人工智能的现实与科幻

DZ1933026 王国畅

2020 年 1 月 9 日

摘要

人工智能的发展过程既是机器智能化程度不断上升的过程，也是人们对人工智能的科幻展望不断增加的过程。机器的智能化在现实中产生了各类社会影响，同时，关于人工智能的讨论也不断增加，典型的两个论题是：1. 人工智能是否会对人类产生威胁；2. 人工智能是否具有智能。本文将明确人工智能的概念本身和两大论题，简介人工智能原理，并基于原理对以上两个论题展开讨论。

关键字

人工智能；机器学习；人工智能威胁论；

1 引言

人工智能（英语：Artificial Intelligence，缩写为 AI）亦称智械、机器智能，指由人制造出来的机器所表现出来的智能。通常人工智能是指通过普通计算机程序来呈现人类智能的技术 [1]。人工智能的产生和发展是信息系统智能化的过程。智能化是信息科学技术发展的主流趋势。人工智能的出现极大的改善了计算机的性能，对社会生活的各个方面产生了巨大影响（例如：医疗，政治，军事……）。与此同时，关于人工智能的讨论也从未停止。一方面，科幻小说和网络争论使得人工智能威胁论不断发酵；另一方面，哲学家和心理学家关于人工智能的存在和智能本身存在哲学争论。

它一方面是现实的，所有贡献都有严密的数理基础，另一方面关于它的讨论又是极具科幻色彩的。

由于人工智能本身是一个较为复杂的概念，为了更好的以上两个论题进行讨论从而获得对人工智能的正确认识，人工智能的定义需要进一步被明确。同时，两个论题本身的来源和各个观点的主要推动者的背景也有着有趣的聚类分布。本文将在**背景**章节从两个维度介绍人工智能的定义，并分别从论题，论点和其主要支持者来介绍人工智能威胁论和人工智能哲学争论。接着在**原理**一章介绍当前人工智能的运行原理，最后基于此原理在**讨论**一章对两大论题展开讨论。

2 背景

2.1 人工智能的定义

人工智能是一个复杂的概念，它的定义包含两个维度：1. 人工智能的研究领域；2. “智能”的含义。根据对以上两个维度的概念的不同理解，会产生不同的认知。

广义上讲，人工智能的研究领域包括以下子领域 [1]：

- 演绎、推理和解决问题：早期的人工智能研究人员直接模仿人类进行逐步的推理，就像是玩棋盘游戏或进行逻辑推理时人类的思考模式 [2]。
- 学习：指的是机器学习，其经典定义是“利用经验来改善计算机系统自身的性能”[3]。其主要目的是为了让机器从用户和输入数据等处获得知识，从而让机器自动地去判断和输出相应的结果。这一方法可以帮助解决更多问题、减少错误，提高解决问题的效率。
- 知识表示法：人工智能领域的核心研究问题之一，它的目标是让机器存储相应的知识，并且能够按照某种规则推理演绎得到新的知识。有许多需要解决的问题需要大量的对世界的知识，这些知识包括事先存储的先验知识和通过智能推理得到的知识。在采用机器学习技术的情况下，它的定义与数据挖掘一致，即“识别出巨量数据中有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程”[4]。大体上看，数据挖掘可以视为机器学习和数据库的交叉。
- 规划：又叫代理，实质是其余子领域的综合调用。它统筹一个任务，并将任务划分为各个子领域问题，交给对应的人工智能组件进行处理。
- 自然语言处理：探讨如何处理及运用自然语言，自然语言认知则是指让电脑“懂”人类的语言。自然语言生成系统把计算机数据转化为自然语言。自然语言理解系统把自然语言转化为计算机程序更易于处理的形式。
- 运动和控制：主要包括机器人运动硬件和驱动软件的设计，运动硬件使得机器人可以进行各类动作，驱动软件使得运动硬件可以通过编程来运作。
- 知觉：研究计算机从现实世界获取信息的方法，如通过红外线感知人体温度，通过摄像头采集信息进行人脸识别，通过语音识别将语音转换为可用于学习的数据等。
- 社交：主要指人机交互及其体验优化。机器需要能感知人类输入并进行反馈，使得交互过程对人类更为友好。
- 创造力：代表了理论（从哲学和心理学的角度）和实际（通过特定的实现产生的系统的输出是可以考虑的创意，或系统识别和评估创造力）所定义的创造力。
- 伦理管理：于诸多社会人士对于人工智能的潜在威胁的担忧而建立，通过对人工智能进行条约和伦理定义进行行为约束。
- 经济冲击：从经济的角度通过统计的手段衡量人工智能对社会产生的冲击，如机器生产替代了手工生产从而导致局部失业率上升等。
- AI 对人类的威胁：来自诸多社会人士的争论，主要分为悲观派和乐观派。

不难发现，以上子领域有鲜明的区别，并可以聚集为三类：1. 理论类；2. 社会影响类；3. 哲学类。其中，理论类包含：演绎、推理和解决问题，规划，学习，自然语言处理，知识表示法，运动和控制，知觉和社交。社会影响类包含：经济冲击和 AI 对人类的威胁。哲学类包含：创造力。

人工智能理论可以划分为两个层次：1. 与现实世界的交互，包括知觉，社交，运动和控制；2. 对信息的处理，包括演绎、推理和解决问题，学习，自然语言处理。其关系可以用动物的身体结构来类比：知觉对应五感，是从现实世界信息到大脑信息到映射，运动和控制对应骨骼和肌肉，接收大脑的信息进行活动，社交是骨骼和肌肉运动形成的一系列和世界交互的行为，而规划好比中枢神经，是大脑对外发布指令和向内采集信息的出入口，演绎、推理和解决问题，学习，自然语言处理，知识表示法等对应大脑功能的不同分区。在一个理论的人工智能系统中，" 骨骼 ";" 肌肉 "，" 中枢神经 " 的作用是平凡的，他们仅仅负责信息的映射和传递，真正对系统优化产生作用的是 " 大脑 "，即演绎、推理，解决问题，学习等模块。

社会影响类和哲学类的人工智能研究都依托于人工智能理论而存在，但其对于人工智能的假设与人工智能理论并不一致。

人工智能理论目前的核心领域是学习和知识表示法，推动人工智能进步的核心研究点是机器学习和数据挖掘。如前文所述，数据挖掘可以看作机器学习和数据库的交叉领域，人工智能理论的问题求解过程依赖于机器学习。人工智能理论的基本假设是：机器可以 " 利用经验来改善计算机系统自身的性能 "[3]。

人工智能社会影响的出发点在于人工智能的飞速发展对社会产生了巨大影响，人们担忧人工智能的能力会超出人类的掌控。其基本假设是人工智能会遵循科技发展的加速度理论，即人工智能会以指数增长的速度快速发展。

人工智能哲学的中心是什么可以被定义为智能，以及在此基础上的机器是否具有智能。其智能定义的基本分歧在于是随着机器的智能化，机器是否可以被认为具有 " 意识 "。其基本假设的分歧是：在对人类智能有充分理解的情况下，设计出与人类具有同等或以上 " 智能 " 的系统是否是可行的，即人工智能是否可以和人类智能同质并完全超越。

狭义上讲，人工智能的研究领域只包括人工智能理论，而不包含其社会影响和哲学思辨。

另一个维度上，基于人工智能哲学的两个基本假设，人工智能可以划分为强人工智能和弱人工智能。强人工智能即认为人工智能可以拥有于人类等同的智能，即可以拥有自己的意志，创造力，并在推理计算等方面超越人类。弱人工智能认为人工智能并非拥有智能，只是计算机系统进一步提高了计算和推理能力，因此无法拥有自我意识。

2.2 人工智能威胁论

人工智能威胁论是在网络产生并逐渐发酵的争议。其基本问题是：人工智能是否会最终导致人类灭亡。

" 史蒂芬 · 霍金、比尔盖茨、马斯克、Jaan Tallinn 以及 Nick Bostrom 等人都对于人工智能技术的未来公开表示忧心，人工智能若在许多方面超越人类智能水平的智能、不断更新、自我提升，进而获取控制管理权，人类是否有足够的能力及时停止人工智能领域的 " 军备竞赛 "，能否保有最高掌控权，现有事实是：机器常失控导致人员伤亡，这样的情况是否会更加扩大规模出现，历史显然无法给出可靠的乐观答案。特斯拉电动车马斯克（Elon Musk）在麻省理工学院（MIT）航空航天部门百年纪念研讨会上称人工智能是 " 召唤恶魔 " 行为，英国发明家 Clive Sinclair 认为一旦开始制造抵抗人类和超越人类

的智能机器，人类可能很难生存，盖茨同意马斯克和其它人所言，且不知道为何有些人不担忧这个问题。”[1]

对于人工智能的潜在威胁，目前主要分为悲观派和乐观派。

悲观派的观点是 [1]：

- 人工智能会遵循科技发展的加速度理论
- 人工智能可能会有自我改造创新的能力
- 人工智能进步的速度远远超过人类
- 人类会有灭绝的危机存在

其代表人物是天文物理学家史蒂芬霍金、比尔盖茨和特斯拉首席执行官伊隆马斯克等。

乐观派的观点是 [1]:

- 人类只要关掉电源就能除掉机器人
- 任何的科技都会有瓶颈，“摩尔定律”到当前也遇到相当的瓶颈，AI 科技也不会无限成长，依然存在许多难以克服的瓶颈。
- 依当前的研究方向，电脑无法突变、苏醒、产生自我意志，AI 也不可能具有创意与智能、同情心与审美等这方面的能力。

其代表人物 Google、Facebook 等 AI 的主要技术发展者。

在人工智能威胁论中，悲观派对于人工智能的认知和基本假设来自于人工智能的社会影响，而乐观派对于人工智能的认知和基本假设来自于人工智能理论。在支持者方面也产生了鲜明的聚类：人工智能技术的主要发展者为乐观派，而悲观派的代表人物并非人工智能领域的直接相关开发者。

2.3 人工智能哲学讨论

人工智能哲学试图回答下列问题 [5]：

- 机械可以有智慧的运作吗？可以解决人透过思考处理的所有问题吗?
- 人类智慧和人工智能是一样的吗？人脑本质上是电脑吗?
- 机械可以如同人类一般拥有自己的精神、心理状态和意识吗？能感知东西是如何吗?

这三个问题分别反映了人工智能开发者、语言学家、认知科学家和哲学家的不同兴趣。透过讨论“智能”、“意识”的定义，和“机械”究竟是什么，来寻找这些问题的科学答案。

基于对此问题从不同角度进行回答可以得到人工智能哲学的若干论断：

- 图灵的礼貌公约：如果一个机器的行为可以如人类一般，那么它就和人类拥有同样的智能。[6]
- 达特茅斯提议：“如果可以精确的描述学习的所有面向，或智能的任何其它特征，一个机器就可以被制造来模拟它。” [7]

- 塞尔强人工智能假设：“只要电脑有适当的程式让能够正确的输入和输出，就能拥有与人类精神相同意义的精神。”[5]
- ……

这类论断可以看作是从智能是否可判定，人类智能认定和机器智能认定的角度对人工智能和人类智能同质性的尝试解答。

人工智能哲学讨论主要对象是强人工智能，而狭义上的人工智能通常指弱人工智能。故人工智能哲学通常不与人工智能理论发生交叉。参与者通常来自哲学，社会学，语言学等领域。值得一提的是以上第一条论断来自图灵，被认为是计算机和人工智能的伟大先驱之一，但图灵的论断并非基于强弱人工智能的分类，而是忽略“意识”，仅从行为能力上对智能进行判别。

不难看出，如果忽略“意识”的定义和判别这一纯粹的哲学问题，人工智能的哲学讨论分歧仅在于：人工智能是否可能拥有“意识”。

3 原理

如前文所述，机器学习是人工智能系统的“大脑”，即“智能”的核心。人工“智能”的“原理”核心即是机器学习的原理。机器学习经过长期的发展，子领域众多，详细分析其原理和区别对于本文讨论并无裨益。因此为了便于讨论，这里大致总结为三类：

- 符号学习：符号学习的基本原理是将现实世界的信息进行符号化表示，例如将西瓜的颜色，大小，敲击响声符号化，通过查看若干个西瓜的颜色，大小和敲击响声，可以总结出一条符号化的规则，如：如果一个西瓜颜色深绿，个头大于直径 20 厘米且敲击响声浑浊，那么这个西瓜是好瓜。
- 统计学习：统计学习由符号学习演变而来，由于符号学习是对现实信息的表示，很容易面临信息量过大而计算机无法处理的情况，统计学习即是在符号学习的基础上通过统计学原理降低符号学习的复杂度，同时尽可能保留符号学习的表示能力。
- 神经网络：神经网络的基本原理是设计若干计算单元（称为神经元）相互连接，每个计算单元有输入和输出，当输入满足一定条件（到达阈值）时，神经元才进行输出，例如当输入小于 1 时则不输出，否则输出。在进行学习时，当不断告知计算单元输入和输出，就可以不断调节神经元的阈值，从而得到一个具备知识的神经网络，例如，当一个计算单元为若 x>=a 则 y=1 否则 y=0，那么当不断向计算单元输入对应的 x 和 y，就可以逼近 a 从而获得阈值。即，大量地告知神经网络什么是好瓜，什么是坏瓜，通过神经网络图就可以通过图算法不断更新每个神经元的阈值，这些阈值构成了神经网络的知识。

可见机器学习的三大流派在根基上殊途同归，即都是数学可理解的。人工智能理论的根本是数学。

4 讨论

从人工智能的概念上讲，人工智能可分为强人工智能，弱人工智能，狭义人工智能和广义人工智能。人工智能的现实是基于狭义人工智能的人工智能理论构筑的弱人工智能，而在人工智能的科幻中：

人工智能威胁论是在广义人工智能下针对弱人工智能的讨论，人工智能哲学则是在广义人工智能下对强人工智能的讨论。

人工智能的“科幻”之所以科幻，是因为其论点并非建立在数理基础上，而是建立在某些假设上。

人工智能威胁论的根本假设是：1. 人工智能会继续依照摩尔定律快速发展；2. 人工智能会超出人类的控制。人工智能哲学的假设是：3. 强人工智能是可以实现的。

以上三个假设均不符合当前现实：1. 摩尔定律并非对物理社会的观测所得，而是英特尔创始人之一预估的电子芯片晶体管数量密度增长速度，而晶体管数量密度并非持续高速增长，而已经放缓。弱人工智能受限于计算机的运算能力，同时其理论研究与其他科学一样，会进入瓶颈期，因此摩尔定律在人工智能的智能化程度提升上并不适用；2. 人工智能的计算能力依托于物理机器，即人工智能的宿主总是在人类的控制范围内。3. 强人工智能目前并无数学理论支撑和可行的技术实现，并且不存在人工创造力或是人工意识的理论模型。

但假设 2 和假设 3 目前并不能证伪：对于假设 2，当人工智能的决策能力达到一定程度，必然可以具备动态扩展的能力，即同一人工智能系统有能力运行在广泛分布的机器上，并有能力动态调度自身的任务，由于机器的分散，通过对物理机器进行管理的难度会进一步上升，同时，由于机器学习存在错误的概率，因此在越重要的任务里，人工智能出错造成的风险就越大。对于假设 3: 目前人类对自身的意识和智能都知之甚少，因此无法断言人类的智能与机器并非同质，或机器的智能无法与人类同质。

由此，即使假设 2 和假设 3 将来成立，也可以在已有的应对方式下得到处理：对于假设 2，只需对人工智能的出错恢复做良好的管理，例如采用中心化的结构管理所有人工智能，人类无需掌握所有的人工智能宿主，只需掌握中心机器即可，同时，只需将实际决策保留于人类，人工智能的决策结果作为参考，即可确保人工智能无法越过人类作出决策。对于假设 3，需要解决的是强人工智能出现后的伦理问题，这一点需要人类在强人工智能诞生前夕达成伦理共识。

尽管经过归纳总结，人工智能的讨论似乎可以给出定论：人工智能是可靠的，人们担心的问题目前并不存在，即使存在也有办法解决。但事实上假设 2 和假设 3 的处理方式同样不可靠，因为其约束在于人类。假设 2 得以解决的前提是人类不会通过战争等方式自我毁灭，假设 3 得以解决的前提是如果强人工智能证明可行，并可能对人类造成存在威胁，人类达成共识决定不制造强人工智能，则强人工智能永远不出现。

前者与广义上的科学与技术的双刃剑并无不同，后者则是因为人类智能本身的不可知而难以判别。

归根结底，任何科学与技术的发展都依赖于“物质”的现实土壤，但人工智能与哲学意义上的“意识”产生了交织，于是开阔了科幻的天空。

参考文献

[1] 人工智能. https://zh.wikipedia.org/wiki/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD.

[2] Stuart J. Russell. *Artificial intelligence: a modern approach*. Pearson, 2016.

[3] Tom M. Mitchell, Richard M. Keller, and Smadar T. Kedar-Cabelli. *Machine Learning*, 1(1):47–80, 1986.

[4] Usama Fayyad. *Data Mining and Knowledge Discovery*, 2(1):5–7, 1998.

[5] 人工智能哲学. https://zh.wikipedia.org/wiki/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E5%93%B2%E5%AD%B8.

[6] A. M. Turing. I.—computing machinery and intelligence. *Mind*, LIX(236):433–460, Jan 1950.

[7] Daniel Crevier. Expert systems as design aids for artificial vision systems: a survey. *Intelligent Robots and Computer Vision XII: Algorithms and Techniques*, 1993.