【大模型及其发展趋势】

人工智能的发展、大模型，最重要的基石是数字化。ChatGPT 的出现代表了三个方面：

- 首先，这是第一次有一个智能体通过了图灵测试。多年来，计算机和人工智能领域的科学家都希望能通过图灵测试来验证技术成果，而 ChatGPT 首次实现了这一目标。
- 其次，ChatGPT 开启了通向通用人工智能的亮光。目前，我们距离实现真正的通用人工智能还有很长的路要走，可能需要十年甚至二十年的时间。不过，ChatGPT 等先进技术的出现为我们指明了前进的方向和道路。
- 最后，这代表着人工智能的操作系统正式出现。随着大模型的兴起，我们已经迈入了一个全新的人工智能时代，这类似于 PC 时代的 Windows、Linux 操作系统，以及移动时代的 iOS、安卓系统，都标志着技术发展的里程碑。

AI 大模型的六大发展趋势：

- 跨模态，多模态，和多尺度大模型：新的大模型包括自然数据（语言文字、图像、视频），也包括从传感器获取的信息（比如无人车中的激光雷达点云、3D 结构信息、4D 时空信息，或者是蛋白质、细胞、基因、脑电、人体的信息）；
- 新算法框架：我们需要新的算法来提升当前的大模型效率。当前的大模型稠密激活，计算效率远低于人脑，且商用成本高昂，甚至模型用的越多亏损越多。人脑是效率最高的智能体，它有 860 亿个神经元，每个神经元有几千个突触，却只有不到 3 斤重，耗能 20 瓦。从这个角度来看，人脑的储存量，计算量和能耗效率之高，是目前任何大模型都无法比拟的。我们需要新的算法体系，稀疏激活网络、效果更优的小网络等来提升模型使用效率；
- 自主智能：模型正在成为一个代理（Agent），自主规划任务、开发代码、调动工具、优化路径、实现目标，包括 N+1 版本的自我迭代、升级和优化；
- 边缘智能：大模型需要很多算力和资源，如何在边缘和设备终端实现高效率、低功耗、低成本、低延时地部署是一大关键问题；
- 具身智能：大模型正在被用到无人车、机器人、无人机和工厂，交通、 通讯、电网、电站和其他物理基础设施；
- 生物智能： 大模型正在被用到人体、人脑、医疗机器人和生物体和生命体的连接和控制。

大模型发展的几点建议：

- 要建立分级体系，特别是对前沿大模型。对于一般的人工智能大模型不需要太多监管，但是对像 Sora 这样的大模型，一定要有监管，包括场景的约束，需要有评估体系。
- 实体的映射。从事技术研究的科学家们往往热爱创新，并倾向于自主管理。然而，我们必须认识到，对于前沿的大模型技术，需要加强治理，这包括对 AI 生成内容进行明确标识。无论是在搜索引擎还是广告平台上，都有责任让用户知道哪些内容是由 AI 生成的。尽管广告标识的可见度逐渐降低，但仍需坚持进行标识。对于数字人和 AI 生成的内容，同样需要明确标识，这是必须遵守的基本原则。

- 以后做智能体机器人，无论软件或者硬件，一定需要有一个对应的主体。如果AI犯了错误，出现问题，一定要能追溯到主体。
- 做前沿大模型的公司，包括国家基金会、科研机构，要把投资的10%用来做AI风险的研究。不仅仅是政策研究，还有学术研究和技术研究，技术人员对AI风险的研究必须现在就开始，使研究者和企业、政府共同前进，而不是对立的关系。

【加强AI风险安全监管力度】

目前，技术界存在一种想法，认为应该技术先行，先把AI大模型、架构、算法等技术做完后，再让政府部门监管。这是不对的，应该一开始就让政府部门参与，大家一起促进技术发展，否则等到技术完善后才监管，可能就来不及了：

- 第一个风险是信息方面的风险。由于大模型具备高度逼真的声音、图片和视频模拟能力，因此存在潜在风险。以美国总统竞选为例，已有人模仿拜登和川普，这可能对选举的公正性造成影响。当前技术的发展已到达一个关键时刻，我们必须高度关注大模型所带来的潜在风险。
- 第二个风险，当信息智能拓展到物理智能、生物智能的时候，如果大模型失控或者是被坏人所利用，会造成很大的风险。随着大模型作为操作系统和工具在各个领域的广泛应用，包括金融、军事和决策系统等，风险将呈指数性增长。
- 第三个风险，生存风险。人工智能大模型可能带来与核武器、流行病相当的风险。我们必须首先树立风险意识，尽管目前人工智能研究和产品仍处于早期阶段，但我们有机会通过多种方法来改变其发展方向。如果我们缺乏这种意识，将无法有效应对和降低这些风险。

在政府监管中，建议对超过某些能力阈值的人工智能系统，包括其开源的副本和衍生品，在建立、销售与模型使用上进行强制注册，为政府提供关键但目前缺失的对新兴风险的可见性。我们同时建议，应规定一些明确的红线，并建立快速且安全的终止程序。一旦某个人工智能系统超越此红线，该系统及其所有的副本须被立即关闭。各国政府应合作建立并维持这一能力。

【产业趋势分析】

以互联网的发展脉络为喻，目前的人工智能相当于处于网景时代的初期，已经构建起了基础系统，并使得公众能够开始使用。随着网景之后互联网世界中IE浏览器、门户网站、社交媒体、电子商务和搜索引擎的兴起，互联网行业才真正迎来了繁荣。展望未来，人工智能有望像互联网一样，对所有行业进行全面的优化升级。

每次产业平台的更迭所产生的效应都是数量级的增长。从过去来看，移动互联时代的产业机会比PC时代至少大10倍，人工智能时代的产业机会比PC时代至少大100倍，比移动互联时代大10倍或更高。从互联网历史看，PC互联网、移动互联网到了后期时，中国公司的规模创新、平台创新都很多。PC时代我们几乎都是抄美国，到了移动互联网时代，中国的移动互联产品比如支付、短视频、微信，以及O2O，都比美国做的要好。

【人工智能场景突破】

过去，人工智能面临的一个主要挑战是常识知识的匮乏，导致每个具体任务都需要开发专门的模型来完成。然而，随着人工智能常识能力的不断增强，它正朝着通用人工智能（AGI）的方向发展，这是一个非常显著的进步。此外，人工智能不仅仅模仿单个人的能力，而是在模仿全人类的智慧和知识。通过积累人类最优秀的成果，人工智能有望形成超越个体的超级智能。随着这种趋势的持续发展，人工智能将能够涵盖人类所学的所有知识和常识。

同时，人工智能的另一种发展形式——生物智能，也将逐步重构社会。如果有人存在视力、听力方面的问题，或是其他身体方面的缺陷，生物智能都可以帮他修复。在 AI 帮助下，人们可以用脑电波、心电波去控制某个物体。比如，现在人们弹钢琴，需要手指在键盘上操作。十年后，人们可以用脑电波控制机械弹奏钢琴。类似操作可以发生在很多场景，如操纵假肢写字、倒咖啡、握手，完全像正常人类的手，还可以让假肢跑步、爬山、攀岩，它可以比人跑得更快。当前，一些科学家已经在做生物智能方向的科研。十年后，这些科研技术会变成产品，并真正进入人类社会。

从更长远的视角来看，五十年或百年后，人工智能有可能催生出一种新的生命形态，这种生命形态将是硅基与碳基生命的融合，实现人类与机器的深度整合。到那时，人工智能将极大地增强人类的能力，使我们在某种程度上成为“超人”。虽然这听起来似乎非常神奇甚至有些不可思议，但我们绝不能低估人工智能的潜力。我们当前所见识到的人工智能，仅仅是其巨大潜能的冰山一角，尽管其已经展现出了惊人的实力，但它仍然处于非常初级的阶段。

【提升大模型能耗】

大模型现在的效率依然很低。目前的大模型效率比人脑差的太多了，可能至少差 1000 倍。从规模和耗能来看，人类的大脑是效率最高的智能体，它有 860 亿个神经元，每个神经元有上万个突触，却只有不到 3 斤重，耗能不到 20 瓦，人脑的储存量和效率之高，是目前任何大模型都无法比拟的。

所以现在需要提高大模型效率，让能耗更小。而且在人脑中，你问一句话时，它并没有调动所有神经元，只调动一小部分，越聪明的人调动越少。可大模型不一样，向大模型提出任何的问题，它可能都要调动所有的资源，这无疑是一种巨大的浪费。
在研究和产品开发方面，所需的时间是不同的。对于大模型而言，其能耗的降低并非一蹴而就，而是一个逐步的过程。不同的研究可能会分别实现 20%或 30%的能耗降低，这需要在多年时间内持续进行并取得累积进展，才能最终达到一个成功的节点。目前，尽管大模型已经具备了一定的可用性，但由于其高能耗问题，微软、谷歌等公司在大模型方面的业务尚未实现盈利。

【自动驾驶机器人场景落地】

在北京亦庄地区，目前已经有无人车在路上进行测试，但其中大多数仍然配备了安全驾驶员以确保行车安全。未来十年内，随着技术的不断发展和完善，真正实现无人驾驶的无人车将成为道路上的常态。同时，人形机器人将逐渐进入一些家庭，它们能够监测居住者的身体状况并提供陪伴和交流的功能。此外，我们所居住的社区将引入机器人担任保安职责，甚至有可能出现机器人警察。在医疗领域，一些医院已经开始使用机器人来读取患者的影像资料，辅助医生进行诊断。预计在未来十年内，机器人的数量将超过人类，并且它们的应用范围也将逐渐扩大，尽管初期可能应用不够广泛，但随着技术的不断进步，这些 AI 场景将逐渐实现并普及。

研究工作需要秉持长期主义原则，研究者应保持心态平和，避免急于求成。

【聚焦前沿方向、开发创新算法】

清华大学智能产业研究院（AIR）已明确三个科研方向，这些方向在人工智能领域未来 5 至 10 年内具有巨大影响力。具体包括：机器人与无人驾驶技术、智慧物联网（涵盖绿色计算及小模型部署到端等）、以及智慧医疗（包括 AI 驱动的新药研发等）。当前的创新机制可分为三个阶段：从 0 到 1 的原始创新，从 1 到 100 的应用发展，以及从 100 到无穷大的规模化推广。学校科研的使命在于推动更多原创性、从 0 到 1 的实验室研究成果。我们期待在中国、特别是在清华大学涌现出新的算法和架构。目前，研究层面仍存在大量从 0 到 1 的工作亟待完成。

【持续长期投入 AI 风险研究】

人类有两种智慧，一个是发明技术的智慧，一个是引导它走向正确道路的智慧。近期，AIR 在人工智能治理和风险研究领域取得了显著进展。随着模型规模不断扩大，特别是达到万亿参数级别的前沿模型，我们对大模型带来的风险和治理问题需给予更多关注。这对于开源和商业闭源领域同样重要。建议最优秀、最聪明的人才投身于治理研究，开发先进的治理技术。目标是让人工智能不仅超越人类智能和能力，更要具备善良和创意，以符合我们的价值观并避免犯下重大错误,打造善良的 AI。

2023 年 10 月， Stuart Russell，我和两位图灵奖获得者 Yoshua Bengio、姚期智先生在英国进行了为期三天的首届“人工智能安全国际对话”(International Dialogue on AI Safety)，部分与会者签署了一份联合声明：呼吁“在人工智能安全研究与治理上的全球协同行动，是避免不受控制的前沿人工智能发展为全人类带来不可容忍的风险的关键。”

今年 3 月 10 日我出席了我国首个 AI 安全高端闭门论坛“北京 AI 安全国际对话”，与图灵奖得主 Geoffrey Hinton、Yoshua Bengio 等 30 余位专家围绕国际 AI 安全技术前沿研究、产业应用实践、政策引领等话题，在为期两天的对话中展开深入探讨，共同拟定并签署了《北京 AI 安全国际共识》，提出人工智能风险红线及安全治理路线，同时呼吁“在人工智能安全研究与治理上的全球协同行动，是避免不受控制的前沿人工智能发展为全人类带来生存风险的关键。”