超级智能的超级隐忧

刘永谋，系中国人民大学哲学院教授

一、智能治理与社会模拟器

近年来，伴随智能体技术（如被称为“通通”的智能体雏形）以及国外“Human AI”“社会模拟器”等概念的发展，一种以个体行为建模为基础的社会治理研究路径逐渐兴起。其核心逻辑在于将社会成员抽象为可计算的 AI 智能体，通过大规模多智能体仿真，对各类政策与社会治理方案进行模拟与效果验证。

从历史视角看，此类尝试并非首创。包括心理学、机械力学乃至化学等多个学科均曾尝试运用其学科范式解释甚至模拟社会运行。例如，化学领域曾出现通过合成激素调节公众情绪以维持社会稳定的设想。智能治理在本质上可视为技术治理思想在智能时代的最新延续。

尽管超级治理对提升治理水平具有积极的推动作用，然而必须指出，其本身存在着根本性的内在缺陷，因此不应将这一模式绝对化：其一，行为主义的理论局限。已有大量研究指出，人类行为的丰富性、意图性与复杂性难以被完全数字化还原，基于行为主义的模拟无法涵盖人类意识的深层维度。其二，环境参数的无限性。构建高保真社会仿真所需输入的参数近乎无限，不仅涉及物理环境变量，更包括个体精神状态、心理活动等难以穷尽且量化的复杂因素。其三，“实然”与“应然”的哲学分野。模拟器仅能反映社会系统“实际如何运行”，无法回答社会“应当如何发展”的规范性命题。社会演进方向本质上是集体选择与价值判断的结果，无法通过技术模拟予以决定。其四，数据与技术层面的现实困境。用于建模的人类数据本身可能存在偏见、污染或错误，而与一个有缺陷的 AI 系统互动，可能引发难以预料的社会后果。

二、超级智能与超级伦理

支持发展超级 AI 的一种常见论据是“超级治理”愿景，即认为极端复杂的全球性挑战（如气候危机、跨国协调等）需要超级智能的介入方能有效应对。倘若由超级 AI 主导全球治理，其发展轨迹可能呈现出如泰格马克在《生命 3.0》中设想的多种情景：它可能将人类视为低等物种并予以灭绝或圈养，也可能选择服务于人类、自我放逐或与人类隔绝。尤其值得警惕的是，超级 AI 可能因无法真正理解人类价值体系，在追求某一工具性目标的过程中，无意间导致人类文明的毁灭或衰退，此即所谓“AI 文明危崖”风险。

尽管如此，需要客观评估的是，在当前的全球风险优先级排序中，超级 AI 的威胁通常并非位列最前。主流风险评估往往将气候变化、核战争与新型病毒等视为更紧迫的生存性威胁。一种合理的推测是，在超级智能具备灭绝人类的能力之前，人类社会可能已因其他原因而陷入严重的生存危机。

三、超级治理与人的机器化

“超级治理”愿景的背后，潜藏着“科学人”理念的兴起——即人日益被理解并建构为一个可被数字化、可被治理、可被控制的智能机器。当“人是什么”

这一根本性问题的解释权，从哲学家、艺术家及思想家手中，全面转移至物理学家、数学家与 AI 专家时，“人的机器化”进程便已悄然启动。

该进程主要体现在两个层面：其一是人的测量化。通过智能手环监测生理指标、利用身体质量指数（BMI）评估健康水平等实践，反映出个体日益习惯于以数据来定义和管理自身。其二是认知模式与心理结构的重塑。例如，许多人已难以在脱离 PPT 等视觉化工具的情况下维持长时间专注；技术工具所提供的即时信息检索，正削弱人类生物记忆的必要性；而创造性思维过程也越来越多地被“外包”给算法与智能程序。长此以往，人类在依赖外部技术延伸认知能力的同时，其内在的思维习惯与心灵结构亦面临着根本性的变迁。

四、如何看待超级智能

目前并无明确证据表明超级智能必然会出现。在此背景下，可将“超级智能”概念部分地理解为一种特定的话语建构或社会修辞。其主要功能体现在两方面：其一，作为资源动员的工具。AI 领域历来有通过新概念炒作维系领域热度的倾向。在一轮技术发展高潮过后，“超级智能”此类极具冲击力的术语，能有效吸引社会关注、撬动资本与政策资源的持续投入。其二，作为人文批判的方法。人文学者通过设想此种极端情境，得以深入剖析并预警其可能带来的颠覆性风险，从而推动社会进行前瞻性研究与防范。

有鉴于此，有必要对 AI 专家群体中可能存在的过度宣传与技术鼓吹进行适度的规范与引导，以确保公共讨论建立在审慎与理性的基础之上。