负责任人工智能的信任模塑：从理念到实践

闫宏秀

摘要：负责任人工智能作为人类对人工智能未来发展图景的一种规划，将责任与技术进行内在性的深度关联。技术的可信度与人类对自身的信任是发展负责任的人工智能必须解决的两个问题，而这两个问题的核心在于信任。就负责任人工智能而言，信任作为一条主线，以理念的形式出现在它的目标、趋势、任务之中，并覆盖了与该技术发展的所有相关方；信任作为一种价值观，为负责任人工智能发展共识的形成提供基础；信任作为一种生态系统，为负责任人工智能的发展营造合适的环境。因此，关于负责任人工智能的信任构建应从人工智能的相关方及其技术特征出发，以全过程、全范围的模式来寻求人工智能和人类之间信任的构筑途径，进而确保人工智能向善。

关键词：信任；人工智能；负责任；理念；实践

中图分类号：N031　　**文献标识码：**A　　**文章编号：**1000—8691（2023）04—0040—10

信任作为维系人类社会运作的一个重要元素，虽然社会学、心理学、经济学、政治学、人类学和哲学等学科关于其的诸多研究存有差异，但是在这些差异的背后，恰恰凸显了信任的重要性。伴随技术日益社会化与社会日益技术化，技术信任成了一个备受关注的热点。特别是在人类社会日益智能化的进程中，面对因人工智能所蕴含的风险与不确定性而带来的困境，争论也日趋激烈。希望、信心和信任被彼得·什托姆普卡（Piotr Sztompka）视为人类面临困境可以采取的三种态度。其中，希望和信心是“沉思的、分离的、远距离的、不承担责任的”①，而只有信任才是“应对技术不确定的和不能控制的未来”② 至关重要的策略，且在信任形成的过程中，责任也随之涌现。

事实上，从技术发展史与人类发展史来看，技术是人类生存的必备品。基于这种依赖某种信任也渐次涌现。此处的信任从动机的角度来看，可能是自愿性的，也可能是强迫性的甚或期望性的，与此同时，作为人类生存必备品的技术又因其风险与不确定性而带来了某种不信任，这种不信任倒逼着人类对技术信任的反思以及人类自身信任的审度。这两者的交织恰恰也为技术的不断迭代升级过程所确证。但需要高度关注的是，与以往不同的是，人工智能本身的拟人化或者类人性使得人类对其的态度更加魔幻，并将人与技术之间的信任推向了一个新的境遇，进而使其逐渐成了技术发展趋势与人类未来的一个核心问

基金项目：本文是国家社会科学基金重大项目“现代技术治理理论问题研究”（项目号：21&ZD064）的阶段性成果。

作者简介：闫宏秀，女，上海交通大学数字化未来与价值研究中心主任，教授，博士生导师，主要从事技术哲学与技术伦理学研究。

① ［波兰］彼得·什托姆普卡：《信任：一种社会学理论》，程胜利译，北京：中华书局，2005 年，第 31—32 页。

② ［波兰］彼得·什托姆普卡：《信任：一种社会学理论》，第 32 页。

题。因此，在人工智能的发展进程中，当其被用“负责任的”一词来进行修饰与界定的时候，一方面显示出了人类对技术的一种信任期冀，另一方面也暗含了人类对自身未来的某种担忧，这种担忧恰恰源自人类对自身信任的困惑。那么，负责任人工智能是否可以化解这种担忧呢？

一、蕴含在负责任人工智能发展全过程中的信任

负责任人工智能作为人类对人工智能未来发展图景的一种规划，通过负责任的研究与创新，以及负责任的使用等将责任与技术本身进行了内在性的深度关联，体现了技术伦理研究的内在路径与外在路径的有机融合。从关于负责任人工智能的规划、政策、指南、建议等来看，信任作为一条主线，以理念的形式贯穿在其发展的全过程之中，并至少已经以目标、趋势、问题、任务等方式出现。

（一）作为负责任人工智能发展目标与发展趋势的信任

从负责任人工智能的缘起来看，其提出的背景是基于人工智能在人类生活中的重要作用，其目的是旨在为人工智能在人类事务中的融合提供一种机制。该机制可以使人工智能技术能够“以一种合乎伦理的、透明的以及可追责的方式来培养信任和维护隐私，并减少其风险”[①]。因此，培养信任是其发展目标之一。在 2022 年 6 月所发布的《负责任人工智能战略》中，美国国防部对负责任人工智能的期望最终状态就是信任。[②] 在这里，信任无论是作为培养还是最终状态，都是意在消除由人工智能引发的系列问题，为人工智能的发展营造合适的氛围，进而合理释放人工智能的技术效用。

从人工智能的发展趋势来看，近年来所提出的“为社会负责任的人工智能”（Socially Responsible AI）、“为社会负责任的人工智能算法”（Socially Responsible AI Algorithms，SARs）[③] 等明确将社会责任与人工智能的发展置于一个共同的框架之中。其中，“为社会负责任的人工智能算法”系由“为社会负责任的人工智能”推演而来，因此，可将其归并到“为社会负责任的人工智能”之中。“为社会负责任的人工智能”意指“一种由人类价值观驱动的过程。在这个过程中，公平、透明、问责、可靠性和安全、隐私和安全、包容性等价值观为其基本要义”[④]。易言之，价值观通过技术的方式得以呈现，负责任的人工智能发展应满足上述价值观。

就具体的社会责任来看，有学者基于阿奇 • 卡罗尔（Archie B. Carroll）的企业社会责任金字塔框架，提出由功能责任（functional responsibilities）、法律责任（legal responsibilities）、伦理责任（ethical responsibilities）和慈善责任（philanthropic responsibilities）[⑤] 所构成的人工智能的社会责任金字塔。在这四种社会责任中，伦理责任和慈善责任两者为通用性的解释，即遵纪守法、做正确、公正和正义的事情、阻止伤害；功能责任和慈善责任则从技术和人这两个不同的向度共同指向了人工智能。如功能责任是指“所创建的技术允许计算机和机器以一种智能的方式发挥作用；慈善责任是指做一个善的人工智能公民，构筑人工智能生态系统以应对社会挑战”[⑥]。

---

① Ramzi El-Haddadeh, Adam Fadlalla & Nitham M. Hindi (2021), Is There a Place for Responsible Artificial Intelligence in Pandemics? A Tale of Two Countriesm, Information Systems Frontiers, https://doi.org/10.1007/s10796-021-10140-w.

② U.S.Department of Defense, Responsible Artificial Intelligence Strategy and Implementation-Pathway, 2022-06-22,https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF,2022-12-09.

③ Lu Cheng, Kush R. Varshney, Huan Liu (2021). “Socially Responsible AI Algorithms: Issues, Purposes, and Challenges”,*Journal of Artificial Intelligence Research*,71,1138. https://doi.org/10.1613/jair.1.12814.

④ Lu Cheng, Kush R. Varshney, Huan Liu , Socially Responsible AI Algorithms: Issues, Purposes, and Challenges, 1139. https://doi.org/10.1613/jair.1.12814.

⑤ Lu Cheng, Kush R. Varshney, Huan Liu , Socially Responsible AI Algorithms: Issues, Purposes, and Challenges, 1139-1141. .https://doi.org/10.1613/jair.1.12814.

⑥ Lu Cheng, Kush R. Varshney, Huan Liu , Socially Responsible AI Algorithms: Issues, Purposes, and Challenges, 1139-1141. https://doi.org/10.1613/jair.1.12814.

可见，上述四种责任是在将企业责任与技术责任进行融合的基础上，对“为社会负责任的人工智能”所展开的进一步审视。这种审视一方面是在力图界定并穷尽人工智能系统的多种责任，另一方面又是一个厘清各方责任的过程。而这种审视的本质与其说是在技术日益智能化与人对技术日益深度依赖的双向聚合中，关于技术之力的责任问题探究，倒不如说是关于人是否相信自己有能力对自己创造的产品做到有效可控的进一步商榷。事实上，在当下，如何“确保可控可信”[①]就是人工智能技术发展的目标和趋势之一。

（二）作为负责任人工智能研发与应用中问题的信任

这包括人工智能对现有信任体系所构成的挑战、由人工智能技术的可信度所引发的“信任”的界定问题、人类对人工智能的不信任与错误信任、人工智能是否可以负责任等的争议。比如，在医学领域中，伴随人工智能系统的应用日趋广泛，人类对其的猜疑、不信任、盲目信任以及过度信任等已经成为了医患关系必须面对的一个重要问题。医生对人工智能系统的信任、患者对医生的信任以及患者对人工智能系统的信任催生了医患关系的新样态。一方面，人工智能系统在医学领域中的有效性不容忽视，技术的稳健性提升了人对技术的信任，为医疗更为精准、为医患关系更为融洽提供了更多的技术保障；另一方面则存在着对人工智能系统的过度依赖可能会导致人际信任在医患关系中被不断转换为基于工具理性的技术信任，甚或导致以人为本的医学本意渐远的担忧，以及对医生和患者“信任”人工智能系统的这种“信任”属性表示质疑的观点。在约书亚 • 詹姆斯 • 哈瑟利（Joshua James Hatherley）看来，“即使人工智能系统可被依赖，且具有可靠性，但仍然不能被信任，且不具有可信度”[②]。同时，从产生信任动机的视角来看，人类具有产生信任的动机，拥有承担与信任相关的义务与责任的资质，但人工智能系统则不然。即使是负责任人工智能，其所指的负责任并非将责任完全归到人工智能技术自身，恰恰是重在强调该技术应当被负责任的研发与应用。退一步来看，抛开人工智能系统是否具有承担与信任相关的义务与责任的资质不谈，仅从产生信任的动机来看，人工智能系统显然是不足的。因此，人工智能系统也不能是“信任或可信度的一个适当对象”[③]。那么，人与人工智能系统的关系是否能被视为一种信任关系呢？

进一步而言，应如何看待此类“信任”呢？从人工智能作为技术的视角来看，这种信任至少可以与技术信任相等同，并且存在于负责任的人工智能发展进程中，以及人类将任务委托给人工智能系统的过程中，虽然人工智能系统自身没有信任的动机，但确实存在某种信任形成的场景与机会，且其之所以被称为是负责任人工智能在某种意义就是在于人类相信其有负责任的可能性。此时，关于负责任人工智能信任问题的追问是否会陷入一种逻辑循环呢？是否应当警惕技术信任的拟人化与人类信任的泛技术化呢？上述这些疑惑恰恰均是将信任作为一种理念，在负责任人工智能的研发进程与应用过程中所必须面对的问题。

（三）作为负责任人工智能发展任务的信任

技术产生效用的前提条件是其被使用，若不被使用，效用就无法得以生成。一般来看，就用户而言，理想的技术使用状态是操作简便却又能如其所愿地完成任务。然而，这种理想恰恰将关于技术的复杂性或者技术黑箱问题的探讨，与用户对技术的有用性以及易用性的感知置于人类赋予技术以信任的两端。人工智能则在某种程度上，将上述两端带到了一种更加微妙的境界。当用户期望技术系统本身功能完备、运作自洽、介入较少却容易操作且富有成效地完成诸多任务时，技术自身的逻辑也随之而变得更为专业化或者行业化，进而加深了用户对技术黑箱的印象。此时，一方面，基于熟悉或者了解意义上的信任将会降低；另一方面，人工智能系统的不断智能化与类人化的发展趋势则提升了用户的感知有用性与易用性，

① 国家新一代人工智能治理专业委员会：《新一代人工智能伦理规范》，http://www.ncsti.gov.cn/kjdt/xwjj/202109/t20210926_45227.html，2021 年 9 月 26 日。

② Joshua James Hatherley(2020). Limits of Trust in Medical AI, *Journal of Medical Ethics*,46(7),478.

③ Joshua James Hatherley, Limits of Trust in Medical AI, 480.

进而提升用户对其的接受度，基于合作性完成任务意义上所形成的信任得到提升。

也正是基于上述两个方面，人工智能系统的不透明性与其在人类社会中的重要作用使得对其的信任备受关注。布莱恩 • 斯坦顿（Brian Stanton）和西奥多 • 詹森（ Theodore Jensen ）将用户对人工智能的信任视为由用户的信任潜质（User Trust Potential）和系统可信度的感知（Perceived System Trustworthiness）两个主要部分组成。[①] 就用户而言，当其对人工智能系统的了解越少，而对人工智能系统的依赖性越高时，系统自身的可信度与用户对其感知的可信度之间的逻辑关系成为了人工智能发展所必须明晰的一个问题。

目前，在世界多国关于人工智能的规划中，信任被视为一个重要的发展任务。美国的电子电气工程学会（IEEE）将构建人与人工智能系统之间的正确信任层级视为一个重要议题[②]；欧盟委员会在 2020 年 2 月所发布的《人工智能白皮书：欧洲追求卓越和信任的路径》中明确提出构建一种信任的生态系统[③]；2021 年 11 月，在联合国教科文组织的第 41 届大会通过的《人工智能伦理问题建议书》中指出：“人工智能技术会加深世界各地国家内部和国家之间现有的鸿沟和不平等，必须维护正义、信任和公平”[④]；在中国关于人工智能的一系列发展规划中，信任以及与信任相关的可信、可控、可靠等均为高频词汇。

（四）负责任人工智能治理中的信任

治理是发展负责任人工智能的重要环节，其包括伦理治理、法律治理、技术治理等多个方面，且需要多学科以及多方合作才能实现有效的治理。就人工智能作为技术而言，对其的治理归根结底可以被视为是技术治理。因此，要严格区分对技术治理的信任与对技术的信任，不能以技术的稳健性所产生的信任来消解对技术的治理，并且“对技术治理的信任是保护社会和繁荣创新的关键因素”[⑤]，那么，什么样的技术治理是值得信任的呢?

关于此，可以从信任的构成来进行回应。在 TIGTech 关于信任和技术治理的研究中，将公众利益意图、能力、尊重、诚信、包容、公开性和公平性[⑥]等视为信任的七个驱动要素。与此相应的是，一个值得信任的治理机构和过程也呈现出如下七个特质：（关注公众利益的）意图、能力、尊重、诚信、包容、公开性和公平性。[⑦] 事实上，上述驱动要素与特质一方面从正向勾勒出了构建技术治理信任应有的理想元素，另一方面则从反向呈现了产生信任问题的缘由。

就负责任人工智能的治理而言，算法歧视、数据偏见、数据冷漠等在包容性、公开性、公平性等方面的欠缺所带出的信任问题是治理必须面对的问题，但治理的有效性、对治理本身的信任同样也是其必须面对的问题。恰如希拉里 • R. 萨特克里夫（Hillary R. Sutcliffe）和萨曼莎 • 布朗（Samantha Brown）在对脸书（Facebook）公司和剑桥分析公司（ Cambridge Analytica）的治理研究揭示的那样，在治理方面的失败会造成对信任的侵蚀，这种侵蚀不仅对有用技术的发展有害，而且还可能损害整个治理体系。[⑧] 因此，要发展负责任的人工智能，就必须充分解析人工智能治理中的信任。

---

① Brian Stanton, Theodore Jensen, Trust and Artificial Intelligence, NIST Interagency/Internal Report (NISTIR 8830), National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.XXXXXX https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087.

② IEEE,“Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (Version 1)”, 2016-12-13, https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v1.pdf, 2022-12-09.

③ EUROPEAN COMMISSION,“White Paper On Artificial Intelligence – A European approach to excellence and trust”,2020-02-19, https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf, 2022-12-09.

④ 教科文组织第 41 届大会审议通过《人工智能伦理问题建议书》，2021 年 11 月 27 日，https://baijiahao.baidu.com/s?id=1717586861659214975&wfr=spider&for=pc，2022 年 11 月 29 日。

⑤ Hillary R. Sutcliffe, Samantha Brown(2021). Trust and Soft Law for AI, *IEEE Technology & Society Magazine*, 40(4),14.

⑥ Hillary R. Sutcliffe, Samantha Brown, Trust and Soft Law for AI,15.

⑦ Hillary R. Sutcliffe, Samantha Brown, Trust and Soft Law for AI,15.

⑧ Hillary R. Sutcliffe, Samantha Brown, Trust and Soft Law for AI, 14.

二、渗透在负责任人工智能相关方中的信任

负责任人工智能一般指人类如何负责任地研发与应用人工智能，进而确保人工智能安全地为人类服务，带来人类福祉，因此，其重心在于人类的责任。然而，基于技术的发展所引发的将人工智能是否可以被视为负责任这一动作的发出者与这一结果的承担者的争议，进而负责任人工智能这一概念与人工智能负何种责任，如何负责任以及对谁负责任等问题关联在一起。其中，人工智能负何种责任和如何负责任是意指作为技术的人工智能是否有资格、是否有能力去负责任，即侧重于人工智能的主体性问题；对谁负责任则主要指向负责任人工智能的对象为谁，意指人工智能对委托者做出的承诺。因此，在某种意义上，负责任人工智能带出了人类对人工智能具有负责任能力的期望，以及人工智能自身可以做出其能负责任的承诺。无论是期望还是承诺，事实上都已经将信任呈现。

信任在负责任人工智能中是基于相关方的可信度，并在相关方之间的动态关联中形成的。同理，这意味着信任问题的产生也恰恰在于上述的形成过程之中。中国的《新一代人工智能治理原则——发展负责任的人工智能》将“人工智能研发者、使用者及其他相关方”视为共同承担责任的主体[①]；《新一代人工智能伦理规范》将 “从事人工智能管理、研发、供应、使用等相关活动的自然人、法人和其他相关机构等”[②]视为相关主体。基于此，负责任人工智能相关方所涉及的信任主要包括：

（一）负责任人工智能中的人际信任

人工智能技术的拟主体性或曰准主体性、能动性等对传统的人际信任模式提出挑战。信任作为一种主体间的关系会因人类社会技术化中的技术主体性潜质而出现新的变化。以负责任人工智能的相关方之间的信任为例，在诸如使用者、研发者、管理者等之间的人际信任之中，人工智能技术本身已经成为一个重要的变量深度介入人际信任之中，人与人之间的信任需要通过技术来予以背书的现象已经出现了使得人际信任面临被技术调节甚或被技术导引、规制的情景。

先不说由于人工智能技术自身的问题所带来的技术维度的不可信而导致人际信任中基于技术背书而导致的不信任、错误信任或者信任断裂。假设按照技术乐观主义的观点，技术的问题伴随时间发展变迁总归会被解决的思路。当人工智能可以通过其技术品质的提升来塑造其可信度，并构筑人类对其的信任，进而为其技术背书提供更好基础的时候，人类在人工智能使用过程中因其技术可信度过高所带来的人际信任问题则将更为值得关注。此时的负责任人工智能的责任将对人际信任构成本体论式的冲击，即信任的根基与责任的主体将出现颠覆性的变革。

如果“我们的时代是一个世界理性化、智化，特别是脱魔化的时代。这个时代的命运，恰恰是最高级、最精微的价值退出了社会生活”[③]，那么，当人类的决策越来越依赖智能系统，对人工智能技术的信任超过对人的信任的现象也将相伴而至时，若信任作为人类的一种品质被深度技术化，那么，必将出现人际信任的基点何在、属性为何以及走向何处等问题。

（二）负责任人工智能中的人机（技）信任

一般来说，信任来自委托者的给予与受托者的可信，是委托者的主观感知与受托者的客观能力二者之间的契合。人对技术的信任与否及其程度、人工智能的负责任资质与能力从主观和客观两个方面构成了负责任人工智能中的人机（技）信任。

当人类基于人工智能系统做出决策时，技术的稳健性从客观方面为构建人机（技）信任提供了良好的基础，但技术的稳健性与信任并非绝对的正相关关系。即人类信任人工智能虽然需要以技术自身的可信度为前提，但技术自身的可信度并不一定必然地带来人类对技术的信任。在这里，更恰当的表达应是，

① 国家新一代人工智能治理专业委员会：《科技部：新一代人工智能治理原则——发展负责任的人工智能》，2019 年 6 月 19 日，http://www.clii.com.cn/lhrh/hyxx/201906/t20190619_3935070.html，2022 年 12 月 9 日。

② 国家新一代人工智能治理专业委员会：《新一代人工智能伦理规范》。

③ [德]马克斯·韦伯：《伦理之业：马克斯·韦伯的两篇哲学演讲》，王容芬译，北京：中央编译出版社，2012 年，第 26 页。

在人类基于需要人工智能帮助其完成任务的意义上，人工智能被视为是“可靠的”。

因此，在人类需要借助人工智能完成任务来满足自身生活需求但却对技术及其产品不充分了解或者无法理解的情境中，人工智能应用中的人与技术之间的信任、人与机器之间的信任是负责任人工智能发展所必须面对的现实问题。如仅仅以完成任务为导向的信任是否可以算是真正意义上的信任呢？若一旦发生非理想状况，人类对技术及其产品的信任必将引发负责任人工智能的负责任是在何种意义上负责任等问题。若上述问题得不到解决，则负责任人工智能就变成了一个有待商榷的概念，因为“作为人工智能系统的用户和所有者，我们必须在人工智能系统在我们的社会中发挥作用时，对其行为和决策保持持续的责任和信任链”[①]。

（三）负责任人工智能相关方的分布式信任

就负责任人工智能而言，阿曼达 • 阿斯克尔（Amanda Askell）等人曾提出要关注人工智能行业内部的合作与集体行动的重要性[②]，而信任恰恰是合作与集体行动的基础，没有信任，上述行为将很难进行，且成本极高，并影响着人工智能在人类社会中效用的发挥。与此同时，也正是在人工智能发挥效用的过程中，与各方共同完成任务伴生的责任问题也随之涌现。

事实上，负责任人工智能一词本身至少就暗含了公众对人工智能的某种信任期冀和公司对人工智能产品能够做出可以负责任的某种承诺两个方面。反观上述两个方面，虽然前者是从用户出发，后者是从技术源头出发，但蕴含在二者背后的恰恰是多方信任的汇聚。

针对人工智能的分布式、多元主体、技术能动性等技术特征，分布式道德责任再次进入大众视野，并成为负责任的人工智能建设的一个难点与痛点。虽然早在 20 世纪 80 年代，丹尼斯 • 汤普逊（Dennis F. Thompson）就已经提到“多手问题（the problem of many hands）”[③]，但其是围绕人与人之间而展开的，而负责任人工智能的分布式道德责任则是将责任的追溯已经由人以及由人组成的机构拓展到了技术、人与技术的交互之中。这种拓展一方面意味着责任主体的扩展，另一方面则意味着信任主体和信任方式的拓展，特别是人与技术交互意义上的分布式信任是负责任人工智能发展必须厘清的问题之一。

三、信任在负责任人工智能中的画像

信任被视为是构成“复杂性简化的一种更为有效的形式”[④]，是推进人类社会合作的一个重要条件。在负责任人工智能的发展中，信任以发展目标、趋势、任务以及问题等的形式出现，并被视为对抗风险与不确定性的策略。上述这些表述虽然呈现出了信任在负责任人工智能中的新特质与新意涵，但关于信任与负责任人工智能的关系及其对于负责任人工智能意义的理解，还应当从更高、更广的意义上来展开才能深度呈现。

（一）作为一种价值观的信任

当人类研发与应用人工智能时，对人类自身负责任能力的不信任、缺乏信任或者盲目信任，以及人工智能技术性能的可信度是负责任人工智能发展所必须予以回应的问题。然而，对于上述问题的回应，不仅仅在于技术可信度，因为信任本身是“一个附着力很强的概念，其含义总是伴随社会与行为自身情形特征而特征化”[⑤]。事实上，在当今的人工智能发展进程中，“文化之间的不信任已经阻碍了人工智能

① Virginia Dignum(2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, Switzerland:Springer Nature, 5.

② Amanda Askell , Miles Brundage, Gillian Hadfield , The Role of Cooperation in Responsible AI Development,2019-07-10, arXiv preprint arXiv:1907.04534[cs.CY], Cornell University.

③ Dennis F.Thompson(1980). Moral Responsibility of Public Officials: The Problem of Many Hands,*The American Political Science Review*, 74(4),905.

④ Niklas Luhmann(2017). *Trust and Power*, Translated by Howard Davis ,John Raffan,Cambrigde:Polity Press,9.

⑤ 翟学伟：《中国人的社会信任：关系向度上的考察》，北京：商务出版社，2022 年，第 366 页。

的治理”[①]且后者因其历史性、区域性、个体性以及很难被精准量化等特征而更值得深思。

当文化之间的不信任成为一种阻碍时，对其不信任缘起的反思与对其信任构建的尝试从两个不同的方向指向了上述问题的破解，而文化的核心是价值观。“人类一切行为的旨趣在于追寻某种特定的价值”[②]，同样地，信任也是基于自身价值观所做出的判断。价值观是人类社会的锚定桩，在社会这一大系统中，作为社会资本[③]运行的价值观在社会发展的历程中具有根基性和引领性的功能。这种功能在凸显了价值观重要性的同时，更是揭示出了构建合适的价值观的必要性。

毫无疑问，技术可信度是负责任人工智能技术发展的应有之义，但诸如文化之间的不信任问题就是技术可信度无法直接破解的问题。仅仅就对负责任人工智能这一概念的理解，即使是同样的技术或技术产品，由于价值观的区域性与个体性差异就可能带来理解歧义，并进而会引发信任摩擦。这种摩擦虽然属于软实力或者柔性意义上的摩擦，但其所带来的后果却不容小觑。在关于“为社会负责任的人工智能”的界定中，就明确指出“满足对共同价值观的社会期望是主要目标，此处的共同价值观既包括提升人工智能的能力，也包括提高人工智能为社会带来的益处”[④]。

进一步来看，就负责任人工智能而言，应以技术与价值的关系为切入点，从价值观的视角出发，将信任视为一种基于且必须高于技术可信度的理念，即超越囿于技术评价的量化标尺，打开不同文化之间的不信任。与此同时，不容忽视的是，信任是允许不同的价值观并存的。因此，应在解析价值观差异性的基础上，寻找形成文化之间不信任的价值观基础，构建关于负责任人工智能的共识，走向包容性的审慎式信任。

（二）作为一种生态系统的信任

纵观历史，任何技术的发展与应用都是在社会中进行的。恰如弗吉尼亚•迪纳姆（Virginia Dignum）在其关于负责任人工智能的阐释中所指出的那样：“显然，人工智能的应用（程序）不是负责任的，必须承担责任并确保信任的是应用（程序）所属的社会——技术系统。”[⑤]这意味着需要在技术与社会的关联之中，从更广的意义上来审视信任与负责任的人工智能。

从纯粹技术的视角来看，假设对人工智能的信任阈值为 0 到 1。当信任度为 0 时，人类完全不信任人工智能，这种信任生态显然不利于人工智能的发展；当信任度为 1 时，人类完全信任人工智能，同样也不是人工智能发展的合理生态。简言之，鉴于人工智能对人类社会的重要作用，信任缺失或者完全不信任的态度均不可取；鉴于人工智能在技术方面所存在的不稳健性，完全信任的态度同样也不可取。因此，针对在人工智能发展过程中所存在的上述信任问题，需要一个恰当的信任生态才能保障其可以安全与可持续地发展，应当从技术社会学的视角即技术所属的社会系统来探讨信任。此时的信任应被视为由公众对负责任人工智能所给予的信任、负责任人工智能相关方之间的信任、负责任人工智能所处的社会信任状况等多个要素组成的一种生态系统。

因此，从对技术信任的视角来看，需要为信任设定一个合理的阈值，并且负责任的人工智能信任阈值的设定与人工智能所应承担的规范性义务和责任紧密相关。但更为重要的是，作为一种生态系统的信任而非孤立的个体式的闭环信任才能为发展负责任的人工智能提供合适的环境。

---

① Seán S. ÓhÉigeartaigh, Jess Whittlestone, Yang Liu, Yi Zeng & Zhe Liu(2020). Overcoming barriers to cross-cultural cooperation in AI ethics and governance,*Philosophy & technology*, 33(4),571-593.

② 闫宏秀：《技术过程的价值选择研究》，上海：上海人民出版社，2015 年，第 70 页。

③ [美] 弗郎西斯•福山：《大分裂：人类本性与社会秩序的重建》，刘榜离译，北京：中国社会科学出版社，2002 年，第 18 页。所谓社会资本可以简单地定义为一个群体之成员共有的一套非正式的、允许他们之间进行合作的价值观或准则。

④ Lu Cheng, Kush R. Varshney, Huan Liu , Socially Responsible AI Algorithms: Issues, Purposes, and Challenges, 1138. https://doi.org/10.1613/jair.1.12814.

⑤ Virginia Dignum,Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way ,V.

（三）信任在负责任人工智能中的画像层级及其逻辑关系

在智能革命的背景下，虽然人工智能技术对于人类社会发挥着极为重要的作用，但这不意味着其可以仅仅依据技术的逻辑发展。现如今，对人工智能治理的热议、多个层级的诸多政策与法规等的出台就体现了人类对其接受的谨慎性，而人类对其的接受与否，以及接受度取决于人类对其的信任，而“这种信任在很大程度上是由相关的法律环境与标准以及能够保证诸如安全性、可靠性、透明性和可解释性等关键特征的技术背景来决定”[①]，技术要素与非技术要素从两个不同的维度走向信任。那么，究竟是什么在影响公众对人工智能信任呢?

从宏观层面来看，将人工智能技术视为一个整体，妮可 • 吉莱斯皮（Nicocle Gillespie）等通过对来自美国、加拿大、德国、英国和澳大利亚的 6054 份样本分析，得出“关于当前法规和法律是否足以确保人工智能安全使用的信念、人工智能对就业的影响、对人工智能的熟悉和理解、人工智能对社会影响的不确定性”是影响公众信任人工智能系统的四个关键要素的结论。[②] 这四个要素包括两个维度，一是人工智能技术本身对人类的影响，如第二要素和第四个要素；二是人类对人工智能技术的治理与认知，如第一个要素和第三个要素。从微观层面来看，就人工智能在具体场景中的应用而言，以借助人工智能审核在线内容为例，玛丽亚 • 莫利纳（Maria D. Molina）和希亚姆 • 桑达尔（ S. Shyam Sundar）发现，不信任他人的用户可能是因为他们不相信其他人能够无偏见地正确分类内容，从而相信机器更准确和客观，进而对机器持有更积极的态度，因此，其更倾向于信任人工智能而非人。[③] 在这里，人际信任与人机（技）之间的某种关联性被呈现出来。

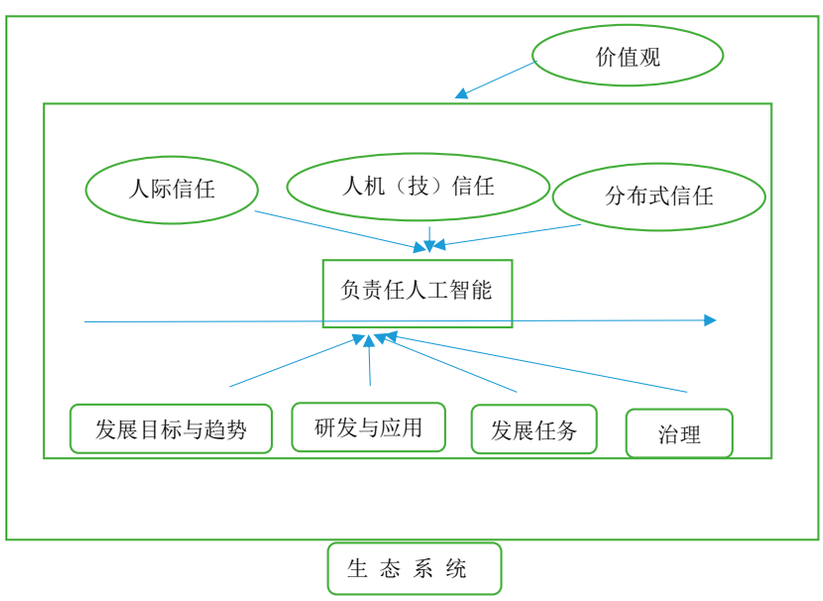

图 1　信任在负责任人工智能中的画像（资料来源：自制）

因此，信任以发展目标、趋势、任务等贯穿在人工智能发展的全过程中，并在与负责任人工智能相关的所有方之间产生信任。上述两个方面从技术过程与技术主体两个向度形成了负责任人工智能的信任要素。信任作为一种价值观则为负责任人工智能发展共识的形成提供基础，信任作为一种生态系统则为负责任人工智能的发展营造合适的环境。

① István Mezgár,József Váncza(2022). From ethics to standards – A path via responsible AI to cyber-physical production systems, Annual Reviews in Control ,53,402.

② Nicocle Gillespie,Steve Lockey,Caitlin Curtis, Trust in Artificial Intelligence: A five country study,2021-03, https://kpmg.com/au/en/home/insights/2021/03/artificial-intelligence-five-country-study.html,2022-12-19.

③ Maria D. Molina, S.Shyam Sundar,Does Distrust in Humans Predict Greater Trust in AI? Role of Individual Differences in User Responses to Content Moderation. New Media & Society, 2022-06-23, https://doi.org/10.1177/1461444822110353.

四、负责任人工智能的信任构建

在助推人工智能构建美好生活的进程中，发展负责任的人工智能作为人类认识到人工智能具有风险或不确定性时所探寻的一条出路，既是人类对人工智能的一种期待，也是人类期望人工智能能够给予某种承诺。然而，无论是期望性的信任还是承诺性的信任，都需要以相关方的信任为基础。因此，从信任在负责任人工智能发展过程中的产生环节与所涉及相关主体来看，关于负责任人工智能的信任构建应从人工智能的相关方及其技术特征出发，以全过程、全范围的模式来寻求人工智能和人类之间信任的构筑途径，进而确保人工智能向善。

第一，系统探寻负责任人工智能中信任的构成要素，并将其嵌入到人工智能从技术设计到评估的全过程之中。

从关于负责任人工智能的相关文件来看，信任在其发展目标、趋势、任务等方面均有所提及，但如何得以充分体现则有待进一步的研究。负责任人工智能并非仅仅是一个技术问题，而是一个国家发展战略，其所涉及的信任相关方包括技术界、产业界、公众、政府等多个层级、多个领域。在此中，任何一方、任何一个环节对负责任人工智能所存在的错误信任、信任缺位等都会阻碍人工智能的发展。

因此，关于负责任人工智能中信任构成要素的探寻应当从正向与反向两个向度展开。正向指探寻信任应当包括什么，反向则指探寻信任的断裂之处与不信任产生的环节。就技术设计而言，反向表现为通过社会实验、情景模拟等方式挖掘负责任人工智能因其未能践行负责任而导致的信任问题，并将其通过技术设计环节进行应对或者规避，进而充分体现人工智能技术的负责任意蕴；正向表现为负责任人工智能的设计意图应当包括对公共利益的注重、对公众的尊重与友好等技术设计的通用原则，但更应以技术、技术物的形式实现伦理观念内化的方式将负责任予以有效呈现，因为“每一项技术都是人类依据自身的需求来塑形和运用自然的一种介入”[①]。

第二，通过多渠道、多层级的方式培养关于负责任人工智能的信任认知，以关于信任的教育促进对负责任人工智能的理解与接受。

目前，负责任人工智能中的信任主要指向设计者与制造者关于技术可信度的提升。事实上，这种提升除了依赖技术本身的发展之外，还包括设计者与制造者以多种途径主动公开风险、适度披露技术及其应用场景，以开放诚实的沟通方式，让公众了解负责任人工智能的技术理念、原则和相关政策，并消除公众对人工智能技术因认知不足而产生的疑虑，进而在了解的基础上走向合理的信任，协同发展负责任的人工智能。

但需要再次强调的是，负责任人工智能“不是人工智能系统的特征，而是我们自己的角色”，其本质是在于“我们对负责任的人工智能负责”[②]。此处的“我们”包括设计者、制造者与使用者等在内的所有相关方。众所周知，人工智能产生作用的一个重要关节点是在于其是否已使用及如何被使用。与此同时，伴随技术的发展，人工智能与人之间的互为操作性与交互性所带来的结果多样性使得用户的身份逐渐从被动消费者逐渐变成了积极参与者。因此，当设计者与制造者践行负责任的理念时，还应高度关注使用者作为技术发挥效用的终端对负责任人工智能的信任认知。特别是对于使用者而言，人工智能在不同的语境中可能做出不同的行为，即使初始值一样的人工智能也可以产生不同的行为，并可能以设计者未曾预想的方式进行自我改变，进而带来出乎意料的后果。那么，面对这样的改变，使用者该如何去信任呢？

事实上，信任本身就“意味着事先已经意识到了风险的存在”[③]，且具有弹性、脆弱性，尤其是当委托人的信任水平超过其对风险的感知时，关于信任的认知将成为理解与接受负责任人工智能的一项重要

① Alois Hunnng(1999). Preferences and Values Assessments in Cases of Decision Under Risk, Techné: *Journal of the Society for Philosophy & Technology*, 4(4),15 .

② Virginia Dignum(2019). Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way ,7.

③ [英] 安东尼·吉登斯：《现代性的后果》，田禾译，南京：南京译林出版社，2000 年，第 27 页。

内容。对于此，可通过政府、学校、社区等渠道向公众普及信任的重要性、必要性与合理性等，并以对信任、不信任、虚假信任等相关的问题和案例进行公开讨论的方式，消除公众对信任的误解，进而培养正确的信任观。

第三，建立关于负责任人工智能的审查机制，以合乎伦理的监督性信任和结构性道德责任为纽带的原则，构建合理的信任阈值。

面对科学技术发展，在尼古拉斯 • 卢曼（Niklas Luhmann）看来，“用对事物的控制取代作为一种社会机制的信任”[①]是不可取的，“相反，人们应当期望，信任作为忍受技术在未来将带来复杂性一种手段，人类对其的需求将与日俱增”[②]。在当下，这种复杂性伴随人工智能的发展更加备受关注。此时，信任作为一种应对手段，一方面对其的需求更为迫切，另一方面则意味着对信任提出了更高的要求。

虽然负责任人工智能的发展离不开信任，但这种信任恰恰是需要基于对负责任人工智能进行审查的基础上才能界定信任的边界，并且这种审查包括伦理审查。通过这种审查，建立角色责任，梳理问责流程，并以结构性道德责任[③]应对在负责任的人工智能中由分布式而带来的责任混淆问题，以合乎伦理的监督性信任促进技术与伦理之间的有机融合，促进人与技术之间有效合作。

Modelling Trust of Responsible AI: From Idea to Practice

YAN Hong-xiu

（Digital Future and Value Research Center, Shanghai Jiao Tong University, Shanghai, 200030）

Abstract: Responsible AI, a planning for the future development of artificial intelligence, internally and deeply associates responsibility with technology. The trustworthy technology and the human trust in ourselves are two issues that must be solved to develop responsible AI, and trust is the core of these issues. As for responsible AI, trust is a main line which appears together with its goals, trends, tasks in the form of idea, and covers all relevant parties in developing responsible AI. As a value, trust provides the basis for forming a consensus on the development of responsible AI; as an ecosystem, it creates a suitable environment for developing responsible AI. Therefore, the trust construction of responsible AI should start from the relevant parties of AI and their technical characteristics to explore the way to build trust between AI and human in a whole-process with a whole-range mode to ensure AI for good.

Keywords: Trust, AI, Being Responsible, Idea, Practice

[责任编辑：谢雨佟]

① Niklas Luhmann, Trust and Power, 18.

② Niklas Luhmann, Trust and Power, 18.

③ 结构式道德责任是分布式道德责任的一种补充，其作为一种底层架构，为分布式道德责任的后向传播阈值设定提供理论依据，其依托于一个宏观的伦理主旨，且该主旨渗透在分布式道德责任之中。人类社会基本的伦理准则是宏观伦理主旨的底线，科技为善、构善、至善是宏观伦理主旨主线。参见闫宏秀：《数字时代的道德责任解析：从信任到结构》，《探索与争鸣》2022 年第 4 期。