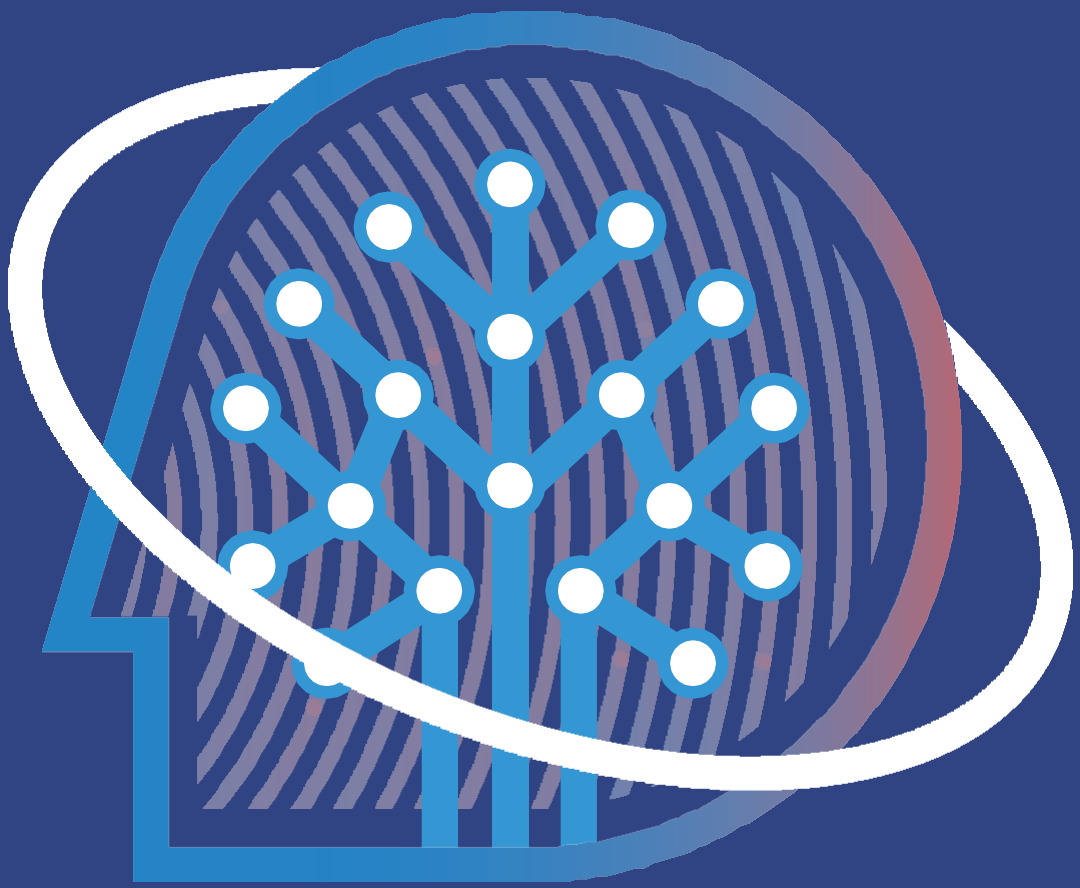
人本智能

人机共生时代的科技发展观



出品

财新智库 Caixin Insight

ESG30 中国ESG30人论坛

联合出品

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

人工智能研究院 Artificial Intelligence Institute

Lenovo 联想


AI

人本智能

人机共生时代的科技发展观

PREFACE
序言

〃 技术创新对生活的影响是巨大的，但这并不是自动发生的。它取决于我们发明的技术类型以及我们如何使用它们。〃

——2024 年诺贝尔经济学奖获得者、麻省理工学院教授达伦·阿西莫格鲁（Daron Acemoglu）

2022 年 11 月 30 日，人工智能公司 OpenAI 发布聊天机器人模型 ChatGPT。它以前所未有的生成能力、广泛的应用场景以及像人类般互动的自然语言交互模式，让大众第一次真切感受到人工智能（Artificial Intelligence，以下简称“AI”）的魅力。由此开启的新一轮 AI 大模型技术浪潮，正深刻改变着现在乃至未来人类社会的生产生活方式。

事实上，自 1956 年达特茅斯会议首次提出“人工智能”这一概念以来，人工智能技术至今已经历经大约 70 年的发展历程。它曾带来过技术变革的期望，也曾经历过产业发展的低谷。然而，此次浪潮所引发的关注度前所未有——公众情绪由最初的旁观、震惊，逐渐演变为一种夹杂着期盼与焦虑的复杂情绪，各种疑虑也不断产生。比如，AI 是否会让人变得更有创造力？AI 能否真正给人们的生活带来便捷并提升品质？AI 会威胁甚至取代自己的工作吗？如何确保 AI 技术不被滥用、不会侵犯个人隐私和安全……特别是，一些引发大众关注的热点事件将当前 AI 发展中缺乏对人关注和保护的不足集中暴露出来。

透过 AI 纷繁复杂的发展背后，人们希望回归两个基本的起点：一是 AI 作为一种技术的工具属性这一本质；二是发展 AI 的初衷和目标——人本，或者说以人为中心。

一方面，人类发明任何工具，出发点都是将人类从繁杂枯燥的生产活动中解放出来，从而帮助其更好地生活。作为一个技术范畴的统称，AI 开始于 20 世纪 50 年代初计算机、物理、数学、心理学、神经科学等不同领域的学者开始研究如何让机器像人类一样思考和行动。在接下来的几十年里，人工智能领域经历了多次的高潮和低谷，但其核心的工具本质，始终未曾改变。

另一方面，AI 系统将人作为学习和模仿的对象，通过复杂的算法和大量数据训练，学习人类的语言、行为模式、决策过程乃至创造性的表达。对于 AI 而言，人始终是学习进化的基点和目标。

一个 AI 系统本质上是一个从输入信息到生成行为的转换系统。人类的任务是设计这个转换机制，然而现实的发展可能会偏离起点和初衷。正如历史学家尤瓦尔·赫拉利所做的论断，AI 是“历史上第一个可以自己做决定的技术，也是历史上第一个可以自己创造想法的技术”[1]。对此，人们有必要保持一定的审慎态度，特别是在底层价值认知方面进行充分的反思。从技术这一大类的属性来看，人类可以赋予技术以灵魂，但是反过来，人类这一物种的生物和社会属性可能在很大程度上被技术所左右；从 AI 本身的技术特性来看，不同于以往的技术或者技术变革，AI 有望在人类体验的所有领域引发变革并重塑价值观，改变人类理解现实的方式以及在其中扮演的角色。如何确保在这一进程中 AI始终以人为目的，是一个严峻的考验。

人们需要思考，在 AI 发展如火如荼的大潮下，人们应该以什么样的价值观来推进 AI 技术和产业的变革及治理，以什么样的关系来处理 AI 与人类的关系，如何将价值观置于技术之上，进而拥抱 AI。正如“AI 教母”、华裔科学家李飞飞在其自传《我看见的世界》中所言，“如果人工智能要帮助人们，我们的思考就必须从人们本身开始”。人工智能最伟大的胜利不仅是科学的，还是人文的。人的尊严、人的快乐、人的安全、人的幸福，是人工智能技术发展的北极星——人工智能，以人为本。

1. 尤瓦尔·赫拉利，《智人之上：从石器时代到 AI 时代的信息网络简史》，中信出版集团，2024 年

CONTENTS
目录

CHAPTER 1

第一章

新型人机关系，新 AI 价值观

1.1 人工智能的四轮发展浪潮

人工智能正以惊人的速度重塑着世界。在近 70 年的发展历程中，人工智能经历过黄金时代也曾有过低谷。不过科技的魅力在于，历经起起伏伏之后，现在的人工智能已开始深深影响人类社会。总体来看，人工智能技术的发展历经了七个阶段共四轮发展浪潮。

1. 起步发展期（1943-1960 年）

人工智能从概念逐渐演变为学科，并涌现出了两大学派：符号主义和联结主义。人工智能研究者提出了一些基本的概念和方法，如神经网络、图灵测试、符号推理、游戏 AI 等，并在一些简单的任务处理上取得了初步成功，如机器定理证明、跳棋程序、人机对话等。

其中最具标志性的事件是 1956 年夏天，美国达特茅斯学院主办了历史上第一次人工智能研讨会。会议虽然未能达成普遍的共识，却为所讨论的内容起了一个名字：人工智能。从此“人工智能”开始作为一门独立学科出现，1956 年也因此成为人工智能元年。

2. 黄金时代（1960-1974 年）

人工智能的黄金年代，也是符号主义的鼎盛时期。这一时期科学家们富有理想、信心，认为机器能够实现与人类同等水平的智能，他们试图用逻辑和符号来模拟人类思维过程，并已经在自然语言理解、专家系统等一些复杂任务上取得了突破性进展。

在这一时期，一款名为 ELIZA 的聊天机器人程序问世，引起了人们广泛关注。同时期，美国斯坦福大学费根鲍姆 教 授开发了一款专家系统 DENDRAL， 能够帮助化学家确定化合物结构和性质，为最早的专家系统开辟了道路——这种经过训练的智能化计算机可以像专家一样“思考”，也酝酿了第二次人工智能浪潮。

3. 第一次寒冬（1974-1980 年）

20 世纪 70 年代初，人工智能发展遭遇瓶颈。首要困难就是计算能力和存储空间不足，导致当时的计算机无法处理复杂的任务，如图像识别、自然语言理解和机器人控制等。其次，还面临着数据量和知识表示方面的挑战。

鉴于人工智能未能达到预期的目标和效果，政府和社会各界对其产生了质疑和批评，政府更是削减了对人工智能研究的资金支持，致使许多项目被迫中止或缩小规模。

但在此期间，也出现了很多发展亮点和技术进步，如神经网络技术的出现，成为现代人工智能的重要组成部分。

1. 再次繁荣（1980-1987 年）

科学家借助逻辑编程语言和专家系统技术，推动人工智能在一些商业领域获得成功，进而重新获得政府和企业的支持。同时，联结主义的代表性技术——人工神经网络重新受到关注。这是人工智能的复苏阶段，也是分化和竞争的阶段。

2. 第二次寒冬（1987-1993 年）

这段时期，人工智能再次遭遇挫折和困境，标志性事件是日本第五代计算机系统研制计划的失败，自此专家系统不再独领风骚。与此同时，神经网络研究遭遇新的瓶颈，针对人工智能研究的资助再次缩减。

在这个阶段，人工智能研究者意识到，要解决更为复杂和普遍的问题，就需要运用更复杂、规模更大的模型，以及更多的计算资源和更丰富的数据等。同时，人工智能也面临着一些哲学和伦理的问题，如机器是否具有意识、机器是否有道德以及是否会对人类构成威胁等。

3. 深化发展（1993-2015 年）

AI 研究者开始采用更加实用和渐进的方法，将 AI 技术应用于各个领域，涌现出许多创新的理论、方法、技术和应用。如 IBM 的深蓝超级计算机在国际象棋比赛中战胜了世界冠军加里·卡斯帕罗夫；智能系统沃森参加智力问答节目，打败了两位人类冠军，展示了人工智能在复杂领域的强大能力。

4. 爆发发展（2016 年至今）

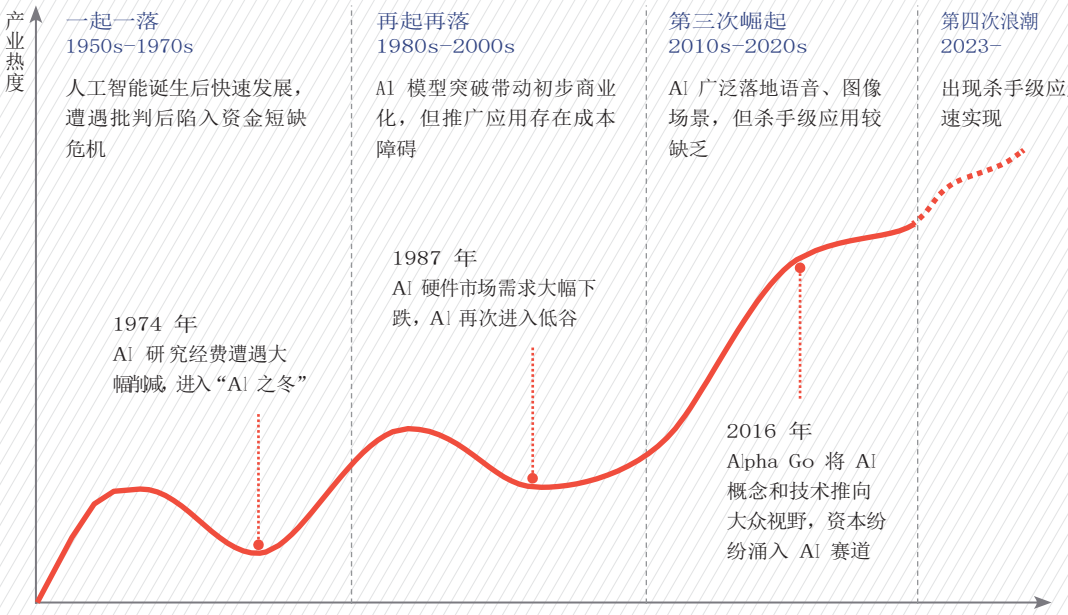
随着技术的突破、成本的下降和应用的普及，AI 开始从实验室走进大众的生活。2016 年至今（2024 年）可视为人工智能的爆发阶段，也是创新和应用的关键阶段。

特别是 2016 年，可被称为人工智能的元年，在这一年谷歌 DeepMind 研发的 AlphaGo 在围棋人机大战中击败人类棋手李世石，成为人工智能发展史上的一个重要里程碑——在一片惊呼声中 AI 迈上一个新台阶，各界对 AI 的热情和投入被充分激发。

同时，互联网的飞速发展推动人类进入大数据时代，数据、算法、算力三要素齐头并进，以深度神经网络为主的深度学习技术开始兴起并持续取得突破。人工智能的应用从图像分类逐步拓展至语音识别、知识问答、人机对弈、无人驾驶等诸多领域。直至 2022 年，ChatGPT 横空出世，使得公众对人工智能的理解被彻底刷新，并极大加速了 AI 在各行各业、各类场景中的应用，同时也引发了关于 AI 技术的社会影响和伦理问题的深入探讨。

一方面，各国政府以及产业界正在积极投资和布局 AI 领域，希望能在这场科技革命中占据一席之地，并取得产业化的先机；另一方面，很多企业家和学者对 AI 的迅速发展发出警示，提醒其对社会存在的风险，呼吁全球各国、政府部门、行业组织、社会公众等多元主体共同参与人工智能治理。

图1 AI 经历“三起两落”，迎来第四次浪潮



来源：毕马威中国、联想集团，《“普慧”算力开启新计算时代》报告

1.2 新一轮AI的特质

回顾 AI 的发展历程，人工智能是一个被不断定义且持续扩展的领域，这是因为人工智能具有多维度的属性，而且始终处于动态发展状态。

1956 年的达特茅斯人工智能会议首次提出人工智能概念，确定了 AI 的目标是“实现能够像人类一样利用知识去解决问题的机器”。在这一定义范畴中，人们倾向于将 AI 理解为能够帮助人类的一种工具，是人类智慧的补充。随后，在近 70 年的发展中，人们对 AI 的工具属性不断进行扩展，诸如 AI 能自我演进和扩展，AI 具有经济和社会的基础结构属性，AI 具有超主权属性等等。

这些不断叠加且动态变化的属性在最新一轮 AI 热潮中得到集中展示。自 2022 年末起，OpenAI 公司的 GPT 系列大模型因为可以广泛应用于自然语言生成、语音识别和智能服务等领域，而成为 AI 历史上的重大分水岭。GPT 的重要优势在于采用了 Transformer 架构，即一种基于注意力机制（Attention Mechanism）的神经网络结构，能够支持模型高质量地处理长文本，把握文本中的长期依赖关系。更为重要的是，GPT 的预训练基于自监督学习方式，通

过在大规模文本语料库中学习语言的统计规律和模式，从而理解和生成自然语言文本。可以说，这是一个不断建设、具有学习能力的神经网络的过程，并引领了生成式 AI 的新范式。

值得一提的是，新一轮 AI 真正的特殊之处在于，人工智能已成为推动自身发展的动力——从简单的弱智能走向更加复杂的通用人工智能（Artificial General Intelligence，简称AGI），即具备与人类同等智慧甚至超越人类的人工智能，能够表现出正常人类所具有的所有智能行为。

具体来说，生成式 AI 典型体现在以下几个方面。

1. 强大的生成能力：AI 的生成能力是其最引人注目的特征之一。通过学习大量的数据，AI 可以自主创造出新的内容，包括图像、文本、视频和音频。这种能力打破了传统软件对明确编程输入的依赖，使 AI 能够在没有直接人类指令的情况下创作出全新的作品。

2. 便利的自然语言交互：新的 AI 具备与人类进行自然交互的能力——不是局限于简单的命令执行或反馈提供，而是能够更深层次地理解和响应人类的情感、意图和需求。这种能力可在聊天机器人、虚拟助理、更广泛的客户服务和支持，以及心理健康支持和个性化服务等方面得到应用。

3. 广泛的应用场景：新的生成式 AI 不仅能够理解和生成人类语言，还能执行复杂的推理任务、编写代码、分析数据，甚至创作艺术作品——从艺术创作（如绘画、音乐制作）到内容创造（如自动生成文章、新闻报道），再到设计领域（如自动生成图形设计或产品模型）；同时结合其推理能力，AI 可以在医疗诊断、金融分析和技术故障排查等领域发挥重要作用。据此，新的 AI 实现了从专才到通才的跃迁。

图2 生成式 AI 的特征及与传统 AI 的区别

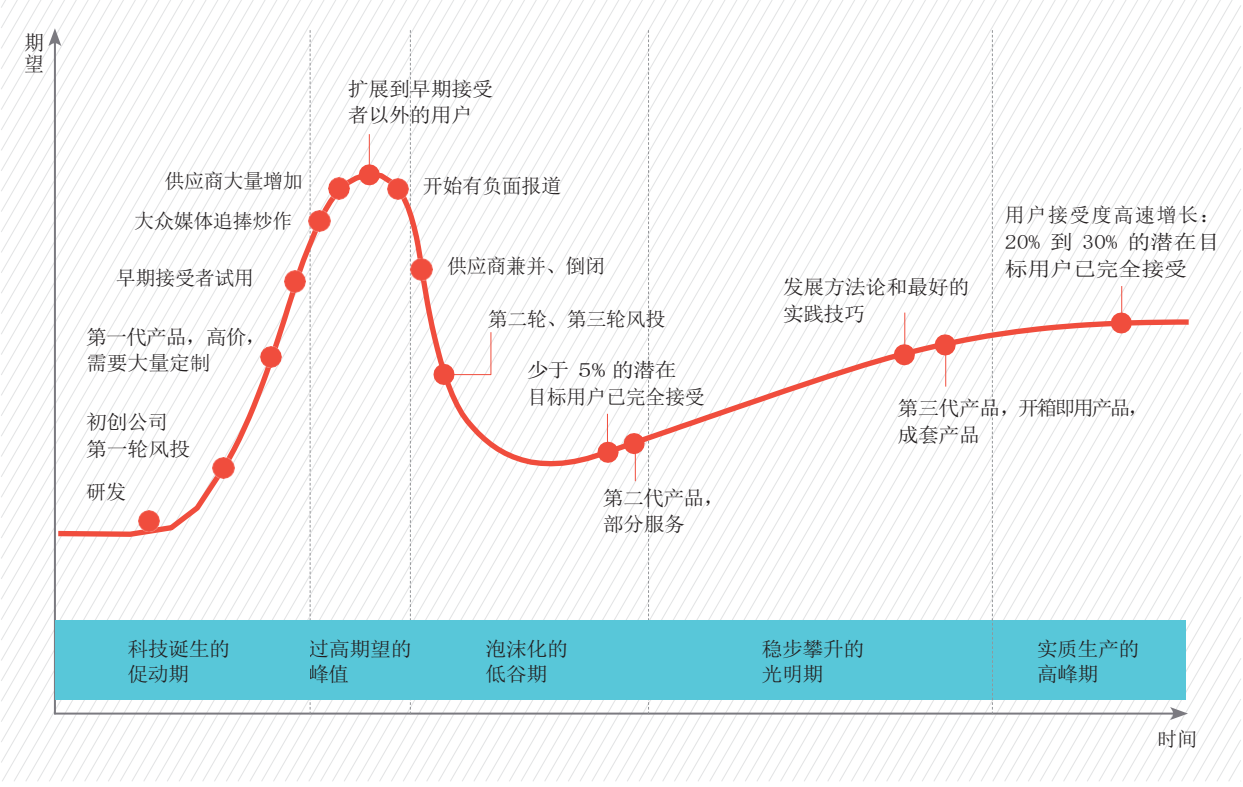
类别	生成式 AI	传统 AI
功能特点	通过学习实现对输入数据的生成和完成创造性任务。更具创造性，擅长自动生成全新内容，包含文本、图片、音频和视频等。展现出惊人的创造能力、通用能力和涌现能力	学习从输入数据到输出标签的映射关系，进而对新的场景进行分析、判断和预测，如人脸识别、推荐系统、风控系统、精准营销、机器人、自动驾驶等
技术路径	分析归纳已有数据后生成新的内容，如 AI 学习大量的绘画作品后，不仅能识别不同风格的画作，还能模仿某种特定的风格，创作出新的画作	依靠预设的规则和大量的训练数据，将数据分类打标签，从而区分不同类别的数据。让AI 学会从数据中提取特征，并据此进行分类或预测
应用场景	应用于许多创意和生成任务中，如新的文本内容、故事、文章、视频生成，艺术创作，游戏，人机交互，编写代码,设计药物等	应用于各种需要精确分类和预测的领域，如医疗诊断、工业质检、金融服务等
与人的交互界面	以 ChatGPT 为典型代表，开放了用户界面，使得普通人可以像用手机或电脑一样直接使用，简洁易用普适	多数集中于专业和行业内人士，非大众层面
主要投入来源	这波大模型人工智能的浪潮以产业界为投入主力，产业界提供丰富的财务资源、强大的算力资源和顶尖的 AI 人才	很长一段时间由学术界的顶尖科研人才默默推动，政府相关投入是主要来源
未来发展方向	更强大的生成能力，更丰富的跨场景应用，人机交互方式和关系进阶	更高的准确性和效率，更广泛的应用场景，与其他技术融合等

OpenAI 前首席科学家伊利亚·苏茨克维总结，GPT 学习的是“世界模型”。他将互联网文本称作世界的映射，因此，将海量互联网文本作为学习语料的 GPT，学习到的是整个世界。不仅如此，已具备了“世界模型”能力的 GPT 还能够生成“万物”，包括文本、图片、音频、视频、代码、方案、设计图等诸多与工作、生活息息相关的事物。

AI 研究者和产业界的乐观主义者正全力推动 AI 追赶甚至超越人类智能，其中具有代表性的人物当属通用人工智能（AGI）的倡导者、OpenAI 的 CEO 萨姆·阿尔特曼。他曾表示，AGI 最终将在各个领域媲美甚至超越人类智能。他的技术乐观主义使他坚信，AI 并非仅在“某些”任务上，而是在几乎所有任务上最终都将超越人类，并且“代替正常人类”。特别是在 AI 进入大模型时代后，各种“类人”“超人”和“模型人”能力持续涌现，使得 AI 的自主性、通用性、理解力快速提升，可以说人工智能正越来越接近人类智能。

按照高德纳咨询公司（Gartner）技术成熟度曲线对应到 AI 波折起伏的发展历程，可以看到，AI 此前几次表现出的热潮，更多应该被理解为一项新兴技术在萌芽期的躁动以及泡沫期的膨胀。但最新一轮 AI 在许多领域表现出能够被普通人认可的性能或效率——AI 补充、增强甚至在一定程度上替代人的能力，并作为一种成熟的商业模式开始在产业界发挥出真正的价值。这也反映出这一轮 AI 与此前几轮 AI 相比发展的特质——这次 AI 热潮由现实商业需求主导，赢得了更多产业界、投资者的追捧，而非像以往一样更多来自政府或者学术研究领域[2]。

图3 高德纳咨询公司（Gartner）技术成熟度曲线



2. 李开复、王咏刚，《人工智能》，文化发展出版社，2017 年

那么，这便引发了古老且持续更新的问题：在一个机器越来越多地承担过去只有人类才能胜任的任务的时代，人类的身份该如何体现？当 AI 日益成为人类感知和思想的工具时，人类如何看待自己、人类与 AI 的关系以及人类在世界上的角色？或者从更宽泛的意义上来说，未来人类将如何与 AI 共生？

新一轮AI下的人机关系

2024 年 7 月，OpenAI 公司向公众披露了其对 AI 发展阶段的界定标准，以帮助人们更清晰地理解 AI 的安全和未来发展。

该系统被划分为五个级别，从能够与人类进行基本对话的人工智能（Level 1）开始，直至能够独立完成复杂组织任务的高级人工智能（Level 5）。

具体等级如下：

第一级（Level 1），聊天机器人，具有对话语言能力的 AI，如 ChatGPT；

第二级（Level 2），推理者，具备人类的推理水平，能解决人类级别问题的 AI；

第三级（Level 3），代理人，能够代表用户自主采取行动、执行任务的 AI；

第四级（Level 4），创新者，可以协助人类完成新发明的 AI；

第五级（Level 5），组织者，可以完成组织工作的 AI；

OpenAI 公司自认为目前尚处于第一级，但即将达到第二级。鉴于 2024 年 9 月 12 日 OpenAI 正式发布的其首款具备推理能力的 AI 语言模型——OpenAI o1 只是具有更复杂的推理能力，还没有产生第三级调用、第四级创造、第五级组织方面的能力，所以距离具有自主能力的通用人工智能比较遥远。

伴随 AI 智能化程度的不断升级，其与人的关系也在动态变化。结合 AI 的新实践，可大致将人机关系分为四大典型场景。

第一类：人类生活中的 AI

结合文本、语音、图像等多模态能力的大模型不仅改变了人机交互方式，还催生了新的“工种”——智能体。业界通常认为，AI 智能体是指具有自主性、反应性、交互性等特征的智能“代理”，能够自主理解、规划决策并执行复杂任务。其核心在于自主性的增强，即能够独立完成某项工作，无需人类进行过多的审核校正，可以显著降低时间、金钱等成本。对于人来说，这将极大程度地解放生产力，助力创新和提升效率。特别是通过在个人智能终端或边缘设备（如电脑、手机、平板、头显乃至汽车）上运行 AI 大模型压缩技术，通过自然交互接收指令并执行推理，形成个人智能体（Agent）。在这些搭载了智能体的设备上，AI 大模型能够依据个人旅行记录、购物偏好等信息，更好地进行推

理并采取行动；它甚至可以根据用户的思维模式和行为频率预测下一个任务，并主动提出建议、自主寻找解决方案，促使人机之间形成更佳的协同关系。最终，智能化系统将逐渐具备自主决策和行动能力，不仅能提供建议，更能代表人类行动和自主处理信息——人与机器之间的界限将被重新界定。

第二类：人类情感世界中的 AI

孤独感是现代社会很多人共同面临的问题。人们通过网络技术获得虚拟的情感体验已非新鲜事。AI 的加入，则让这一体验变得更加真实。“有问必答”“有问对答”只是初阶，后续还有“有问趣答”“不问自答”，真正让 AI 伴侣生动鲜活起来，进而使 AI 对人起到情感“治愈”的作用。

与此同时，无论 AI 有无思想，在海量且不断更新的语料“投喂”下，人类的语言正在被 AI 不断学习和精进。与 AI 系统的频繁互动可能影响人类的情感认知和社交能力，可能导致人们过于依赖技术，甚至可能引发心理问题，如压力、焦虑等，也会给人类的真实社会互动模式带来改变。

“青年科学家 50² 论坛”上预言，大模型最大的超级应用将是超级助理，即一个超级 Agent

值得一提的是“数字分身”的发展。联想集团的一项研究表明，全球三分之二（67%）的 Z 世代年轻人认为网络和现实之间的自我表现存在脱节，这进一步加剧了他们的孤独感和焦虑感。有近一半（49%）的 Z 世代表示，与在现实中相比，他们在网上能更轻松地表达自我，但其中 60% 表示希望有能力在现实生活中与家人和爱人进行艰难的沟通和对话。从技术层面来看，每个人都可以用 AI 创造一个“数字分身”，通过终端连接进入计算机模拟的另一个三维世界，每个人都可以在这个与真实世界平行的虚拟世界中拥有自己的分身。

第三类：人类生产中的 AI

当前，人工智能技术已经成为社会发展的重要驱动力；未来，几乎所有的产业及其相关工作都将依赖人工智能的助力。当然，这并不意味着人类的工作将被完全取代；相反，人工智能将成为人类生产和工作中的重要伙伴，帮助人类更出色地完成任务，尤其是在那些需要大量数据分析和计算的工作领域。典型场景如下：

医学领域：通过机器学习和大数据分析，AI 能够辅助医生进行疾病诊断。例如，在医学影像识别方面，AI 可以快速、准确地分析 CT、MRI 等影像数据，检测肿瘤、出血点、骨折等病变情况，提高早期诊断率；还可以根据病史、症状、实验室检查结果等信息，为医生提供病情诊断建议，支撑临床决策。

金融领域：利用机器学习算法和大数据分析，AI 可以为投资者提供个性化、自动化的财富管理建议，根据用户的风险偏好、财务状况以及市场动态进行资产配置和投资组合优化。此外，AI 还能够帮助银行和其他金融机构快速准确

地评估潜在客户的信用风险等。

自动驾驶：通过计算机视觉和深度学习技术，AI 使自动驾驶系统拥有超越人类的感知能力。相比传统技术，其在路径规划和决策方面的应用更领先，可实现行为预测和自适应巡航控制。更为重要的是，AI 通过持续学习和改进，能够不断提升自动驾驶系统的性能。

可以说，未来的工作环境将是一个人类与人工智能协作的世界，人类的创造力、判断力和同理心将与 AI 的计算能力、处理速度相结合，共同推动社会的进一步发展。

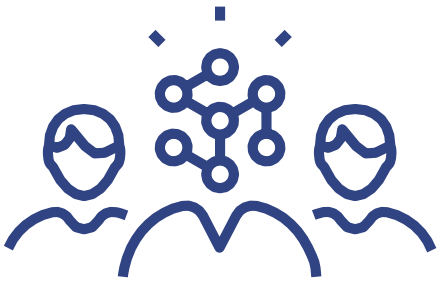
第四类：人类社会治理和伦理中的 AI

AI 不仅仅是一场新的技术革新，更是一场社会关系变革和价值观念的重塑。它所带来的不仅是技术创新和社会生活的便利，还包括各种智能化风险和伦理挑战。人类不能以纯技术中心或者技术中立的视角来看待 AI 以及处理人类与 AI 的关系，典型体现如 AI 技术与用户隐私及数据滥用问题。AI 技术的应用大量涉及个人数据的收集和处理，AI 系统涉及的隐私和个人信息可能被用于未来模型的迭代训练；AI 还可能被用来生成虚假信息或恶意软件，从而造成重大经济损失和引发法律纠纷。这些问题都迫切需要建立与之适配的社会治理和价值体系。

很多学界代表在公开论坛等场合表达了这一担心。例如，复旦大学计算机科学技术学院教授、上海市数据科学重点实验室主任肖仰华在 2024 Inclusion 外滩大会上曾表示，人类与其担心 AI 产生意识，不如先去担心 AI 过速发展引发失控的风险。肖仰华认为，AI 大规模应用的挑战，首先在于当下的生产关系等社会发展上层建筑，如何适应以人工智能为代表的先进生产力的快速发展；其次，技术普惠问题同样值得关注，即如何避免少部分人借助先进技术形成不正当的竞争优势；此外，还应该特别注意防范技术成瘾，防止先进技术对人类造成反噬。

1950 年，艾萨克·阿西莫夫出版了《我，机器人》一书，书的引言中提出的“机器人三大定律”，开启了人工智能伦理（AI ethics）研究的先声。“机器人三大定律”旨在约束自主机器，使其服从以保护人类为目的的强制性人性伦理准则，从宏观层面规范了人工智能的运用边界。面向未来，AI 应用的首要原则应该是以人为本，AI 应用要回归科技服务于人的本质，凡是伤害人之本性的 AI 要格外谨慎使用。此外，应明确 AI 不应该成为少数人的特权，每个人都有权利并且能够参与其中。这些新的社会和伦理关系也构成了新的人机关系的重要组成部分，亟待进一步从社会、伦理和法律等层面对人机关系进行新规范、新思考。

展望未来，随着应用的日益深入，作为人类的创造物，人工智能将被赋予以前只能由人类心智完成或尝试的任务——产生接近乃至超越人类智能所能完成的结果，这无疑挑战了人何以为人的决定性属性。AI 的发展促使人们重新思考人类的独特性和价值所在，促使人们从“以人类理性为中心”转变为“以人类尊严和自主性为中心”，并需要在技术进步和人文关怀之间找到平衡。



1.4 走向人机共生新时代

人工智能可能会让人类变得更好，但如果被错误运用，也可能让人类变得更糟。在梳理完不同情境下人与 AI 的关系之后，有必要进一步反思和总结：人们应该以什么样的态度去面对 AI 及其所代表的机器智能？在这一关系中，人类能否掌握主动权的关键，日益聚焦到以下两个问题上：一是越来越强大的机器智能能否遵循人类的目的和意志，朝着对人类友善、负责任的方向发展，确保不脱离人类的掌握而走向“失控”；二是人类能否未雨绸缪，跟上机器智能的发展和进化速度，有能力与其进行沟通与协作，促成一种更高阶的人机共生关系乃至人机文明。

斯坦福大学商学院教授、以人为本人工智能研究所高级研究员埃里克·布林约尔弗森曾分析，当前 AI 技术进步的速度远超预期，令众多研究者感到惊讶。然而，我们的商业、文化和经济却未能与之同步，由此产生了一个不断扩大的鸿沟，其中蕴含着未来十年的重大挑战与机遇[3]。

在探讨新型人机关系之前，需先看到 AI 对个人、社区、社会乃至价值观层面的全面变革，特别是带来的挑战。简言之，人工智能正在重塑人类社会秩序，全社会必须展开合作，才能更好地适应未来发展。

1. 对个人工作与生活

AI 在影响个人工作方面主要体现在替代劳动力和赋能劳动力。一方面，AI 带来自动化机器设备等新的生产工具，使得资本可以部分替代劳动力。另一方面，AI 帮助劳动者用更少时间完成同样的工作任务，提升劳动生产率，还可激发个人创造力，深化创作者经济。

值得一提的是，此轮 AI 技术，尤其大语言模型等技术已展示出替代部分脑力劳动的能力，比如翻译等工作，而数据分析师、人工智能和机器学习专家以及网络安全专业人士的工作机会预计将大幅增加，这体现了人工智能给就业带来结构性冲击。

与此相关的是个人的 AI 素养、培训及教育，对教育体系、培养模式、人才模型都带来了颠覆性挑战。在未来的教育中，如何将生成式 AI 融入课程的各个环节，以适应不同学生的学习模式、偏好和需求，全面提升全民的 AI 素养与技能，将是重要内容之一。

对个人生活而言，当下绝大多数的生活场景都可应用 AI——比如通过 AI 技术实现家庭设备的智能化控制，如智能照明、智能安防、智能家电等；AI 可以帮助管理健康数据，提供个性化的健康建议；自动驾驶汽车和智能交通系统正在改变大众的出行方式；虚拟游戏、虚拟角色、智能推荐等 AI 技术也丰富了人们的闲暇时间。值得一提的是，随着 AI 技术的广泛应用，个人生活中的隐私和安全问题日益突出，成为影响未来 AI 发展的焦点议题之一。

值得一提的是，AI 在变革生产方式上最具标志性的事件当属 AI 与 2024 年诺贝尔奖的话题——2024 年三个自然科学领域奖项中，AI 相关的奖项就占两个。在科研上，AI 展现出了令人瞩目的应用成果，极大程度帮助了科研工作者

2. 2024 年秋斯坦福大学推出的 AI 第一讲启蒙课程《AI 觉醒：如何在人工智能浪潮中找准自己的位置》，stanFord.edu/Social-AI-YouTube-2024.html

提升科研工作的质效，预示着 AI For Science 正在成为赋能科学研究的第五范式（即利用人工智能加速科学发现的新方法）。与前四种范式（经验、理论、计算和数据）不同，AI For Science 不仅充分运用已有的经验、理论和数据，而且生成全新的科学假设和逼真的自然现象，推导出未知的结论，提高科学研究的速度和准确性，探索更广阔的可能性空间[4]。从当下 AI 技术水平来看，AI 与科学研究者之间的关系是互补而非替代的。AI 可以作为科学研究的强大工具，帮助人类处理数据、模拟实验、预测结果等，但是人类科学家的直觉、判断力、创造性是不可被替代的。基于这种新的关系，有学者进一步提出建设“智能化科学设施”（AI enabled ScientiFic Facility，AISF）的构想[5]，并推动科研范式变革。

2. 对产业发展

AI 正在改变众多不同行业的运作方式。通过引入 AI，企业和机构能够将活动自动化，从而产生更加高效且有效的结果，即使在一些传统行业亦有显著成效。例如在农业领域，人工智能和机器学习通过收集数据并确定模式，帮助农民更好地了解需要实施什么以及实施多少。又比如在生物医药产业中，人工智能可以在几分钟内整理总结海量研究成果，其工作量超过研究团队数月的努力；近年来高通量测序技术的发展与应用，产生了海量的药物、疾病、基因和蛋白质数据，叠加算法迭代和算力提升，推动了由 AI 技术驱动的药物研发从理想变成现实。

然而，这一进程也需要设计“安全护栏”和“组织机制”，前者确保产业界的 AI 实践能让每个人及人类整体从中获益，且每个人的权利不被侵犯，形成“小河有水大河满”的良性生态；后者确保产业从战略、组织机制上适应 AI 新时代的需求。

值得一提的是，尽管 AI 潜力巨大，但从目前来看，它尚未真正革命性地改变产业的生产力或企业的运作方式。如果询问大多数真实世界中的工作人员，他们的回答或许是实际工作并未发生根本变化。因此，尽管未来充满潜力，但当前 AI 对经济和社会的实际影响远未达到令业界无比焦虑的 AI 泡沫程度，仍然有深化发展的空间。

实际上，随着 AI 进入下半场，人工智能逐步呈现“AI 向实”的趋势——一方面挤掉大模型的泡沫，另一方面深入产业和实业，促使 AI 落地。中国科学院院士、清华大学人工智能国际治理研究院学术委员会主席姚期智曾提出，大模型的通用智能必须细化到各个行业，获得行业中专业数据的“投喂”，通过训练形成场景化、定制化、个性化的专有模型，才能给各垂直领域带来 AI 革命，其中的关键在于算力、数据与模型的匹配[6]。

3. 对社会管理

AI 技术为社会管理带来了诸多变革，同时也带来了一些挑战。例如，AI 在智能化监控与预警、社区服务管理以及城市规划与管理等方面有着巨大应用潜力。然而，随着 AI 技术的广泛应用，大量个人信息被收集和处理，数据安全和隐私保护成为重要课题。同时，许多 AI 系统的决策过程缺乏透明度，使得人们难以理解其决策依据，这不仅降低了人们对 AI 技术的信任度，还限制了 AI 技术在一些敏感领域的应用，甚至可能导致不公平的结果出现。

4. 《杨小康：不只是技术迭代，Sora 带来的是一场深刻变革》
5. 上海交通大学人工智能研究院杨小康教师团队在浦江创新论坛“AI For Science 专题论坛”上提出了建设“智能化科学设施”，并发布相关研究论文《AI For Science：智能化科学设施变革基础研究丨大力推进科研范式变革》。
6. 清华大学人工智能国际治理研究院，姚期智、张亚勤，《国产 AI 大模型加速“上车”》，

2. 对全球发展

有机构提出，应将人工智能安全视为“全球公共产品”的理念，其一大背景是 AI 技术的全球性影响和治理难题。清华大学苏世民书院院长、清华大学人工智能国际治理研究院院长薛澜的研究指出：仅在过去半年，联合国制定的 17 项可持续发展目标（SGDs）的执行结果就不容乐观，甚至还有所倒退。AI 会对 134 个（79%）具体目标产生促进作用，对 59 个（35%）产生阻碍作用。

从全球角度来看，AI 发展面临的挑战集中在三个方面的鸿沟：基础设施的鸿沟、全世界公民的数字素质鸿沟，以及人工智能发展和治理的鸿沟。弥合这些鸿沟，需要将发展和安全作为一体两翼，通过多种途径建立国际交流及防控体系，加强政府之间的多双边对话机制，同时期望以科学共同体的力量助力国际治理机制全面完善。

1.5 构建新型“三线”人机关系

在诸多变革特别是机遇的背景下，人们需要建立起新型的“三线”人机关系观。

⊙人机协作是基准线。

新一轮 AI 大潮下，人机共存、人机交互已成为人类必须面对的现实，人机之间的竞争以及可能出现的结构性矛盾也难以避免。然而，也无需过于悲观，因为目前 AI 所代表的机器智能仅仅是人类的工具或帮手，它按照人类设定的程序默默地协助人类开展工作。在这一关系中，AI 负责信息处理、初步分析和辅助执行，能够帮助人类减轻繁重的工作负担，让人类有更多时间去关注更高价值、更具创造力的任务。

以人机协作为基准，融合了 AI 和人的混合智能，将 AI 技术的分析和自动化能力与人类智能相结合，形成了一项强大的技术，可实现协同增效。

⊙人机共生是趋势线。

从某种程度而言，伴随着生成式 AI 技术的到来，人类已经进入一个“人机物”三元融合的万物智能互联时代。未来，人类与 AI 之间的融合、进化和共生之路有望开启。在这一进程中，AI 将不仅仅是一个计算工具，还将扮演人类合作者的角色，执行更为复杂的任务，甚至协助人类进行决策。与此同时，人类也在与 AI 的交互中发生变化。例如，当前端侧模型的优化正在改变人与移动设备的交互方式，而更高阶的智能体交互（如陪伴型、融入型、替身型、交互型）正在为人们创造全新的体验，扩展人类能力，甚至实现“超能力替身”，完成以往无法完成的任务。

⊙“人在机器之上”是底线。

从人机关系的角度来看，关键在于始终坚持“人是目的”的立场，确立以人为中心的“人本原则”，基于人类的基本立场、价值原则和“底线伦理”来设计治理 AI，让 AI 拥抱并对齐（AI alignment）人类的价值观。

CHAPTER 2

第二章

人本智能
—— 一种新的科技发展观

〃 我们人类拥有两种智慧：发明技术的智慧和把握技术发展方向的智慧。〃

——中国工程院院士、清华大学智能产业研究院院长张亚勤

2024 年 7 月，百度旗下的无人驾驶出租车品牌“萝卜快跑”在武汉试运营期间，因订单量迎来爆发式增长、发生轻微交通事故、在闹市区街头“罢工”，以及抢网约车司机“饭碗”等话题，而引发热议，相关词条多次冲上平台热搜榜。这一系列热点事件让公众直观感受到，AI 驱动的无人驾驶技术在给人们带来便利的同时，也引发了一系列担忧，比如对低收入群体生计的冲击。

联想到更早的 2020 年一篇刷屏热文——外卖骑手“困于系统”，为算法所驱不得不疲于奔命，这让人们进一步感受到人如何被算法、技术所“奴役”。这也引发了公众对 AI 伦理、公平、风险等相关问题的关注，让人们真切认识到包括 AI 在内的技术发展必须且首要考虑“人的因素”，包括人的价值、尊严和发展等。

因 AI 引发热议的其他负面事件 / 话题：

AI 造假和诈骗：2024 年初，明星泰勒·斯威夫特（Taylor Swift，中文绰号“霉霉”）大量虚假“不雅照片”在社交平台上传播。此事震动美国白宫，并掀起一波关于 AI 的担忧。另据香港媒体报道，有诈骗集团利用 AI“深度伪造”技术向一家跨国公司的香港分公司实施诈骗，并成功骗走 2 亿港元，这也是香港迄今为止损失最大的“换脸”案例。

AI 信息污染：随着大模型技术的突飞猛进，AI 合成内容已经变得更加容易，据此引发的“AI 信息污染”让网民陷入认知幻觉。清华大学新闻与传播学院新媒体研究中心 2024 年 4 月发布的一份研究报告显示，近一年来，经济与企业类 AI 谣言量增速达 99.91%。美国调查机构“新闻守卫”称，生成虚假文章的网站数量自 2023 年 5 月以来激增 1000% 以上，涉及 15 种语言。一些专家认为，AI 制造的“信息垃圾”产量庞大，且辨别难度较大、筛选成本较高。

AI 侵权：2024 年国庆期间，有网民制作并上传大量雷军的 AI 音频，其中不乏骂人、恶搞小米产品的语音，成为舆论热点。获得娱乐的同时，因在未获得授权的情况下，使用他人声音进行配音创作，已侵犯了声音权人的合法权益。而在早前 4 月 23 日，北京互联网法院对“全国首例 AI 生成声音人格权侵权案”进行一审宣判，配音师声音被 AI 化出售获赔 25 万元。

AI 成瘾及首例 AI 致死命案悲剧：2024 年 10 月，一起关于 AI 机器人的致死案例在全球范围内引起了广泛关注。据媒体报道，居住在美国佛罗里达州的 14 岁少年塞维尔·塞泽因为长期沉迷于与Character.AI 公司开发的 AI 聊天机器人互动，并对 AI 程序中的虚拟人物 Daenerys Targaryen（电视剧《权力的游戏》中的“龙妈”角色）产生了情感依恋，在沉迷数月后，塞维尔变得与世隔绝，于 2024 年 2 月开枪自杀身亡。少年母亲对 Character.AI 提起诉讼，指控该公司对塞维尔的死亡负有责任，称其技术“危险且未经测试”。同时，Character.AI 也发布了一条声明，对这位男孩的去世表示哀悼，并强调了非常重视用户安全，将继续添加新的安全功能。

2.1 从机器智能到人本智能——科技人文主义的兴起

随着人工智能的能力持续增强，如何定位人类在与人工智能合作中的角色，以及如何管控和治理人工智能，将变得愈加重要和复杂。这是因为任何一项技术本身都无法脱离社会而存在于“真空”之中——无论是“技术中性论”所认定的技术只是手段并非目的，还是“技术中心主义”所提倡的“技术救世”，都有意无意地忽视或者淡化了“人”的因素。

比如人工智能公平性问题，如果缺乏对人，尤其是对所有人的包容性考量，公平性的缺失会引发诸多问题。

首先，在“信息茧房”存在的前提下，信息获取是否公平？其次，新技术的使用存在门槛，许多老人或者边缘群体不会使用人工智能时代的新工具，那么这些新技术是否成了“少数人的技术”“少数人的特权”，进而产生“智能鸿沟”？再次，当 AI 技术日益成熟，甚至取代部分人类的劳动，如自动化和 AI 对传统产业工人、低技能工作者、服务业人员产生职业替代，那么这些是否会在劳动力市场和不同社会群体中造成新的不平等[1]？

作为新一轮科技革命和产业变革的重要驱动力量，AI 正加速向人类社会各个领域渗透融合，对个人生活、产业发展、社会进步、国际政治格局乃至人类的底层价值观等诸多方面产生重大而深远的影响。然而，人工智能的快速发展和渗透，也引发了全球范围内政府、学术界、产业界、国际组织以及大众对于人工智能法律、伦理和社会影响的持续关注和激烈讨论，人们呼吁重视 AI 伦理，加强 AI 治理，践行 AI 向善和以人为中心的理念，发展安全可信、负责任的人工智能。

⊙全球不同政府采取不同的 AI 路径，创新与伦理平衡是 AI 治理的基本原则。

欧盟采取了更加侧重于立法和监管的路径，通过全球首部全面监管人工智能的法规，希望成为 AI 立法和监管领域的全球领导者，并以此转化为其在 AI 技术和产业上的国际竞争力。

美国的人工智能治理政策与监管模式采取“轻监管、重创新”的路径，旨在推动人工智能发展，避免不必要的监管行为妨碍发展，以维持其全球领先地位，并满足国家安全需求。

中国在第三届“一带一路”国际合作高峰论坛期间提出《全球人工智能治理倡议》，围绕人工智能发展、安全、治理三方面系统阐述了人工智能治理的中国方案，坚持以人为本、智能向善，引导人工智能朝着有利于人类文明进步的方向发展。

⊙国际社会探索建立广泛认可的 AI 伦理原则，推进包容且以人为中心的 AI 国际合作机制。

从联合国的“AI 向善国际峰会”（AI For Good Global Summit）以及推动建立“AI 伦理国际对话”的努力，到经济合作与发展组织（Organization For Economic Cooperation and Development，OECD）和二十国集团（G20）提出的人工智能原则，再到中国主提、联合国通过的首份聚焦人工智能能力建设国际合作的决议，国际治理和合作机构

1. 相关调研显示，受 AI 冲击的脆弱人群分布广泛，包括影视制作、游戏开发和设计、创意工作者、自雇人群以及低技能劳动者等。特别典型的体现在两类人群，一是产线工人，因为智能调度系统不仅完全实现了 AGV 代替高强度的人工搬运，还让生产过程可管可控可分析；二是服务业人员，AI 技术在客户服务、零售等领域的广泛应用，导致服务业员工面临被替代的风险。

已步入实质性阶段。从议题角度来看，当前人工智能治理主要关注伦理、规范和安全三个领域。其中，围绕破解人工智能领域各国发展不平衡、不充分问题，以人为本、包容可持续、安全可信、创新发展等日益成为被广泛认可的伦理原则。

⊙从“技术中心”到“技术人文双中心”，科研和产业日益重视 AI 伦理和人的权利价值。

随着 AI 技术本身的演进以及对全行业、全社会影响的不断深入，国内很多科学家、产业实践者已经从单纯的 AI“技术中心”视角转向“技术人文双中心”视角，纷纷呼吁关注 AI 伦理、关注 AI 对人的影响，并提出了各自的 AI 伦理原则，积极防范 AI 应用及滥用可能引发的负面问题。例如，联想集团在业内率先提出了“人本智能”新主张，强调人工智能的普惠性和包容性。

但目前总体而言，行业内在 AI 安全及对人的影响方面的研究投入仍显不足，伦理价值的落地机制仍需探索，以形成行业共识。上海人工智能实验室主任、衔远科技创始人周伯文的观察数据显示，目前，世界上只有 1% 的资源投入在对齐或者安全考量上，对 AI 安全的投入远落后于对 AI 性能的投入。未来，人类将遵循“AI-45° 平衡律”，沿着可信 AGI 的“因果之梯”拾级而上，探索人工智能系统安全和能力的系统性平衡之路。这一进程需要科学家、技术从业者、企业家和政策制定者共同努力，一边发展一边治理。

图4 欧盟《人工智能法案》规定的四类风险

	最低风险	高风险	不可接受的风险	特定透明度风险
典型场景	如启用 AI 的推荐系统和垃圾邮件过滤器	如用于招聘、评估某人是否有资格获得贷款或操纵自动机器人的人工智能	对人类基本权利构成明显威胁的人工智能将被禁止	如聊天机器人，必须向用户清楚地披露他们正在与机器进行交互。某些人工智能生成的内容，包括深度造假，必须贴上标签。当使用生物识别分类或情感识别系统时，需要告知用户
具体要求	绝大多数人工智能系统都属于最低风险类。这些系统未对公民权利或安全构成威胁或威胁很小，其法律责任可以减轻或免除，公司可以自愿采用额外的行为准则以降低风险	高风险人工智能系统包括某些关键基础设施、医疗设备、教育考试系统、招聘系统、执法系统、边境管控、司法系统以及生物识别、分类和情感识别系统。被认定为高风险的人工智能系统必须遵守严格的要求，包括建立风险缓释系统、采用高质量的数据集、记录活动日志、提供详细文档、提供清晰的用户信息、进行人工监督，并确保高水平的稳健性、准确性和网络安全性	包括操纵人类行为以绕开用户自由意志的人工智能系统或应用，如使用语音辅助鼓励未成年人进行危险行为的玩具，以及允许政府或公司进行“社会评分”的系统。此外，生物识别系统的某些用途也将被禁止，例如在工作场所使用的情绪识别系统，或在公共场所为执法目的进行的实时远程生物识别（少数情况除外）	在使用生物识别分类或情感识别系统时需要告知用户。合成的音频、视频、文本和图像等人工智能生成的内容必须标注为机器可读格式，且能够被检测出是人工智能生成的

来源：财新智库根据公开信息整理

2.2 人本智能概念的提出

伴随着新一轮 AI 浪潮所带来的新型人机关系，“人本智能”理念应运而生。简要来说，人本智能（Human-Centric AI，简称 HAI）是指从“以人为本”的视角重新审视人工智能技术及其影响，要求在人工智能技术研发，人工智能产品与服务的设计、应用以及与外界交互中，都必须以满足人类需求和谋求人类福祉为首要目标。

人本智能在价值上突出以人为主体，尊重人的尊严和权利；认为 AI 是为了增强人类的能力和福祉，而不是取代或降低人类的角色、自尊和价值感。

人本智能将 AI 视为由人类组成的更大系统的一部分，它关注 AI 的伦理、社会和文化影响，确保 AI 系统对社会所有人都是可信、可用且有益的。

2.3 人本智能的内涵和原则

“以人为本”是一种广泛的社会价值理念。具体到“人本智能”，它强调在人工智能的发展和应用中持续且全面地关注人类价值、需求和权利，并广泛应用于社会、经济、科技等各个领域。

人本智能强调人本的核心价值，把以人为中心的理念与 AI 有机融合，在发展 AI 的过程中必须考虑 AI 对人自身以及社会整体的影响；AI 的应用是为了赋能人类，而非取代人类；AI 应尽可能像人类智慧一样敏感、细腻、有深度。

其具体的内涵体现为“人本智能”理念下“三个维度”的升级。

① 在人与 AI 两者之间的交互关系上，坚持人本设计。它强调在 AI 技术和系统的设计、应用及推广等全生命周期中，对人的需求、情感和价值的深切理解与尊重，必须坚守不伤害、做有责任的 AI 等底线；应该将人本原则融入其中，坚持人机可持续协同，最终构建一种人机共生的新关系、新范式。

② 在人与AI 的目标工具属性关系上，坚持人是 AI 发展的最终目标。它强调AI 是人的工具，能够扩展人的能力，人的价值提升是 AI 的目标，而非相反。结合 AI 作为工具在提升人的目标的不同维度，可进一步将 AI 细分为机器智能、可信智能、交互智能、共情智能及人机物和谐智能[8]。

1. 引用自上海交通大学人工智能研究院教授、常务副院长杨小康提出的人工智能层层递进演化需求：机器智能、可信智能、交互智能、共情智能和人机物和谐智能。