·构建中国阐释学（十三）：人工智能阐释·

可解释人工智能：本源、进路与实践

闫坤如

【内容摘要】 人工智能技术具有颠覆性，给人类生产、生活带来革命性影响，但人工智能系统的不透明性影响用户信任，也难以保障技术的安全性和可靠性。阐释人工智能技术、揭示人工智能的运行逻辑至关重要。人工智能技术的风险性、不透明性是人工智能系统的固有属性，通过立足于内在主义与外在主义相结合的技术伦理进路以及规范性和描述性相结合的应用伦理进路，可以揭示可解释人工智能的因果相关关系，构建基于因果相关性的可解释人工智能模型，发展负责任的人工智能。

【关 键 词】 可解释性　可阐释性　可解释人工智能　负责任的人工智能

【作　　者】 闫坤如，上海大学马克思主义学院院长、教授。（上海 200444）

【基金项目】 国家社会科学基金重大项目“负责任的人工智能及其实践的哲学研究”（21&ZD063）；国家社会科学基金项目“现代技术风险的认识论问题研究”（18BZX047）

自图灵（Alan Turing）发表《计算机器与智能》之后，人类开始追求机器人如何智能地行动，进而把机器能够思考作为追求目标，并尝试让机器像人类一样具有智能。人工智能（artificial intelligence）试图用数字计算机模拟智能行为，以机器方式再现人类智能。经过几十年的发展，人工智能技术在生产和生活中得到广泛应用，但在与人类互动时，人工智能技术的有效性又受到缺乏解释力的限制。正如张江教授在《中国阐释学构建的若干难题》一文指出的，算法语言就是阐释，在模拟人脑思维表达上，算法语言的理解与阐释呈何种状态，依据何种机制运作等，是我们必须面对的问题。[①]在人工智能领域，输入数据、模型，经过算法系统加工，最后输出决策结果，这一系列操作中的不透明性会影响用户信任，也难以保障人工智能系统的安全性，特别是在医疗、交通等生命攸关的场景中，阐释人工智能技术、揭示人工智能的运行逻辑至关重要。

何谓可解释人工智能

人工智能技术对人类生产、生活具有革命性影响。中国、美国、日本等都把发展人工智能作为国家战略，人工智能发展水平已成为决定国家科技实力的重要技术基础。但在人工智能技术中，存在着算法黑箱问题，即使决策结果正确，如果内部机理和因果机制不可解释，那么用户就不能理解计算机如何决策和运行，从而对人工智能技术不信任，这会阻碍人工智能的进一步发展。因此，可解释性是人工智能在实际应用中面临的主要障碍之一。复杂算法、模型及系统普遍缺乏决策逻辑的透明度和结果的可解释性，这些将导致其场景落地缺少合理性，例如无人驾驶汽车、医疗诊断等都需要增加透明性来保障人类安全，因此，需要通过捍卫可解释性从而打开人工智能技术黑箱。对可解释人工智能概念的界定也就成为学界关注的主题。

可解释人工智能英文一般翻译为“explainable artificial intelligence”（XAI）。“explanation”和“interpretation”都可以翻译为“可解释性”，但两者的内涵是不同的。“interpretation”的内涵是某人为使某事清晰或易于理解而给出的细节或理由，[②]也翻译为“阐释”。“interpretation”是将抽象概念（如预测的类别）映射到人类可以理解的领域。“explanation”是可解释领域的特征集合，这些特征有助于给定示例产生决策，[③]它是事物或者事件出现或者发生的原因，它除了具有解释项和被解释项之间的因果关系外，有时也会涉及解释的语境，需要关照解释者，因此，“explanation”也需要给出用户或者被解释者易于理解的理由，从这个方面来讲，“explanation”包含“interpretation”。因此，我们这里谈到的“可解释人工智能”具有两个面向：一是需要揭示人工智能系统的因果性；二是对解释者或者用户来说具有可理解性。

可解释人工智能概念的起源要追溯到 1983 年，斯沃图特（William R. Swartout）建议充分解释程序行为，以解决代码合理性问题。[④]1991 年，索斯威克（Richard W. Southwick）区分了作为解决问题过程的阐释和解释推理过程的阐释。他认为阐释的有用性取决于用户的期望，而用户的期望又取决于用户背景和系统环境等，系统必须向用户解释其推理。[⑤]他把解释者的理解或者信念添加到了可解释的人工智能系统中。2004 年，范伦特等（Michael van Lent）首先运用英文缩写 XAI 表示可解释人工智能，可解释人工智能的概念由此诞生。[⑥]

之后，国内外学者对“可解释性”概念进行了界定和澄清。米勒（Tim Miller）等将“可解释性”定义为“展示自己或其他主体作出的决定所依赖的原因”。[⑦]贡宁（David Gunning）等把可解释性界定为人工智能技术产生可解释的模型，并保持高水平的学习性能，即预测或者决策的准确性。[⑧]

2016 年，美国国防部高级研究计划局（DARPA）设立可解释人工智能（Explainable AI）项目，该项目从可解释的机器学习系统与模型、人机交互技术以及可解释的心理学理论三个方面，全面开展可解释性 AI 系统的研究。欧洲《通用数据保护条例》（General Data Protection Regulation，GDPR）中提出用户在受算法决策影响后“获得解释的权利”，从而降低“黑箱”带来的侵犯隐私等问题的风险。[⑨]在我国，2017 年国务院印发的《新一代人工智能发展规划》提出，要“重点突破自适应学习、自主学习等理论方法，实现具备高可解释性、强泛化能力的人工智能”。2022 年 3 月，中共中央办公厅、国务院办公厅印发《关于加强科技伦理治理的意见》，也提出了“保持公开透明”的科技活动伦理准则。

综合来看，可解释性已成为衡量人工智能透明度的重要标准，可解释人工智能是能够以人类

可理解的方式阐释其决策结果的人工智能算法及系统，它的核心是人工智能系统行为的可解释性。

可解释人工智能的本源

为了提高人工智能决策的合理性，使人工智能技术安全合理地使用，一方面人工智能系统的性能要足够好，即能够弥补人类在数据处理上的低效性；另一方面需要人工智能系统的鲁棒性能够适应或优化不同使用环境。AI 自身的安全性也是人工智能发展的重要因素。

（一）人工智能技术的风险本质

风险是客观实在性与主观建构性的统一。风险是技术的内在属性，任何技术都具有不确定性，都具有风险属性。人工智能技术给人类带来福祉的同时，也可能带来伦理关系失调、伦理规范失控、伦理价值失衡、伦理行为异化等风险，亦即人工智能伦理风险。一方面，人工智能嵌入经济社会生活，其催生的伦理风险危机，如公共安全、就业替代、算法歧视、隐私侵犯等会影响社会和谐稳定；另一方面，人工智能技术的风险属性取决于不确定性程度，如算法错误等伦理风险，均可能偏离技术创新的初衷。人工智能在发展的同时带来了一系列伦理挑战，如人工智能设计阶段可能产生道德算法风险和道德决策风险，人工智能使用阶段会产生人类主体地位被削弱、数据隐私泄露、算法歧视、道德标准固化等风险。

这些风险既具有客观实在性，是人工智能技术的本质属性，也与相关主体的价值取向和价值选择相关，体现了不同利益相关者的价值倾向和利益考量，它折射出的是人类自我道德的风险问题。如算法歧视、数据泄露会导致丧失公平的伦理风险，危害到人类的数据权利和隐私权等，也会导致人的自由的丧失等主体性异化风险，危害到人类的自由、解放等终极目标。风险是人工智能的本质属性，要想规避风险，保证人工智能的安全可控，就要保持人工智能的透明性，保证人工智能系统可以解释。

（二）人工智能系统的固有属性是不透明性

可解释人工智能系统向使用者提供人可以理解的辅助性内容，来提升系统决策过程的透明度，增强人对系统输出的信任。可解释人工智能技术则通过增强系统的透明性、可解释性，让人类更容易理解机器的行为，以及机器作出任何判断背后的逻辑，使机器以人类可理解的方式，获取用户的信任。

计算机科学最初认为任务是从输入到输出的转换。人工智能具有自我推理和自我决策能力，在输入和输出之间具有决策的“不透明性”特征。2009 年，保罗·汉弗莱斯（Paul Humphreys）对“不透明性”进行了详细分析。虽然汉弗莱斯的工作主要涉及科学学科中的计算机模拟，但他对“不透明性”的分析也是讨论人工智能中黑箱问题的重要借鉴。根据汉弗莱斯的分析，计算系统“在 t 时刻相对于认知主体 X 不透明，以防 X 在 t 时刻不知道（系统）的所有认知相关元素”。[10]根据汉弗莱斯的观点，计算系统本身从来都不是不透明的，只是对于某个特定的主体而言是不透明的。这种不透明性是一种认知属性，它涉及行动者缺乏知识。

透明度（transparency）强调人工智能系统能够给出自身工作原理的程度，是用户理解机器的基础。为了让不透明的计算系统变得透明，许多不同种类的解释可以而且应该被认为是合理的。人工智能系统本身就是一个黑箱，这个黑箱并不可怕，如医疗、人脑等都是黑箱，但并不妨碍医

疗机器人或者人类作出正确决策。但考虑到机器具有自主性和自我决策能力，算法的可行性和结构的透明性对于人工智能的善用具有重要作用。

（三）可解释性人工智能的疑难

人工智能的自主性和涌现性让人类难以对人工智能系统进行全面认识。人工智能已经从基于符号和逻辑的专家系统逐渐转为采用统计和逻辑推理技术的混合系统，这也让模型和机器学习技术变得愈发复杂且不透明。一方面，人工智能系统具有自主性。人工智能之所以被认为具有“智能”，是因为其具有自我决策、自我推理能力。这种能力在某些方面不受人类控制，是不被人类认识和把握的。另一方面，人工智能系统具有涌现性，即人工智能系统具有人工智能结构中单个要素不具有的特征。人类无法把握 AI 所有的决策和推理过程。特别是不透明决策系统的兴起导致可解释人工智能遇到困难，如深度神经网络（DNNs）、深度学习（DL）等不透明模型的经验源于高效的学习算法及其与巨大的参数空间的结合。该空间包括数百层和数百万个参数，这使得 DNNs 被认为是复杂的黑箱模型，可信度成为一个可解释的人工智能模型的主要目标。[11] 然而，根据模型诱导信任的能力来声明模型是可解释的，可能并不完全符合模型可解释性的要求。模型可解释性必须与数据隐私、模型保密性、公平性和责任性等相关的要求或约束一起解决。人类的智慧能够针对一个给出的结论进行推理和论证，但无法解释我们得出一个特定结论的复杂的、隐含的过程。这就是说，人类思维也具有复杂性，因果链未必完全确定，从这个意义上来讲，对人类智能进行模拟的人工智能也难保障完全透明。

可解释人工智能的哲学进路

（一）外在主义进路与内在主义进路相结合的技术伦理进路

外在主义进路把人工智能技术作为独立实体，对技术的负面社会后果进行人文的、伦理的反思、批判与治理。外在主义进路的学者往往把技术作为一个整体，关注人类命运，关注人类生存境遇，反思技术对人类社会和人本身的影响。内在主义（internalism）进路与外在主义进路不同，其把伦理与技术的关系从“对立”转换为“合作”，把伦理的职责从外在的“监督”转变为内在的“介入”，把关注的焦点从下游的“应用”转移到上游的“设计”，利用技术对伦理问题进行治理。可解释人工智能需要在设计之初就嵌入人类价值。维贝克（Peter-paul Verbeek）提出“道德物化”（materialization of morality）的观点，即把抽象的伦理准则嵌入人工智能设计中。通过提高人工智能的安全可靠性，或者说通过对人工智能设计的透明性的追求来保障人工智能的可解释性。荷兰代尔夫特理工大学的克罗斯（Peter Kroes）和埃因霍温理工大学的梅耶斯（Anthonie Meijers）等人提出技术哲学的经验转向，尝试打开技术黑箱，通过分析结构–功能关系关注具体的技术人工物，而不是把技术作为一个整体来看待，只有从内在主义进路对人工智能技术应用进行全面反思，才能打开技术黑箱，增强人工智能系统的可解释性。

（二）规范性和描述性相结合的应用伦理进路

人工智能技术伦理属于应用伦理学的范畴，不同于传统伦理学，它是描述伦理学，需要以问题为导向进行研究。描述伦理学先有生活形式与人工智能实践，才有可解释人工智能的规范和原则，可解释性的伦理准则是从实践中总结出来的。可解释人工智能通过设计把人类价值嵌入人工智能系统中，把伦理作为技术的内生因素，发挥伦理的积极作用。同时，要融合进传统伦理学经

典理论，例如，义务论内核是康德的道德律令，它奉行规则至上，以自律为理论前提，功利主义又称为目的论或者后果论，它以他律为理论前提，以结果为评判标准。可解释人工智能既要从出发点，也要从结果去思考。由此引出的解释包括两个维度——过程和结果，解释既是模型或者人工智能决策的结果，也是一个过程。[12]解释首先是认知过程，是对确定事件的原因的探寻，是选择这些原因来阐释的认知过程，也是解释者和被解释者之间传递知识的过程，目标是被解释者有足够的信息来理解事件的原因。解释的过程还包括可解释人工智能建模前的可解释性分析，以及构建可解释性模型的规则、特征等，还包括模型构建后的可解释性评估等内容，既要研究可解释性的理论基础，也要研究事后可解释性以及模型的解释力等问题。不可解释算法黑箱意味着只能看到结果，无法了解作出决策的原因和过程，难以建立人与机器之间的信任。

人工智能就是用计算机模拟人类思维，实现“输入－算法－输出”的过程。算法具有自主性特点，计算系统是不透明的，很难知道它们为什么做这些事情或者它们是如何工作的。可解释人工智能研究计划旨在开发分析技术，以使不透明的计算系统变得透明。

可解释人工智能的实践路径

考虑可解释性作为人工智能技术的发展因素，原因有三：一是可解释性有助于确保决策制定的公正性，即检测并纠正训练数据集中的偏差；二是可解释性通过强调可能改变预测的潜在对抗性扰动来促进鲁棒性的提供；三是可解释性可以确保只有有意义的变量才能推断出输出，即保证在模型推理中存在潜在的真实因果关系。[13]可解释人工智能就是在于对算法结构的原因的探求，在于针对多元主体建构因果相关的解释模型。

（一）解释是对原因的探求

解释往往指的是对事物原因的探求，解释可以看作是对“为什么－问题的回答”。我们称被解释的事件为“被解释项”（explanadum），用于解释它的事件为“解释项”（explanans），解释就是找到解释项和被解释项之间的关系。例如，为什么天空是蓝色？因为光的散射。上述语句“为什么天空是蓝色”是被解释项，“光的散射”是解释项。亚里士多德在“物理学”中提出“四因说”，即质料因、形式因、动力因和目的因。质料因，即构成事物的原始质料；形式因，即构成事物的样式和原型；动力因，即推动质料变成形式的外部力量；目的因，即事物产生和运动变化所追求的目的。亚里士多德认为，只要把一个事物的四个原因都解释清楚，也就认识了这件事物，他的“四因”中的每一个原因都是对“为什么”问题的回答，从这个意义上来讲，“亚里士多德是第一位科学哲学家，他通过分析科学解释的某些问题而创立了这门学科”。[14]1948 年，亨普尔（Carl G. Hempel）和奥本海姆（Poul Oppenheim）合写了《解释的逻辑研究》，[15]把对“为什么”问题的回答放在一个规律中。在他们看来，解释的哲学基础是它的逻辑形式，在解释中定律或者规则起着至关重要的作用，解释本质上是运用形式化语言对解释项与被解释项之间进行逻辑重构，解释就是论证。1984 年，萨尔蒙（Wesley C. Salmon）提出统计相关模型（The Statistical-Relevance Model），[16]费茨尔（Jomes H. Fetzer）提出因果相关模型（The Causal-Relevance Model），[17]萨尔蒙和费茨尔都不把解释作为论证，他们认为解释实质上是指出和辨别现象背后的原因和因果关系。这些解释模型都是从语义学维度提出的，关注解释项和被解释项之间的关系，不关注解释者，但解释需要关涉解释者，范·弗拉森（Bas. C. van Fraassen）和阿欣斯坦（Peter Achinstein）从语用学

视角提出了新的科学解释模型，他们认为解释需要针对被解释者的知识状态或语境的需要做调整。语用学解释观点包括范·弗拉森的语境的解释理论（The Contextual Theory of Explanation）[18]和阿欣斯坦提出的考虑到解释者意图的以言行事行为模型（The Illocutionary Act Model），[19]他们都关注解释者的背景知识和行为，关注解释者的可接受性。从对解释相关的发展历程可以看出，解释不仅仅要揭示解释背后的因果关系，还要关注解释者的解释行为和解释的可接受性。刘易斯在《因果解释》中把解释作为因果，“解释一个事件就是提供一些关于其因果历史的信息。在解释的行为中，一个拥有一些关于某个事件的因果历史的信息（我称之为解释性信息）的人试图把它传达给其他人”。[20]解释就是揭示事物的因果规律，也就是打开算法黑箱，保持人工智能算法的透明性，从而保证算法公平、避免算法偏差，赢得用户信任等。

（二）基于因果相关性的多元解释模型

可解释人工智能是多元主体性模型，解释是多元主体解释。可解释人工智能系统向用户提供可以理解的内容，来提升系统的透明度，增强人对系统输出的信任。可解释人工智能技术则通过增强系统的透明性、可解释性，让人类理解机器的行为以及机器作出决策背后的逻辑，实现机器以人类可理解的方式，获取用户的信任。在人工智能技术系统中，包括设计者、管理者、使用者、政府、企业等不同的利益相关者，不同的主体具有不同的知识结构和认知能力。因此，解释需要针对多元主体，涉及多元主体、主体知识结构差异以及理解能力差别导致解释具有复杂性，未必能找到一个可以让所有人理解的可解释人工智能模型。从这个方面来讲，我们只能把人工智能系统的可解释性作为工具，是更好地使用人工智能系统的工具，而不是作为我们发展人工智能技术的目标。可解释性是手段，并不是发展人工智能的最终目标。

解释是因果相关关系模型，可解释人工智能可以分为多种类型，例如，基于案例（case-based）模型、语境（contextual）相关模型、对照（contrastive）模型、反事实模型（counterfactual）、模拟（simulation-based）模型等多种类型。[21]虽然可解释人工智能具有多种类型，但因果相关关系模型是最根本的模型，珀尔（Judea Pearl）在《为什么：关于因果关系的新科学》中指出，以数据为导向的机器不可能具有智能，因为“我们仍然无法教会机器理解事情的前因后果”。[22]只有挖掘机器内部的因果关系才能让机器具有智能。从这个意义上来讲，因果关系比数据之间的相关关系更可靠。珀尔把“因果关系之梯”分为三个层级，第一层级是“关联”因果关系，第二层级是“干预”因果关系，第三层级是“反事实推理”因果关系。因果关系是比大数据更基本的东西，因果模型是比数据更真实的逻辑。珀尔提出大数据分析和深度学习都在因果关系之梯的最低层级，是永远不可能透过数据看到世界的因果本质的。“深度学习具有独特的优势，但这类程序与我们对透明性的追求背道而驰。即使是 AlphaGo 的程序编写者也不能告诉我们为什么这些程序能把下棋这个任务执行得这么好。”[23]按照珀尔的观点，缺少因果判断的人工智能只能算人工智障，其不能透过数据看到世界的因果本质。可解释人工智能分为两个层次：一是自省与可解读性，即机器与人类达成共同语言表达；二是自辩的能力，即机器要向人类解释其计算的机理与过程。

可解释人工智能就是找到输入–输出之间算法的因果链，明确人工智能程序中的知识，“解释知识就是关于因果机制的知识”。[24]对因果关系通常有两种不同的理解。第一种对因果关系的理解是认识论意义上的。所谓“A 是 B 的原因”只表示在正常情况下 A 类事件后面总是伴随着 B 类事件，它反映的是两类事件之间的恒常联系，这种恒常联系是我们对经验现象的概括，可能被新的经验

所修正。任何因果解释都是演绎解释或概率解释的简略形式，人们只是把预先假定的定律省略了。第二种对因果关系的理解是本体论意义上的。因果关系所反映的是自然的必然性，科学的解释本就是在揭示各种事物之间的因果关系以及世界的因果结构。解释是找出和辨别现象出现的原因。解释通过展示被解释事件如何适合于世界的因果构造来获得解释值。

汉森（Norwood Russell Hanson）认为解释是依赖语境（context dependent）的。汉森把弗雷格的语境原则用来研究因果语词，他提出几个原则：(1) 拉普拉斯式的决定论因果观根本不考虑因果语词意义的依赖性；(2) 关注 x 的原因的主要理由是为了解释 x，只有把 x 嵌入关涉别的事件 y 和 z 的概念框架中，我们才能对 x 作出解释，嵌入方式不同，则对 x 的解释就不同；(3) 并非所有的假说、解释和推理都是因果性的，但无论什么样的假说、解释都依赖于特定的语境。[25]范·弗拉森的解释模型的核心是语境，他认为一个解释就是对“为什么”的回答，正确的答案依赖语境，回答者的兴趣也包含在语境中。阿欣斯坦 1977 年在《美国哲学季刊》上发表的《解释是什么？》(*What is an Explanation?*) 提出解释关系是“对于某人 P 来说 X 解释 Y”，解释是对于解释者而言的。

可解释人工智能除了需要找到输入–输出之间的因果链，还要针对主体的知识背景与理解能力进行不同的解释。符号主义人工智能主要基于逻辑推理或者符号系统的智能模拟方法，具有良好的可解释性。联结主义人工智能运用神经网络的联结机制，把人的智能归结为人脑的高层活动，强调智能的产生是由大量简单的单元通过复杂的相互联结和并行运行的结果，联结主义的人工智能的可解释性需要对人脑单元进行科学分析。行为主义人工智能需要对人的心理活动进行分析并加以解释，才能保证人工智能的可解释性。

可解释人工智能需要针对不同知识背景用户，以用户能理解的方式给出不同的解释方法。可解释人工智能除了找到输入–输出的因果链之外，还需要增强用户的理解，满足用户差异化的需求，用户的理解需要采用不同的表征形式，不同的用户具有不同的价值追求，例如，一般用户在使用人工智能中需要可信赖的人工智能才能保证使用的安全，对一般用户来讲，专业的因果链并不重要，也难以理解；对于政府部门来说，需要公正、公平的人工智能，这样才能保障人工智能的合理应用，探求人工智能的科学原理不重要。但对于专业的技术人员来讲，可解释人工智能就是需要揭示真正的人工智能系统中的因果链。

（三）走向负责任的可解释人工智能

对人工智能信任起着决定性作用的是人工智能的可解释性问题。负责任地使用人工智能对于避免模型中缺乏公平性、可问责性和透明度所带来的风险至关重要。纠正数据、算法和社会偏见对促进公平至关重要。人工智能系统应该是可分析的，其透明程度应该是可理解的，以便用户对人工智能及其关键任务应用程序的预测有信心。可解释人工智能可以增强理解和信任，坚持公平、负责和透明的原则是至关重要的。

负责任的人工智能设计中需要借助于技术手段，把人类社会共识性的价值以编码的形式内在地嵌入智能机器中，通过机器学习实现负责任人工智能系统的自主决策。可解释人工智能设计既要关注设计标准，也要关注设计者的规范、设计的人工物的规范、考虑设计过程本身的动态性，以及道德是否可以计算、能否转变成算法等众多问题，解决了这些问题，才可以解决道德机器的可能性问题。可解释人工智能把人类价值转变为人工智能设计的技术要求，把其嵌入人工智能体中。

综上所述，负责任的人工智能实践的价值体系和信任机制构建是联系人工智能理论和实

践之间的“中介”和“桥梁”，由于人类社会的价值观念和价值体系存在多元化的特点，未来要发展具有可解释性的人工智能，要针对不同国家和地区、不同历史文化传统和发展阶段的影响，加强人机之间的理解，让人工智能技术跳出“黑箱”，成为可解释、可理解、可信任的人工智能。

注释：

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。

② Elizabeth Walter, *Cambridge Advanced Learner's Dictionary*, Cambridge: Cambridge University Press, 2008.

③ Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, “Methods for Interpreting and Understanding Deep Neural Networks,” *Digital Signal Processing*, no.73, 2018, pp.1-15.

④ Willian R. Swartout, “XPLAIN: A System for Creating and Explaining Expert Consulting Programs,” *Artificial Intelligence*, vol.21, no.3, 1993, pp.285-325.

⑤ Richard W. Southwick, “Explaining Reasoning: An Overview of Explanation in Knowledge-Based Systems,” *Knowledge Engineering Review*, vol.6, no.1, 1991, pp.1-19.

⑥ Michael van Lent, Willian Fisher, Michael Mancuso, “An Explainable Artificial Intelligence System for Small-Unit Tactical Behavior,” The Nineteenth National Conference on Artificial Intelligence, California, 2004, pp. 900-907.

⑦ Tim Miller, Piers Howe, Liz Sonenberg, “Explainable AI: Beware of Inmates Running the Asylum or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences, 2017,” DOI:10.48550/arXiv.171200547.

⑧ David Gunning, Mark Stefik, Jaesik Choi, et al, “Explainable Artificial Intelligence（xAI）,” Tech. rep., Defense Advanced Research Projects Agency, DOI: 10.1126/scirobotics.aay7120.

⑨ Bryce Goodman, Seth Flaxman, “European Union Regulations on Algorithmic Decision-Making and a ‘Right to Explanation’,” *AI Magazine*, vol.38, no.3, 2017, pp.50-57.

⑩ Paul Humphreys, “The Philosophical Novelty of Computer Simulation Methods,” *Synthese*, vol.169, no.3, 2009, pp.615–626.

⑪ Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1135–1144.

⑫ Tania Lombrozo, “The Structure and Function of Explanations,” *Trends in Cognitive Sciences*, vol. 10, no.10, 2006, pp.464–470.

⑬ Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, et al, “Explainable Artificial Intelligence（XAI）: Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI,” *Information Fusion*, no.58, 2020, pp.82-115.

⑭ 约翰·洛西：《科学哲学历史导论》，邱仁宗等译，武汉：华中工学院出版社，1982年，第6页。

⑮ Carl G. Hempel, Paul Oppenheim, “Studies in the Logic of Explanation,” *Philosophy of Science*, vol.15, no.2, 1948, pp.135-175.

⑯ Wesley C. Salmon, *Scientific Explanation: Three Basic Conception*, Princeton: Princeton University Press,1984, p.293.

⑰ James H. Fetzer, *Philosophy of Science*, New York: Paragon House, 1993.

⑱ Bas. C. van Fraassen, *The Scientific Image*, New York: Oxford University Press, 1980.

⑲ Peter Achinstein, *The Nature of Explanation*, New York: Oxford University Press, 1983.

⑳ David Lewis, *Causal Explanation*, 1986, DOI:10.1093/0195036468.003.0007.

㉑ Shruthi Chari, Daniel M. Gruen, Oshani Seneviratne, et al, “Directions for Explainable Knowledge-Enabled Systems,” 2020, DOI: arxiv-2003.07523.

㉒㉓ Judea Pearl, Dana Mackenzie, *The Book of Why:The New Science of Cause and Effect*, New York: Basic Books, Hachette Book Group, 2018, p.287, p.288.

㉔ Philip Kitcher, Wesley C. Salmon, *Scientific Explanation*, Minnesota: Unversity of Minnesota Press, 1989, p.128.

㉕ Norwood Russell Hanson, *Patterns of Discovery*, Cambridge: Cambridge University Press, 1958, p.54.

编辑　张　蕾

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