Advances in Philosophy 哲学进展, 2026, 15(4), 413-418
Published Online April 2026 in Hans. https://www.hanspub.org/journal/acpp
https://doi.org/10.12677/acpp.2026.154181

Hans汉斯

德雷福斯人工智能批判思想研究

吴小翠

广西大学马克思主义学院，广西 南宁

收稿日期：2026年3月20日；录用日期：2026年4月10日；发布日期：2026年4月23日

摘　要

本文旨在梳理德雷福斯人工智能批判思想的研究历史和现状。德雷福斯基于海德格尔现象学对早期符号主义人工智能观展开了批判，起初被人工智能专家排斥而后接受，促使该领域产生了海德格尔式人工智能转向，深刻地影响和推动了该领域的研究和发展。德雷福斯的批判是成功的，但也有一定的局限。

关键词

人工智能，符号主义，德雷福斯，海德格尔

A Study of Dreyfus's Critical Thought on Artificial Intelligence

Xiaocui Wu

School of Marxism, Guangxi University, Nanning Guangxi

Received: March 20, 2026; accepted: April 10, 2026; published: April 23, 2026

Abstract

This paper aims to review the research history and current status of Dreyfus's critique of artificial intelligence. Drawing on Heideggerian phenomenology, Dreyfus critiqued the symbolic AI approach. Initially rejected by AI experts, his critique was later accepted, prompting a Heideggerian turn in the field and profoundly influencing and advancing its research and development. Dreyfus's critique was successful, yet it also has certain limitations.

Keywords

Artificial Intelligence, Symbolism, Dreyfus, Heidegger

文章引用: 吴小翠. 德雷福斯人工智能批判思想研究[J]. 哲学进展, 2026, 15(4): 413-418.
DOI: 10.12677/acpp.2026.154181

1. 引言

本文旨在梳理德雷福斯人工智能批判思想的研究现状。近年来生成式人工智能聊天机器人 ChatGPT 和文生视频 Sora 的横空出世，引起全世界的讨论狂潮，其中不乏对充满想象力的通用人工智能 AGI 的讨论。但纵观人工智能的历史，不难得出一个结论：人工智能似乎一直在存在者的层面上、在逻辑和运算的角度模拟人类，科幻大片中的机器人仍然只是一个想象。所谓“智能”究竟是什么？人类智能是否等同于逻辑和计算的智能？这就是德雷福斯所要批判和思考的问题。对此，德雷福斯认为海德格尔现象学所揭示的人类在世界中存在的整体性和情境性、人类的生命体验也必须被人工智能考虑在内，应追求身心合一的完整的智能而非逻辑和运算的片面的智能。德雷福斯的批判在人工智能领域产生了巨大的影响，促使专家们开始研究一种具身认知式的海德格尔式人工智能。然而德雷福斯的批判也有一定的局限，即过于强调海德格尔的身体上的“操劳”，而忽略了表征主义和非表征主义之间的关联地带。

2. 人工智能(AI)的开端：符号主义

人工智能一词最初由麦卡锡于 1956 年在达特茅斯会议上提出，旨在通过机器模拟人类智能，以实现机器的智能功能。符号主义是人工智能哲学发展的首个阶段，其研究者普遍认为，通过精心设计的程序算法，利用符号进行的计算能够实现全面的智能。因此，这一阶段人工智能的核心特征是利用符号进行计算与推理。以数字为例，儿童在学习数字“3”时，实际上是在学习这个符号，一旦他们理解了“3”这个数学符号可以代表任意三个物体，他们就能进行数据处理。同样地，推理也是基于符号进行的，例如“如果天下雨，那么地面会湿”，计算机只要理解了“天下雨”和“地面会湿”之间的符号关联，就能够像人类一样进行推理。

符号主义持有一种乐观观点，即人工智能在各种领域中都可以归结为复杂的计算过程，如爱因斯坦的相对论、足球运动员的射门技艺或某些手工技能等都不例外。通过为计算机设计精确的程序和算法，可以实现全面的智能化[1]。这一信念植根于西方哲学史上长达两千多年的理性主义传统，坚信理性能力足以诠释人类所有智慧，其起源可追溯至毕达哥拉斯和柏拉图，并一直占据主导地位。然而，正如德雷福斯所认为的那样，符号主义是西方理性形而上学的顶点与极限，因此也是符号主义路径的人工智能技术的极限。符号主义的确很快遭到了巨大的挑战：如何在有限的步骤和资源下，寻找到最优解。以下象棋为例，机器下象棋需将规则转换为由“0”和“1”组成的程序；尽管走法千变万化，但都是基本符号的组合。计算机需要搜索所有可能的选项以找出最佳走法，这些选项也是符号的组合，因此所有可能的选项构成了一个庞大的集合。

随着问题复杂度的增加，计算量呈指数级增长，因为符号间的潜在组合数量也在急剧增加，导致计算机处理这类问题时容易遇到性能瓶颈。例如，计算机在处理中国象棋时表现出较高智能，但面对国际象棋则稍显吃力。历史上，“深蓝”战胜国际象棋大师卡斯帕罗夫曾引起关注；但围棋的求解复杂性远超象棋，传统符号主义方法难以应对，围棋人工智能程序的表现并不理想。

3. 德雷福斯对符号主义的批判

德雷福斯基于现象学对这种计算推理路径的符号主义人工智能观展开了批判。德雷福斯是美国杰出

的现象学学者，对海德格尔与梅洛·庞蒂的哲学体系均有着深入的研究。然而，这位深受欧陆人本主义哲学熏陶的学者，为何会选择涉足人工智能领域，将现象学与这一前沿科技领域联系在了一起？

1962 年，德雷福斯在麻省理工学院任教期间，就曾有学生向他直言不讳地指出，哲学家对于人性的思辨已然陈旧，因为明斯基等人工智能领域的科学家预测，工程学的方法不久后便能够全面实现人类智能的各种功能。这一观点令德雷福斯深感震惊与疑惑，因此，他决定在次年前往兰德公司进行深入的调研。在仔细分析与评估了当时由艾伦·纽厄尔和赫尔伯特·西蒙主导的“认知模拟”研究项目后，德雷福斯对后者的符号主义智能观提出了质疑。1965 年，他撰写了报告《炼金术与人工智能》，该报告后来扩展成为专著《计算机不能做什么》。在这份报告中，德雷福斯从现象学的视角出发，对符号主义人工智能的哲学理论基础进行了深刻的批判。

德雷福斯在理解“什么是智能”的问题上与符号主义产生分歧，其背后仍然可以说是亲数理科学的分析哲学与具有浓厚人文色彩的欧陆哲学之间的分歧。符号主义汲取了笛卡尔的认识论以及霍布斯“理性就是计算”、莱布尼茨的“普适”特征等对人类意识的理解，把理性主义进路的哲学成果应用到了人工智能领域，认为智能就是人的计算和推理能力，人类的智能不管多么复杂，归根到底都是由符号计算来实现的。而对于德雷福斯来说，人的智能显然不能够被程序和算法所穷尽。他站在海德格尔现象学的立场，认为符号主义的认识论哲学依据即笛卡尔的人作为独立的主体用心灵表征外部世界的二元对立的认识论已经被海德格尔的“在世存在(being-in-the-world)”现象学所超越，人不是与世界分离的独立的智者，而是被卷入整个情境之中、沉浸在世界之中。他指出，人类智能并非算法化的，而是随着时间推移与世界的互动所形成的某种嵌入式的理解[2]。笛卡尔的与世界分离的立场，就像是一名学习开车的初学者，把自己看成操作一台机器的某个零部件，处在一种“现成在手”的状态；而当这个初学者开车越来越熟练，进入一种专家状态时，就会忘记自己在操作一台机器，他与机器融为一体，他被卷入了世界，感受到的是自己正在去某个地方，这时的他处于一种“上手状态”。德雷福斯的这种批判给予当时的人工智能研究以巨大的冲击。

4. 德雷福斯遭到的拒斥及原因

为回应德雷福斯的批判，人工智能领域的权威人士西蒙·派珀特(Seymour Papert)于 1968 年撰写并发表了题为《德雷福斯的人工智能：谬误之集》(The Artificial Intelligence of Hubert L. Dreyfus: A Collection of Fallacies, 1968)的反驳报告，其中的观点也得到众多人工智能领域科学家的广泛认同。以下是派珀特的反驳观点。

4.1. 关于语言意义的“理解”

德雷福斯认为符号主义 AI 在上下文环境中依赖情境排除歧义的能力不足并提出质疑和批评，麦卡锡回应道这种局限性并非机器与人类之间的本质差异，因为即便是人类也无法完全避免认知的不清晰性和理解的歧义性。比如在法律领域，法令本身就可能存在潜在的模糊地带，而立法者、律师和法官在事先也未必能洞察所有的模棱两可之处。在人工智能的研究当中，完全可以摒弃这种偏见，只要依赖形式化的非单调推理，使计算机能够像立法者、律师和法官一样妥善处理问题，那么就是实现了人工智能[3]。

此外明斯基认为，在讨论机器的“理解”能力问题上，研究者陷入了语言带来的误导中，因为语言都说不清“理解”到底是什么。“我觉得没有义务去定义‘意义(Mean)’和‘理解(Understand)’这两个词，因为其他人已经尝试了几千年！”[4]当人们试图弄清楚什么是“意义”并希望与人工智能研究联系起来时，并没有帮助智能研究获得好的想法，而是使智能研究陷入困惑并妨碍了研究的进展，是语言的歧义与误用本身导致了界定机器的理解问题时的困难，而非机器本身的理解能力不足。问题应该在于系

统是否通过使用技巧来回避“真正的意义”，关于理解的问题应该回归到技术探索上。

4.2. 应分析智能而不是追问前提

相较于德雷福斯对人工智能哲学假设前提的追问，人工智能领域的专家们更多地从分析哲学的角度出发，认为对智能进行深入分析更加重要。德雷福斯对人工智能认知研究的批评在于其过于依赖“人与机器智能均源于符号操作”的假设，而非深入探究该假设的可能性；声称有大量证据表明人类大脑处理基本信息的过程与计算机处理信息的方式高度相似，然而并未提及这些证据的具体来源和细节。因此，符号主义人工智能在智能研究上的方法存在循环论证的问题。然而在人工智能专家看来，不断追问前提并没有什么实际意义，远不如分析智能来得更加重要。现象学所探讨的概念较为抽象和模糊，因此德雷福斯对人工智能的批判只是以虚幻批判现实。

派珀特也对德雷福斯所采取的无限后退的质疑方式进行了反驳。他认为，德雷福斯将下棋程序排除在智能范畴之外，仅仅因为其由编程实现；同样将塞缪尔的机器学习视为非智能，仅因其学习过程受到形式化条件的限制。这种简单的“X 不是智能”的论断缺乏实际意义。原因有二：其一，人工智能领域目前仍处在持续发展的量变阶段，计算机的潜在能力尚未被充分了解。德雷福斯过于关注对智能终极目标的批判，而忽视了这一领域的广泛探索与不断演进的现状。其二，鉴于我们对人类心灵的理解仍相当有限且对于智能这一概念仍存在诸多争议，我们总是能找到计算机尚未能实现的智能表现，这种无限倒退的悲观追问方式不仅对人工智能科学的发展无益，反而可能阻碍其进步。

4.3. 哲学不应插足科学

派珀特认为德雷福斯的批判有很多不成立的地方：德雷福斯认为人工智能在人类行为模拟上连打乒乓球都做不到，但是派珀特认为机械臂的设计是一个常规性的技术问题，根本不需要哲学的介入，这种批判只是自作多情的浮想联翩。德雷福斯对科学技术的理解也不充分，以至于无法分辨技术发展情况的本质特征和偶然特征，因而陷入“业余科学家综合征”的局面。

德雷福斯一再申明对人工智能专家的批判是哲学的批判，而不是他们的技术性工作，并没有贬低他们在技术研究上的意义和价值，但还是引发了科学家们的排斥。对于派珀特来说，人文学科应该在政治领域发挥作用，而科学要做的事情是建模－实验－证明，科学才是推动历史发展的动力，而这样理解智能的本质的研究往往会被哲学家矫揉造作地批判。面对这种科学对哲学的排斥，德雷福斯深感无奈，他一再解释他是在反思重建人工智能，而不是一个破坏者。

德雷福斯之所以遭到这样的拒斥和批判，背后根本原因是分析哲学和欧陆哲学的裂痕，也表现在二者对科学带来的风险的理解。科学乐观主义的人工智能专家认为，最大的危险不来源于技术而是来源于人，最大的限制也在于人，这种最大的危险是：人文学者因为担心技术发展威胁社会结构、传统和文化价值，而放弃理智上负责任的、可靠的研究传统。AI 专家们认为，充满恐惧的人文主义者总是担心技术发展会威胁到我们的社会结构和文化价值观，但是放弃为理智负责的危险要大得多。现象学家们试图限制计算机进一步侵入他们认为独特的属于人类的活动领域，这是一种怯懦[5]。由此可见双方论战的激烈程度之深。

德雷福斯在审视人工智能技术时希望建立一种新的智能观，以克服海德格尔提到的技术最大的危险：“原子时代即将到来的技术革命浪潮会如此吸引人、迷惑人、使人眼花缭乱，以至于有一天，计算思维(Calculative Thinking)可能会被接受并被实践为唯一的思维方式。”[6]对于德雷福斯所遭到的批判，明斯基的学生，前符号主义者特里·温诺格拉德认为这是一种现代科学的威望和成功造成的偏见。温诺格拉德指出人工智能专家的批判源自视理性主义为纯科学和应用科学的基础的思维模式，这种思维模式被他

们视为思考和智能的典范。在研究思维的过程中，人们主要关注规则的形式以及逻辑地应用这些规则的过程。正是这一基本认知框架，导致了德雷福斯的人工智能观受到批判。学者江怡认为二者哲学基础的分野是一种错误的策略，“造成这个分野的主要原因完全在于分析哲学家们的学术自负和认识盲点，而不在于他们与欧陆哲学家之间的观点差异。”[7]

5. 德雷福斯的批判对 AI 研究的影响

德雷福斯对符号主义人工智能观的批判，起初虽然引起了 AI 专家们激烈的反对，但随着时间推移，现代人工智能的许多研究者开始重新审视并采纳他的某些理念。

首先，由于强调人类直觉、情境感知和身体经验的重要性，德雷福斯推动了具身认知(embodied cognition)理论的发展，即认为认知过程是与身体状态、环境交互以及行为紧密相连的。这一理论后来影响了许多 AI 研究者去探索如何将身体因素和环境互动纳入 AI 系统的设计中。德雷福斯的批判还促使人工智能的研究从单纯的符号处理和逻辑推理转向更注重如何使 AI 更好地学习、适应和发展，例如深度学习、强化学习以及机器人技术的发展都受到了他思想的影响。现在被理解为人工通用智能(AGI: Artificial General Intelligence)的研究汲取了德雷福斯的批判，转向以海德格尔思想为基础的人工智能研究，出现了至少三种新的研究范式：罗德尼·布鲁克斯(Rodney Brooks)的行为主义方法，菲儿·阿齐(Phil Agre)的实用主义模型，以及瓦尔特·弗里曼(Walter Freeman)的动态神经模型[8]。

其次，德雷福斯的批判还引发了关于框架问题的讨论。在传统符号主义人工智能沦为“好的过时的人工智能(简称 GOFAI)”从而走向衰退后，新型的联结主义的神经网络人工智能开始发展。神经网络人工智能是一种基于仿生学原理构建的人工智能，其灵感来源于人脑中神经元网络的工作机制，这种人工智能不再单纯处理符号和逻辑，而是对神经元网络的一种模拟。不过和符号主义人工智能一样，神经网络人工智能仍然没有克服 AI 面临的框架问题，此问题是麦卡锡提出的，被称为“魔鬼问题”，即如何在不明确和不断变化的情境中有效选择并处理相关信息。对于框架问题，德雷福斯曾发问道：“计算机对当前的世界状态进行表征，如果世界中的某个东西发生了改变，那么程序怎样确定，它表征的事实中哪些可被看作保持不变的，哪些必须更新？”[9]面对动态变化的世界，计算机如何关注那些重要状态的变化，同时忽略那些无关紧要的状态变化？又如何在众多信念中提取并修订那些在特定情境下相关的信念呢？对此德雷福斯认为海德格尔的具身嵌入范式有望合理地解决或消解框架问题。然而，在实际操作层面，现象学的研究确实存在一定的局限性。徐英瑾指出，现象学家们在处理“身–心”关系问题时，即大脑与意识的关系，显得力不从心，他们难以深入探讨心灵的因果关系以及自由意志等深层次问题。这也正是德雷福斯在批判人工智能时所忽视的维度，使得他的批判在某种程度上显得“不精准”[10]。

最后，虽然最初不被理解，但德雷福斯的工作最终还是激发了哲学家和技术专家之间关于人工智能本质和可能发展方向的持续对话，为理解智能的本质和未来 AI 的发展提供了哲学视角。言而总之，德雷福斯对人工智能的批判不仅质疑了当时的主流研究方向，也催化了新的研究思路和发展路径，从而在学术和实践层面都对人工智能领域产生了持久而积极的重塑作用。因此，德雷福斯也被称为“AI 领域的黑骑士”，人们肯定他作为一位持久而坚定的批判者，不断挑战着传统人工智能范式，并推动了该领域向着更全面和深入的方向发展。

6. 德雷福斯人工智能批判的局限

李日容认为，德雷福斯对于人工智能的批判虽然触及了人工智能发展过程中的核心问题，即自主性和创造性的挑战，但并未为这些问题提供有效的解决方案。其主要原因在于，德雷福斯的人工智能批判过度依赖于梅洛·庞蒂的身体现象学资源，却未能充分领悟海德格尔关于人类智能本质的深刻洞察。海

德格尔认为，人的存在是一个时间性的整体，而非仅仅局限于德雷福斯所强调的日常操劳与应对技能的维度[11]。因此，德雷福斯的理论在这一方面受到了广泛的批评。

经由海德格尔的时间性此在所揭示的人类智能的本质，既不属于形式化范畴，也不属于非形式化范畴，而是介于这两者之间。换言之，其存在之根基在于此在作为时间性的整体存在，且此在的存在亦需基于其对存在本身的归属性来理解。然而，德雷福斯的人工智能批判过度强调了人类智能的非形式化特性，将非形式化等同于非概念或非表征，却忽视了此在的生存同样具有形式化特征，即包含概念与表征元素，尽管这些元素可能并不总是以显性方式呈现。

德雷福斯的重要性在于，他通过现象学的视角与方法揭示了人类智能的非形式化特征，进而推动了海德格尔式人工智能的复兴与发展，并揭示了传统有效人工智能形式化路径的本质局限性。然而，人工智能发展的正确路径应当在于深入认识和理解人类智能形式化与非形式化特征之间的内在融合与统一，并基于此为解决人工智能当前面临的自主性与创造性瓶颈问题提供有效方案。这种内在融合与统一并非简单地两者相加所能实现，而是需要深入探索形式化与非形式化之间的关联领域。这构成了人工智能发展所面临的真正挑战，也是推动其不断前进的关键所在。

本文认为德雷福斯的批判及局限体现的是人类的逻辑思维和感性直观的深刻矛盾，逻辑在先还是感性直观在先的问题是分析哲学和欧陆哲学争论的焦点，二者都从自己的角度出发解决康德遗留的二元论鸿沟问题，从历史上看，这个问题来自理性主义和经验主义的鸿沟，归根到底又始于柏拉图和亚里士多德，具体反映在科学和人文之间的分野。本文认为人文科学对数理科学的态度应是超越和包容，而非对立，毕竟二者从同一条道路上走出，这同一条道路就是我们自身。

参考文献

[1] 玛格丽特·博登. 人工智能哲学[M]. 王汉琦, 译. 上海: 上海译文出版社, 2006: 11.

[2] 成素梅, 姚艳勤. 哲学与人工智能的交汇——访休伯特·德雷福斯和斯图亚特·德雷福斯[J]. 哲学动态, 2013(11): 102-107.

[3] Papert, S. (1968) The Artificial Intelligence of Hubert L. Dreyfus: A Budget of Fallacies. Project of MAC, Massachusetts Institute of Technology. https://wellcomecollection.org/works/w4ktp96z/items

[4] Minsky, M. (1982) Why People Think Computers Can't. *AI Magazine*, **3**, 3-15.

[5] 蔚蓝, 孙小淳. "一堆谬误"——人工智能专家对德雷福斯的批判[J]. 科学文化评论, 2023, 20(1): 100-114.

[6] Dreyfus, H. andWrathall, M. (2002) Heidegger Reexamined: Vol. 3. Art, Poetry, and Technology. Routledge, 164 p.

[7] 江怡. 重新审视分析哲学与欧陆哲学的分野[J]. 哲学动态, 2018(1): 64-69.

[8] 段似膺. 海德格尔式人工智能及其对"意识"问题的反思——兼与何怀宏先生商榷[J]. 社会科学文摘, 2019(3): 24-26.

[9] 朱清华. 德雷福斯与海德格尔式人工智能[J]. 哲学动态, 2020(10): 72-79, 128.

[10] 徐英瑾. 译者前言: 英美心灵哲学到底在何处异于欧陆现象学? [M]//塞尔 J. 心灵导论. 上海: 上海人民出版社, 2008: 12.

[11] 李日容. 海德格尔的时间性此在与人工智能发展的自主性难题——兼论德雷福斯人工智能批判的局限性[J]. 陕西师范大学学报(哲学社会科学版), 2022, 51(0): 46-56.