将价值观放在技术之上拥抱 AI（上）

——专访清华大学智能产业研究院（AIR）院长张亚勤院士

本刊特约记者　钟秀斌

如果说 2016 年 AlphaGo 战胜人类围棋最强手的影响力还局限在中国和亚洲，那么 ChatGPT 则是席卷全球，深入各个领域各个层次。ChatGPT 究竟是何方神圣能全球吸睛？它将给 AI 时代带来什么？它给人类惊喜的同时，又存在哪些隐患？中国将面临怎样的机遇和挑战？为此，本刊专访了清华大学智能产业研究院（AIR）院长、清华大学“智能科学”讲席教授、中国工程院院士张亚勤。

张亚勤，中国工程院院士，清华大学讲席教授、智能产业研究院（AIR）院长

1978 年，12 岁的张亚勤从山西考入中国科大少年班，是当年全国最耀眼的一名大学生之一，成为 20 世纪七八十年代十年寒窗苦读者的集体偶像。1985 年，年仅 19 岁的他从中国科大研究生毕业，赴美国乔治华盛顿大学深造，1989 年获得博士学位。1999 年回国出任微软中国研究院首席科学家，一年后担任微软中国研究院院长。2001 年微软中国研究院升格为微软亚洲研究院，张亚勤担任首任院长。2004 年他被擢升为微软公司全球副总裁，2007 年执掌微软中国公司（董事长），兼任微软亚太研发集团主席。2014 年出任百度公司总裁，2019 年从百度公司退休后，来清华创办 AIR。

AI 时代的操作系统

张亚勤认为，过去 30 年，IT 行业最重要的变革是实现数字化。人工智能的三要素：算力、算法和数据，过去 10 年算力增加了 10 万倍，远超摩尔定律。数字化经历三个发展阶段，数字化 1.0 是内容数字化，包括文本、音乐、图片、视频、语言，直接产物是消费者互联网；数字化 2.0 是企业数字化，即企业信息化，如 ERP、CRM、商务智能、企业智能等，直接产物是云计算；现在是数字化 3.0 时代，信息、物理、生物世界的数字化，比如车、机器、城市、道路、家庭、大脑、身体器官、细胞、分子、基因都在数字化。1.0 到 2.0 是从原子到比特，3.0 是比特和原子分子的相互映射，直接产物是海量数据，数据比 1.0 和 2.0 时代大很多数量级，这是 AI 发展最重要的基石。

AI 的鼻祖是图灵。图灵测试可验证人类是否能够分辨出交互对象是人还是机器。ChatGPT 是人类第一个通过图灵测试的智能体。之前很多年所做的聊天机器人属于分析式、决策式、预测式、鉴别式 AI，在专有领域如智能客服等方面表现很好，但跨领域效果一般，因为还不是通用类生成式 AI。ChatGPT 走向通用 AI。大模型用不同的微调策略，执行不同的任务。因此，从产业方面讲，GPT+ 等各种大模型是 AI 时代的“操作系统”。

钟秀斌　我们多数人没有 AI 背景知识，特别期待您

这样的权威学者科普一下，如何认识 ChatGPT 这一事物?

张亚勤 ChatGPT 及其所引领的大语言模型和生成式 AI，不是横空出世的，而是随着人工智能技术发展进化而来的。大基础模型 Foundation Model （FM）正在成为人工智能时代的操作系统。

说到操作系统，人们自然想到 PC 时代的 Windows，移动互联网时代的安卓和 iOS。现在有两种操作系统，一种是在云平台上，一种是在用户终端上。GPT 所代表的大模型是云平台里的操作系统。云平台有 IaaS（基础设施层）、PaaS（中间层即平台层）和 SaaS（软件服务层）三层。GPT 可以理解为 MaaS（Model as a Service），模型即服务，属于 PaaS 层。

云和端的说法比较形象。在端这边，是一个更小的系统，会具体到手机、机器人、无人车等物体上。这个系统较小，建立在现有的体系之上，比如手机终端建立在安卓和 iOS 上，或建立在机器人操作系统 Ross 之上。

操作系统会撬出一个大的产业生态，包括硬件（芯片）、软件和服务体系。大语言模型是 AI 时代的平台级技术，将形成比移动互联网规模更大的产业生态。移动互联网产业生态比 PC 时代要大一个数量级，ChatGPT 为核心的 AI 产业生态比移动互联网还要大一个数量级。现在的 AI 芯片不同了，芯片早期是 CPU x86，后来是 ARM，现在 AI 芯片多是 GPU、ASIC 等。AI 时代，大模型的算法也不同，整个计算架构完全变了。

冯诺依曼架构过去 60 年一直是计算机架构的主体，目前需要新的计算架构。大模型需要大计算，无论 GPU，还是数据中心，其内部通信功能很强而形成大连接大计算。深度学习算法需要大量的向量 Tensor 运算，稀疏矩阵和布尔代数逻辑。数据处理的这些变化，需要新的芯片架构、新的指令集、新的框架、新的工具链，在此之上的应用也得重写。比如从电脑到手机，手机上的 APP 是新的，各类平台上有 APP store，商业模式同样发生变化。这是我从产业方面理解所带来的新体验与新变化。

AI 三大要素算力、算法和数据变化了，基于 ChatGPT 或者 FM 上的应用和产业模式也随之改变。因此，它在成为 AI 时代的操作系统。人们现在感受还没那么深。就像我们所经历的互联网发展一样，刚开始只在行业内扩散，然后走向通用，在各个垂直领域里深入融合。ChatGPT 也是这么一个过程。现在为什么 ChatGPT 的震撼这么大，因为它有用户体系，界面有接口，用户可以直接应用它。过去大模型基本都有 API（应用程序接口，Application Programming Interface），像 Windows 一样呈现一个界面，用户直接用，但后台复杂的计算处理，用户不知道。ChatGPT 像 Office 一样，用户可以直接用。目前尚是一个最简单的接口，今后它会发展成横向通用模型，会有各种垂直模型，应用在 to B（行业）的垂直领域，机会非常大。

钟秀斌 从计算机到人工智能，工具或者技术能力越来越强大。比如 1998 年 IBM 深蓝计算机打败国际象棋高手，2016 年 AlphaGo 战胜人类围棋高手，2021 年 AlphaFold 成功预测了人体蛋白质结构，意味着 AI 开始攻克生物科学和医学领域的重大难题。这些成果令人惊叹，AI 在计算和逻辑方面比人更具优势。ChatGPT 完全可能融会贯通人类有史以来的所有知识，它是否会突然“涌现”出一种新的能力和价值来，比如在数学、物理、化学、生物等基础研究，甚至在药物研发、蛋白质解析、基因编辑、智能驾驶等应用领域，有较人类更出色的能力，去解决人类未曾解决的难题?

张亚勤 AI 生成式大模型正在朝这一方向走，这些都是未来必然的发展趋势。深度学习 AI 可以分

成两个阶段，第一阶段是解决专用问题，比如说 AlphaGo 下围棋，AlphaFold 解析蛋白质，已经超过人类。无人驾驶其实也已超过人类，只是大规模商用还需要点时间。语音识别、人脸识别能力 AI 超过人类。AI 在每个专业任务的能力基本上与人类差不多，过去的 5 年 AI 发展很快，在每一个垂直领域和解决某一类问题的能力还会继续提高。

ChatGPT 开始具备一定的通识能力，把人类的知识融合在一块，变成生成式 AI 大模型，它就可以写文章、作画、编程序。目前也能做些简单的数学题，考试和人类平均水平差不多，甚至更强。但是要在数理化领域做基础科研，它还需要更大的数据，培养更多的能力。为什么 ChatGPT 能写程序？是由于 Open AI 把 Github 等开源社区上 10 亿行代码，拿来训练后而形成的能力。今后，如把所有的数学知识（包括数学公式）都训练之后，ChatGPT 数学肯定做得很好。只要有足够的物理学数据，它经训练之后，物理水平也能很高。有家公司在训练 ChatGPT，尝试批阅高考试卷，据说批得还挺准的。可能有很多公司在每一个垂直领域去做 AIforScience 这些事。5 年之后 AI 数理化做题水平都会超过人类。它可以证明人类有些证明不了的数学、物理公式或猜想。它不仅可帮人类证明数学题，还可能发明新公式，它每天都在学习，会不断进步。甚至了解各种化学反应后，会推导新的化学反应式。人的精力和能力是有限的，而 AI 不一样，它学会了数理逻辑和各种定律后，综合能力完全可能强过人类。

当然，人类有些能力 AI 是没有的。如果把人的能力分成三个层次，第一层感知能力，像听说交互或图像识别，AI 和人类差不多，甚至比人还强；第二层逻辑能力，AI 的推理、判断能力，将来发展和人类也差不太多；第三层情感和意识能力，AI 没有。人的感情和意识，AI 虽然可能会学习，但是我们目前并不知道人类的情感和意识是怎么产生的。因此，人可能会对 AI 有情感，但它对人可能并没有情感，或者不是真正的情感。对于通用 AI 或者超智能体，大家可能觉得它们会有自我意识。就像好莱坞科幻电影里的机器人，最后控制了人类。我觉得这不太可能。

钟秀斌 您能肯定 AI 不会有情感，对人类也不会有情感？

张亚勤 我觉得有时候这是一种哲学或者是一种信仰，我也不肯定，因为这是个悖论。人类情感和意识是怎么产生的？到底这是碳基生命的特点，还是比如说人类记忆力达到一定程度，或者推理能力达到一定程度之后就有意识了？我们并不知道。既然不知道，我们只能选择相不相信。我是选择不相信。因此，这没有对错，而是每个人的判断或者信仰问题。

像对待核武器一样重视 AI 风险

2019 年张亚勤到清华大学创建 AIR，出任掌门人。几十年来引领 AI 研究的他，将目光聚焦在应用研发上，致力于 AI 与产业融合。他给 AIR 立了规矩——做负责任的 AI，提出 3R 原则，即通过 Responsive（积极响应）、Resilient（适应发展）、Responsible（坚守价值）三大准则，来推动 AI 赋能行业发展。

如何做负责任的 AI？张亚勤认为，要了解不同行业的底层基础，分析技术将产生的影响和后果，通过技术创新，国际合作和治理，广泛应用 AI 推进第四次工业革命。因此，AIR 选择智慧交通、智慧物联、智慧生命作为主攻的三大方向，以人的生存环境、生命健康和人的价值为主题，发挥 AI 效能，造福人类。张亚勤多年领导世界顶流公司和顶级研发团队，比常人更深谙技术的两面性。他热爱 AI，拥抱 AI，他确信 AI 将给人类带来无法限量的

价值，是人类第四次工业革命引擎。同时，他深信 AI 存在尚未为人所知的不确定性风险，因此，当国际上有学者将 AI 风险和核战争、新冠疫情相提并论，需要引起人类高度重视时，他和老东家比尔·盖茨在第一时间回应，并在这份声明上签名，表明心迹：一定要把人的价值、价值观和责任放在技术之上。

钟秀斌 人类科技史表明，人类将一项新技术使用到一定程度后越用越熟练，效用也越大。反过来，一旦失控，它也可能对人类伤害越大。比如核能技术，既能绿色应用，也能制造核武器。现在的社交网络成为人们生活的标配，可网络诈骗层出不穷。基因编辑可以在医疗上造福人类，但同样也可在科学伦理不允许的地方犯错。化学造福人类满眼皆是，但也有人为了牟利，往食品里添加不宜食用的化学品。正因为 AI 的超能表现，令人不免想到如何使它与人类的价值观和目标对齐的问题。您如何看这一问题?

张亚勤 这一问题我 20 年前就开始谈了。每当 AI 有进展，这个问题就会回来。这也自然，任何技术能力越来越强时，它的风险也会越来越大。早期人们不太相信 AI 的能力，好多人觉得像是吹牛。现在看起来，当年科幻电影里的 AI 场景和内容，现在已越来越近了，正一步步成为现实。人类拥有两种智慧：发明技术和控制技术走向，二者要均衡，目前后者稍微落后了些。要解决 AI 和人类价值观对齐问题，第一，做技术的人要把技术和研究放到对齐上面，先要让机器理解人的价值，追随人的价值。对齐（alignment）这个词很好，其实这不仅仅是伦理的问题，还有如何实现的问题。做技术和研究的人要致力于实现对齐任务，不能只开发能力，不着力对齐问题，这是相当重要的问题。现在有门新学科 AI Security Research，即 AI 安全研究，就像航天有门学科 RocketSafety Engineering，专门研究火箭安全工程。AI 也需要有人专门研究安全问题。

第二，要制定和坚持一些基本原则。20 世纪 50 年代美国科幻作家阿西莫夫定义了机器人和人类的三原则。2017 年一批科学家又制定了《阿西洛马人工智能 23 条原则》（Asilomar AI Principles），我认为这是人和机器的基本原则。机器永远是从属体，人类是主体。不管机器、软件，或是机器人也好，它是从属的，其主体可以是人，也可以是公司，或者我们目前的实体。因此，机器产生的任何行为，主体要负责任，并从法律层面明确主体责任。比如说无人车出车祸，用户和制造无人车软件的技术公司，加上保险公司等主体共同承担主体责任，就像现在司机出车祸，可能是司机问题，也可能是车的问题，由用户或车厂和保险公司承担事故责任。AI 也一样，今后也得购买保险。总之，AI 本身不能独立成为主体。

第三，AI 不能有自己独立的伦理和价值系统。它服务人的系统，它的价值就是人的价值，它的伦理体系就是人的伦理体系。我们要让它服从这一体系，实现这一体系。此外 AI 要可信任，具备安全性和可控性，这点也非常重要。这类问题涉及技术、伦理道德和法律层面。当前人们在这方面所做工作还不多。最近欧洲刚签署一个新规则，中国网信办起草了《生成式人工智能服务管理办法（征求意见稿）》，工信部出台《工业和信息化领域数据安全管理办法（试行）》。技术研发、道德伦理、立法监管等合力并进，才能让 AI 发展更健康。

我们欢迎政府立法监管，监管才能使 AI 方向正确。哪怕走得慢一点，也需要监管，以确保方向正确。有时政府政策法规出台相对慢些，就像互联网一样，刚开始野蛮生长，发展到一定程度，才能出台法规加以规范。因此，正如我之前所说的两种智慧要平衡，技术往前跑，监管来规范。信息社会技术发展快，人的意识形态、政策法律体系仍然按

2023 年 7 月 6 日至 8 日，上海世博展览馆，2023 世界人工智能大会展示 AI+ 智慧医疗服务成果。 中新社 陈玉宇 摄

工业时代的节奏，自然会滞后些。

钟秀斌 信息时代面临的互联网治理问题越来越突出。比如个人隐私信息保护，类似健康码、行程码，在特殊时期，人们为公共价值而让渡一些个人权利。但如果国家相关立法没有跟进，这类工具就可能被滥用。现在网购已成为人们日常生活的一种方式，但一些平台利用大数据算法，可把用户浏览过的一些物品，无限地推送给用户，干扰用户购物体验。AI 时代类似的问题肯定不会少。人们在享受 AI 带来的便利的同时，也必然会受到负面影响。正如您所说的，技术发展和风险控制要平衡，齐驱并进。如何让主导技术发展的科学家和工程师们有清醒的认识和坚定的原则，哪些是要鼓励发展的应用，哪些是要时刻警惕的技术？如何培养他们科技为善的价值观？

张亚勤 我在清华大学智能产业研究院（AIR）强调，做研究或者做技术，一定要把人的价值、价值观和责任放在技术之上。因此，AIR 选择了在未来五年十年 AI 具有巨大影响力的三个方向，研究课题（包括与公司合作项目）都与这一理念相关。一是智慧物联，面向双碳（碳达峰碳中和）的绿色计算、小模型部署到端等，节能减排。物联网应用广泛，可以做许多东西，但 AIR 选择围绕双碳做文章。二是智慧交通，机器人和无人驾驶，安全第一。无人驾驶安全性增加 10 倍以上，现在 90% 交通事故都是人为事故。AI 驾驶可以排除人工驾驶中的失误，大大增加安全性。同时，低碳节能减排，各种应用无缝衔接，效率高。三是智慧医疗，AI 新药研发、生物技术，服务人的生命健康。

清华 AIR 的选题围绕人的生命健康、生存环境和人的价值，基于计算机学科基础 ABCD（Artificial Intelligence, Big Data, Cloud, Device），面向世界科技前沿、经济主战场、国家重大需求和人民生命健康的方向，开展相关研究工作。这是 AIR 社会责任所在。我们要有这样的认识，但确实存在挑战。这一挑战不仅仅是人工智能，而是面对所有的技术。

每项技术就如核技术，如果人类有选择，也许最好不去找像铀或镭这些放射性物质。核磁共振医学应用造福人类，而核武器却可以毁灭人类。像化学和生物，早先有生物战和化学武器，后来因为世界大国之间达成共识，立法禁止使用生物化学武器。核武器出现后，也就相应立法禁核。现在的基因编辑技术，世界各国也有明确的立法，不能用于改变物种，尤其是针对人类。需要有一些清晰的基本规则，来规范和约束技术的发展与应用。AI，尤其现在到大语言模型程度之后，由于具备生成式能力，具有不可预测性，它能生成什么人们并不能完全预知。

过去 AI 工作主要帮助人做分析、决策和预测，但现在它完全可能创造出新的东西，不加控制，未

必是好事。尤其是银行金融或者具有关键使命的系统，应用 AI 时，小心保守些为好。AI 有些能力的因果关系目前人类还不清楚，比如说智能有限问题、AI 怎么达到智能、黑盒子问题或者透明性问题，人们并不明晰它的因果关系。我们有时了解 what 而不太清楚 why，可能了解 what 百分之三四十就可以做，知其然不知其所以然。其实 Why 很重要，我们现在处在不清楚 AIWhy 的情况下，在应用到物理系统或关键使命体系时，更得小心保守。

钟秀斌 今天处于 AI 时代，人们不得不面临如何确保 AI 与人类价值观和目标对齐的问题。您作为权威 AI 科学家，更能感受控制技术风险的重要性和迫切性。

张亚勤 现在越来越感受到问题的重要性，越来越多地讨论这类问题。以达沃斯论坛为例，我十多年前参加达沃斯论坛的时候就在讨论 AI 对第四次工业革命或者社会变革的深刻变化和深远影响。但从 2018 年开始，达沃斯论坛的议题开始主要谈论 AI 发展所带来的风险及其管控，在刚刚闭幕的天津夏季达沃斯论坛上，80% 议题都谈风险控制。早年我参加达沃斯论坛时，大家多在探讨 AI 的能力，对它具有什么特别功能并不确信，认为人工智能就是个软件，所能做的事情有限，不可能对产业有什么深远影响。

之后人们越来越多地谈论 AI 相关话题。先是谈数据，担心掌握大数据会形成数据垄断大公司。我 2019 年以百度总裁的身份参加达沃斯论坛。当时有二三十家大公司 CEO 都在讨论大公司大责任大担当的话题。特别是当时脸书（Facebook）数据泄露和操纵选举事件发生后，人们觉得公司掌握数据越多，AI 能力越强，需要承担的责任也更大。这几年大家的风险意识增强。之前国内媒体很少谈论 AI 风险，像你这样问我的人很少，一般都是关注 AI 怎么改变产业，投资机会在哪里，中国和美国企业如何竞争等等。

我们拥抱 AI，希望它走得更好。但目前对于 AI 可能的风险问题，已经引起越来越多的人关注和担心。最近关于 AI 风险有两份著名的公开声明，一份由特斯拉创始人、SpaceX 公司 CEO 埃隆·马斯克（Elon Reeve Musk）和美国未来生命研究所（Future of Life Institute）创始人、《生命 3.0》作者迈克斯·泰格马克（Max Tegmark）发起的，呼吁暂停训练比 GPT4 更强大的 AI 模型至少 6 个月。

另一份由剑桥大学助理教授戴维·克鲁格（David Kreuger）发起，包括多伦多大学教授杰弗里·辛顿（Geoffrey Hinton）和深度学习之父、图灵奖得主、蒙特利尔大学教授约书亚·本吉奥（YoshuaBengio），比尔盖茨、OpenAICEO（主导 ChatGPT）山姆·奥特尔曼、谷歌 DeepMind 首席执行官戴米斯·哈萨比斯（Demis Hassabis）等多位顶级研究人员、工程师和 CEO，就 AI 对人类可能构成的威胁发出最新警告，超过 350 位相关领域人员共同签署了这份只有 22 个单词的声明：“减轻 AI 带来的灭绝风险，应该与流行病和核战争等其他社会规模的风险，一起成为全球优先事项。”把 AI 可能带来的风险，用了十分显眼的词——灭绝风险，将 AI 风险等级和核武器、流行病相提并论。

第一份声明我没签名，因为我觉得科研很难停得下来。一个企业可以暂停研发，其他企业未必会暂停；或者一个国家可以暂停相关研究，但并不能阻止另一国家继续。第二份声明我签名了，我认为做人工智能研究要是没有这样的风险意识，就不会重视，如果 AI 研究一旦失控就会带来灾难性的风险。有了风险意识之后，政府、企业、研究院校、社会各方就会像对待核武器、新冠疫情一样，时刻警惕，强化监管，使技术走在正确的道路上，从而达到发展和风险的平衡。