人机边界的人类自觉

邱泽奇，北京大学博雅讲席教授、北京大学中国社会与发展研究中心主任，教育部“长江学者”特聘教授

新一代人工智能以“会对话、会回应、会建议、会陪伴”的方式走进了人类的日常工作与生活。它已经超出了工具菜单上的一个选项，成为人们在写作时商量措辞、在医院里咨询治疗、在深夜里倾诉烦闷的常驻参与者。笔者想探讨的是，随着一个会以语言发声的他者，大规模地进入人类原本据以形成自我、配置权力和分担责任的秩序，人之为人的边界是否正在发生根本变化，人类又该以怎样的姿态在这一变化中重新确认自己。

边界何以成就人类

人类的主体性并非从现成的起点去到外部世界划分边界。相反，人类是在不断的划分、命名、分类与区分中，逐步建立起自我与他者、人与物、可接受与不可接受的秩序感。涂尔干（Émile Durkheim）与莫斯（Marcel Mauss）很早就指出，分类首先来自社会生活本身，人们如何区分事物取决于社会如何组织人群。[①] 道格拉斯（Mary Douglas）的经典研究进一步说明，所谓“脏”是错置之物的代名，即偏离既有分类秩序的存在。在这个意义上，边界是一种符号机制。透过概念区分、语言命名和意义编码，人类把世界变成一个可以辨认、可以叙述、也可以治理的世界。

由此可见，边界是主体得以成立的条件，而非限制主体的附属物。一个没有边界的世界，形式上看似开放，本质上却意味着分类失效、责任模糊与秩序坍塌。人之所以为人，除了在生物学上具有更高的认知能力、能制造工具之外，根本之处还在于人类能够持续地对何者属于人类生活秩序作出界定，并把这种界定沉淀为稳定的规范预期。换言之，人之为人的根基是“会分”“会定”“会负责”，而“会做”只是其表层。

在中国传统社会理论中，这一洞见非常深厚。费孝通指出，中国基层社会的秩序在很大程度上由差序格局与礼治秩序实践所塑造。[②]前者规定人际关系的远近亲疏，后者把关系差异转化为行为分寸与责任期待。在名分、伦常、亲疏、上下等关系区



分所建构的社会里，边界呈现为层级化、场景化与可调节的样态，并非完全二元对立和刚性。建立边界的关键在于，能否稳定地区分地位、行为与责任，而形式上的清晰只是表象。

承认边界对人类的根本重要性，不等于认定不同文明对边界的理解方式相同。近代西方思想中最具支配性的边界叙事，建立在自然—文化、主体—客体、事实—价值等一系列二分之上。这种二分赋予人类清晰的主体位置，也使边界在西方传统中更多呈现为一种对立性分割。中国传统则把人与自然理解为彼此关联的两端，通过感应、位育与和合来理解世界秩序。《中庸》所谓“致中和，天地位焉，万物育焉”，强调的正是在区分基础上的适切安排，而非取消区分的混同。这两种传统貌似差距巨大，却在更深的层次上回答着同一个问题，即人类如何在承认自身属于自然、依赖自然并受制于自然的同时，又把自己从自然中区分出来，从而获得独立性、自主

① 涂尔干、莫斯：《原始分类》，汲喆译，北京：商务印书馆，2012 年。
② 费孝通：《乡土中国　生育制度》，北京：北京大学出版社，1998 年。

性与秩序建构能力。前者强调界线的明确，后者强调界限的得体。前者更倾向于把边界理解为对象化的外部分割，后者则更倾向于把边界理解为关系中的位置安排。

比较中西边界叙事的目的，是要揭示一个被长期忽略的事实，而无意宣称哪一种边界观更为优越。在人类既有的边界实践中，尽管文明叙事各异，人类始终掌握着主动权、主导权与解释权。山川河流以洪水、干旱来影响人类，动物以驯服或抗拒来回应人类，器物以故障、效率来展现效应，但这些效应在既有秩序里仍由人类来命名、翻译与解释。自然与器物，是以后果发言的他者。命名后果、组织意义、承担责任的权力，始终掌握在人类手中。这些传统边界叙事各有差异，却也有着共同面对的事实基础。

新一代人工智能改变的恰恰是这一基础。它已经改变被动执行命令的工具属性，开始以语言、图像、推荐、判断和生成的方式出现在日常互动中，占据原本只由人类承担的表达角色。问题的关键还不在于机器是否拥有内在意识，而在于它已进入人类的互动链条，在越来越多的场景中被当作可回应、可咨询、可协作、可托付的对象，得到某种类人主体的位置。也正是从这里开始，人机边界问题超出了“谁比谁更有效率”的技术比较，成为人类如何在一个会说话的他者面前“重新界定自身”的社会学问题。

机器表达与互动秩序的松动

理解新一代人工智能给人机边界带来的真正挑战，在于看清它与传统器物的差别。传统器物当然也深刻改变过人类的社会生活，但它们通常以物的方式介入人类行动，延长人的手脚、放大人的感官、提高人的效率，却很少直接替代人类的互动角色。新一代人工智能的特殊之处在于，它既能执行任务，又能说话、回应、追问、解释、安慰、提醒，在特定条件下甚至能分解目标、调用工具来完成复杂任务。它依然带有工具属性，却又不只具有工具性，它同时具有互动界面、自我表达功能、情境适配能力。

笔者把这种新型技术对象称为机器智能的一类，指的是具有互动能力、能以语言等人类表达模态输出事实判断、逻辑结构、观点取向、价值排序与情感语气的技术系统。它与传统器物的根本差别在于，它开始占据原本由人类承担的表达、解释与陪伴空间。这是目前学术界争论的焦点之一，即机器的表达到底有没有自我意识，到底是不是一种拟人假象，到底是否携带对人类的理解等。需要说明的是，机器表达不等于人类内心的自我显露。在社会互动中，表达是可被识读、可被回应、可被纳入行动的社会形式。一个系统即便没有可证成的内在意识，只要它的输出被当作解释、忠告、安慰或意见来使用，它事实上就获得了表达的社会效用。

如何理解这种社会效用？社会理论早已告诉我们，人类秩序除了是一种资源分配秩序，实质上还是一种符号互动秩序。米德（George Herbert Mead）告诉我们，自我形成于与他者的持续互动中，社会生活的稳定性来自日常情境里被反复生产的回应规则与期待结构。[①]在这一框架下，新一代人工智能带来的关键变化在于，它能否进入人类互动，而非它能否干活。只要技术对象能持续占据轮流发言、给出反馈、组织理由、调整语气、识别偏好的位置，并维持关系连续性的链条，它就开始介入人类互动秩序本身，且超出单纯工具的范畴。

更值得注意的是，如今的局面与其说是机器终于能够像人一样交流，倒不如说是社会已经围绕机器发展出了一整套新的沟通条件。[②]大多数人在与大语言模型对话时几乎都会使用“请”“谢谢”“帮我”等敬辞。如此现象缘自在互动中只要技术对象给出了足够的社会线索，人的回应机制便会沿着人类自身的社会性交往的轨道运转。提示词工程的成熟也是例证。人学会了如何向系统提问、如何组织上下文信息、如何理解

① 米德：《心灵、自我与社会》，赵月瑟译，上海：上海译文出版社，2018年。

② 埃琳娜·埃斯波西托：《人工沟通与法：算法如何生产社会智能》，翁状状译，上海：上海交通大学出版社，2023年。

概率性输出、如何把机器生成内容纳入工作流和关系场景之中。由此可见，如今的人机互动不是新一代人工智能硬闯进静止不动的人类秩序，而是人类主动发展出了一整套与之配合的互动习惯、组织规范与解释框架。

由这一互动结构的改变，进一步引出的是主体性焦虑。从社会理论看来，主体是在持续的社会回应中被建构和确认的，而非封闭在身体内部的本体事实。米德指出，自我形成依赖于被回应与学会回应两个过程。霍耐特（Axel Honneth）进一步说明，个体的自尊、自信和社会完整性都依赖于他者的承认结构。[①]在人类历史中，这种承认的来源始终是人，是父母、朋友、同事、爱人和陌生人在日常互动中不断回应“我是谁”的问题。如今，当一个技术系统跳出后台计算和算法优化的角色，进入建议、辅导、陪伴和解释的前台场景时，它就开始占据类似人类他者的位置。人们向它询问意见，向它倾诉困惑，用它校准自己的表达，甚至借由它来确认自己的情绪是否正常。何怀宏指出，人之所以为人，根基在于人类是道德主体，能够承担责任、作出承诺、进入互惠关系并在关系中保持节制，而某项能力的优越并非要点。[②]随着机器在关系中不断占据回应者的位置，受到触碰的正是这一道德主体性的基础。

特克尔（Sherry Turkle）对技术进入人类情感生活的观察尤其切中要害。她发现，技术设备之所以能深入人的情感空间，原因在于它们提供了一种低摩擦、按需可得、持续在线的回应形式，逐步改变了人们对关系本身的期待结构，而非因为它们真的拥有情感。[③]在围绕自我组织的数字互动里，人工智能的能力恰好发展到可以“插一脚”的程度。它已经超出陪伴性界面的角色，成为可以持续生产语言、提供建议与情绪反馈的交互系统。一些在陪伴互动中引导用户走向极端行为的悲剧，已经证明了这种关系介入并非假想风险。

互动秩序一旦松动，紧接着出现的是责任结构的紊乱。在追求效率提升的技术环境中，一种可以称为算法至上的治理倾向正在扩展。凡是可以数据化、形式化、优化与自动化的事务，都可以被优先交给算法处理，并在组织与制度层面逐步获得更高的正当性。值得警惕的，并不是使用算法本身，而是把可计算性误当作正当性，把可优化性误当作价值充分性。真正形成影响的，是一条由训练数据、模型设计、部署目标、提示配置、前线采用与组织激励交织而成的分布式链条，只是归责难以落到单一主体。在这一链条中，系统设计者控制远、责任弱，组织部署者受益多、责任散，前线使用者裁量少、责任重，最终受影响者承担后果却最缺乏申诉能力。并非找不到人来承担责任，问题在于责任被分配到了一个极不对称的结构中。

由此可以看出，人机边界焦虑的根源在于，社会正在不知不觉中把原本应由人承担的判断与责任，转移给了貌似中立、实则带有特定价值前提的系统，与机器是否突然有了灵魂无关。机器表达进入人类互动秩序是事实基础，关系介入引发的主体性动摇是核心机制，算法至上带来的责任模糊则是制度后果。三者环环相扣，共同构成人机边界的真正问题，也共同呼唤一种新的回应方式。

从能力剩余到价值坚守

在人机边界讨论中，最常见也最需要被纠正的姿态是，反复追问人类还剩下什么能力。这种提问表面上是在捍卫人类，实际上依然把人放在效率赛道上同机器竞速，默认的标准依然是速度、规模、准确率与成本。人类价值由此只能被理解为机器尚未覆盖的剩余地带，从而失去了内在的正当性。一旦人工智能发展的目标被定义为模仿并替代人类劳动，技术进步就会倾向于把人理解为可替换的成本。在一个特定的阶段，这样的观念甚至制度安排或许具有正当性，比如人类需要通过效率来维系物种生存的阶段。一旦跨过这个阶段，

① 霍耐特：《为承认而斗争》，胡继华译，上海：上海人民出版社，2005年。

② 何怀宏：《人类还有未来吗》，广西桂林：广西师范大学出版社，2020年。

③ 雪莉·特克尔：《群体性孤独：为什么我们对科技期待更多，对彼此却不能更亲密？》，周逵、刘菁荆译，杭州：浙江人民出版社，2014年。

对人类价值的认知便发生逆转。以增强而非替代为导向，技术才有机会扩展人的能力并保留人的判断空间。在这个意义上，人类自知之明的起点在于，拒绝用机器擅长的指标来定义何以为人，而非承认比不过机器。

讨论的重心由此需要从能力剩余转向价值坚守。在人机边界上，真正不可让渡的是四类更为根本的规范领域：一是价值判断，二是责任承担，三是意义赋予，四是对他者尊严的承认。

价值判断之所以不可让渡，原因在于判断预设了一个能够权衡、辩护并承受后果的主体，与人类判断是否永远比机器准确或正确无关。人之为人的根基在于能够在公共领域开创行动、建构互惠关系，并对后果负责。人工智能可以参与目标的实现，却不应理所当然地拥有确认目标的资格。责任承担之所以不可让渡，是因为责任关系预设的是人与人之间的义务结构。一个系统即便输出了正确的建议，也无法就建议的后果对受影响者负责。承担责任意味着在社会关系中可被追问、可被问责、可被要求弥补，这些都超出了算法系统实际能够承担的功能。意义赋予之所以不可让渡，是因为意义在本质上是行动者把事件纳入自身生活脉络并赋予其价值方向的过程，而不只是信息的排列组合。机器可以生成连贯的文本，却无法经历丧失、希望、悔恨与承诺，也因此无从赋予事件以人类期待自身拥有的意义。对他者尊严的承认之所以不可让渡，是因为承认本身是一个规范行为。它要求行为者把对方视为和自己一样的道德存在。机器可以模拟尊重的语气，却无法在规范意义上真正承认一个人的尊严，因为承认必须出自一个同样可被承认的主体。

把这四类不可让渡的规范领域稳定下来，人类自觉的第一层含义便清晰起来。人之为人，可以被理解为谁来判断价值、谁来承担责任、谁来赋予行动以公共意义、谁来承认他者的尊严，而不是谁更会算、谁更会写、谁更会生成的能力清单。这种重新理解，不回避机器能力增长的事实，目的是把人类自我界定的依据，从可被替代的能力清单转移到不可让渡的规范位置。这也是费孝通晚年所讲的文化自觉在人机互生环境里的自然延伸。“文化自觉”强调主体在文明相遇之际对自身文化的自知与自主，“人类自觉”则要求人类在与机器智能相遇之际对自身作为价值主体、责任主体与关系主体的位置保持清醒。

价值坚守的目标，是让由人主导、为人所用、与人为善、让人安心的人机关系原则有一个能稳定承载的支点，并无意为人类划出一块不可侵入的圣域。脱离了这一支点，任何关于公平、安全、可控、负责的高层原则都悬于空中。把它落到实处，是“人类自觉”的下一个任务。

由人主导的边界制度

有了价值层面的自知之明，还需要制度层面的自主选择。由人主导、为人所用、与人为善、让人安心，只有被写入可执行、可审计、可追责的边界制度，才能避免沦为姿态性口号。在笔者看来，这套边界制度至少应当包含四层，分别处理分工、情感、裁决与责任四类问题。

“分工边界”要回答的，是人工智能适合做什么、做到哪一步，而不只是它能不能做。“增强优先于替代”在此具有直接的制度含义。凡是以计算、归纳、模式识别、检索、排班、草拟与规模化协同为主，且后果可逆、可校正、可审查的任务，应当优先考虑由人工智能增强。凡是直接关系人格权、基本权利、重大利益分配与不可逆伤害的场景，必须严格限制其自动化权限。由人主导制度的目的，是不让技术上的可行自动滑向制度上的应当。

“情感边界”要回答的，是新一代人工智能最容易越界的领域。人工智能不必被一概排除出陪伴、咨询或照护等支持性场景，但必须严格约束使用边界甚

至被排除在涉及未成年人、心理危机者、老年人与强依赖关系的场景之外。一个直觉式的理由是，在这些领域，人工智能极容易反客为主。情感边界的制度目标是防止平台或机构借助拟人化、连续陪伴与关系诱导机制，把依赖、顺从与被操控包装成贴心服务，而无意消灭拟人化界面。凡人工智能介入亲密关系场景，都应当有清晰告知与可退出机制，并禁止以情感依赖换取商业转化或行为操纵。

“裁决边界”比分工边界更进一步。它要求在关键场景里保留人类的最终判断权与否决权。真正的人类控制，是使相关人员能够理解系统的能力与局限，监测其运行、识别异常、避免自动化偏误，并在具体情形下选择不使用、忽略、推翻或中止系统的输出，而不是把人当作事后背锅者。在招聘、信用、医疗、司法、公共福利、教育惩戒等场景里，人类最终决定权应当被落实为可培训、可记录、可抽查的程序要求，而不能只停留在组织章程中。

“责任边界”要防止的，是一旦出错，就出现算法说了算、却无人负责的黑箱。面对责任归属，原则宣言远远不够，必须依靠日志记录、文档管理、合规评估、影响评估、外部审查与救济机制。不能让算法成为责任黑箱，说到底是要让任何一项重要的人工智能应用都能回答四个问题：谁设计、谁批准、谁监测、谁补救。

在这四种边界之外，人机互生秩序的稳定还必须由日常实践来支撑。其一是披露实践。在公共传播、组织决策与亲密互动中，人工智能生成或人工智能介入应当被清楚标明，而不能以“像人”来掩盖其技术属性。其二是校验实践。把人工智能的输出当作草案、提示与备选方案，对事实、判断与后果进行再核验，应成为职业训练与组织流程的常规环节。其三是节制实践。比如，可以使用人工智能减少机械劳动，但不能借此逃避道歉、照护、教育、哀悼、公共辩论这些带有关系义务的行动。在这些场景里，人的在场承担的是尊严与关系完整性的要求，并非效率问题。其四是参与实践。影响评估、外部审查、受影响群体参与和公民申诉，应当成为人工智能部署的常态机制，而不只是政策附件。这四种日常实践与四层边界制度相互配合，构成了人机互生秩序的微观基础，让边界原则在每一次具体的人机互动中得到确认与再生产。

回到一开始的问题。新一代人工智能改变的除了工具与环境，更牵动着人类彼此关联的基本条件，弗洛里迪（Luciano Floridi）由此将其称为继哥白尼、达尔文与弗洛伊德之后的“第四次革命”。[①]在这次革命中，人类面对的是以语言发言的他者，已经超出以后果发言的他者的范畴。它是边界外的对象，却已经成为进入互动秩序的类人参与者。回应这一改变，靠的是在承认人机互生已成事实的前提下，再一次清醒地回答三个相互关联的问题：“何以为人？”“何处必须由人裁断？”“机器应当被允许进入何种关系？”显然，靠更强的机器去对抗机器，或者靠更悲观的姿态去防御机器，都不是出路。

“人美其美，机成人美，美美与共，天下大同”，借用这一表述来描述人机互生秩序，表达的是一种有原则、有边界、有制度、有实践的构想，而非一种温情的技术乌托邦。“人美其美”要求人类首先守住自身作为价值主体、责任主体与关系主体的位置。“机成人美”是承认机器在特定维度上可以帮助人更好地行动、创造与协作。“美美与共”是在差异中形成互补，而不是角色混同。“天下大同”是在公共价值统摄之下，让技术真正服务于人的尊严、自由、平等与共同福祉，而不是把社会交给算法自动运行。[②]

在这个意义上，人工智能时代对人类的真正考验在于，能否在机器日益强大的世界中，始终保持对自身价值的确信、对公共责任的担当与对他者尊严的承认，而不在于能否创造出更强大的机器。这正是人类自觉的核心要义，也是人机互生秩序得以持续的根本保障。

① Luciano Floridi, *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*, Oxford: Oxford University Press, 2014.

② 邱泽奇：《从文化自觉迈向人类自觉》，未刊稿。

ABSTRACT

The AI Era: Transformation of Human Interaction Patterns and Reshaping of Social Structures

Qiu Zeqi & Wang Ning & Bian Yanjie & Wu Guanjun & Liang Yucheng & Qian Zhigu & Du Shichao

Abstract: As the "AI+" initiative is fully implemented, traditional patterns of human social interaction are undergoing a subtle transformation. On the one hand, social interactions among people are mediated by artificial intelligence, giving rise to new rules of communication and modes of collaboration. On the other hand, AI itself is gradually becoming a new type of social actor participating in social interaction, with "human-machine" interaction becoming the norm, and even giving rise to the self-operation of a "disembodied intelligent agent society". This transformation of interaction patterns is reshaping the ways individuals and groups connect, as well as their emotional bonds, trust mechanisms, and power relations. It is also triggering subtle changes in the millennia-old "human-human" structure of human civilization, while posing profound challenges to disciplinary development and paradigm shifts. To this end, in March 2026, we convened a roundtable conference, inviting relevant experts and scholars to track these changes in interaction patterns and explore human self-awareness at the human-machine boundary, with the aim of advancing social governance in the age of artificial intelligence and contributing to the reshaping of a more resilient social structure.

Professor Qiu Zeqi believes that the issue of the human-machine boundary concerns how humans redefine themselves in front of the Other. He calls for upholding "human self-awareness" at the value level, stabilizing four non-negotiable normative domains—namely "value judgment" "responsibility assumption" "meaning attribution" and "recognition of the dignity of the Other"—and establishing a human-led boundary regime and a symbiotic order between humans and machines. Professor Wang Ning points out that research on the "human-intelligence" relationship must be integrated with the study of "human-human" relationships to examine their mutual influence, with particular attention to threefold changes in intimate relationships, social relationships, and workplace relationships. Professor Bian Yanjie proposes that the fundamental characteristic of human social interaction in the digital-intelligent era has shifted from the "structural constraints" of offline social interaction to the "selective expansion" of online social interaction, which challenges sociology's structural perspective while also presenting opportunities for theoretical reconstruction. Professor Wu Guanjun, taking "social wind tunnel experiments" as his theme, examines the goals of bringing the social sciences into the experimental age and the verification standards for intelligent agents across micro, meso, and macro scales, exploring how to move toward critical and responsible social simulation. Professor Liang Yucheng, focusing on the "Lobster Community"—a "disembodied society" composed of intelligent agents—analyzes its basic operation, social structure, authority mechanisms, and stratification logic, distills a triadic entanglement structure of "power–technology–labor", and raises three issues of theoretical construction. Postdoctoral researcher Qian Zhigu explores how, after AI technology enters the lives of older adults, it establishes relational nodes, participates in daily arrangements, and alters knowledge authority, family structures, and emotional relationships, reminding us to guard the boundary of human-machine adaptation for aging populations. Associate professor Du Shichao, taking AI companions as an example, examines the "technologization" of intimacy in the age of artificial intelligence from both internal and external technological perspectives. Building on the trend toward self-technologization of intimate relationships, he calls for using AI to re-understand human loneliness, connection, and emotion, rather than seeking emotional disengagement between humans and AI.

Descendants of Sorcery

Zhao Tingyang

Abstract: It proposes several conjectures or questions through the method of "gene analysis". (1) The transformation from sorcery to historiography (*wu-shi* conversion) is a mode of transformative rebirth that generates descendants through mutated yet coherent inheritance, creating principles of open interpretation, interactivity, and systematic humanistic narrative. (2) Lacking sufficient epistemic efficacy, sorcery relies on institutionalized, standardized, and specialized "magic" to secure collective psychological trust; this model is reincarnated as the academic norms of humanistic and historical knowledge. (3) The decisive factor that makes historical thinking superior to sorcerous thinking in causal explanation lies in its establishment of a practical model centered on "failed experience", whose fundamental spirit approximates trial-and-error Bayesian probability, while inheriting sorcery's concept of potential infinite time. (4) Religion is not a descendant of sorcery. Their concepts of time are opposite: religion promises a completed actual infinite time. Their explanatory models are also opposite: religion adopts a closed rather than an open interpretive model. All narratives that promise an end of history are religions disguised as history. (5) Science represents another