

论作为主体的机器

——基于机器学习的不可理解

杨子辰

宁夏大学法学院, 宁夏 银川

收稿日期: 2025年1月18日; 录用日期: 2025年2月12日; 发布日期: 2025年2月20日

摘 要

随着人工智能的快速发展和广泛应用, 关于机器主体性的讨论逐渐成为哲学与技术交汇的重要议题。本文分析了机器学习生成的内容独立于人类且不可直接理解, 并以此为全文的出发点和立足点, 在此基础上重新审视“中文屋”及他心问题, 提出智能机器在人的认识中已经被看作一种特殊的主体, 并通过探讨人机关系从工具论、家畜论到奴隶论的转变和人机主奴关系的倒转, 指出在人机关系中机器主体性的上升并根据此趋势展望未来人机关系。本文强调作为主体的机器并非遥远未来的设想, 而是科技发展和应用下的现实, 旨在推动对机器主体性及人机关系变化的深刻理解, 呼吁从哲学角度对此进行深度反思和建构以应对人工智能在伦理和法律等领域带来的挑战、为未来的人机关系发展提供哲学指引。

关键词

机器学习, 主体, 他心问题, 人机关系

On the Machines as Subjects

—Based on the Uninterpretability of Machine Learning

Zichen Yang

Law School, Ningxia University, Yinchuan Ningxia

Received: Jan. 18th, 2025; accepted: Feb. 12th, 2025; published: Feb. 20th, 2025

Abstract

With the rapid development and widespread application of artificial intelligence, discussions on the subjectivity of machines have gradually become a critical issue at the intersection of philosophy and technology. This paper begins by analyzing the independence and uninterpretability of content generated by machine learning, establishing this as its foundation and point of departure. On this basis,

it reexamines the “Chinese Room” thought experiment and the problem of other minds, arguing that intelligent machines are increasingly regarded as a unique kind of subject in human understanding. By exploring the transformation of human-machine relations—from instrumentalism to domestication, and ultimately to enslavement—and the potential reversal of master-slave dynamics, the paper highlights the rise of subjectivity within these interactions. Furthermore, it forecasts the future trajectory of human-machine relations under this trend, emphasizing that the conceptualization of machines as subjects is not a distant futuristic scenario but a present reality shaped by technological advancement and application. The study aims to deepen understanding of subjectivity and the evolving dynamics of human-machine relations, calling for profound philosophical reflection and construction to address ethical and legal challenges brought about by artificial intelligence, and to provide philosophical guidance for the development of future human-machine interactions.

Keywords

Machine Learning, Subject, Problem of Other Minds, Human-Machine Relations

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来,人工智能相关话题随着人工智能技术的发展和在各个领域的广泛应用而备受关注,关于机器主体性的讨论和研究也在广泛开展并取得了巨大进展,人们对相关问题认识有了长足进步。不过人工智能的概念本身比较含混模糊,相关的技术细节又非常复杂,现有哲学领域的相关研究也常常只是泛泛地讨论人工智能而难以涉及现实的具体技术。此外,现有的讨论多关注强人工智能、超图灵机这些对遥远未来的抽象设想,对现实中已经获得广泛应用并产生重大影响的革命性技术重视程度不高。

实际上人工智能是包含了诸多子领域的交叉科学,而机器学习是人工智能的核心子领域,也是使得作为主体的机器真正成为可能的关键技术。本文将聚焦于机器学习的不可理解这一关键点,指出机器已经开始在不同的意义上作为主体,并讨论作为主体的机器何以可能。

2. 机器学习的不可理解

参考机器学习领域权威专家 Mitchell 的说法,机器学习就是“计算机利用经验自动改善系统自身性能的行为”[1]。即机器学习是研究计算机识别现有知识、获取新知识、不断改善性能和实现自身完善的方法,使得计算机能够学习大量输入数据样本的数据结构和内在规律,从而对新样本进行智能识别,或实现对新样本的预测,甚至在此基础上进一步进行决策。

普通的计算机程序只是人类逻辑的算法从自然语言的形式最终转化为机器语言的形式,每一段指令都来自人类,不论是在怎样的平台实现,其本质都是人的逻辑的不同形式。而机器学习的革命性正在于出现了并非来自人的指令,这些内容不再是由人的逻辑直接转化而来的,而是从数据中学习得到的,是真正的机器自己的逻辑。虽然正如人的逻辑可以转化为编程语言一样,机器学习的模型的逻辑可以通过编程语言表述,而对于具备专业技能的人而言编程语言又是可理解的,所以如果是在内容不多也不复杂的情形下人类依旧可以通过直接阅读具体内容以理解这样一种并非来自于人的逻辑。但这种理想情况在现实中却显然是不可能的。OpenAI 曾在 2020 年详细介绍了 GPT-3 的架构、训练过程和性能,提到了 GPT-3 模型的规模为 1750 亿个参数[2]。面对如此复杂又抽象的内容,通过阅读程序代码让人对程序进行

直接理解已经成为了天方夜谭。

底层结构复杂且不透明的机器学习模型，其输入和输出之间对人类来说真正成为了“黑盒(Black Box)”，没人能确切地知道黑盒中的逻辑，更遑谈确保作为输出的识别结果或决策可靠。正是由于建立在机器学习之上的人工智能系统背后的运作逻辑难以理解，或许会存在不某些确定的风险，使得用户难以完全信任经过黑盒得到的结果。这种现实应用上的困难推动了机器学习的可解释性研究。

但必须要清楚，可解释性研究并不能从根本上彻底解决黑盒不可理解的问题。可解释的人工智能 Explainable Artificial Intelligence (XAI)并不意味着可直接理解的人工智能。此意义上的可解释性(explainability)指使用解释作为人类和智能模型之间的中介，作为模型代理能够被人类所理解，倾向于指代原模型不可理解，而需要构造事后模型进行解释。从作用上讲，可解释性是人类与模型之间的接口，它既是那些不可理解的模型的代理，又是人类所可以理解的[3]。这种可解释性是出于实践应用的需要，在“黑盒”之外利用解释方法或者构建解释模型作为中介进行一种科学的推测。虽然这种可解释性的探索取得了很大进展，也已经发挥出了在实践中的极大作用，一定程度上提高了用户对模型的理解和信任，进一步的拓展了应用领域。但从逻辑上看，解释和原模型的符合得不到完全的确证，从认识论看，这样通过中介理解的只是中介，永远也不能超越中介通达直接的理解。

3. 机器在人的认识中作为主体

约翰·塞尔(John Searle)提出了“中文屋”(the Chinese room argument)，认为即使人工智能可以同人类进行文字交流，也正如一个不懂中文的人在房间内通过查阅中文语句间的对应规则用自己拼凑的中文语句回应房间外传递来的中文语句，这并不意味着人工智能真的拥有了理解能力。

首先要指出，约翰·塞尔的中文屋实际上是把黑盒类比为有一个内部有规则手册和一个不懂中文的人的房间。但是如前文所说，由于机器学习的不可理解，实际上我们无从得知黑盒内部的结构。故而黑盒和中文房间的类比本身只是对黑盒的一种预设，不能真实反映黑盒的内部结构更不能以此解释人工智能的运作机制。其次，中文房间内不懂中文的人在黑盒内不必然有逻辑结构上的对应者，因为不懂中文的人只是中文房间内的一个部分，没有作为独立的个体和房间外直接接触，而黑盒内部结构却不为人知。所以，真正和人工智能黑盒在逻辑结构上处于同一位置的是中文房间整体而不是中文房间内不懂中文的人，传入房间的纸条对应黑盒的输入，传出房间的纸条对应黑盒的输出结果，而房间外的人也正如我们对黑盒一无所知一样，对房间内也一无所知。

建立起新的对应关系之后，中文屋的问题可以被表述为一种对图灵测试的诘问：只根据输入和输出判断机器是否具有智能何以可能？不难看出，这正是他心问题的一个翻版，或者说是不仅涉及到了他人的心灵，也涉及到了那些非人对象的心灵的广义的他心问题[4]。对于机器学习的黑盒，我们只能了解到黑盒之外位于其两端的输入和输出，正如对于他心我们也只能了解外在于其的环境和行为。在该问题的语境中，机器学习的黑盒同他心的黑盒具有逻辑结构上的一致。

虽然有一种观点认为，人工智能的应用会改变他心问题的困局[5]，不过这主要是指通过脑机接口或其他一些技术手段结合人工智能进行“读心”“读脑”，神经科学、脑科学等科技领域的他心问题会随之获得进展。事实上对于哲学领域而言，机器学习的出现和广泛应用甚至会使得他心问题涉及的范围扩大，甚至变得更加复杂。

首先，这种特殊形式的他心问题即机器的他心问题被提出、被讨论这一现象本身实际上标志着在现实中智能机器已经开始被看作主体。因为他心问题的提出和讨论实际上建立在将他人看作主体的事实上，时间上人们先将他人看作另一主体然后才反思地提出他心问题，逻辑上人们处理他心问题是尝试为将他人看作主体的现实寻找理论的基础。类似的，上世纪“中文屋”的提出也以图灵测试(通过对话判断机器

智能)相关理论的出现和流行为前提。当然在此意义上的将机器看作主体不意味着被看作主体的机器实际存在,这还只是假言中的存在、想象的存在而非现实的存在。但是能够通过图灵测试的机器在今天已然实现,至少是已经接近实现,正如近来的一项实验中,有 54%的参与者将 GPT-4 误认为是真人,而真人被正确辨认的比例则为 67% [6]。

更突出的问题是现有的处理他心问题的理论难以将机器和人区分开。例如类比推论方案(Analogical Inference Solution, AIS):我的某一特定心理活动导致了我做出某一特定行为,又可以观察到他人做出了和我特定行为类似的行为,由以上两点推出他人也具有和我类似的心理活动[4]。如果承认该推理模式,即由自己某一特定心理活动和某一行为间的因果联系可以推知有类似行为的他人也具有类似的心理活动,就不得不承认:按照相同的逻辑,有类似行为的机器也具有类似的心理活动。即使像维特根斯坦那样认为人们在语言游戏中并不关心使用语言所谈论的东西,否认谈论心理的公共语言是对个体内在的直接经验的描述,采取否定私人语言的立场,从而将存在某种心理活动的问题看成是无意义的以最终取消他心问题,恐怕也很难将机器和人区分开来。因为根据维特根斯坦的理论,语言游戏无法将机器排除。语言游戏就在于语词在活动中的使用,使用语词取代自然表达更进一步消解了私人经验的参与,所以实质上这一过程与是否产生内在的心理过程无关,故而机器与人进行对话的过程实际上就是机器与人进行语言游戏的过程[7]。甚至纵使采取彻底或温和的怀疑论立场,直接怀疑他心存在或者或否认他心可知,也已经是将他人与机器的存在或他人与机器的可知性一视同仁地怀疑和否定。

针对他心问题,我们不应去追求那种既不存在,也是不可能的绝对无误、具有普遍必然性的知识[8]。本文也无意比较各种处理他心问题的不同方案间的优劣,而是希望通过上述讨论指出,目前在他心问题上区分机器与人是困难的。时至今日各种图像识别系统、决策系统的输入常常正是人类所处的外在环境的数字化,而包括聊天机器人在内的各类机器人的输出也已经与人高度重合。机器学习的黑盒内部不可理解,黑盒外两端又和人的外在环境和行为高度相似,无论采取什么样的立场,现有的处理他心问题的理论恐怕都难以将机器学习的黑盒排除到他心之外。换言之,不论是采取怀疑论的立场还是反怀疑论的诸立场,都只能将机器学习的黑盒作为一类特殊的他心处理,从而在该问题上赋予机器和人一样的地位。如上文提到的,将他人看作另一主体已经是一种事实,此外实践也要求我们将他人看作主体,即他人“应该”被看作主体。那么在他心问题上和人有同样地位的机器、实际上已经在认识中开始被看作主体的机器是不是也应该被看作主体?是不是应该又如何成为道德、法律上的主体?这一系列问题成为真正具有现实意义的问题恐怕不需要等待遥远的“奇点”到来,在今天这些问题已经亟待讨论。

4. 机器在人机关系中作为主体

传统的工具论将机器视为一般工具,理解为扩展身体器官行为的方式的纯粹客体,这显然已经不再适用。在计算机出现之前,还可以说机器不过是为了某种特定目的根据物理规律设计并使用的复杂工具,甚至在机器学习出现之前的计算机也只是对人的逻辑设计进行了物理执行和呈现,其每一条指令都是来自人类的逻辑,只是经由自然语言转换为编程语言并最终落实到机器语言的物理实现。换言之,包括以往的机器和早期计算机在内的工具的设计完全内在于人,逻辑上是先经历了人的构思然后才被实现,或者说是被发明出来的。此外,工具一般意义上的使用则是经历了人结合目的对工具进行分析后才实现的。所以传统工具论中,机器一方面是完全由人构造的,一方面又是被所人理解的,对于发明者是内在自知的,对于使用者也是可知的。但如前文所说,一方面机器学习的原理使得程序中出现了不直接来自人的逻辑,另一方面这些不来自人的逻辑往往数量极其庞大又极端抽象复杂。所以,人类无法再对(智能)机器像一般工具那样从人类设计的角度采取内在透明化的理解,也无法根据人类使用的需要采取外在分析通达彻底的理解。

由于传统工具论难以适用到智能机器，一种将人机关系类比为人与动物关系的机器家畜论被提出和广泛讨论。家畜论认为弱人工智能与人类的关系更像是动物(尤其是家畜)与人的关系，将人类与动物的关系也被视为人类与人工智能关系的样板[9]。家畜论承认了上述智能机器与一般工具的不同，但是在机器家畜论中，机器的工具地位并没有改变。换言之，智能机器从根本上来讲如家畜一样是以其功能的工具作用在和人的关系中存在的。不过今天的智能机器显然又与动物不同，家畜论也难以准确描述现在的人机关系。归根结底是因为在人和动物的关系中，二者之间明显的鸿沟使得人们可以用语言的使用、复杂工具的制造或是理性能力等诸多单一标准轻易地确立人的优势地位，从而否认动物的主体性，以此可以为人和动物之间的不平等地位进行理论上的解释。而排除智能机器将人从万物中凸显出的单一标准在今天几乎难以寻找，无论是语言还是逻辑亦或学习能力都不再为人所独有，甚至在诸多单一领域人的表现早已远远不如专用人工智能。

在这样的境地下将人机关系类比为奴隶关系的机器奴隶论或许更能描述现阶段的人机关系。主人之所以是主人，奴隶之所以是奴隶，正是因为主人对奴隶的支配和奴隶对主人的依附，是因为其处于这种关系中，而非只是由于奴隶和主人本身的如理性能力、语言能力的差别。人类是机器的主人，机器是人类的奴隶，这正是现实的人机关系的一种写照，这一方面体现了现阶段人机关系中人的主导，另一方面也承认了机器的智能和相对的独立性。把机器看作人类的奴隶正是因为机器已经开始被看作主体，因为奴隶虽然在主奴关系中不具有独立主体性，但是其本身被作为能思维、能认识也能实践的主体认识。机器身为奴隶在人的支配之下，仍然有其工具作用的一面，机器的行动是人的意志的贯彻事实上也是人的行动，但是机器完成人支配的行动的过程本身也是一种独立意识的体现。所以机器虽然在现实的和人的关系中沦为奴隶，服从于人的支配而不具有独立的主体性，但是已经有了一种在人机关系中获得主体性的可能性。今天在金融决策、交通管理等各领域，智能机器的优势愈发凸显，人类逐渐依赖机器来执行操作和进行指导甚至做出决策。另一方面，人类在互联网获取的信息是智能机器通过数据采集和分析后经过筛选才提供给人的。可以看到，人类对信息的获取、一些复杂问题的决策都或多或少被机器所掌控。人类通过智能机器接触世界，通过机器进行认识也通过机器开展实践，人类生活的方方面面已经在很大程度上依赖着机器变得间接化，而机器却在实实在在地“亲自”认识世界、改造世界。在现实中人对智能机器的依赖越来越高，智能机器已经开始作为中介阻隔人和世界的直接联系，人已经在一定程度上开始被机器架空了。所以说不需要等到所谓的“强人工智能”或是“超图灵机”问世，人和机器像黑格尔主奴辩证法里那样的主奴关系倒转今天就已经悄然开始了。

在这种人机主奴关系的基础上，人机关系又将会何去何从？有悲观的观点认为，机器最终会完全取代人的主体地位，人类的命运将取决于超级智能机器。这实际上是人机主奴关系的彻底倒转，机器直接认识世界改造世界，最终在人机关系中获得独立的主体性；而人类渐渐只是享受机器运行的结果，只是接受机器提供的信息，最终变得只能服从于机器，彻底沦为奴隶甚至是家畜亦或最终走向物种的终结。对于这样的结果还有不同的理解，有种超越人类中心主义的观点认为“智能机器人以人类‘继承者’的身份自觉或非自觉地登上历史舞台，成为虚实结合的宇宙中的新一代主宰；人类则在残酷无情的生存竞争中败下阵来，沦为生物进化历程中的一个阶段性的辉煌。”[10]但无论如何，这些观点都认为机器将在人机关系中获得完全独立的主体地位，同时彻底剥夺人的主体性。也有更乐观的观点认为，机器确实会以人类无法企及的速度进步并在各个领域获得优势地位，所以机器必然会在人机关系中获得更多的主体性，但这并不意味着只有机器具有主体性，并不意味着人的主体性将彻底丧失。也就是说，机器会获得主体性但不会获得无限的主体性，其“半主体”地位最终会得到确立，但永远居从属地位，或者是机器最终和人获得同等的地位，达成平等、互助的人机关系。这种机器主体性上升的界限依赖于机器会永远进行自我约束的假设，但由于机器学习的不可理解，这种假设也不可能得到确证。此外，还有一种常出现于

科幻作品中的设想，那就是随着人机联系愈加紧密，最终人和机器不再对立，而是会通过技术手段融为一体，人和机器合一成为共同的主体。

未来人机关系会如何发展？机器能在关系中取得怎样的主体性？这在今天是值得我们深思的问题。但只有不再泛泛将智能机器等同于更复杂的传统工具，直面机器学习不可理解的黑箱并承认现实人机关系中机器主体性的不断上升，才能对这近在眼前的问题进行有意义的思考和讨论。

5. 结语

由机器学习生成的、不来自于人的逻辑真正赋予了人工智能以可能，但这种机器学习的不可理解也带来了认识上的黑盒。从机器学习的运用开始，智能机器便不再以透明化的姿态出现在我们眼前，而是成为了一种本身不能被直接认识的“他心”，所能被测试的终究只能是智能的表现而非智能本身。所以从人的认识看，智能机器早已作为假言的主体被讨论，而从输入和输出端判断机器智能的图灵测试已经开始被突破，机器似乎已不得不在认识中开始被当作现实存在的主体。同时机器学习的不可理解也使得智能机器超越了一般工具，现有的智能机器在多个领域上甚至有着远超前人的表现，当今社会人对机器的依赖更是前所未有。从人机关系看，智能机器的发展同时也伴随着机器主体性的上升，智能机器不像一般工具那样完全不具备主体性，今天机器已经可以被看成是具备了一定主体性的奴隶，同时智能机器今天仍然在高速发展也意味着其主体性还会进一步上升。

近些年人工智能高速发展和得到广泛应用后，一系列相关问题接踵而至，如自动驾驶、智能决策和信息茧房的伦理及法律问题等都亟待解决，而在这些问题上，形而上的基础建构还落后于现实的需求。本文旨在强调作为主体的机器不是想象中遥远的强人工智能或超图灵机时代的特有物而是自上世纪机器学习运用以来就随着不可理解的黑盒现实存在着，并且其主体地位还会获得更进一步的确立，只有认识到并立足于这一当下现状和发展趋势才能开展有助于解决问题的哲学讨论和哲学建构。

参考文献

- [1] Tom M. Mitchell. 机器学习(英文版) [M]. 北京: 机械工业出版社, 2003: 2.
- [2] Brown, T., Mann, B., Ryder, N., et al. (2020) Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, **33**, 1877-1901.
- [3] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2018) A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, **51**, 1-42. <https://doi.org/10.1145/3236009>
- [4] 王晓阳. 为他心辩护——处理他心问题的一种复合方案[J]. 哲学研究, 2019(3): 89-100.
- [5] 肖峰. 人工智能与认识论新问题[J]. 西北师大学报(社会科学版), 2020, 57(5): 37-45.
- [6] Jones, C.R. and Bergen, B.K. (2024) People Cannot Distinguish GPT-4 from a Human in a Turing Test.
- [7] 郑炳楠. 基于维特根斯坦哲学的人-机语言游戏[J]. 东北大学学报(社会科学版), 2023, 25(1): 7-13.
- [8] 费多益. 他心感知如何可能? [J]. 哲学研究, 2015(1): 119-126+129.
- [9] 韩敏, 赵海明. 智能时代身体主体性的颠覆与重构——兼论人类与人工智能的主体间性[J]. 西南民族大学学报(人文社科版), 2020, 41(5): 56-63.
- [10] 孙伟平. 智能时代的新型人机关系及其构建[J]. 湖北大学学报(哲学社会科学版), 2023, 50(3): 18-25+168.