨论。

结语：

什么是理解？永恒的哲学难题

理解不仅是符号操作，而是意识、经验与意义的统一

超越形式

真正的理解超越了形式符号的操作，涉及意义的把握和概念的形成

意识基础

理解依赖于意识的主观体验，这是纯计算过程无法捕捉的维度

经验依赖

语义理解根植于主体的生活经验、身体感知和世界互动

意义统一

意识、经验和意义形成不可分割的统一整体，构成理解的完整图景

未来展望：超越中文屋的智能探索

1. **体**现认知整合

   发展具有真实身体和感知能力的智能系统，实现与环境的真实交互

2. **神**经启发架构

   借鉴生物神经系统的组织原理，设计更接近自然智能的AI架构

3. 语义理解突破

   探索真正具备语义理解能力的模型，超越纯统计的模式匹配

4. **意**识探索

   在严格的哲学和科学框架下，审慎探索机器意识的可能性边界

哲学视域中的中文屋实验总结

40+	1000+	3
影响年限	学术引用	核心领域
自1980年提出至今，持续影响AI哲学研究超过40年	相关论文被引用超过千次，成为最具影响力的思想实验之一	深刻影响哲学、认知科学、人工智能三大核心领域的发展

中文屋实验作为经典思想实验，其深远影响超越了时代和学科界限，持续激发着人类对智能本质、意识现象和理解能力的深刻哲学思考。

谢谢

欢迎讨论与提问