·构建中国阐释学（十三）：人工智能阐释·

本刊 2020 年第 1 期开辟“构建中国阐释学”专栏以来，学界围绕相关话题展开了多维度、开放性的深度讨论和争鸣。由此我们看到，中国学术传统中有着丰富的阐释学思想和经验，这是构建当代中国阐释学的可靠资源和坚实基础。今年，专栏将重点邀请学界同仁将中国阐释学传统与经验融会于当代各学科，尤其是前沿学科的学术创造与实践。本期推出“人工智能阐释”专题，其中徐英瑾教授聚焦影视作品对人工智能的误解，引导读者去思考人文主义精神的文艺作品在阐释科学产品时可能遭遇到的一些一般性问题；闫坤如教授针对目前人工智能技术的不透明性和公众的普遍隐忧，阐释了可解释人工智能的必要性与实践路径。以此阐释人工智能技术，打开论域的多维面相，为中国阐释学注入“现代元素”。

——编者按

阐释的偏差：科幻影视对于人工智能的误读及其社会影响

徐英瑾

【内容摘要】 科幻电影对于人工智能的形象刻画，具有明显的双刃剑效应。一方面，人工智能的确通过相关的科幻电影的广泛传播而获得了更为广泛的公众知名度，并因为这种传播学效应间接获得了在商业与行政方面的更多的支持；另一方面，在特定艺术规律与心理学规律指导下的人工智能形象刻画，也往往会对人工智能的技术实质作出错误的阐释。尽管从实际情况看，目前的人工智能既未必具有人形机器人的外观，也不具备主流科幻电影所赋予它们的那些能力，这些误解在电影之外的外溢效应已经扭曲了公众对于人工智能的认识，使公众形成对于人工智能的不必要的期望或不必要的恐慌。

【关 键 词】 人工智能　科幻电影　阐释　人性机器人　人格化效应

【作　　者】 徐英瑾，复旦大学哲学学院教授。（上海 200433）

【基金项目】 国家自然科学基金项目“探索研究 AI 伦理对科研环境的影响”（L2124040）；教育部哲学社会科学研究重大课题攻关项目“新一代人工智能发展的自然语言理解研究”（19JZD010）

人工智能（artificial intelligence，以下简称为“AI”）是一个具有高度技术集成性的学术领域，而其商业运用的范围却非常广泛。由于两者之间的信息不对称，AI 在专业领域内的“内部形象”

与其在公众（包括政界与商界精英人士）心目中的“外部形象”之间往往有巨大的落差。要减少这种落差，阐释活动的重要性就不容低估了——特别是基于多学科的阐释活动，是“学术创造与创新的原始动力”。[①]从信息哲学的角度看，优秀的阐释往往能够有效减少沟通双方的信息差；从语言哲学的角度看，成功的阐释往往能够将被阐释对象的概念结构顺化为阐释接收方的理解能力所能把握的新概念结构。但需要指出的是，在人类的信息传播历史上，不少诠释方案也确实造成了不同知识背景的受众的更多的误解，而对于 AI 概念的误解就是这方面的显著案例。

① 张江：《中国阐释学建构的若干难题》，《探索与争鸣》2022 年第 1 期。

严格地说，“artificial intelligence”这个词是在 1956 年才进入英语词汇的，最早想到这个英文词组组合的是人工智能的元老级人物麦卡锡（John McCarthy）——至于这个词汇本身，则在同年举办的美国达特茅斯会议（Dartmouth Conference）被学界普遍确认。与之相较，与 AI 相关的公众形象竟然是抢在 1956 年之前已经进入了民众的视野。譬如，1818 年出版的西方第一部科幻小说《弗兰肯斯坦》（*Frankenstein*）就设想了用电路将不同的尸体残肢拼凑成人工智慧体的可能性。在 1920 年上演的科幻舞台剧《罗梭的万能工人》（*Rossum's Universal Robots*）中，“人造人”的理念再一次被赋予形象的外观。至于在 1927 年上映的德国名片《大都会》（Metropolis）中，一个以女主人公玛丽娅面目出现的机器人，竟然扮演起了工人运动领袖的角色。而科幻作家阿西莫夫的名篇《我，机器人》（*I, Robot*）也是在 1950 年出版的（其中有些篇章是 1940 年代写就的），其时间也要早于给“AI”予以正名的达特茅斯会议。

为何对于一种技术样态的面向公众的诠释方案，反而会比该技术样态本身更早出现呢？这一貌似奇怪的现象，其实是由 AI 自身的特殊性所导致的。AI 的技术内核虽然艰深，但“模拟人类智慧”这一理念本身却并不晦涩。因此，该理念就很容易被一些敏锐的思想先驱者转化为一些艺术形象，由此形成对于技术形态本身的“抢跑”态势。此外，专业的 AI 科学家其实本来也就是普通公众，他们之所以能够对 AI 产生兴趣，在相当程度上便是受到了大众文化对于 AI 的想象的激发。然而，需要注意的是，此类想象所带来的惯性，却在 AI 真正诞生之后继续引导大众对于 AI 的认识，并在相当程度上偏离了 AI 业界发展的实际情况。由此所导致的情况是：直到今天，不少公众对于 AI 的认识都是由关于 AI 的科幻艺术作品所带来的。而此类科幻艺术作品本身对于 AI 技术实质的有意或无意的误读，则进一步扩大了专业的 AI 研究圈与外部公众之间的信息不对称性。

本文就试图对在全世界范围内比较有影响的一些科幻影视作品对于 AI 的错误阐释进行归类，并对此类误解对于公众的误导进行大致的评估（至于为何要以影视作品为主要聚焦点，也主要是因为在今天科幻影视的受众影响力要大于科幻文学作品）。而此类研究也能为我们提供一些具体而微的案例，让我们发现基于人文主义精神的文艺作品在诠释科学产品时所可能遭遇到的一些一般性问题。

以 AI 为主题的科幻影视作品的基本构成要素之间的内在张力

在本节中，笔者试图对以 AI 为主题的科幻影视作品的一般特征进行分析，以便为后续的讨论提供基础。从概念上说，以 AI 为主题的影视作品具有三个属性：（甲）作为影视作品，它们必须满足一般剧情片所应当满足的一些形式条件；（乙）作为科幻作品，它们应当承担起一定的科普任务；（丙）它们必须将 AI 视为故事的核心要素。不过，笔者将试图指出，这三个要素之间其实

是有逻辑冲突的。

先来看（甲）。众所周知，今日的影视剧无疑是传统戏剧的直接后裔，因此，影视创作的基本规律其实是脱胎于戏剧理论的。按照亚里士多德在《诗学》中提出的观点，戏剧本身是为抽象的理念提供了具体的展现方式。譬如，莎翁笔下的李尔王就为“心肠虽善，却不分忠奸”这一抽象标签提供了感性的展现。很显然，戏剧人物生动与否是这种感性的展现能否成功的关键，否则，过于单薄的人物设计会造成“纸片人”的观感，最后使得观众无法共情。这一文艺创作普遍规律对于科幻作品依然适用。优秀的科幻作品往往能够将特定戏剧人物的形象刻画得丰满动人，如《火星救援》（*The Martian*）中孤身留在火星的宇航员马克·沃特尼的复杂心理活动，《重返地球》（*After Earth*）中本有情感裂痕的父子在充满怪兽的地球上重新修补亲情的过程，以及《盗梦空间》（*Inception*）中盗梦师柯布对于妻子之死的沉重负罪感，都给观众留下了深刻的印象，并由此改善了观众的观影体验。

再来看（乙）。虽然科幻作品本身分为“软科幻”（即受到的科学知识的约束较少的作品）与“硬科幻”（即受到的科学知识的约束较多的作品），但“与特定的科学设定相关”依然是科幻作品的本质规定性。正是这一规定性将其与一般文艺作品区隔开来。需要注意的是，由于科学原理相对比较抽象，将其加以感性阐释的难度也就比较大。一般会用到的影视阐释技巧，是通过剧中科研人员的叙说来展现相关科学原理，如在《侏罗纪公园》（*Jurassic Park*）中，科学家向公众解释通过剪辑两栖类动物的基因片段修复恐龙基因的可能性，或将某个科学理论所预报的自然过程在荧幕上完全展现出来，如《后天》（*The Day After Tomorrow*）对于全球灾变的令人难忘的视觉重构。

最后来看（丙）。从表面上看来，与 AI 相关的科幻影视作品仅仅是将科幻影视的主题置换为 AI 而已。殊不知，恰恰是这样一种置换，导致了因素（丙）自身与因素（甲）（乙）之间的复杂连锁反应。具体而言如下。

第一，（丙）与（乙）之间产生了冲突。AI 的内部原理相当数理化，无论是符号 AI 系统的内部推理过程，还是深度学习算法的内部架构，都牵涉很多无法被影视化的技术细节。同时，此类细节也很难被镶嵌到人物的对话之中，成为剧情的有机组成部分（举例来说，即使是一个能够大致听懂基因编辑技术的观众，恐怕也无法忍受一个影视人物花五分钟去解释“卷积神经网络”算法）。一个可以与之类比的例子是：讲述博弈论大师纳什故事的传记电影《美丽心灵》，几乎也没怎么谈他的数学研究本身，而是花费大量时间去讲述主人公是如何与精神分裂症做斗争的。从这个角度看，AI 自身科学内容的艰深性，本就使得 AI 的硬核内容不容易成为科幻电影的科普内容。然而，也正是（丙）与（乙）之间的这种冲突，反而促进了下一个问题。

第二，（丙）与（甲）产生了合流。前面已经提到，AI 有两个面相：硬核的技术面相与大众心目中的 AI 形象。前一个面相涉及的是那些繁复的算法与精密的电路，而后一个面相所涉及的则是机器人的外观。既然将前一个面相镶嵌到影视作品中是不合适的，那么，影视剧创作者就会立即转向对于 AI 的公众形象的挖掘。需要注意的是，由于机器人与真人之间高度的可类比性，这种转向就会导致一个在以 AI 为主题的科幻影视作品中屡见不鲜的现象：机器人本身成为一个戏剧人物，并且被赋予了一些特定的性格特征甚至价值观。然而，这样的操作就立即会导致下面的问题。

第三，（丙）与（甲）之间的合流反过来加大了（丙）与（乙）之间的冲突。这也就是说，由于被赋予人类特征的 AI 戏剧“人物”在技术设定上完全脱离了目前的 AI 实际研究水平，这样的影

视作品的实际科普价值已经变得非常可疑。同时，也正因为这一点，此类作品对于公众的 AI 认知的误导也变得非常之大。

下面，笔者就将向读者展现：究竟在哪些方面，以 AI 为主题的主流影视作品对 AI 的技术状态做出了哪些错误的阐释。

主流科幻影视作品对于 AI 的三大误解

笔者将主流科幻影视作品对于 AI 之技术实质的误解分为以下三类。

误解一：AI 的典型出场样态是人形机器人。比如，在电影《人工智能》（*AI*）中，主人公小戴维就是一个标准的人形机器人，其外貌与一般的美国小朋友没有任何两样。在电视剧《西部世界》（*Westworld*）中，整村、整镇的机器人都被做成了美国西部牛仔的样子。日本电影《我的机器人女友》（僕の彼女はサイボーグ）亦是按照类似的思路将机器人设计成了一个美女的模样。

从影视创作的角度看，将 AI 设计成人形机器人有三点好处：（甲）这样做可以让真人演员直接扮演机器人，由此省略制作真机器人的道具成本；（乙）人形机器人的表情与动作更容易引发观众的共情；（丙）人形机器人更容易与真人产生戏剧冲突，由此推进剧情发展。但从 AI 的技术实质上看，这样的做法是多少有点误人子弟的 。相关的误解是建立在如下三个预设之上的。

预设（甲）：AI 与机器人是一回事。澄清：AI 与机器人本就是两个不同的学科领域，遑论人形机器人研究。严格来说，AI 的研究任务是编制特定的计算机程序，使得其能够模拟人类智能的某些功能——譬如从事某些棋类游戏。很显然，这样的智能程序完全可以仅仅具备一般商用计算机的物理外观，而未必具有人形机器人的外观。与之相比，机器人的建造是“机器人学”的任务，而机器人学所涉及的主要学科是机械工程学、电机工程学、机械电子学、电子学、控制工程学、计算机工程学、软件工程学、资讯工程学、数学及生物工程学等。在其中，AI 并不扮演核心角色。当然，AI 与机器人技术的彼此结合往往会导致更有趣的工程学应用案例，但是这并不意味着两者在概念上是一回事。

预设（乙）：机器人就应当采纳人形机器人的形式。澄清：即使是机器人，也往往不采用人形机器人的外观设置。以世界上第一台全自动机器人“Unimate”为例，该机器人在美国新泽西州尤因镇的内陆费舍尔向导工厂（Inland Fisher Guide Plant）的通用汽车装配线上承担了从装配线运输压铸件并将这些零件焊接在汽车车身上的工作。在经过特定调试后，这个机器人也能将高尔夫球打到杯子里，会倒啤酒。但这台机器人并没有类似人类的眼睛、嘴与皮肤。该机器唯一类似人类肢体之处，仅仅是一个机械臂，以及臂端的一个简易抓举设施。由此不难想见，就像我们可以设想只具有一个机械臂的机器人一样，我们自然也可以设想机器人被做成很多别的形状，比如鱼形与鸟形。

预设（丙）：人形机器人是智能或者灵魂的天然载体。澄清：人类其实是具有“万物有灵论”的心理投射倾向的，即会将很多具有动物或者人形的非生命体视为有灵魂者。孩童喜欢对卡通塑料人偶自言自语便是明证，而这种心灵倾向也在成人的心理架构中得到了保留。在心理学文献里，这种心理倾向被称为“人格化”（personification）或者“人类化”（anthropomorphization）。已经有文献指出，这一心理倾向有助于那些缺乏真实社会关系的人通过对于物体的“人化”而获得代偿性的虚拟社会交往方式，由此克服孤独。[①]而利用这一心理机制的广告商也能由此将产品包装的

① Nicholas Epley, Scott Akalis, Adam Waytz and John T. Cacioppo, “Creating Social Connection through Inferential Reproduction: Loneliness and Perceived Agency in Gadgets, Gods, and Greyhounds,” *Psychological Science*, vol. 19, no. 2, 2008, pp. 114-120.

外观设计得具有人形外观，以获取消费者的好感。需要注意的是，激发人格化的心理倾向运作的刺激门槛是很低的，只要对象看上去有点像人就可以了。这也就意味着，就科幻影视的观影体验而言，只要影视主创方将相关的机器人设计得像人，这样的视觉输入就会顺利激发观众的人格化倾向，由此自主赋予这样的机器人以智慧与灵魂。但这样一种讨巧的做法却在实际的 AI 研究完全行不通。具体而言，要赋予一个实际的 AI 体以任何一种实际的操作功能，都需要编程者在后台付出巨大的努力。很显然，以 AI 为主题的科幻影视往往会忽略这些辛劳的存在，好像 AI 体的智慧是某种唾手可得之物似的。这自然就会在相当程度上使得公众对 AI 的技术实质产生误解。

误解二：AI 可能具有人类所不具备的全局性知识，即所谓的“上帝之眼”，以此实现对于个体人类的压迫。譬如，在系列科幻电影《生化危机》（*Resident Evil*）中，保护伞公司的幕后操控者竟然是一个叫“红皇后”的超级 AI 体“她”（之所以叫“她”，是因为该 AI 体在片中被赋予了一个小女孩的外观）。“她”能够预知以主人公爱丽丝为首的人类团体的行动。她为了保护伞公司的利益，会毫不犹豫地杀死大批无辜的群众。无独有偶，在电影《机械公敌》中，也有一个叫“薇琪”的超级 AI 体“她”（这又是一个具有女性外观的超级程序）经过反复计算之后，得出了一个令人感到恐怖的结论：只有消灭一部分人类，才能使人类的整体得到更好的发展。她甚至还将这个骇人听闻的计划称之为“人类保护计划”。

从戏剧冲突的角度看，AI 体在这些影视作品里所具有的全局性的冷酷视角，与人类个体所具有的局部性的（但同时却更具温情的）视角构成了鲜明的对照。这种对照本身就具有了很强的戏剧要素。同时，相关影视主创人员对于 AI 体的设想也满足了一部分观众对于 AI 的设想：AI 虽然缺乏情感，但是在计算能力方面却是超越于人类的。所以，AI 能够比人类更清楚何为“大局”——尽管这并非是个体人类所希望接受的“大局”。

但上述印象是建立在对于 AI 的很深的误解之上的，因为超强的计算能力并不意味着对于全局知识的把控。实际上，任何一个智能体若要把握这样的全局知识，还需要一个针对所有问题领域的超级知识图谱，该知识图谱往往是人类智慧的结晶。例如，如果你要计算一枚导弹在各种复杂的空气环境中的轨道变化情况，你首先要有一个合理的空气动力学框架（该框架无疑是来自学术共同体的长期知识积累），并在该框架中设置大量的参数。至于如何计算这些参数，则是下一步才要考虑的问题。而在开放式的问题解决场域中，若要建立一个合适的知识图谱，此项任务对于人类建模者来说也是充满挑战的。譬如，在面对所谓的“电车难题”[①]时，任何一种比较稳妥的解决方案都需要问题解决者预设一个特定的规范伦理学立场（功利主义的，或是义务论的，或是德性论的）。众所周知，我们人类暂且并没有关于这些立场之短长的普遍一致的意见。这也就是说，不存在着一个用以解决“电车难题”的统一观念前提，遑论在这一前提下去构建统一的知识图谱。从这个角度看，作为人类的智慧转移形态，任何 AI 体也不能超越人类目前的智慧上限，就所有问题的解决方案给出一个毫无瑕疵的知识图谱。

基于上述分析，我们不妨再来审视《机械公敌》里“薇琪”的结论，即杀死一部分人类以保护人类整体的利益是合理的。她得出这一结论的推理过程是：人类的过渡繁衍已经影响了地球的安全，所以必须清除一部分人类以便为更多的人类预留出生存空间。很显然，她得出这个结论的知识框架是建立在某种粗暴的计算方式之上的。她将每一个人都视为一个消费者，并且以此为分母，让其平分世界既有的资源总量，最后得出了“资源不够分”的结论。而在这个知识框架中被忽略的因素有：（甲）人类不仅是消费者，同时也是生产者，因此，人类的有目的的劳动能够使得

① 该思想实验的原始提出者是福特（Philippa Foot），相关文献有：Philippa Foot，“The Problem of Abortion and the Doctrine of Double Effect,” *Oxford Review*, vol.5，1967，pp.5–15。

世界的资源总量被增量；(乙) 即使目前世界人口太多，也推不出未来人口会继续增多，因为我们必须考虑人口老龄化所导致的人口萎缩问题；(丙)“人类”内部有复杂的社会共同体结构分层（国家、民族、地方共同体、家庭等），因此，不存在某种将所有人的乡土背景信息加以销声后的针对全人类的生存机会再分配方案。反过来说，如果有人硬是要将所有这些参数都放在一个超级平台上予以思考的话，他就必须放弃某种全局式的上帝视角，而不得不在种种彼此冲突的立场中进行选择（譬如，在基于不同民族国家利益的不同出发点之间进行选择）。但这样的计算方案显然会固化特定人类团体的偏私，并由此激化不同人类团体之间的既有立场冲突——而不会像主流科幻电影所展现的那样，仅仅激化毫无社会背景的全体 AI 体与全体人类之间的冲突。

误解三：AI 很容易就具备与人类进行顺畅语言与情感沟通的能力。从表面上看，这一误解的具体内容是与上一条相互矛盾的，因为根据前一条误解的内容，AI 应当是缺乏情绪的。但需要指出的是，由于在主流科幻影视中 AI 已经被赋予了人格化，所以，就像影视剧中的人类角色有善、恶之分一样，AI 角色自然也就有善、恶之分。对那些“善良”的 AI 角色来说，预设其具有与人类共情与交流的能力，是主流科幻影视的标准操作模式。比较典型的案例有：在电影《人工智能》中，机器人小戴维不但能够立即学会英语，而且还热烈地渴望能够得到人类母亲的真正的爱；在动画电影《超能陆战队》(*Big Hero 6*) 中，充气机器人大白成为人类小伙伴最值得信赖的“暖男”；在电影《她》(*She*) 中，男主人公竟然在与 AI 系统 OS1 聊天的过程中爱上了这个聊天软件；在系列电影《星球大战》(*Star Wars*) 中，礼仪机器人 C-3PO 的人际交流能力甚至要远远超过人类，按照剧情设定，它能够翻译 3 万种星际语言，并凭借这个本领帮助人类主人在复杂的星际外交活动中游刃有余。

在科幻影视的场景中预设 AI 体具有流畅的人–机交流能力，显然对推进剧情大有裨益。不过，从客观角度看，以上影视作品所呈现的人机一家的美好图景，已经远远超出了目前主流 AI 所能实际提供的技术产品的水平。相关评判理由有二。

第一，机器与人类之间的顺畅交流能力，显然首先是建立在“自然语言处理”(natural language processing, 简称 NLP) 技术之上的。目前，这种技术最重要的商业应用是机器翻译 (machine translation)。不过，目前主流的建立在深度学习路径上的 NLP 技术并不像主流科幻电影描述的那么成熟。传统的深度学习程序采用的是监督式的学习方式。这种学习方式需要程序员对所有的语料进行辛苦的人工标注，编程成本很高（人工标注的意义在于计算机能够借此了解到语料处理的标准答案究竟是什么）。近年来随着互联网上语料的增多，NLP 的研究更加聚焦于非监督式学习和半监督式学习的算法。不过，虽然这些算法能够大大减少人工标注的工作量，但由于失去了人类提供的标准答案的校准作用，此类系统最终输出的错误率也会随之上升。而要弥补这一缺陷，除了提高输入的数据量之外，是别无他法的。由此不难看出，主流的 NLP 产品的技术水平的提高，是高度依赖训练数据量的扩容的。这也就反过来意味着：这种技术无法应对语料比较少的机器翻译任务，特别是对于缺乏网络数据支持的方言语料与某些个性化的口头禅的处理任务。然而，根据人际交往的常识，对于特定方言与口头禅的熟悉，是迅速在对话中拉近人际关系的不二法门。按照现有的技术，我们是很难做出一个能够像《她》中的 OS1 系统那样可以自由地切换各种英语口语而与人类进行交谈的软件的，遑论像《星球大战》中的 C-3PO 那样精通三万种语言的机器语言学家（实际上，目前的 AI 技术甚至都很难应对地球上的很多缺乏相关网络数据的冷门语言）。

第二，虽然情绪交流是人际交往的重要方面，但要在 AI 体中实现一个可以被算法化的情绪机制，其实是非常困难的。此项工作需要 AI 专家从认知心理学那里先去提取出一个足够抽象的关于情绪生成的理论，然后再设法将其实现于计算机的载体。其中，哪些关于情绪的心理学要素是仅仅对人而言才有意义的，哪些要素是能够通用于 AI 与人类的，这些需要逐项鉴别。实际上，目前主流 AI 能够做的事情，并不是让自己变得具有情绪，而是鉴别人类的情绪。比如，从 1995 年开始，美国麻省理工学院就开始了一个叫“情绪计算”（affective computing）的项目，[①]其主要思路是通过搜集从摄像机、录音笔、生理指标感知器中得到的关于人类行为的种种数据，判断相关人类究竟处于何种情绪之中。不过，计算机借以作出这种判断的算法基础依然是某种样式的深度学习机制。就深度学习的有监督学习版本而言，人类标注员需要对每张人脸图片的实际情绪状态进行语言标注，然后以此为样本，慢慢训练系统，使其也能掌握将人脸与特定情绪标签联系的一般映射规律。不过，需要注意的是，完成训练的系统即使能够精准地对人脸的情绪进行识别，他们自身也是没情绪的：一台能够识别出快乐表情的机器人没有一天自己是快乐的，而且，它们也不知道人类为何会感到快乐。这样的 AI 产品是很难产生与人类之间的真正共情的，遑论在理解人类的真实情感动机的前提下与人类展开深层的精神交流。

① Rosalind W. Picard, “Affective Computing: Challenges,” *International Journal of Human-Computer Studies*, vol. 59, no. 1–2, July 2003, pp. 55-64.

从本节完成的讨论来看，以 AI 为主题的主流影视作品其实是向观众全面掩盖了这样一个真相：现在的主流 AI 技术其实并不能支持那些在影视中展现出来的信息处理能力。当然，对于未来科技的适当幻想是科幻影视作品的天然权利。但需要注意的是，几乎所有的 AI 为主题的主流影视作品都没有向观众解释清楚，未来的 AI 专家究竟将沿着怎样的技术路径才能兑现影视主创者在影片中提出的技术许诺。与之相较，以生物学为主题的科幻电影（如《侏罗纪公园》）以及以生态学为主题的科幻电影（如《后天》），对于所涉及的科学主题的介绍要深入很多，遑论像《地心引力》（*Gravity*）与《火星救援》这样的大量基于真实宇航科技知识的“硬科幻”作品。我下面就要证明：恰恰是因为真正的 AI 知识在主流科幻作品中是如此稀疏，这些影视作品的传播，其实是加剧了公众对于 AI 的种种错误理解。

科幻影视对于 AI 的误读在影视圈外的“外溢”现象

由于影视剧在现代传媒体系中的优势地位，以 AI 为主题的主流科幻电影，在相当程度上塑造了公众对于 AI 的印象，并使得相关影视主创者对于 AI 的错误阐释被外溢为全社会的误读。这些误读包括以下方面。

第一，与科幻电影创作者对于人形机器人的青睐相对应，关于人形机器人的出镜新闻亦更容易得到公众的高度关注。以汉森机器人公司（Hanson Robotics）设计的名噪一时的人形机器人索菲亚（Sophia）为例，这是一台以好莱坞大明星奥黛丽·赫本的外貌为基准进行外貌设计的机器人。“她”能够与人类进行对话，并展现出相对自然的人类表情。因为其可人的外表所激发的公众的“人格化心理效应”，“她”在 2016 年第一次在公众面前露面之后，迅速获得了主流媒体的大量报道。“她”还在 2017 年获得了沙特阿拉伯的公民权，由此成为世界上第一台获得主权国家之正式公民权的机器人。

尽管如此，索菲亚的技术实质，无非就是人形机器人技术与下述技术的结合。（甲）针对人类的语音输入的语音识别技术——此项技术的实质，便是将人类的带有各种口音的话语都处理成机

器能够处理的标准格式。（乙）针对已经被标准化的文本信息的“聊天盒”（chatterbot）技术——该项技术的实质，就是一个通过文字信息的交换而与人类进行聊天的 AI 程序。（丙）语音综合技术——此项技术的本质，便是将在上述环节中被处理过的文本信息重新变成抑扬顿挫的语音信息，并将这些信息从机器人的扬声器中发出。

需要注意的是，大多数对于索菲亚的新闻报道，都没有如实呈现以下事实，即如上三项技术都未达到真正人类智能的水平。譬如，对于（甲）技术与（丙）技术的主要实现方式是深度学习。换言之，系统必须通过对于大量关于标准文本与相关语音之间关系的样本的学习，才能自行对相关数据进行合理的标注。与之相较，人类则可以通过简单的学习迅速把握某种标准语的方言（譬如，一个北京人不需要太多的训练就能听懂四川话；一个东京人不需要太多的训练就能听懂大阪方言；一个纽约人不需要太多的训练就能听懂英式英语；等等）。至于技术（乙），其实在 AI 发展中有着漫长的历史，至少可以上溯到 1964—1966 年研发的程序“ELIZA”。而此项技术的传统实现方式所自带的缺陷也是很明显的：相关系统只能对特定范围内的话题进行应对，而无法应对“超纲”的话题（比如，如果系统只准备好了与用户谈天气的程序，用户就不要指望它能与你讨论哲学）。在这个问题上，人类的表现显然要好于“聊天盒”（因为人类可以通过某种更一般性的推理方式来获得新领域的知识）。

按照上述分析，沙特政府授予索菲亚以公民权的做法显然只有“博噱头”的意义，因为索菲亚完全不具备一个合格的公民所需要的理解力、推理力、感知力与行动力。但是，媒体对于此类事件的大量报道，却掩盖了索菲亚的真正技术实质，由此促使公众错误地低估了仿人机器人技术与真正的通用人工智能技术之间的技术差距。而这种误导机制，与以仿人机器人为主角的科幻电影对于公众的误导机制，是如出一辙的。

第二，与科幻电影对于 AI 全局思维力的夸张相对应，现在的新闻界都会倾向于将 AI 在某些方面对于人类某些方面的胜利描述为机器对于人类的胜利。以 Deep Mind 公司的产品 Alpha Go 程序为例：AlphaGo 是第一个击败人类职业围棋选手、第一个战胜围棋世界冠军的 AI 程序，由谷歌(Google)公司旗下 DeepMind 公司开发。2016 年 3 月，通过自我对弈数万盘进行练习强化，AlphaGo 在一场五番棋比赛中 4∶1 击败顶尖职业棋手李世石，成为第一个不借助让子而击败围棋职业九段棋手的电脑围棋程序，引发了媒体广泛报道。五局赛后韩国棋院授予 AlphaGo 有史以来第一位名誉职业九段的称号。2017 年 5 月 23—27 日在我国乌镇围棋峰会上，最新的强化版 AlphaGo 和世界第一棋手柯洁比试，并配合八段棋手协同作战与对决五位顶尖九段棋手等五场比赛，获取全胜的战绩。在与柯洁的比赛结束后，中国围棋协会授予 AlphaGo 职业围棋九段的称号。

媒体对于 AlphaGo 的高度关注，显然与人类对于 AI 的某种期待有关——既然 AI 被认为是对人类智慧的模仿，那么，最先进的 AI 就应当有能力模仿人类智慧中最具代表性的一些部分——而下围棋的能力便是一种在东方文化中被高度推崇的理智能力。需要注意的是，虽然与前面提到的索菲亚不同，AlphaGo 并不具备人形的身体（并因此似乎并不与主流科幻电影的 AI 主题的演绎方式直接相关），但是 DeepMind 公司对于此项技术的呈现方式依然带有很强的“影院效果”。譬如，该公司安排全网直播柯洁在 2017 年与 AlphaGo 于乌镇对决的场面的做法，本身就具有很强的戏剧因素。虽然观众不可能看到 AlphaGo 的表情，但是，通过观察柯洁的表情，观众的“人格赋予倾向”会自然被激活，好像柯洁与之斗争的是一个真正的人，或至少是人类智能体。

然而，从技术实质上看，AlphaGo 在本质上无非就是深度学习技术与“蒙特·卡洛树形搜索”（Monte Carlo tree search）技术的结合。“蒙特·卡洛树形搜索”技术本身是一种相对传统的逻辑空间搜索技术，而深度学习在这里扮演的角色是系统能够通过它来模拟人类棋手对于大棋局的宏观感知能力，由此指导上述搜索机制在特定的逻辑空间中更仔细地进行棋局搜索，最终节省系统的运作资源。不过，尽管如此，与一般的深度学习技术一样，AlphaGo 所使用的深度学习技术也很难在不被重新编程的情况下自动拓展到与围棋无关的新的数据领域——与之相比较，人类的大脑却能在非常不同的领域之间做到举一反三。因此，作为棋类专用系统的 AlphaGo 是很难被升级为真正的通用人工智能系统的。此外，由于棋类活动本身是一种被高度定义化的活动（譬如，关于何为输赢，各种棋类游戏规则都有清楚的规定），我们也很难说这种技术是否能够被运用到那些缺乏清晰定义的人类活动之中。然而，正如主流科幻电影没有清楚地向公众展示现有 AI 技术的种种不足之处一样，主流媒体对于 AlphaGo 技术的上述缺陷的报道也是相对不足的。

第三，与科幻电影对于“AI 压迫人类”的戏剧化场面的刻画相平行，现在有不少人都对全自动开火的 AI 武器抱有过分的担心——对于此类担心的最典型的建制化产物，便是在国际上小有名气的“禁止杀人机器人研发运动”（Campaign to Stop Killer Robots）。[①]不过，从观影体验丰富的科幻电影观众的视角视之，此类运动的支持者对于“杀人机器人”的一般理解方式，无非就是依据电影《遗落战境》（*Oblivion*）中的全自动化无人机的展示形态来进行的：与现有的遥控无人机不同，这些无人机可以在人类操控员不介入的情况下完成对于目标的侦察、识别与攻击，而这些特征自然就使得那些被其追杀的人类目标很难逃脱噩运（耐人寻味的是，在这部电影中，此类无人机是为邪恶一方服务的，而代表正义的人类反抗军则缺乏与之对标的装备。由此可以看出电影主创者对于自动开火无人机的消极态度）。无独有偶，在反对杀人机器人研发运动支持者的标准叙事方式中，此类 AI 技术与军事技术的企图也被打上了负面的道德标签。在他们看来，此类结合所导致的技术产物已然完全排除人类的自由意志在“扣动扳机”这一最后环节中所起到的作用，而这一点就必然会导致机器对于“人性”的压迫——因为用纯然的机器杀死人类这一做法本身，就是纯然泯灭人性的。

然而，依笔者之见，这样的推理，其实已经是将具备自动化开火能力的 AI 系统的“自动化”程度予以了片面的高估，并在这种高估的基础上片面夸张了人类与机器的对立。实际上，即使未来的无人机达到了能够自动开火的技术高度，其飞行范围与候选攻击目标依然是人类拟定的，因此，我们切不能说人类指挥官的意志因素已经在此类兵器的开火因果路线中完全缺失了。而且，在现代战争中，无人机首先要摧毁的目标毕竟是敌军的装备（如坦克、装甲车、火炮、雷达站等），之所以这种打击往往会带来人员的伤亡，也仅仅是因为这些装备本身往往是有人操控的。因此，“杀人机器人”这个名号本身也是多少有点误导人的，“装备毁伤机器人”这个名号恐怕会更名副其实一点。从这个角度看，即使未来能够自动开火的 AI 军事装备真的问世，这些装备也会被整合到特定人类武装组织的整体架构中去，而不会另成一类，与人类整体进行对抗。当然，上述分析无法在逻辑上排除下面这种担忧，即某些装备了此类先进武器的国家会在国际军事竞争中获得过大的优势，由此导致国际军力的失衡。这种担忧所涉及的，毕竟还是人类群体之间的关系，而非人与机器的关系。此外，即使是这种担忧，也不能被过度放大，因为除了能够自动开火的武器之外，能够拉大军事强国与弱国之间技术差距的武器种类何止上百种（如高性能卫星、高超音速导弹、电子战干扰设备等），因此，将批判的注意力过多集中在能够自动开火的武器之上，也缺乏基于

① 该运动的网站是：https://www.stopkillerrobots.org/。

扎实的军事装备研究的充分理由。

另外，需要注意的是，类似电影《生化危机》中“红皇后”那样的能够通过全局推理而自动开启各种武器的超级 AI 系统，完全超越了目下的 AI 发展现状，因此，当前我们根本不用担心它们的出现可能带来的伦理问题。在可以预见的未来，我也很难设想有任何组织会有动力去研究一种完全不在人类控制范围之内的自动开火系统。

结语

电影媒介对于 AI 的形象阐释，具有明显的双刃剑效应。一方面，AI 的确通过相关的科幻电影的广泛传播而获得了更为广泛的公众知名度，并因为这种传播学效应间接获得了更多的在商业与行政方面的支持。但在另一方面，在特定艺术规律与心理学规律指导下的 AI 形象刻画，也往往会偏离 AI 的技术实质，引发公众形成对于 AI 的不必要的期望，或是激发公众对其产生不必要的恐慌。不过，这一结论也并不意味着以 AI 为主题的电影就肯定无法兼顾故事性与科学性。以电影《点球成金》（*Moneyball*）为例，该电影描述了棒球队教练比利·比恩如何通过算法进行队员遴选，然后使得原本成绩平常的球队在全美棒球联赛中一路披荆斩棘的故事。这部电影虽然在通常被归类为体育电影（而不是科幻电影），但是，其对于计算机算法在体育决策中发挥的作用，却有相对准确的描述，因此，也算是对相关的计算机知识进行了一定的普及。而且，电影本身的商业成绩也是不错的，获得超 11 亿美元票房（相关拍摄费用只有 0.5 亿美元）。很显然，类似的制片思路，亦完全可以被移植到真正的科幻语境中予以展现。当然，要做到这一点，除了需要科幻影视的创作者更虚心向 AI 业界学习之外，更需要有艺术才华的 AI 业界人士主动参与此类艺术作品的创作，并以自己的专业知识保证相关作品的学术底色不走样。当然，未来是否能够有一定数量的此类作品真正问世，就需要看各种机缘的配合了。

而影视界对于 AI 科技的艺术阐释活动，本身也提供了一个契机，以便帮助我们重新理解张江先生在《中国阐释学建构的若干难题》一文中所提出的有关于阐释活动与跨学科思维之间关系的评论。张江先生指出，德国哲学家狄尔泰对于“精神科学”与“自然科学”之间的差异的强调虽然有一定的道理，但若由此过分区分文理思维之间的分野，则会由此使得论者忽略自然科学对于人文科学的借鉴意义——特别是忽略了自然科学对于精确性的追求以及对于人文科学基本方法论的启发意义。[①]笔者认为这一评论是非常准确的。以本文所涉及的主流电影作品对于 AI 科技的艺术诠释活动为例，很多艺术创作者都以文艺创作的一般规律为指针，强行安排 AI 科技产品在影片架构中扮演的“戏份”，却由此丧失了对于此类产品的技术实质的精准刻画。这样的科幻作品并没有真正实现文理思维的真正融合，而仅仅是将一些科技要素作为外在的符号拼贴在了文艺作品的架构之上，并由此使得很多粗心的观众误认为自己已经了解到了相关科技产品的实质。从这个角度看，科幻作品的创作者本身就需要在剧本构思阶段具备融合“精神科学”与“自然科学”之分野的觉悟，并从意识深处培养对于自然科学既有成果的敬畏之心。唯有如此，才能做到科幻作品成为科学知识的可靠的诠释渠道，而非误解之来源。

编辑　张　蕾

① 张江：《中国阐释学建构的若干难题》，《探索与争鸣》2022 年第 1 期。

gain sustained and effective universal recognition and consensus on rights protection, legal trust, belief in the rule of law and government credibility. Only when the public authority treats the right seriously and treats the right well, can it embody its rational political morality. Only when the public power treats every individual in the society fairly, can the law be accepted and respected by people, and then make people believe in the law, respect the law, believe in the rule of law, and maintain the credibility of the government.

Keywords: uncertain state; take rights seriously; legal validity; belief in the rule of law

Restoration or Innovation: Consensus and Reflection on Historical Politics

Xu Yong & Yang Yang & Li Lifeng & Yang Guangbin & He Donghang & Wang Xiangmin & Tan Huosheng & Ma Xuesong & Liu Wei

Abstract: It has been more than 40 years since the restoration and reconstruction of Chinese politics. With the development of Chinese politics, the awakening of the subject consciousness of political science research and the introduction of interdisciplinary vision, more and more political scholars have realized that Chinese politics needs to move from "introduction" to "independence" and "creation". At the same time, due to the continuity of Chinese history and the complexity of Chinese politics, the research object of Chinese politics is far more than contemporary issues, but to analyze long-standing political phenomena based on the historical view of "long period and great history", and establish a political theory system with both local explanatory power and universal significance. In recent years, historical politics has become a topic of great concern in Chinese political circles. A group of scholars have carried out rich research around the basic position, major issues, methodological orientation of historical politics research. On May 28, 2022, the Political Scientist Official Account and the Historical and Political Science Research Center of Renmin University of China held an academic seminar on "Historical Politics: Consensus and Reflection". The participants had an in-depth discussion on the consensus that has been formed in historical politics and the issues that need to be reflected.

The Deviation of Hermeneutics: Misreading of AI in Science Fiction Films and Its Social Impact

Xu Yingjin

Abstract: Science fiction movies have obvious double-edged sword effect on the image depiction of artificial intelligence. On the one hand, AI has indeed gained wider public awareness through the wide dissemination of relevant science fiction films, and has indirectly received more support in business and administration because of this communication effect; On the other hand, the image depiction of AI under the guidance of specific artistic and psychological laws often makes a wrong interpretation of the technical essence of AI. Although from the actual situation, the current artificial intelligence may not have the appearance of humanoid robots, nor do it have the capabilities conferred by mainstream science fiction films, the spillover effect of the above misunderstanding outside the film has indeed distorted the public's understanding of artificial intelligence, caused the public to form unnecessary expectations for artificial intelligence, or stimulated the public to have unnecessary panic about it.

Keywords: artificial intelligence; science fiction film; hermeneutics; human robot; personification effect

Explainable Artificial Intelligence: Origin, Approach and Practice

Yan Kunru

Abstract: Artificial Intelligence (AI) technology is subversive, which has a revolutionary impact on human production and life, but the opacity of its system affects user trust, and it is difficult to ensure the security andreliability of AI technology. It is very important to explain AI technology and reveal the operation logic of AI. The risk attribute and opacity derived from AI technology are the inherent attributes of AI system. Through the technical ethics approach based on the combination of internalism and externalism, and the application ethics approach based on the combination of standardization and description, we can reveal the causal correlation of explainable artificial intelligence, and build an explainable artificial intelligence model based on causal correlation.

Keywords: explainable; interpretability; explainable AI; responsible AI