人机融合社会中的系统调控

王芳¹, 郭雷²

(1. 山东大学 数据科学研究院, 济南 250100; 2. 中国科学院 数学与系统科学研究院, 北京 100190)

摘要 科学技术的深入发展和广泛应用正在深刻改变人类社会的生产和生活模式, 尤其是信息技术的快速发展, 推动人工智能 (AI) 与人的关系由“辅助”到“互助”和“融合”不断深化。相应地, 社会调控对象向着能力不断增强的 AI 系统以及 AI 赋能的人和组织演变, 社会系统将呈现“信息 - 物理 - 社会”的高度耦合, 调控环境也相应转换为人机融合的复杂社会系统。这一显著变化将倒逼法律体系的变革, 对社会系统调控能力提出更高要求, 迫切需要人文社科与科学技术相关方法的交叉融合。论文基于对人机融合社会系统发展的研判, 在分析法律体系、控制论、人工智能等所面临挑战的基础上, 从法律规制和系统控制的角度, 提出将调控工程系统的控制科学与调控社会系统的法律充分结合, 在系统科学框架内研究处理人工智能时代社会中的系统调控问题; 并进一步归纳探讨了以人为本、优势互补、系统分析、量化支撑、算法规制、博弈控制等若干调控原则和可尝试的前瞻领域, 以及需要深入研究的问题。这一思路可望量化表达公平正义等法学基本问题, 优化调控系统技术指标, 促进人机融合社会系统实现其安全稳定、公平正义、民主自由、和谐发展等价值目标。

关键词 人机融合; 智能系统; 社会系统; 信息 - 物理 - 社会系统; 道德伦理; 法律量化; 复杂系统; 控制论; 博弈论; 系统论

Regulation of systems in man-machine integration society

WANG Fang¹, GUO Lei²

(1. Data Science Institute, Shandong University, Jinan 250100, China;

2. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China)

Abstract The rapid development and wide applications of artificial intelligence (AI) are changing the structure of human society profoundly. The AI has developed from auxiliary tools to good consultations, and deeply integrated with human society. Furthermore, the objects of social regulation are being evolved into powerful AI as well as organizations and individuals with capabilities significantly enhanced by AI, and the environment is being transformed into complex systems of man-machine integration. This change will inevitably bring various uncertain risks and challenges in the man-machine integration systems (MMIS), and it is of great importance to build a social regulatory system that conforms to such value objectives as safety and stability, equity and justice, democracy and freedom, and harmonious development, etc. That will need the combination of related methods in both humanities/social sciences and natural/technological sciences. A possible path is to combine the science of law for social systems with the control theory for engineering systems, which may likely enable us to secure basic ethical and legal issues, and at the same time to realize the technical objectives of regulation for MMIS. There are some basic principles and methodologies that could be followed, for examples, human-oriented AI, advantageous complementariness, systems analysis, quantitative support, algorithm regulation, and control of game-based systems.

收稿日期: 2020-05-26

作者简介: 王芳 (1980-), 副研究员, 博士, 研究方向: 系统与计算数学, E-mail: wangfang226@sdu.edu.cn; 郭雷 (1961-), 中国科学院院士, 研究员, 研究方向: 系统与控制科学, E-mail: Lguo@amss.ac.cn.

基金项目: 国家自然科学基金 (11688101)

Foundation item: National Natural Science Foundation of China (11688101)

中文引用格式: 王芳, 郭雷. 人机融合社会中的系统调控 [J]. 系统工程理论与实践, 2020, 40(8): 1935-1944.

英文引用格式: Wang F, Guo L. Regulation of systems in man-machine integration society[J]. Systems Engineering — Theory & Practice, 2020, 40(8): 1935-1944.

Keywords man-machine integration systems (MMIS); AI systems; social systems; cyber-physical-social systems (CPSS); ethics; legal quantitative; complex systems; cybernetics; game theory; systems theory

信息技术,特别是互联网、移动通信、超级计算、大数据、人工智能与区块链等技术的飞速发展与广泛应用,正在推动形成耦合程度越来越强的“信息-物理-社会系统”(cyber-physical-social systems)。这一系统以人在回路(human in the loop)为基本特征,以各类信息的感知、监测、收集、传输、分析、融合和利用为基础,广泛影响人类生活的方方面面。可以预见,随着信息技术的深入发展,各类人造机器将拥有越来越强的自主性和智能性,推动社会生产力和生产关系产生重大变化,并衍生出新的社会特征:人的决策与智能算法高度融合,技术指标与价值目标密切关联,算法规制与社会规范合为一体。这一新形态将远远超出人类传统经验,更超出有任何一门学科的研究范围。面向人机深度融合的社会系统,如何有效调控并实现其理想的价值目标,是人类共同面对的深刻议题。

1 人机融合不断加深的社会系统

1.1 信息时代加速人机融合

工业革命以来,科学技术的发展创造了巨大的生产力,表现出惊人的工具价值。自动化、生物、信息、材料、认知技术等人类增强技术在不断塑造、改良、增强人类自身的同时,将逐渐与人形成物质融合的统一整体,60年前提出的赛博格(Cyborg)概念不再仅具有象征意义^[1]。目前,纳米机器人已经能够诊断和修复人体疾病和缺陷,人工耳蜗、假肢、心脏起搏装置等人体植入设备也越来越普及化。人体器官3D打印技术的发展,甚至将使人机器的物质界限变得模糊。AI技术的不断突破进一步推高这一工具价值。智能手机、智能导航、智能搜索、社交媒体、智能金融交易工具等AI产品,在辅助人类提高生活质量的同时,正引导着人类社会从人的行为向“人人-人机-机机”的融合与互动转型。

随着人机协同水平的不断提高和人机融合深度的持续增强,AI由人类的辅助和工具发展为参谋助手,与人类形成了优势互补的“互助关系”。一方面,人机融合可以弥补单纯AI在理性与感性、逻辑与直觉的“博弈均衡”方面的不足。模仿人类记忆能力的AI记忆系统技术,在未来有可能提供全生命周期和广度的经验及其他信息的知识的完整存储、回溯和取回^[2]。另一方面,AI也为人类配备了更为强大的感知、计算、推理、识别与控制等多种能力。脑机接口技术能够帮助严重脊髓损伤患者恢复手部触觉和运动能力^[3],无创脑机接口技术业已实现意念控制下机械臂对运动光标的连续追踪^[4]。全球首个用细胞做成的活体机器人已经诞生^[5],AI技术不再停留在生命与机器的物理关系层面,开始向“人机融合”的有机体发展。

人机融合,是以人工智能为基础的一种人机协同共存的状态。广义理解,从语音识别、定位导航到智慧城市、智能交通,都呈现了人机融合的特点。人机融合也是一个由弱到强不断发展的过程。自动驾驶技术由基本的自适应巡航控制等基本辅助功能向自动驾驶的发展充分体现了这一点。尽管人们对AI的未来发展方向持不同的理解与期望,在AI是否可能替代人类或威胁人类生存等问题上存在争议,但从AI与人类各自的优势与短板(后文将进一步阐述)来看,人工智能系统的不断升级将是发展大趋势,人类和机器将长期处于深度融合与共同发展状态。

1.2 社会组织模式和运行方式发生显著变化

信息技术的发展和AI能力不断增强,迅速改变着人类社会的经济形态、社会交往模式和政治-法律结构,社会运行方式也将随之逐渐改变,“人人-人机-机机”的互补、竞争和协调将成为社会运动的新动力,社会调控的对象也因此演化为能力不断增强的AI系统和AI赋能的人和组织。

这一变化具有重要意义,人机融合互动更加深入,超大规模的社会协作与共享更为可能,人类社会可期进入一个“超级互联时代”,进而形成“信息-物理-社会”高度耦合的复杂巨系统。这种高度耦合带来了“万物互联”的高效与便捷,但同时也加剧了稳定性与安全性风险。一方面,“人人-人机-机机”的协同和协作往往更加困难,如何理解人的认知与行为对物理世界和信息空间的影响和作用机理,理解、预测和调控社会网络的结构和行为等成为社会管理的重要问题。另一方面,许多人机融合系统可能更容易受到攻击。2019年委内瑞拉电力系统因受到网络攻击出现多次大范围停电^[6],就是通过从网络空间到物理空间的攻击最终影

响到社会系统的实例。

2 现有调控方法面临系列挑战

显而易见,人机融合社会中的系统调控既涉及规范自然人与社会组织的法律体系,又涉及对人造系统调控的控制技术,并且人工智能技术的发展对系统调控方法也有影响。然而,现有调控方法仍存在根本局限。

2.1 社会组织模式和运行方式的变化倒逼法律体系的变革

2.1.1 法律体系的适应性变革需求

全球化、网络化、数字化潮流正在塑造和改变着法律所处的宏观社会环境,网络通信和多媒体等技术的迅猛发展,大大突破了社会交往和信息交流的时空局限,传统基于层级的社会调控手段备受挑战。社会结构与运行方式的一系列变化倒逼着法律体系的变革。

以信息收集与融合泛在带来的信息安全和隐私风险为例。现代智能技术裹挟着每个人向“数字化”转变,信息收集无处不在,个人在事实上逐渐失去了对隐私的有效控制,传统的隐私保护思路呈现疲态。为此,欧盟《通用数据保护条例》要求进一步完善设计保护隐私(privacy by design)、默认保护隐私(privacy by default)、知情同意、加密等概念的标准,以保护个人数据隐私权利。同时,多种数据的聚集融合很可能蕴含单一或少量数据所无法预见的功能,我们称之为“1+1>2”的“涌现”^[7]。这种“涌现”的新信息很可能超出一般个人的常识和预测能力。因此,目前信息收集领域所奉行的“告知+同意”机制已经不能满足隐私保护的要求,基于传统单个文件的保密审查制度也对政府信息公开原则下的保密目标带来严峻挑战。

此外,数据资源垄断则有可能带来民主的风险。随着数据量及其覆盖范围的不断扩大,数据已升级为重要的战略资源,并作为基本生产要素参与社会分配^[8],人类社会生产、交易、消费和分配模式都随之改变,传统物权面临着结构性调整。但由于人与机器算法在信息收集计算和存储利用能力上的鸿沟,海量信息可能不断向少数有能力的的数据收集者流动,逐渐形成技术领先者和使用在先者的绝对优势,为数据垄断者形成“超级权力”带来可能。加之信息寡头、技术霸权逐利的本能,这些“超级权力”可能掌控相关社会经济资源的获取和分配,甚至可能影响政府的决策和走向,潜在地带来财富和福祉分配不公,甚至赢者通吃的吸利效应,进而可能使得技术上的弱势群体演变为社会的弱势群体。这将带来信息时代人权体系的扩展,如所谓的“第四代人权”数字人权^[9]。这一“以数据和信息为载体,展现智慧社会中人的数字化生存样态和发展需求的基本权利”^[10]的新型人权,不可避免地要处理与其他权利及公共权力的关系,进而产生了如何通过数据技术兼顾公共安全与隐私保护,如何处理数据使用效用及经济利益与信息保护的关系等一系列新的重要课题^[11]。

事实上,法律体系变革与科技发展是近代社会发展的“双轮”驱动力,它们在相互适应中也互相推动,形成了科技-法律系统在安全与发展动态平衡中“适应演化”的循环。法律应当规制科技“双刃剑”的负面影响,使其始终为人类的福祉服务;而科技的发展也会影响法律调整的边界。从控制论角度看,法律对科技的调控,也可在某种程度上看作是从“鲁棒控制”到“适应控制”的演化过程。一项“新技术”的诞生及应用,首先要通过法治守住伦理与安全的“底线”,规范科学技术的发展应用和社会影响;随后围绕应用中出现的新问题和对不确定性的新认识,进一步适应性调整法律以纠正偏差,推动技术与社会的再发展和再进步。因此,新的法律体系难以简单遵循自上而下的单向形成机制,而更应通过社会系统的自主运行和异质主体间的相互博弈共同演化形成。这将是一个不断试错与适应的过程,法律与科技互为系统与环境,形成两者之间不断变化的结构耦合关系,在大量争议和纠纷的实践中逐步形成满足新型社会需求的法律调控体系。

2.1.2 法律体系技术重构的困难

随着人机融合系统能力的不断增强,确保其在特定规范框架内行为尤为重要。为此,必须实现伦理法律对于AI的“内化于心”和“外化于行”,具体而言,就是要实现伦理法律对AI的嵌入,并在此基础上对由此产生的社会关系进行伦理法律调控。

关于AI的伦理规制,无论在原则层面还是具体规则层面,国际社会已经形成了一定的规范体系(后文将详述)。相较伦理规则而言,法律规则的具体性和规范化为其嵌入AI提供了优势和便利。法律自动执行已经在行政审批等领域得到广泛应用。然而,这并不意味着体系性的法律也具有相同的优势。法律系统在理想状态下是遵循形式逻辑并且相对封闭的“自我指涉系统”^[12],是基于一定逻辑起点,满足同一律、矛盾律、排中

律和充足理由律的规则体系。但事实上,法律系统包罗万象,是各种社会关系规范的总和,任何成文法系统都难以满足严格的形式逻辑,伦理道德偏好以及社会各种利益的平衡是法律规则体系形成的基本驱动力。因此,霍姆斯反对使用纯粹逻辑的方法来构建法律学说,提出“法律的生命不在于逻辑,而在于经验”^[13]。但对于成文法系统而言,至少“逻辑+经验”仍为其内在线索,而其中“经验”的部分难免带来逻辑检验的困难,并因此可能产生法律条文之间的潜在矛盾。若将这种存在潜在矛盾的法律体系嵌入算法,很可能带来算法执行的内在冲突。因此,逻辑检验是成文法律体系完成技术重构、完整嵌入基于形式逻辑建构算法面对的重要困难。

2.2 现有控制理论应用于社会系统的局限

由于智能技术和算法的加入,社会系统调控必然要引入自然科学和工程技术的理论和方法,需要伦理法律治理规范与科学技术方案之间的融合。调控工程系统的控制论与调控社会系统的法学拥有天然的关联。但是,从复杂程度来看,面对异质主体复杂多样并高度关联的人机融合社会系统,现有控制论仍存在局限。

虽然自动控制具有很长的发展历史,但是控制论作为一门现代科学诞生的标志是 1948 年美国数学家维纳出版的名著《控制论 (Cybernetics): 或关于在动物与机器中控制和通信的科学》。书中控制论的研究对象超越了传统自动控制,既包含了机器也包含了作为动物的人。该书密切围绕“控制与通信”这条主线以及“反馈”这一基本原理,从“动物智能”与“机器智能”的若干重要方面展开对比讨论。这是一项划时代的开创性工作,对当今众多科技领域产生了广泛影响,并促进了许多新兴交叉学科方向诞生,“赛博 (Cyber)”语言在当今也广泛流行。1950 年,维纳出版了《人有人的用处——控制论与社会》,该书进一步考虑了以通信作为反熵的手段来研究社会问题,认为法律问题可以看作通信问题和控制论问题,并针对若干具体情形分析了系统所具有的博弈性质。然而,由于生物系统和社会系统的复杂性,控制论中迄今发展最深入和成熟的方向是针对工程系统的控制理论 (control theory),对由人组成的社会系统的调控要比对工程系统的控制复杂得多。

首先,社会系统一般是不可还原、不可逆的,甚至难以预测的。研究者与被研究对象都具有自主意识,互为客体、互相影响,“极难使观察者与被观察对象间的耦合降到最低限度,相反,观察者会对被观察对象施加巨大的影响”^[14]。这是对社会系统调控与对自然或工程对象调控的显著不同,由此福尔斯特 (Heinz von Foerster) 曾于 20 世纪 70 年代提出二阶控制论 (second-order cybernetics) 概念。实际上,社会系统的调控对象不仅是具有自主意识和博弈行为的人,而且在人机融合的社会中,调控对象还应包括智能机器人和人机融合系统。这些异质主体目标各异、复杂多样且高度关联,“人人-人机-机机”之间可能产生各种竞争合作行为和自组织现象。在这一复杂社会系统的调控回路中,被控系统、传感系统、通信系统和执行系统等往往有人的参与,不再是传统的机械装置或工程对象,各个关键环节都可能有自己的利益追求,甚至可能产生诈骗伪造、歧视误导、扭曲政策、本位主义、“上有政策,下有对策”等行为。然而,目前控制理论的调控对象主要是遵循基本物理规律的工程系统,一般没有自己的利益追求,这与社会系统调控对象存在根本不同。因此,不能将工程控制的研究方法直接套用到社会系统上,否则可能会产生异化现象,这已被大量事实证明。

其次,在工程系统的控制中主要采用负反馈控制来消除系统运行中所产生的偏差。然而,在社会系统的调控中除了须保持系统稳定性和秩序外,还要保障系统的动力和活力,因而往往需要正反馈激励与负反馈纠偏等调控措施的有机协同。后文将进一步说明,这也会为社会系统的运行带来更多风险,并使其调控研究更为复杂。

再者,人机融合社会中“人人-人机-机机”多重交互融合,异质主体高度关联,必然带来系统难以预测的“涌现”。基于涌现性,个体的、局部的行为聚集成整体功能从某种意义上能够脱离它的起源。虽然近年来有大量关于复杂动态网络分析、多主体系统集体行为、分布式优化与协同控制、信息物理系统安全性等问题的研究和成果,但目前的相关研究模型和处理方法,远不足以用来合理描述并有效规范人机融合社会中的复杂系统。

2.3 目前基于深度学习的 AI 支撑社会系统调控的局限

目前以大数据与深度神经网络 (DNN) 学习为基础的主流 AI 算法,在支撑社会系统调控方面尚存在明显欠缺。

第一,以大数据与 DNN 为主流的机器学习算法,在可解释性和因果推断等方面存在根本局限。

众所周知,确定因果关系是人类认识世界的基本目的之一。鉴于传统的统计推断方法只能确立相关关系,近年来关于因果关系的研究被认为是“新科学”^[15]。一般来讲,相关关系比因果关系更易于确立,但是社会价值判断最为关键的就是确立因果关系。存在意义变量(真实的因果关系)时,算法的可解释性是保障判断公正性的关键,也是保障决策稳健性的基础。但目前以DNN为基础的机器学习算法在本质上是对“黑箱”的功能模拟,只可能发现它所学习的数据之间的相关性,不足以揭示内部因果关系。从系统学的角度看,“在给定环境条件下,系统的结构可以唯一决定功能,但反之一般不然。这一基本事实,造成了根据系统功能来认知其内部(黑箱或灰箱)结构的困难”^[7]。

目前美国超过一半以上的州运用COMPAS,PSA,LSI-R等再犯风险评估算法辅助量刑。但此类基于一系列变量推断再犯风险的算法的可靠性引发质疑,并由此产生了若干诉讼^[16]。欧盟相较美国更为保守,对基于算法的自动化决策持谨慎克制态度,并在其《通用数据保护条例》第21条和22条即赋予了数据主体免于自动化决策的权利^[17]。

第二,机器学习以“存在即为合理”的大数据为基础,难以对现实世界进行“超我”的改进。

人类拥有在“实然”(is)与“应然”(ought)之间进行判断或行为的能力。但是,机器学习以“实在”的大数据为基础,通常以“存在即为合理”这一假设为逻辑前提。而大数据一般都来自具有各种反馈回路和人为博弈行为的社会复杂系统,来自存在于现实世界的大量未经逻辑规整的经验和带各种干扰因素的量测,其准确性、合理性与正当性都难以证实,一般也不具备数据处理理论所需要的独立性和平稳性等良好统计学性质。因此,一方面,不加甄选地以大数据为基础进行学习认知,很可能导致算法的混乱和偏见。另一方面,大数据学习无法回避“休谟问题”^[18],即使基于现实合理的数据进行学习,也仅为依赖算法对现实世界的认识,无法超越现实并实现向“应然”的演进。比如,对于什么是“正义”,不同的文化、不同的语境,会产生不同的判断,功利主义、自由主义、道德主义等思想的争论从未休止。基于现实数据的机器算法,难以对价值进行超越经验的判断,也就难以对现实世界进行“超我”的改进。

3 人机融合社会系统调控的若干原则和方法

虽然法律是社会调控的基本工具^[19],但是由于人机融合社会系统的技术特质,对其调控必然需要与工程控制思想方法结合。值得指出,本文中的“调控”并不试图涵盖所有的社会治理问题和相关方法,而主要针对需要约束性较强的方法(如法律规制和工程控制等)来进行调控的问题。

人机融合社会系统不同于单纯的人造复杂系统和传统社会系统,如何对其进行科学调控是现代自然科学技术和社会科学共同面对的严峻挑战。面对人机融合社会系统的特殊复杂性和不断增长的智能性,以及人工智能技术本身的不确定性,如何构建人机融合时代的伦理道德秩序和法律秩序,进而实现社会的安全稳定、和谐优化和持续发展?回答这类根本性问题,迫切需要多学科交叉以突破传统调控手段的局限。我们认为,以安全稳定、公平正义、民主自由、和谐发展等为价值目标,将调控工程系统的控制科学与调控社会系统的法律充分结合,在系统科学的框架内研究处理人机融合社会中的系统调控问题,是解决未来复杂社会系统中相关治理问题的一条可能路径。这一路径在大数据、云计算和人工智能等技术的助力下,可望在人机融合社会系统调控中,量化表达公平与正义等法学基本问题,并优化实现稳定与效率等系统技术指标,值得研究重视。具体而言,实现这一路径至少应当遵循以下原则。

3.1 以人为本

以人为本、造福人类是AI发展的前提,这一发展理念获得了国际社会的共识,各国纷纷制定AI发展规划和指南。例如,2019年4月欧盟发布的《可信AI伦理指南》^[20]提出了可信AI的7项关键要求,首先强调了AI不能侵犯人类的自主性,应当有助于人类行使基本权利,支持个体基于目标做出更明智的决定。为此,该指南指出可以建立适当的人的监督机制,例如,human in the loop,指人在AI系统的每个决策周期进行干预的能力(尽管许多情形既无可能也不期望);human on the loop,指在AI系统设计周期进行人工干预并监控系统运行的能力;human in command,即监督AI系统的整体活动(包括其经济社会和法律伦理影响)并决定某一情形是否使用AI的能力。我国亦于2019年出台《新一代人工智能发展规划——发展负责任的人工智能》,旨在更好协调人工智能发展与治理的关系,确保人工智能安全可控可靠,推动经济、社会及生态可

持续发展, 共建人类命运共同体^[21].

3.2 优势互补

“以人为本”的共识涉及人类发展目的和共同命运, 但显然也与对 AI 与人类各自的局限及优势互补这一事实的考量有关. 尽管人造机器在减轻或替代人的体力劳动和脑力劳动方面取得巨大进展, 但是目前尚远不具备人在价值判断、创新思维、态势感知和情感沟通等方面的优势. 发挥 AI 与人类的互补优势是人机融合发展的必然要求.

理性, 代表了人类的至上的知识能力^[22], 是人类认识“实然”世界、通向“应然”世界的阶梯, 是逻辑与非逻辑复杂处理在现实世界的一元表征, 也是人类所独有的特殊优势. 而算法的优势在于对“实然”的发现和解决, 但它在超越“实然”的价值判断和“模拟人类的反思性思考”^[23]等方面目前仍存在根本性的局限. 可以预见, 未来人机融合系统将把人类的优势与机器的优势进一步结合, 人机融合将长期存在. 人处理其擅长的“应然”等价值取向的主观信息; 机器处理其擅长的“实然”数据, 同时在学习人处理“应然”信息的过程中优化算法, 使得“人类独有的悟性可以灵活自如地赋能机器, 解决复杂动态问题内在的悖论”^[24], 从而产生人机融合系统整体优化的效果, 使人工智能成为人类自然的自然延伸和拓展. 另一方面, 新技术的发展(如区块链等)也可以变革传统的社会信任机制, 促进人机融合系统新型信任机制的建立及社会和谐发展.

3.3 系统分析

系统科学为人机融合系统的分析提供了基本方法. 系统科学是关于各类系统的结构、环境与功能的普适关系以及演化与调控的一般规律的科学^[7], 他肇始于 20 世纪 40 年代贝塔朗菲提出的一般系统论, 目前已发展成包括控制论、信息论、运筹学、博弈论、突变论、协同学、耗散结构理论、建模与估计理论、稳定性与鲁棒性、非线性动力系统、复杂适应系统、以及复杂网络等在内的关于系统演化、认知、分析与调控的丰富理论体系^[7].

受控制论与信息论启发, 卢曼 (Luhmann)^[25]提出的社会系统理论认为社会是一个功能分化的系统, 通过“语义”的关联实现“语义系统的自我指涉与自我生产”. 而法律系统作为社会子系统, 具有运作上的封闭性和认知上的开放性^[26]. 这些思想对社会系统的分析调控提供了有益洞见. 哈肯 (Haken) 创立的“协同学”^[27], 对于通过“序参量”来实现人机融合社会中系统的协同调控 (coordination regulation) 具有启发意义. 霍兰 (Holland) 提出的复杂适应系统 (CAS) 理论^[28], 为分析具有学习和适应等功能的多自主体 (agent) 系统自下而上的“涌现”行为提供了一般性框架. 最近 MIT 等机构科学家提出的“机器行为学”^[29]跨学科研究, 着眼于智能机器在与现实环境交互中所展现出的复杂行为, 有望成为人机融合系统研究的基础.

我国系统科学奠基人钱学森曾提出系统科学体系的四个层次划分^[30], 以及处理 (包括社会系统在内的) 开放复杂巨系统的方法论, 即以人为主、人机结合, 从定性到定量的综合集成方法和大成智慧工程思想. 如前所述, 人机融合社会系统的一个重要特征是不可还原性. 钱学森认为, 系统论是还原论与整体论的辩证统一^[30]. 这使得系统论超越还原论成为可能, 具体途径至少有三条, 即“整体指导下的还原与还原基础上的综合相结合, 机理分析与功能模拟相结合, 系统认知与系统调控相结合”^[7].

特别地, 关联分析是系统分析的重要内容之一. 人机融合社会系统是通过组织结构、物理结构和 (或) 信息结构实现复杂关联的. 这种关联一般会增加调控的复杂性, 但也可能增加调控的灵活性, 因为除了可能对人或机器直接调控外, 还可能通过关联来进行调控, 也可能通过调控关联来实现目的. 此外, 人与机器的异构性和异质性决定了无论调控目的还是调控手段, 对人机融合社会系统的调控都与传统社会系统不同. 无论如何, 对人机融合系统研究不能孤立的研究人和机器的各自行为. 有例子说明, 哪怕是两个都稳定的线性子系统, 如果连接不当, 则关联系统也会变得不稳定^[7]. 此外, 由于社会系统中被调控对象的自主性, 不科学的调控策略可能导致“一管就死、一放就乱”等现象. 进一步, 人机融合系统的秩序和活力一般依靠众多正负反馈回路的关联和协同, 但如果反馈不当, 系统可能产生不稳定甚至异化行为. 例如, 不当的负反馈可能会使工程系统的瞬态响应产生震荡, 而不当的正反馈又可能使社会系统出现“奖励的惩罚”^[31]等现象. 进一步, 如果负反馈机制失效, 哪怕是小的扰动或误差, 关联系统也可能在正反馈机制推动下失稳, 从而导致“蝴蝶效应”. 2008 年由美国次贷危机引发的全球金融危机以及近期新冠病毒在全球的蔓延就是例证. 此外, 设计不当的人机融合系统甚至可能隐藏安全性风险. 2018 和 2019 年波音 737 MAX 客机的两次空难就直接源于人机融合

系统的设计缺陷^[32]。另一方面, 如何充分利用人机融合社会系统的结构特点实现更好的调控, 也是重要研究问题。例如, 面对新冠肺炎病毒等突发新型传染病, 如何科学设计并动态优化兼顾疫情防控、权利保障和社会经济发展等不同目标的调控措施, 值得从系统科学的角度进行深入分析研究。

3.4 量化支撑

人机融合系统的分析与调控需要量化支撑。量化, 是打通法律规范和工程调控之间壁垒的桥梁, 是在系统思维基础上实现人机融合系统有效调控的基础。人机融合社会系统中“信息 - 物理 - 社会”的高度耦合性, 将对传统的以定性为主的社会调控方式带来挑战, 科技方法与传统社会规范的结合势在必行。实现对各种社会关系以及规范社会关系的道德和法律的量化, 将促进人机融合社会中系统调控研究的更好发展。

同时, 量化方法的准确运用也需要与研究系统结构与功能关系的系统科学深度结合。定性是定量的基础, 量化可以为定性研究提供证据, 检验定性研究所给出的重要结论。同时, 量化也是发现社会现象之间必然性、决定性联系的重要工具, 甚至可能发现仅通过定性研究而难以推出的“反直觉”结论。Schelling 对种族隔离的研究^[33]即运用了一个经典的计算社会模型。但是, 在基于数据进行某种性质推断时, 系统的结构信息具有不可忽视的重要作用, 否则得出的结论可能不够科学。比如, 统计学中著名的“辛普森悖论”说明, 一个系统的某些整体性质可能与其组成部分的性质完全相反。

大数据时代, 数据收集与融合的泛在性使得量化研究的意义更加凸显。量化研究的相对客观性逐步受到社会科学领域日趋广泛的认可, 在较谨慎引入量化方法的法学领域也有学者尝试量化工具。最高人民法院公布的海量裁判文书, 为司法量化研究提供了数据基础, 为计算法学这一新兴学术领域提供了有力支撑, 并催生了“借助数学方法量化罪刑关系, 通过研究并科学化刑法数量关系, 经由罪行量化解析, 以期达致罪刑均衡之正义境界”的“数量刑法学”^[34]等新兴方向的发展。量化方法在法学研究中的运用, 为立法、司法、法律监督等从定性走向定量, 弥补定性规范到定量执行之间鸿沟提供了可能。通过对社会各方利益充分博弈的共识形成机制、伦理与法律的基本概念、基本理念和调整措施的量化表征的分析和研究, 有望为给出“立法 - 执法 - 司法”之间定量反馈信息以促进偏差校正提供科学性依据, 而相关研究(如 [35, 36])也将为人机融合社会系统的调控提供有益借鉴。

3.5 算法规制

人机融合社会, “信息 - 物理 - 社会”高度耦合, 科技、社会、法律、环境紧密联系。为实现系统的有序运行, 需要一个符合安全稳定、公平正义、民主自由、和谐发展等社会基本价值目标并且适应技术发展的灵活治理框架。在这一治理框架中, 对机器算法的规制是必要基础。法律规范体系“内化于心、外化于行”, 嵌入算法的同时调整“人人 - 人机 - 机器”互动博弈所形成的社会关系, 实现对社会系统的调控。

首先, 需要对机器算法的底层伦理规制。

算法规制, 即对 AI 的伦理规制。这一层面的规制是人机融合社会系统有序运行的基础, 也是“底层”的调控手段之一。“底层”的算法规制, 需要通过伦理的嵌入实现, 即量化表达基本伦理道德问题, 嵌入控制算法或代码, 确保机器能够凭借道德调解机制做出价值判断, 能够在与环境的交互中正常行为并与自然人进行良好协同。这一层面的调控必须通过伦理与算法的内部融合实现, 是一种“内化于心”的内部规制。

算法规制的核心在于其对问题的识别、判断和评估, 即首先识别出伦理问题并进行追溯, 发现事实背后隐藏的价值冲突和道德困境; 随后对识别出的问题进行分析并做出判断; 继而评估被写入算法的伦理规范的有效性, 即是否与现实规范一致、兼容。从伦理规制角度观察, 伦理判断原则主义强调“以两个或更多个没有固定道德排序的原则作为伦理分析的基本框架”^[37]; 决疑主义则假设伦理判断无法由道德原则直接推演而来, 需要以类推的方式对事实进行伦理分析并做出判断。目前的 AI 设计中, 原则主义运用较为普遍, 即对算法设定伦理规制, 要求算法编写者遵守一系列的伦理法则, 并特别限制机器做出涉及人的生命的判断。例如德国交通部伦理委员会伦理指导意见第 8 条规定, “诸如伤害一个人以避免对更多人的伤害这样的伦理难题, 不能通过实现编程来处理, 系统必须被设定为出现这种情况下请求人工处理”。但是, 未来的人工智能不可避免地需要处理诸如“铁轨困境”等正义难题, 更需要面对各种具体伦理场景。如何嵌入伦理规范并准确识别, 如何构建 AI 道德调解机制并引导群体与社会道德的实现, 仍面临着技术重构的多重困难。为此, 欧盟提出了“经由设计的伦理”(ethics by design) 原则, 推动通过算法标准、技术指南、设计准则等方式将伦理转化为 AI

要素, 将其价值嵌入其中。

其次, 需要构建规制算法的法律规范体系。

伦理规制是 AI 规制的一个侧面, 未来能力不断增强的人工智能更需要法律体系 (而不是简单的单部法律或单个规范) 的内部嵌入和外部规制相结合, 以确保其行为不超越人类预期的框架。

算法规制法律体系的首要任务是, 确保算法设计目标符合安全稳定、公平正义、民主自由等人类社会基本价值目标, 确保人类的主体地位, 避免由于技术优势可能产生的利益垄断和贫富差距过大等社会问题, 维护人类的福祉和社会整体的福利。为此, 国际社会出台了系列文件, 如欧盟《可信人工智能伦理指南》《算法责任与透明治理框架》《人工智能白皮书》, 我国《新一代人工智能发展规划——发展负责任的人工智能》等, 对算法伦理、责任等问题进行系统规制。其中, “技术稳健性和安全”(AI 系统的准确、可靠且可被重复实现)“透明性”(确保 AI 决策数据集、过程和结果的可追溯)等规范被反复提及, 以算法可解释性为基础的算法信任机制, 以及算法设计缺陷防范机制成为关注焦点; 在此基础上, 将算法可解释性作为实现算法公平的工具, 建立算法影响评估机制 (AIA), 帮助政策制定者了解算法的使用场景, 评估算法预期用途, 建立算法问责机制。由此, 也衍生出一系列相关权利, 如算法解释权、免于自动化决策的权利等。

在具体策略方面, 考虑人工智能产业发展需求, 算法规制法律体系也需要采取一些针对性的规制策略。例如, 分阶段规制, 即根据规制场景区分设计研究开发、市场营销、应用推广各个环节, 制定规制规则^[38]。比例规制, 即根据技术风险决定规制强度, 以“风险进路”为核心, 在达成既定目标的同时避免不成比例的企业负担, 从高风险的人工智能应用与非高风险的人工智能应用两大方面予以类型化监管^[39]。

算法规制体系无疑仍处于建构阶段。如何将基本价值与算法技术统一起来, 协调不同主体的基本需求, 协调人的主体地位与人工智能发展活力之间均衡关系, 整合零散的制度规范, 形成内在自洽的算法规制体系, 系统思维和分析至关重要, 而博弈控制则是实现算法规制体系建构的一条可能路径。

3.6 博弈控制

从法律规范到系统控制体现了从定性到定量研究的有机结合。算法规制通过伦理法律的嵌入规制和外部治理, 为实现人机融合社会价值目标设定了基本框架。博弈控制则进一步作为宏观调控人机融合系统的“法”, 为人机融合社会系统的安全有效运行提供必要的技术支撑。通过对社会博弈系统控制的量化研究, 在基本治理框架内实现并不断优化调控的具体技术指标。

博弈控制是博弈论与控制论相结合的一个具有层级结构的调控框架, 上层为宏观调控变量 (可以是多元的, 包括前馈、反馈、关联、协同和指标设计等), 下层为相互关联 (包括竞合) 且功能不尽相同的多个主体^[40]。宏观调控的目的是在保障主体基本权益条件下, 使系统整体功能达到所期望的目标。如果下层主体之间的非合作博弈形成了某种不良的纳什均衡 (Nash equilibrium), 那么博弈控制就是对纳什均衡状态的调控; 如果主体之间的博弈失去良好的动态平衡, 则博弈调控可望通过“辨证施治”恢复并保持这种平衡。此外, 通过对人机融合系统性能的技术调控指标影响群体共识和群体行动, 也是近年的研究热点。

博弈控制的框架与博弈论中的机制设计 (mechanism design) 思想可以兼容。机制设计又称为反向博弈论, 其目的是通过设计一定的机制使期望的调控目标转化为博弈的结果。虽然这一机制仍面临诸多困难, 如处理结果的“均衡”与“最优”的矛盾、沟通与互动成本及信息真实性问题、非合作博弈与合作博弈等^[41], 但仍不失为一种有益的参考和“开环调控”手段。此外, 正如控制理论与优化理论的关系一样, 博弈控制框架与斯塔克尔伯格博弈 (Stackelberg game) 之间虽有相通之处, 但两者内涵存在根本差异^[40]。

实际上, 博弈控制最基本的特点是对现代控制理论框架的继承和超越。他继承了控制理论中反馈控制这一核心思想, 也超越了控制理论中的被控对象没有自己利益追求的局限。控制理论中最重要的经验之一是简单的反馈控制算法 (如著名的 PID 控制), 可以对付具有大范围不确定性的非线性被控系统^[42], 这为复杂人机融合系统的定量调控提供了有益借鉴。进一步, 近些年人工智能的蓬勃发展主要依赖机器学习算法, 但多在相对封闭环境和明确规则条件下运行, 并且学习与决策多未形成闭环运行系统。由于人机融合系统是开放的, 并且存在各种随时变化的大范围不确定性, 对其调控具有特殊困难。反馈机制可以在大范围不确定性变化和各種强干扰影响下, 实现对复杂开放系统的有效控制。反馈控制与机器学习一样, 都是基于系统数据对付系统不确定性的有力手段, 两者的作用可以形成有效互补。控制理论中自适应控制方法的广泛应用就是这

种互补性结合的成功范例. 因此, 机器学习算法与反馈控制算法在线结合, 形成“边学习边控制”的自适应闭环反馈回路, 对人机融合复杂开放系统的有效调控具有重要意义. 最后, 由于博弈控制系统的调控对象包含了博弈行为, 这就为人机融合系统的调控提供了更为恰当的理论研究框架, 同时也为现代控制理论发展带来新的机遇与挑战^[42,43].

上述六条原则从目标、系统和调控三个方面探讨了实现人机融合社会系统调控的可能路径. 以人为本是社会调控的根本目标和伦理基础, 优势互补是在人本基础上 AI 与人类相互促进与和谐发展的一条基本技术原则. 系统分析是研究人机融合社会中系统调控问题不可或缺的基本前提, 而量化支撑则是信息时代运用系统方法实现社会调控的必要基础. 算法规制和博弈控制是系统思维指导下社会调控的具体原则和方法, 有望促进法律等传统社会调控方式与现代科学技术的有机结合.

4 结语

人类锲而不舍地探索着“理性”的规律, 寻找社会文明有序的理想形态. 新一轮科技革命将人类与机器的融合发展推向新的高峰, 历史的车轮挟裹着智能技术正将人类社会带往未可确定的未来. 面对不确定的未来, 适时调控, 是唯一的确定. 我们认为, 调控工程系统的控制科学与调控社会系统的法律是两个基本视角, 两者在系统科学框架内的充分结合, 可望为研究解决人机融合社会中的系统调控问题提供可行途径. 在这一思路框架下, 如何发挥人与机器的互补优势形成人与机器的和谐共进, 如何将系统科学的思维和基于大数据的量化方法具体用于社会系统的认识和调控, 如何通过算法规制和博弈控制实现社会的有序运行, 人机深度融合对人类身心和社会文化将产生何种长期影响等, 都是值得进一步深入研究的重要问题. 为此, 我们必须调整固有的经验和认知, 探索法律规范与系统控制深度结合的基本路径, 并在教育、科技、法律等多方面做出前瞻性与系统性的深刻调整, 为走向安全稳定、公平正义、民主自由、和谐发展的人机融合未来社会筹谋布局, 让人工智能真正造福人类和我们赖以生存的家園.

参考文献

- [1] 万百五. 21 世纪控制论综述评论集 [M]. 广州: 华南理工大学出版社, 2018.
- [2] Gil Y, Selman B. A 20-year community roadmap for artificial intelligence research in the US[R]. Computing Community Consortium (CCC) and Association for the Advancement of Artificial Intelligence (AAAI), 2019.
- [3] Ganzer P D, Colachis S C, Schwemmer M A, et al. Restoring the sense of touch using a sensorimotor demultiplexing neural interface[J]. Cell, 2020, 181(4): 763-773.e12.
- [4] Cheng Y L, Wei J Z. Progress in the brain-computer interface: An interview with Bin He[J]. National Science Review, 2020, 7(2): 480-483.
- [5] Kriegman S, Blackiston D, Levin M, et al. A scalable pipeline for designing reconfigurable organisms[J]. PNAS, 2020, 117(4): 1853-1859.
- [6] 委内瑞拉大规模停电事件的初步分析与思考启示 [EB/OL]. [2020-6-28]. <https://www.antiy.com/response/20190316.html>.
- [7] 郭雷. 系统学是什么 [J]. 系统科学与数学, 2016, 36(3): 291-301.
Guo L. What is systematology[J]. Journal of Systems Science and Complexity, 2016, 36(3): 291-301.
- [8] 中共中央关于坚持和完善中国特色社会主义制度、推进国家治理体系和治理能力现代化若干重大问题的决定 [EB/OL]. [2020-6-28]. <http://xinhuanet.com>.
- [9] 张文显. 新时代的人权法理 [J]. 人权, 2019(3): 12-27.
- [10] 马长山. 智慧社会背景下的“第四代人权”及其保障 [J]. 中国法学, 2019(5): 5-24.
- [11] Bellovin S M, Dutta P K, Reitinger N. Privacy and synthetic datasets[J]. Stanford Technology Law Review, 2019, 22: 1-39.
- [12] 卢曼. 法律的自我复制及其限制 [J]. 韩旭, 译. 李猛, 校. 北大法律评论, 1999, 2(2): 446-469.
- [13] 小奥利弗·温德尔·霍姆斯. 普通法 [M]. 冉昊, 姚中秋, 译. 北京: 中国政法大学出版社, 2006.
- [14] 维纳. 控制论 (Cybernetics): 或关于在动物与机器中控制和通信的科学 [M]. 2 版. 郝季仁, 译. 北京: 科学出版社, 2009.
- [15] Pearl J, Mackenzie D. 为什么: 关于因果关系的新科学 [M]. 江生, 于华, 译. 北京: 中信出版集团, 2019.
- [16] 汪庆华. 人工智能法律规制路径的一个框架性讨论 [J]. 现代法学, 2019(2): 54-63.
- [17] General data protection regulation[EB/OL]. [2020-6-28]. <https://gdpr-info.eu>.
- [18] 休谟. 人性论 [M]. 关文运, 译. 北京: 商务印书馆, 2004.
- [19] 庞德. 通过法律的社会控制 [M]. 沈宗灵, 译. 北京: 商务印书馆, 2010.

- [20] Ethics guidelines for trustworthy AI[EB/OL]. [2020-6-28]. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.
- [21] 科技部: 发展负责任的人工智能: 新一代人工智能治理原则发布 [EB/OL]. [2020-6-26]. <http://most.gov.cn/kjbgz/201906/t20190617147107.htm>.
- [22] 康德. 纯粹理性批判 [M]. 李秋零, 译. 北京: 中国人民大学出版社, 2004.
- [23] 李平, 杨政银. 人机融合智能: 人工智能 3.0[J]. 清华管理评论, 2018(7-8): 73-82.
- [24] 杜健荣. 法律与社会的共同演化 —— 基于卢曼的社会系统理论反思转型时期法律与社会的关系 [J]. 法制与社会发展, 2009(2): 109-117.
- [25] 高宣扬. 鲁曼社会系统理论与现代性 [M]. 北京: 中国人民大学出版社, 2016.
- [26] 李忠夏. 宪法学的系统论基础: 是否以及如何可能 [J]. 华东政法大学学报, 2019(3): 21-33.
- [27] 哈肯. 协同学导论 [M]. 徐锡申, 等译. 北京: 原子能出版社, 1984.
- [28] Holland J H. Studying complex adaptive system[J]. Journal of Systems Science and Complexity, 2006, 19(1): 1-8.
- [29] Rahwan L, Cebrian M, Obradovich N, et al. Machine behavior[J]. Nature, 2019, 568: 477-486.
- [30] 钱学森. 创建系统学 [M]. 上海: 上海交通大学出版社, 2007.
- [31] 埃尔菲·艾恩. 奖励的惩罚 [M]. 程寅, 艾斐译. 上海: 上海三联书店, 2006.
- [32] 波音 737MAX-8 客机空难事故初步分析 [EB/OL]. [2020-6-28]. <http://www.cnnnews.com.cn/2019/0319/192195.shtml>.
- [33] Schelling T C. Micromotives and macrobehavior[M]. New York: W. W. Norton & Company, 1978.
- [34] 储怀值, 何群. 论我国数量刑法学的构建 [J]. 中国法学, 2019(3): 186-203.
- [35] Wang F, Guo L. Do confessions contribute to lenient punishments in China? An empirical study based on crimes-of-intentional-injury trials[J]. Tsinghua China Law Review, 2019, 12: 57-85.
- [36] 王芳. 刑事诉讼中积极赔偿对量刑的影响及其合理控制研究 [J]. 法学论坛, 2020(3): 95-103.
- [37] 肖健. 彼彻姆和查瑞斯的生命伦理原则主义进路评析 [J]. 道德与文明, 2019(1): 43-46.
- [38] 王钰, 程海东. 人工智能技术伦理治理内在路径解析 [J]. 自然辩证法通讯, 2019, 41(8): 87-93.
- [39] White paper on artificial intelligence — A European approach to excellence and trust[EB/OL]. [2020-6-28]. <https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust-en>.
- [40] Zhang R R, Wang F, Guo L. On game-based control systems and beyond[J]. National Science Review, 2020. <https://doi.org/10.1093/nsr/nwaa093>.
- [41] 苏宇. 论算法规制的价值目标与机制设计 [J]. 自然辩证法通讯, 2019, 41(10): 8-15.
- [42] Guo L. Feedback and uncertainty: Some basic problems and results[J]. Annual Reviews in Control, 2020, 49: 27-36.
- [43] Zhang R R, Guo L. Controllability of Nash equilibrium in game-based control systems[J]. IEEE Transactions on Automatic Control, 2019. doi: 10.1109/TAC.2019.2893150.