中外法学 Peking University Law Journal
Vol.34, No.2(2022)pp.425－444

国家安全视角下社交机器人的法律规制

李 晟*

摘 要 网络与人工智能技术的发展,以及流量经济的运行机制,催生了社交媒体中的社交机器人。社交机器人的兴起,既可能传播假新闻,也可能在不涉及"真/假"问题的意见表达方面体现出其显著影响。通过社交机器人发表的言论,网络与现实中对意见的认识都会受到干扰,并且可能产生作用于物理空间的社会后果,从而形成对言论自由的挑战。这种挑战在试图影响政治的舆论中更为明显,并可能与国际政治领域的"信息战"相叠加成为国家安全问题。因此,对人工智能语境下社交媒体中的社交机器人言论,应当在总体国家安全观指引下予以重视和规制,并在此基础上重思言论自由的分析框架和思维模式。

关键词 社交媒体 算法 社交机器人 政治机器人 言论自由

引言:流量支配的言论

在当下这个信息爆炸的时代,所有人的注意力不可避免被具有流量的个人或事件所吸引,从而汇聚成为流量的一部分。流量经济的模式,在智能互联网时代表现出强大的优势,构成了一种全新的经济形态。〔1〕尽管公众或许反感流量对自身注意力的支配,希望"流量为王"能

* 中国海洋大学法学院教授。本文系2020年度国家社科重大项目"数字社会的法律治理体系与立法变革研究"(项目编号:20&ZD178)的阶段性成果。感谢"网络法研究新航线"工作坊的各位同仁与匿名审稿专家的宝贵意见。

〔1〕 参见石良平、王素云、王晶晶:"从存量到流量的经济学分析:流量经济理论框架的构建",《学术月刊》2019年第1期,第50－58页。

够向“内容为王”转变;〔2〕然而面对这一强大的潮流,仍然无法抵挡。流量经济与智能互联网紧密结合在一起,整体架构决定了其具有的绝对优势。在网络中被创造和分享的海量信息超出了人类的信息处理能力,造成信息的过载,因而人们更多依赖于一些具有信用品意义的符号来简化自己的认知负担,形成快捷的决策。因此,这些符号成为了流量的基础,聚焦了人们的注意力。而算法的分析能够更精确地将部分符号推送给受众,使其认知的视野被有效的框定。同时,基于算法形成的排名,使公众的认知进一步形成循环,从而导致赢家通吃的局面。处于“头部”的符号因为其流量规模更容易被公众关注到,而公众对其的关注又会进一步提升流量数据,从而不断循环,使排名靠后者更难以获得被认识到的机会。平台获得消费者的行为数据,以此对算法进行完善,再通过这样的算法将消费者更紧密地绑定在平台上。由此,消费者的行为数据成为了核心的生产资料,服务于掌握着平台的资本。在这样一个整体架构中,流量经济得到了有效的运行,个人也难以从中脱离出来。

处于这样一个流量时代,无论是社会还是个人都不可避免地深受影响。在当代智能互联网场景中,公众所发表的言论一旦成为大数据可以处理的信息,也就具有了资源的性质。公共言论中讨论到的人物与事件获得了流量,再将流量转化为影响力,而没有得到讨论和表达的声音,则会逐渐销声匿迹,进而失去对现实世界的影响力。因此,在经典的法学视角中具有重要地位的言论自由,在新的时代背景下就会体现出新的问题形态与实践挑战。本文的问题意识由此引出,但本文并不试图面面俱到地讨论言论自由问题,而是关注其中一个具体问题,即在智能互联网与流量经济的技术与政治经济逻辑推动下,人工智能如何重塑了言论的生成与运作机制,并产生了怎样的影响。这样的研究视角,已经被运用于对“假新闻”的研究,形成了深刻的分析。〔3〕而在新闻之外,公共言论包含了更大空间。因此,本文将超出新闻这一范畴,试图关注与客观事实“真/假”无涉的公共言论,提出主观意见的“真/假”问题对言论自由的挑战。更具体来说,这种主观意见的“真/假”问题并不是指意见本身的不存在,而是指其通过人工智能的运用而得到新的表达效果,在当代互联网中出现的社交机器人(social bots),能够基于算法的设计,以自动化的方式生成言论,并伪装成人的言论进入思想市场。社交机器人及其影响在政治学和传播学领域中已经得到了重视与讨论,但法学界的研究尚处于空白状态。〔4〕围绕这一问题,本文将首先讨论信息技术变迁对网络言论产生的影响,描述社交机器人出现的背景。文章第二部分描述社交机器人的运作机制,指出机器人言论通过什么样的技术条件与

〔2〕 参见喻文益:“‘流量为王’的‘善’与‘恶’——‘质量为王’才是真正的‘王道’”,《人民论坛》2019 年第 6 期,第 124－126 页。

〔3〕 参见左亦鲁:“假新闻:是什么?为什么?怎么办?”,《中外法学》2021 年第 2 期,第 544－559 页。

〔4〕 这一概念的提出最早是 2011 年。See Yazan Boshmaf et al., “The Socialbot Network: When Bots Socialize for Fame and Money”, Proceedings of the 27th Annual Computer Security Applications Conference, December 2011, http://lersse-dl.ece.ubc.ca/record/264/files/264.pdf, lasted visited on 16 February 2022. See further, Emilio Ferrara et al., “The Rise of Social Bots”, *Communications of the ACM*, Vol. 59, No. 7, 2016, pp. 96－104.国内最早的引介,参见蔡润芳:“人机社交传播与自动传播技术的社会建构——基于欧美学界对 Socialbots 的研究讨论”,《当代传播》2017 年第 6 期,第 53－58 页。

社会基础产生影响。接下来,第三部分指出社交机器人改变了传统意义上的思想市场,从而形成对于言论自由的挑战。文章第四部分将进一步分析机器人言论引发的政治后果,首先是对国内民主政治造成的损害,进而体现出对国家安全的威胁。第五部分从现象的描述转入对策的分析,指出在中国语境中治理社交机器人,主要意义和关键难题都在于国际政治维度的信息战,因此需要在总体国家安全观的指引下探讨规制对策。

一、谁的言论:社交机器人的兴起

互联网兴起以来,其中的言论问题就始终备受关注。早期互联网被视为一个高度开放的言论空间,让不同个体获得便捷渠道畅所欲言。当社交媒体成为重要的网络平台之后,广大网民借助于社交媒体形成紧密链接来获得各类信息并交流意见。在这样的交流过程中,由于网络中信息交流成本的大幅度降低,其中的言论表达较之于传统社会中的线下交流产生了显著变化。其中被关注最多的现象,是网络上的虚拟流瀑导致"群体极化",帮助那些分散遥远的人发现持有相同意识形态的伙伴,强化自己的观点,使社会中形成更多持有极端意见的群体,而失去在共同经验基础上的社会粘性,从而造成社会分裂。〔5〕社交媒体平台中的这种言论形态,一度引发关于"谣言"和"辟谣"的热烈争议,也引发了相应的理论讨论。〔6〕

随着互联网技术进一步发展,智能互联网时代中的算法推荐,对言论表达产生了进一步的影响。在群体极化的基础上,个体更进一步被封闭在由自己选择愿意看到的信息所构成的"信息茧房"中。〔7〕如果说在早期的社交媒体中,人们是基于自己主动的搜寻和交流形成了极化的群体,那么在智能互联网时代的社交媒体中,人们实际上并非主动发现与自己同气相求的群体,而是基于算法的个性化推荐被吸引到了一起。大数据分析使个体的兴趣与偏好被更充分细致的分析,算法技术的发展形成了基于内容和网络互动的推送方式,基于个体行为数据,算法对个体偏好进行用户画像分析,从而对个体进行更精准的信息投送,始终以符合其偏好的信息进行引导。于是"网络共和国"被"标签共和国"所取代,标签共和国中的个人被算法推送的标签更有力地支配,并与其他人形成了信息和声誉上更强的虚拟流瀑。〔8〕

无论讨论群体极化还是信息茧房,虽然体现出算法介入的影响,但言论的主体并未改变,仍然是互联网中的"网民"。尽管在早期互联网中曾有一个著名的说法"没有人知道网络对面

〔5〕 参见(美)凯斯·桑斯坦:《网络共和国:网络社会中的民主问题》,黄维明译,上海人民出版社 2003 年版,第 36－37 页。

〔6〕 在"假新闻"2016 年成为美国语境中的热词之前,"网络谣言"在 2011 年已经成为中国语境中的热词,对相关讨论的梳理,参见李晟:"修辞视角中的'思想自由市场'及其影响",《华东政法大学学报》2014 年第 2 期,第 76－86 页。

〔7〕 参见(美)凯斯·R·桑斯坦:《信息乌托邦:众人如何生产知识》,毕竞悦译,法律出版社 2008 年版,第 8 页。

〔8〕 参见(美)凯斯·桑斯坦:《标签:社交媒体时代的众声喧哗》,陈颀、孙竞超译,中国民主法制出版社 2021 年版,第 8－17、159－170 页。

是不是一条狗”,但这只是强调网络账号同线下身份的分离,而不会怀疑言论是否来自人类。前人工智能时代,对互联网的言论问题关注重点在于言论如何得到更为自由的表达,以及表达的社会效果。而人工智能时代的信息技术演进,则在算法推荐的基础上形成了社交机器人这种革命性的变革,导致了言论主体的改变,并进一步改变了互联网中的言论生态。

社交机器人作为一种自动化的软件,能够控制社交网络中的特定账号,以类似于控制账号的人类的行为方式在社交网络中活动,包括主动发布文本、照片或表情包形态的信息,转发或评论其他账号发布的信息,访问其他账号主页以及点赞或申请添加好友。更重要的是,社交机器人隐秘行动,同那些直接表明是自动化程序的账号区分开来,冒充为人类行为,利用对人类社交关系网的分析,潜伏在社交网络中获得有影响力的地位。〔9〕社交机器人的这些行动,重点在于抓住眼球,强化了人类用户与其发布的信息的接触,突出的是流量与“用户粘性”,而不是全面和审慎的信息获取与鉴别。

在早期的网络言论生态中,“水军”就被视为一种干扰力量,对人们造成误导,因而需要对其规制以保障网络中的言论自由。〔10〕而社交机器人的兴起,实际上就是“机器人水军”取代了“人类水军”。早期的机器人水军技术水平较低,技术基础是“群控系统”,即通过系统自动化控制集成技术,把多个手机操作界面直接映射到电脑显示器,实现由一台电脑来控制几十台甚至上百台手机的效果,在此基础上通过批量模拟脚本来模拟许多账号的行为。〔11〕如果只能按照事先设定的言论进行机械的重复,完成简单功能,就很容易被发现,重复的言论也会被作为无效的信息被排除,难以影响人们的认知。而社交机器人的自组织,意味集群个体不依赖某一操控主体对其进行集中的组织管理,而依靠自身遵循一定的行为准则,通过观察其所处的环境,与临近的目标对象进行局部交互,在整体上协同配合达成集群所需要的目标,因此操纵者不需要如同水军的操纵者那样事无巨细地进行实时具体的指导,而可以事先在宏观上制定机器人个体的行为规则,让其依赖人工智能自动运行,涌现演化形成了所有个体相互配合的集群智能传播模式,以达成所需的传播效果。〔12〕

早期的社交机器人非常幼稚,或是只能自动发布预先设定的文本,或是语无伦次因而毫无意义。〔13〕因此,社交机器人的广泛运用,其技术基础在于机器学习能力的提升,能够使其表现出类似于人类用户所控制的账号的行为模式。其通过抓取人类聊天中的关键词进行学习,理解语言背后所表达的意图,并由此不断改进自身语言表达方式,与人类用户在社交网络中进行实质性互动。在人工智能不断提升的过程中,除了对文本自然语言的理解,对图像、表情等表达形式的理解和学习能力也在不断提高。社交机器人通过“情感计算”技术提升了互动能

〔9〕 See Boshmaf et al., supra note 4, p. 1.

〔10〕 参见胡凌:“商业网络推手现象的法律规制”,《法商研究》2011 年第 5 期,第 3—11 页。

〔11〕 参见《微信发文介绍“微信群控”黑色产业链 称将坚决处理》,载网易科技,https://www.163.com/tech/article/CLUL22RT00097U7R.html,最后访问日期:2022 年 2 月 16 日。

〔12〕 参见郑晨予、范红:“从社会传染到社会扩散:社交机器人的社会扩散传播机制研究”,《新闻界》2020 年第 3 期,第 60 页。

〔13〕 See Ferrara et al., supra note 4, p. 98.

力,以及对人的共情能力和情绪表达能力。〔14〕通过这样的技术手段,社交机器人所表达的观点并非来自于事先拟就的文本,而是在交流语境中自动生成,更紧密地同话题联系起来。而且,社交机器人还能够识别出人类的社交关系图谱,分析人类在社交网络中的行为,设法同拥有大量关注对象的用户建立起联系,并利用共同好友关系继续进行扩散。〔15〕当这样的社交机器人活跃于社交网络中时,就不仅仅是作为一个静态数字的"僵尸粉",而是能够形成实际的互动。

社交机器人通过诸多接口可以接入社交网络中,简单的验证码并不能形成屏蔽,社交网络对此的防御相当脆弱。〔16〕虽然检测程序不断发展,试图更好识别出伪装成人类的社交机器人,但机器人也在改进行为特征,使之更难以被检测程序所发现。〔17〕由于社交机器人的设计本身就追求隐秘性,检测程序的运用并不能像过滤垃圾邮件那样轻松地将其在社交网络中排除掉。一系列实证考察都显示了社交机器人在互联网已经发展到一个较高的比例。2012 年,Facebook 的一份报告估计,所有账户中有 5—6%是虚假或伪造的,这意味着大约有 5000 万用户是虚假的。〔18〕而研究者对推特中的社交机器人比例的估计则更高,系统检测分类显示 53.2%的用户是人类,36.2%是人机协同(cyborg),10.5%是机器人。〔19〕如果考虑到社交机器人完全可能比人类在社交网络中更为积极地发布信息,那么来源于社交机器人的信息比例会比用户数据的比例还要高得多,例如由 Distil Networks 公司所发布的《2018 恶意机器流量报告》指出,2017 年 42.2%的互联网流量由机器产生。〔20〕

二、如何发言:社交机器人的运作逻辑

社交机器人的伪装性,使其表现为像人类用户一样正常参与社交网络活动:表达观点、发布信息、关注其他用户。当大量的社交机器人以这样的方式在网络中活动时,就构成了"僵尸网络"(botnet)。〔21〕这样的僵尸网络没有侵入人类用户账号或窃取隐私信息,也不同于传统

〔14〕参见高山冰、汪婧:"智能传播时代社交机器人的兴起、挑战与反思",《现代传播》2020 年第 11 期,第 9 页。

〔15〕See Tim Hwang, Ian Pearce and Max Nanis, "Socialbots: Voices from the Fronts", *Interactions*, Vol. 19, No. 2, 2012, p. 42.

〔16〕See Boshmaf et al., supra note 4, p. 3.

〔17〕See Ferrara et al., supra note 4, p. 100.

〔18〕See Norah Abokhodair, Daisy Yoo and David W. McDonald, "Dissecting a Social Botnet: Growth, Content and Influence in Twitter", Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, Februrary 2015, p. 841, https://arxiv.org/pdf/1604.03627.pdf, last visited on 16 February 2022.

〔19〕See Zi Chu et al., "Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?", *IEEE Transactions on Dependable and Secure Computing*, Vol. 9, No. 6, 2012, p. 822.

〔20〕转引自高山冰等,见前注〔14〕,第 8 页。

〔21〕See Abokhodair et al., supra note 18, p. 841.

的垃圾邮件程序造成的侵扰，而是试图表现为具有实质意义的社交活动，活动形态与人类高度相似。社交机器人只是在表达言论而已，只是这种言论的主体并非人类，而是人工智能。而基于这样的特性，社交机器人的挑战就更加凸显出来。我们必须正视新的言论生态：社交机器人不仅仅在转发新闻，其发表的观点是言论，转发和点赞也是言论，网络空间中存在着大量的机器人言论。有学者指出，社交媒体的生态正在从完全由“人”主导变为“人＋社交机器人”的共生状态，社交机器人已经成为社交媒体的一个有机组成部分，社交机器人产品的不断迭代逐渐消解了人类用户在社交媒体中的唯一主体地位，社交机器人也逐渐成为信息内容生产、观念传播和意义表达的重要参与者，社交媒体成为“人”与“社交机器人”的共生空间。〔22〕 如果从技术上很难清除所有的社交机器人，回到其出现之前的网络空间的话，那么，人们就不能装作看不见社交机器人的影响，而必须在这样的共生状态下理解社交网络，理解网络空间的言论自由。当言论的主体并非人类时，就不能运用对人类的言论自由观念与制度去加以回应。而且，在社交机器人与人类共生的网络空间中，社交机器人试图模仿人类行为，但目标并不局限于模仿人类，而是要改变人类行为。〔23〕 由于社交网络已经成为人类社会的重要组成部分，这种大规模出现的新形态言论的作用不局限于虚拟空间，而是也影响到线下的物理世界，形成具有社会意义的后果。当人类行为确实可能被改变时，这一问题的法学意义也就更为突出，必须加以高度关注。不同于计算机科学对社交机器人的研究更关心什么样的言论，也不同于社会学的研究更关心人机关系，法学的研究必须重视这种社会后果。

社交机器人在网络中的使用，不是个体性的，而是大规模的，通过相互分工合作结合成为机器人组成的僵尸网络。也只有通过大规模的使用，才能形成一个复杂系统的“涌现行为”、形成其自组织，体现出更高的智能。〔24〕 也只有在达到一定规模的情形下，社交机器人问题才有必要进入法学的视野。对社交机器人言论所造成的社会后果，应当重点关注作为大规模群体的社交机器人的协同使用所产生的效果，而不是只考察作为个体的言论，同时也应当注意到协同使用大规模社交机器人的实际控制者。

当社交机器人被大规模协同使用时，也就在社交网络中增加了大量的言论，显著超出人类用户言论的规模。当社交机器人能够以类似于人的行为模式去进行互动时，这些言论生成的信息会进一步强化人们在网络空间中的信息过载，因而导致对信息处理能力的干扰，并且影响到判断和决策。而且，机器人言论相对人类并不仅具有数量上的影响，信息效率也至关重要。当一篇文章首次在社交媒体上发布的最初几秒，机器人可能比之后更大规模地使用，这种早期干预让许多用户更容易看到，增加了文章被病毒式传播的机会。〔25〕 但对人类用户而言，其往

〔22〕 参见张洪忠、段泽宁、韩秀：“异类还是共生：社交媒体中的社交机器人研究路径探讨”，《新闻界》2019 年第 2 期，第 16 页。

〔23〕 See Ferrara et al., supra note 4, p. 96.

〔24〕 关于大规模个体如何通过自组织的涌现行为形成复杂系统，表现出不同于简单的个体行为的总和，参见（英）杰弗里·韦斯特：《规模：复杂世界的简单法则》，张培译，中信出版社 2018 年版，第 22－24 页。

〔25〕 See Chengcheng Shao et al., “The Spread of Low-Credibility Content by Social Bots”, *Nature Communications*, Vol. 9, No. 1, 2018, p. 4789.

往难以在第一时间注意到新的文章或标签，也无法一天 24 小时在线，难以进行病毒式传播。在病毒式传播机制中，社交机器人作为社会传染的传染源，通过按需改变社交网络的动态结构，能够掌控目标受众接触传染源的渠道，并且比人类之间的社会传染具有更高的效率，形成了从社会传染到社会扩散的高效驱动机制。〔26〕 因此，当社交机器人被广泛运用之后，人们在网络空间中所看到的言论，大量来自于机器人。而且，社交机器人往往被用于传播那些低可信度的言论。〔27〕 对这些低可信度的言论作为“谣言”或“假新闻”的讨论已经较为充分，但是，不涉及真假的言论，同样也会造成相应的社会后果。

社交机器人的言论所造成的干扰，最直接的影响就是“社会流瀑效应”，即缺乏足够信息处理能力的人们，因为某些信息相对其他信息更高强度地出现，因而选择相信这些信息。〔28〕 这种社会流瀑效应不仅仅针对关于事实的新闻，也会涉及对观点的取舍，并无前见、未形成坚定立场的中立公众，因为某些观点似乎被更多人表达而作出选择接受这些观点。由于人类认知固有的局限性，“三人成虎”的社会流瀑也并非新鲜现象。但当机器人言论这一变量加入之后，这一现象发生了进一步的变化：机器人言论能够更有效率的增强人们接收到信息的强度，使人们更容易因为看到某些信息的反复出现而倾向于接受。人们过去是基于更多人表达出某种言论而被社会流瀑所影响，但当言论可能出自机器人时，人们更多看到的言论其实并非更普遍的言论，因而受到误导，进入到原本不存在的社会流瀑中。

在这样的社会流瀑效应之下，真正的人类观点可能被压制了。接收到大量机器人言论的网民，即使原本有自己的明确观点，但可能以为自己属于少数派，从而害怕多数派制造的社群压力因此避免表达观点，尤其是在一些存在激烈争议的言论场域中。“信息流瀑”之外，“声誉流瀑”的效果也同样得以强化。〔29〕 这样一来，沉默的大多数变得更为沉默，少数极端的观点反而变得最为强势。而在另一些情况下，即使并不是因为害怕，持有不同于机器人言论观点的个体可能因为频繁接触到对立观点而自我怀疑与反思，从而改变自己的立场。研究者发现，算法能够在社交网络中通过向特定目标人群进行信息推送改变其观点，成功地将一定比例人群的立场扭转到某一阈值之上，从而改变社交网络中两极分化的舆论强弱。〔30〕 通过压制或转变人们所持有的观点，网络社会中的共识也可能改变，甚至影响到社会规范的塑造。而通过表情包、动图、短视频等新型多媒体，信息的传播还可以更具有隐喻效果，由社交机器人创造出文

〔26〕 参见郑晨予等，见前注〔12〕，第 51－62 页。

〔27〕 See Shao et al., supra note 25, p. 4789.

〔28〕 参见左亦鲁，见前注〔3〕，第 547 页。

〔29〕 See Timur Kuran and Cass R. Sunstein, “Availability Cascades and Risk Regulation”, *Stanford Law Review*, Vol. 51, No. 4, 1999, pp. 683－768.

〔30〕 See D. Scott Hunter and Tauhid Zaman, “Optimizing Opinions with Stubborn Agents Under Time-Varying Dynamics”, MIT Workpaper, 2019, https://arxiv.org/abs/1806.11253, lasted visited on February 16 2022.

化传播的“模因”(meme)。〔31〕

社交机器人的言论,使某些信息更高强度地出现,即使并未导致人们接受某些事实或观点,也同样造成了强有力的影响。社交机器人的出现,其基本的经济逻辑就在于互联网的商业模式需要流量,需要注意力资源。为了获取注意力资源,社交网络中除了通过推荐算法让网民更容易看到自己感兴趣的内容,还要不断引导网民被流行趋势所吸引。而社交机器人则是流行趋势的有力建构者,可以在短时间内高效率地生成关于某一特定主题的大量言论,强有力地吸引注意力资源。“标签”在社交媒体上作为公共议题的建构机制,很大程度上影响着人们对公共议题的关注和参与。〔32〕当网民被搜索算法或推荐算法引导接触到某个新建的标签,并注意到这个标签之下大量的言论时,自然也就会更容易认为这是最需要关注的舆论热点,从而将自己的注意力投放进去。普通公众被社交机器人所制造的流量所吸引,从而自己又继续贡献真实的流量,在这样的流量经济运作过程中,社交网络中的议程设定也就由此形成。人们因为社交机器人的言论产生错觉,以为这些话题受到绝大多数人的关心,因而参与到对这些话题的讨论中去,从人类真实观点出发原本可能被关注或更值得被讨论的议题则被忽略了,人们按照社交机器人的幕后控制者所希望的议题去展开公共讨论。

三、思想市场的异化:社交机器人对言论自由的挑战

从对人类信息处理能力的干扰出发,社交机器人有力地强化了某些信息,同时也弱化了另一些信息。从客观事实的角度来说,这些信息可能都是真实的,并不能将其判定为谣言。但即使是真实的信息,通过社交机器人的表达,因为其相对人类言论具有大规模、高效率、全天候、自组织等一系列优势,从而挑战了人类的言论自由。真实的人类言论因为被忽略或隐藏而无法得到有效的传播,这种现象颠覆了思想自由市场这一理想形态。密尔(John Stuart Mill)指出,正是对自由言论和多样性意见的保护使真理得以浮现出来,并获得自我矫正功能,被压制的言论可能是真理,即使不是真理,也可能包括了真理的成分或者通过交锋而磨砺真理。〔33〕当人们普遍接受真理是相对的并且不断发展这一观念,也就自然而然接受了密尔对言论自由价值的这一论证,认为自由进行表达和讨论是接近与发展真理的重要保障。霍姆斯(Oliver Wendell Holmes)则进一步提出了著名的“思想的自由市场”的比喻。〔34〕习惯了市场经济的人们形象地意识到思想的自由竞争类似于实体商品市场上的自由竞争,市场的自由竞争最终会实现优胜劣汰,因此,思想的优劣最终是由消费者的行动作出决定而非由单一的权威事先确

〔31〕 See Nellie Bowles, “The Mainstreaming of Political Memes Online”, on *New York Times*, https://www.nytimes.com/interactive/2018/02/09/technology/political-memes-go-mainstream.html, last visited on 17 February 2022.

〔32〕 参见桑斯坦,见前注〔8〕,第 9—10、90、108—112 页。

〔33〕 参见(英)约翰·密尔:《论自由》,许宝骙译,商务印书馆 1959 年版,第 18—64 页。

〔34〕 See *Abrams v. United States*, 250 U.S. 616(1919).

定的。自由的思想市场可以与商品市场进行类比，两个市场之间没有本质性的差别。〔35〕而在网络时代，这个市场的“基础设施”已经发生了根本变化。〔36〕如果在这个市场上，竞争的对手并非普通人，而是大规模有组织的社交机器人时，这就成为了严重不对等的竞争。普通人的言论自由面对社交机器人，处于压倒性的劣势地位。尽管人类的思想市场从来也不是平等的自由竞争的理想状态，而是如同商品的自由市场的异化一样，更好的商品在异化的市场中难以“优胜”，而劣质者也未必会“劣汰”。〔37〕但社交机器人进入到这种竞争中来，就如同热兵器对冷兵器的代差，造成了言论生态的根本性改变。

社交机器人通过在网络空间中对言论生态的改变，进一步影响线下社会。人类社会中许多行为决策都依赖于社会互动，对他人会如何决策所做出的预测决定个人决策，而这种预测又来自于他人表达的言论所提供的信息。在这种情况下，社交机器人言论就可能导致对他人的错误预测，进而形成错误决策，通过自我实现的预言，对人类预期和决策产生影响。例如，在高度依赖于预期的金融市场当中，机器人言论使投资者认为某一家公司的股票确实得到了许多人的关注，从而大量交易该股票，而这就进一步验证了预言，使市场朝着机器人的控制者所期望的方向波动。〔38〕当然，类似的这种预期也会在其他市场中体现出来，例如市场营销中对社交机器人的使用，以机器人言论伪装为消费者对商品或服务的评价，消费者对品牌的认知也可能以这种方式被掌控，从而影响市场格局。〔39〕而娱乐市场则是这种操纵的更极端体现，无论是“黑”还是“粉”，机器人言论都在制造流量，而流量决定了市场资源的分配，虽然许多消费者并不认同流量明星以及他们的作品，但在这样的市场中却别无选择，离开了流量明星的作品可能根本就无法进入市场。

通过机器人对社交网络中言论的干扰，无论是线上还是线下，人们关于许多社会观念的认识都产生了偏差。即使社交机器人并没有创造出一种虚假的观点，而只是对社会中实际存在的观点进行表达，也会造成显著的误导。人们将机器人表达的观点当成了人类的观点，因此无法把握社会中真实的个体究竟持有什么样的观点，而是按照被扭曲的镜像去看待其他个体，错误地估计社会中与自己相同或相反立场的群体的规模，在这样一种预测的基础上调整自己的立场与表达。这样的偏差造成的最极端的后果，就是社会的分裂。伪装成人类言论的机器人言论，可能导致部分群体对对立观点形成更强的敌意从而紧密抱团，或是以为持有相同观点的群体人多势众而更加激进，导致不同派别的立场更加难以调和。互联网中原本固有的群体极化现象，在社交机器人的影响下日益突出。例如，性别议题固然在社会中始终存在，但其激进极化却是社交媒体上的新现象，当代中国社交媒体上的性别对立显然是在短时间内被明显的

〔35〕 See R. H. Coase, “The Market for Goods and the Market for Ideas”, *The American Economic Review*, Vol. 64, No. 2, 1974, pp. 384－391.

〔36〕 参见左亦鲁：《超越“街角发言者”：表达权的边缘与中心》，社会科学文献出版社 2021 年版，第 189－211 页。

〔37〕 参见李晟，见前注〔6〕。

〔38〕 See Ferrara et al., supra note 4, p. 97.

〔39〕 See Hwang et al., supra note 15, p. 44.

塑造出来,明显出现了社交机器人挑动性别对立的言论,并由于这种群体对立情绪进一步导致了“货拉拉跳车事件”这样的极端个案。在这样的言论生态中,群体极化的言论质量日益下降,言论自由的经典观念所试图实现的理性公共交流越来越难以实现。虽然操纵公众舆论其实在人类历史中始终存在,无论是口耳相传、印刷媒体、广播还是电视都可能传播大量错误信息,但人类的传播与通过机器人的自动传播有着根本的差异,基于人类行为模式所形成的言论自由理念,难以适应机器人言论的兴起。

网络时代中社交媒体的兴起,伴随的是传统媒体的衰落,导致许多传统媒体失去了对舆论的议程设置能力,而只能跟随网络议程设置寻找话题。在这种背景下,一种常见的现象就是传统媒体直接将社交媒体中的热点作为新闻或是公共意见进行报道和讨论,部分传统媒体的典型报道模式,就是某某事件在网络“引发争议”,但这种争议可能只是一种为了话题效应而刻意寻找的。如果不能正常识别和区分社交媒体热点从何而来的话,就会使社交机器人言论制造的错觉通过传统媒体又引入到线下社会中,再进一步影响到同互联网接触较少的群体对公共议题和意见的认知,从而让信息流瀑效应从线上波及线下,通过这样的运作过程,线上与线下的社会舆论其实都被社交机器人所引导。一种真实的观点以不真实的程度得到表现,同样也会造成负面影响。当这些言论进一步被作为“舆情”被政府机构加以认知并且回应时,政治后果也随之产生。如果公共决策的议程设置被社交机器人所创造的舆情所牵引,那么这种决策就偏离了社会的真实需求,原本无需关注的事件得到了重视,而更重要的问题则被忽略,导致公共资源的错误配置。

社交机器人所创造的舆情,不仅仅只是在议程设置这一层面干扰公共决策,还会在更具体的决策中体现出其后果。公共决策机构对舆情的观察,也考虑到社会中不同群体的利益诉求之间的均衡。如果某些群体在社会中占据着舆论的优势,其利益诉求也往往更容易得到回应。在传统的言论市场中,公共决策基于这样的逻辑是合理的,但如果考虑到社交机器人的介入,公共决策机构认识到的舆情并不反映社会真实,做出的回应也就会被误导。即使相关个案在社会中本身比较微小,但如果借助于互联网的社会扩散机制,则有可能从一个局部的个体事件产生全局性的宏观影响。

四、从自由到安全:社交机器人的深层挑战

由于上述影响,社交机器人的运用就很容易超出商业范畴,不仅仅被用于制造流量和吸引用户,还被用于政治目标。社交机器人传播的低可信度信息,很大一部分就是政治信息。〔40〕因此,社交机器人又发展出了“政治机器人”(political bots)这一特定分类,在社交机器人中占据了很大的比例。〔41〕政治机器人的运用机制与一般的社交机器人相同,但用途更为集中于

〔40〕 See Shao et al., supra note 25, p. 4789.

〔41〕 参见张洪忠、段泽宁、杨慧芸,“政治机器人在社交媒体空间的舆论干预分析”,《新闻界》2019 年第 9 期,第 17 页。

政治领域。政治机器人并不只发表政治言论，而是通过类似于人类用户的全面社交活动，形成政治影响。对政治机器人的关注始于2010年的美国中期选举。[42] 如同流量明星需要粉丝一样，政治竞选中的明星也同样需要，而政治机器人首先就起到了这样的功能，诸如罗姆尼(Willard Mitt Romney)、金里奇(Newt Gingrich)、奥巴马(Barack Hussein Obama)等人都被质疑通过政治机器人来制造虚假粉丝获得人气。[43] 在此之后，政治机器人的作用日益突出，2016年的美国总统大选使其更受到前所未有的关注，在选举中政治机器人的运用导致了基于恶意目标的影响力再分配，强化了社会中的群体极化，也助长了假新闻的流行。[44] 以政治和法律为目标的机器人成为了政治战略与通讯技术领域最流行的创新趋势，在社交媒体中广泛且活跃存在，政治机器人在美国的政治沟通中占据的角色虽然微小，但却具有战略意义。[45] 政治机器人作用于竞争性选举，不仅局限于美国，而是形成了更一般性的运作机制。[46]

在竞争性选举之外，政治机器人的影响也不容忽视，不仅通过虚假的人气或假新闻干预选举，更通过伪装的人类观点影响政治决策。同一般的社会话题相比，政治话题本身在社会中就具有更高的争议性，有分歧的公众更容易形成激进的对峙，也更容易形成抱团取暖的极化群体，因而体现出社交媒体更显著的影响。尽管其结果好坏不一，但社交媒体已经成为了许多政治运动的协调工具。[47] 越是极端和激进的观点更容易在社交媒体中吸引眼球，并且强化用户的情感参与、加强互动，达成群体极化。相对于“理中客”的意见，“喷子”的声音更容易在社交媒体中被放大。因此，政治机器人更有效的手段并非传递纯粹虚假的新闻，而是去强化某些观点的传播。政治机器人表达出来的观点可能本身是真实并合法的，但被放大了。当分裂的政治信息被强化时，既可能作为一种有组织的骚扰迫使人们自我审查，也可能作为煽动信息而引发更情绪化和更极端的意见。[48] 通过政治机器人的参与，某些政治意见被高强度地在互联网中展现出来，从而导致关于政治议题的网络舆论被扭曲，使公众与政治家都无法真实把握社会中的立场分化和不同立场的强弱比较。例如，在英国脱欧公投前后，推特上有一系列机器人高度活跃，之后又快速消失，通过机器人网络迅速生成了社会流瀑，其转发的并非严格意义

---

〔42〕 See Boshmaf et al., supra note 4, p. 1.

〔43〕 See Ferrara et al., supra note 4, p. 97.

〔44〕 See Alessandro Bessi and Emilio Ferrara, “Social Bots Distort the 2016 US Presidential Election Online Discussion”, *First Monday*, Vol. 21, No. 11, 2016, p. 11.

〔45〕 See Philip N. Howard, Samuel Woolley and Ryan Calo, “Algorithms, Bots, and Political Communication in the US 2016 Election: The Challenge of Automated Political Communication for Election Law and Administration”, *Journal of Information Technology & Politics*, Vol. 15, No. 2, 2018, p. 85.

〔46〕 See Emilio Ferrara, “Bots, Elections, and Social Media: A Brief Overview”, in Kai Shu et al. (ed.), *Disinformation, Misinformation, and Fake News in Social Media*, Cham: Springer Nature Switzerland AG, 2020, pp. 95－104.

〔47〕 See Clay Shirky, “ThePolitical Power of Social Media: Technology, the Public Sphere, and Political Change”, *Foreign affairs*, Vol. 90, No. 1, 2011, p. 30.

〔48〕 See Elizabeth Dubois and Fenwick McKelvey, “Political Bots: Disrupting Canada’s Democracy”, *Canadian Journal of Communication Policy Portal*, Vol. 44, No. 2, 2019, p. 29.

上的假新闻,而是倾向性非常明显的观点。〔49〕在互联网导致政治参与更加普遍化和扁平化的背景下,草根的力量似乎变得更为强大,但政治机器人的运用使得人们不知道支持者究竟真是草根还是机器人,无法真实观察什么是草根的声音、谁代表着草根的力量,因此导致了对民主的破坏。〔50〕在这样的运作过程中,政治团体通过将政治机器人打造成虚拟的意见领袖,与民众建立起更为稳固的社交关系,借助算法的机器手段来实现政治传播目的,"真人"意见领袖需要一定培育时间,不可复制而且其传播效果还会受到个体行为的影响,但"机器"意见领袖可以实现量产,而且其形象不会受到现实生活影响,更容易塑造完美的意见领袖形象。〔51〕

政治机器人在社交网络中的运用挑战了言论自由,也进一步构成了对民主政治的挑战。言论自由作为一种宪法权利,其正当性基础建立在民主自治的基础之上,即认定自治的公民必须通过充分的信息交流尽可能掌握充分的信息,才能够做出更明智的公共决策,从而实现公民自治。〔52〕思想市场的充分竞争,正是为民主决策提供了保障:"边缘领域要能够满足这种期望,非建制化公共交往网络必须使多多少少自发的意见形成过程成为可能。而这种能形成共鸣的、自主的公共领域,又取决于它植根于市民社会的社团之中、身处于自由主义的政治文化类型和社会化类型之中——一句话,取决于一种合理化的生活世界与其呼应。"〔53〕同时,对言论自由的保障也是为了保障公民之间的平等关系:"公共政治商谈必须拥有一种体面的论辩结构,如果我们希望将它作为一种保持歧见而互相尊重的伙伴之间的交流的话。"〔54〕由于社交网络中人类意见的表达和接受都被机器人强有力的干扰,这样的公共领域不再是哈贝马斯(Jürgen Habermas)与德沃金(Ronald Myles Dworkin)所设想的那种理想形态,而公共领域的这种转型,也意味着通过言论自由的充分交流协商达成公民基于伙伴关系的民主自治的理想日趋衰落。

政治机器人对网络舆论所实施的干扰,可以作用于各类型的政治决策与社会运动,可能对一个国家的政治共识与政治稳定产生重要影响。因此,这使其不局限于国内政治的场景,也能转向国际政治的用途。政治机器人的幕后控制者,不限于国内的政治力量,也涉及了国外的政治力量。例如政治机器人在 2016 年美国大选中的运用,就被认为有俄罗斯政府的介入。〔55〕而在 2017 年的海湾危机中,政治机器人背后浮现出了多个国家。2017 年 4 月,一个机器人网络在推特被建立起来,准备好了一场反卡塔尔的社交媒体运动,就在卡塔尔埃米尔发表言论之

〔49〕 See Marco T. Bastos and Dan Mercea, "The Brexit Botnet and User-Generated Hyperpartisan News", *Social Science Computer Review*, Vol. 37, No. 1, 2019, pp. 38－54.

〔50〕 See Dubois et al., supra note 48, p. 29.

〔51〕 参见张洪忠等,见前注〔41〕,第 23 页。

〔52〕 参见(美)亚历山大·米克尔约翰:《表达自由的法律限度》,侯健译,贵州人民出版社 2003 年版,第 17－20 页。

〔53〕 (德)哈贝马斯:《在事实与规范之间:关于法律和民主法治国的商谈理论》,童世骏译,生活·读书·新知三联书店 2003 年版,第 444 页。

〔54〕 (美)罗纳德·德沃金:《民主是可能的吗:新型政治辩论的诸原则》,鲁楠、王淇译,北京大学出版社 2014 年版,第 131 页。

〔55〕 See Howard et al., supra note 45, p. 90.

后，机器人发起的社交媒体运动立即爆发，其制造的标签和言论都与由沙特、埃及、阿联酋和巴林组成的“反卡塔尔四方”(the Quartet)提出的主张联系在一起。〔56〕美国学者认为，推特上服务于俄罗斯立场的账号中至少10%是机器人，拥有上千名雇员的俄罗斯互联网研究所(Internet Research Agency)掌控着这些机器人。〔57〕

随着政治机器人在社交媒体中的广泛运用跨越了国界，涉及国际政治，被定义为“计算政治宣传”，也就成为一个重要的国家安全问题。〔58〕国际政治视野下，运用政治机器人的计算政治宣传，其手段可能表现为对某一国家国内舆论的干扰和破坏。人工智能通过利用算法自主生成内容“子弹”(自动生成具有诱导性或欺骗性的内容)、实施个性化的“靶向”锁定(利用情感筛选锁定最易受到影响的受众)和密集的信息“轰炸”组合而成的“影响力机器”(the Influence Machine)来操纵他国国内的社会舆论；〔59〕也可能是面向本国或第三方的舆论场域，利用政治机器人对其他国家进行恶意攻击，从而造成对他国的污名化，挑动民粹情绪。例如在新冠病毒溯源以及涉疆涉港议题中，美国社交媒体上针对中国的机器人攻击也表现非常突出。〔60〕

面对政治机器人的攻击，许多国家都可能采取以彼之道还施彼身的对策。由于政治机器人脱胎于商业化的社交机器人，属于一种相对廉价高效的工具，因此并非只有大国才能使用，小国同样也可以借助于雇佣军对大国进行不对称战争形态的舆论干扰。例如叙利亚无疑在政治经济军事方面都是小国，但叙利亚内战中的双方却都充分运用了机器人构成的僵尸网络在推特上展开攻防。〔61〕政治机器人之间的攻防活动，不仅可以对其他国家施加国际舆论压力，也可以通过扰乱其他国家的国内舆论造成社会动荡，从而表现为信息战。信息战不是克劳塞维茨(Carl von Clausewitz)意义上的战争，也不是目前战争或武装冲突法所承认的任何意义上的战争；更确切地说，是敌对或对抗性的心理操纵，具有软实力的内涵(更恰当地说，是软实力和巧实力的结合)：宣传、说服、文化、社会力量、迷惑和欺骗。正如孙子所言：“不战而屈人之兵，善之善者也。”〔62〕通过宣传手段作用于公众心理，对敌方进行煽动或欺骗——信息战在人

〔56〕 See Marc Owen Jones, “Propaganda, Fake News, and Fake Trends: The Weaponization of Twitter Bots in the Gulf Crisis”, *International Journal of Communication*, Vol. 13, 2019, p. 1397.

〔57〕 See David M. Beskow and Kathleen M. Carley, “Characterization and Comparison of Russian and Chinese Disinformation Campaigns”, in Kai Shu et al.(ed.), *Disinformation, Misinformation, and Fake News in Social Media*, Cham: Springer Nature Switzerland AG, 2020, p. 78.

〔58〕 参见韩娜、虞文梁：“国家安全视域下的计算政治宣传：运行特征、风险识别与治理路径”，《公安学研究》2020年第6期，第4页。

〔59〕 参见阙天舒、张纪腾：“人工智能时代背景下的国家安全治理：应用范式、风险识别与路径选择”，《国际安全研究》2020年第1期，第13页。

〔60〕 参见韩娜等，见前注〔58〕，第2页。

〔61〕 See Abokhodair et al., supra note 18, pp. 839－851.

〔62〕 See Herbert Lin, “On the Organization of the U.S. Government for Responding to Adversarial Information Warfare and Influence Operations”, *I/S: A Journal of Law and Policy for the Information Society*, Vol. 15, No. 1－2, 2019, p. 2.

类战争史中自古有之,“用兵者,服战于民心”。〔63〕但只有随着技术的发展与网络的扩散,机器人参与的信息战才得以出现,具体来说,这得益于以下因素:现代信息技术和互联网提供了高连通性、低延迟、高度的匿名性、定制化的信息搜索、弱化了距离和国界、民主化的出版渠道、低成本生产和消费信息内容。〔64〕在全球范围内,出现了有相当规模的“社交机器人部队”参与国外政治运动,计算宣传驱动的全球信息战争,可能引爆新一轮的网络军备竞赛。〔65〕

从国际政治的视角来看,政治机器人不仅是某些政治利益集团的“雇员”,而可能是敌对国家的“军人”。机器人言论所引发的政治后果,就不只是思想市场如何保障民主政治质量的问题,而是敌对宣传扰乱国内舆论引发的国家安全问题。前者影响的是公民关系,而后者影响的则是敌我关系。出于这样两个维度的分析,更凸显了机器人言论的复杂性与对其进行规制的必要性。

五、中国语境的规制:总体国家安全观的视角

基于对社交机器人的分析可发现,即使其言论并不传播客观虚假的信息,但其在社交网络中造成的负面影响仍然不容忽视,尤其是政治机器人,如误导公共决策,扭曲民主过程,扰乱社会舆论,加强群体极化,甚至直接意味着一种新型战争手段发起的攻击。随着信息技术的不断发展,社交机器人不会被彻底消灭,回到之前的状态。社交媒体生态系统的未来可能已经指明了这样一个方向:机器与机器之间的互动成为常态,人类将走向一个主要由机器人组成的世界。〔66〕而且,由于机器人言论本身具有的合法特性,利用和人类用户同样的网络架构发挥作用,因而无法仅仅通过网络安全的技术手段加以处理。更好的网络安全有助于降低对手发动网络战攻击的能力,但却无助于降低对手发起信息战攻击的能力。〔67〕因此,需要形成的是对社交机器人的规制,在其仍然被使用的前提下尽可能削弱负面影响。

在美国语境中,更关注社交机器人尤其是政治机器人对国内民主政治特别是竞争性选举制度造成的威胁。这种倾向性可以从两方面加以解释。一方面,由于社交媒体上的技术和话语优势,美国在国际层面的政治机器人信息战中更多属于攻方而非守方,因此无需对来自于国外的政治机器人攻击做过多警惕。〔68〕另一方面,基于固有的价值观念,其更多强调社交媒体对公民社会和公共领域的支持,因而要求坚持将互联网自由作为总体目标而非实现短期政策

〔63〕 引自《韩非子·心度》。

〔64〕 See Lin, supra note 62, p. 11.

〔65〕 参见罗昕:“计算宣传:人工智能时代的公共舆论新形态”,《人民论坛·学术前沿》2020 年第 15 期,第 30－31 页。

〔66〕 See Ferrara et al., supra note 4, p. 102.

〔67〕 See Lin, supra note 62, p. 7.

〔68〕 虽然美国舆论常常指责其他国家对美国进行这种攻击,但更多只是以此为借口在社交媒体上对其他国家的账号封号,而这些实际上是人类账号。参见韩娜等,前注〔58〕,第 2 页。

的工具。〔69〕而基于这样的预设前提，政治机器人对民主的破坏自然受到高度的关注。因为在选举中需要高度的言论自由来充分激发竞争，只有能选举出更好的代表整体利益的代表，才能体现这种民主制的价值。因此，政治机器人的负面作用主要是竞选中形成僵尸粉的“伪草根”(AstroTurf)运动，以复杂的形式协调竞选战略并传递信息，吸引选民捐赠资金或投入注意力，但对其进行规制则要考虑到第一修正案保护作为政治言论的捐款，这限制了政府的规制强度。〔70〕甚至对机器人的检测以及封禁，都受到政治攻击，被认为是压制言论自由的审查制度。〔71〕除此之外，另一种视角关注言论自由的人权意义，从思想市场的经典理论出发，认为政治机器人的使用使思想市场中的某些人凌驾于其他人之上，因而侵犯到人权，因此要求从人权保护的层面通过限制机器人同人类用户的互动进行规制。〔72〕

从美国的视角出发，其规制目标是更好地实现言论自由以保障国内民主政治的有效运行，基于目标选择相应的规制手段。而从我国视角来看，我们面对的机器人言论问题与美国有着显著差异，因而不能简单套用美国的理论与制度。如果说美国更多是国内政治问题的话，我国则更多是国家安全问题。我国互联网与国外的网络之间具有一定的区隔，在全球社交媒体方面缺乏强势的平台，而国内社交媒体也形成了较大规模，因此面临境外的政治机器人在国外平台攻击和在国内平台渗透的双重压力。同时，我国宪制也有着根本差异，不存在两党制或多党制的竞选政治，也就没有国内明显分裂的政治派别和利益集团，因此出于国内政治动机而使用政治机器人的可能性也微乎其微。在这样的背景下，针对机器人言论，也就应当更多将其作为国家间信息战的一种形式来认识，从总体国家安全观的视角进行规制。随着信息技术的发展和网络与现实社会的深度互动，网络空间中的国家安全问题，已经从信息安全和数据安全进一步推进到算法安全，必须强化算法安全意识，避免将其视为一种技术中立的纯粹工具。〔73〕从国家安全的视角来看，一个更为简单的策略就是针锋相对的机器人对抗，问题即被转化为如何取得技术优势赢得对抗。美军网络司令部(U. S. Cyber Command)所形成的战略就是先发制人，以积极主动的进攻来代替防御，保持其优势以破坏敌方的行动自由。〔74〕尽管中国在许多敏感的政治议题中受到了来自政治机器人的攻击，但在国内外社交媒体平台上，几乎没有发现中国利用自动化宣传来影响话语流的证据，中国没有将自动化作为其国内或国际宣传战略努

〔69〕 See Shirky, supra note 47, p. 30.

〔70〕 See Howard et al., supra note 45, p. 86.

〔71〕 See Kai-Cheng Yang et al., “Arming the Public with Artificial Intelligence to Counter Social Bots”, *Human Behavior and Emerging Technologies*, Vol. 1, No. 1, 2019, p. 56.

〔72〕 See Nathalie Marechal, “When Bots Tweet: Toward a Normative Framework for Bots on Social Networking Sites”, *International Journal of Communication*, Vol. 10, 2016, p. 5024.

〔73〕 参见杨蓉：“从信息安全、数据安全到算法安全——总体国家安全观视角下的网络法律治理”，《法学评论》2021 年第 1 期，第 131－136 页。

〔74〕 See Max W.E. Smeets and Herbert Lin, “A Strategic Assessment of the U.S. Cyber Command Vision”, in Herbert Lin and Amy Zegart (ed.), *Bytes, Bombs, and Spies: The Strategic Dimensions of Offensive Cyber Operations*, Washington, D.C.: Brookings Institution Press, 2018, p. 85.

力的一部分。〔75〕 因此，如何对待社交机器人，就不能只考虑技术手段的对抗，而是需要思考如何建立规制体系，运用制度手段来保障国家安全与主权。

结合国内外视角来看，对机器人言论的规制，需要全球协同治理。〔76〕 但从主权国家之间信息战的层面理解，会发现难以取得治理的共识。例如，在美国看来，最好的结果就是制定规范禁止通过即时通讯与社交媒体干预民主政治。〔77〕 这种所谓的“民主”预设，其实也就意味着一种“只许州官放火，不许百姓点灯”式的治理。正如国际合作不能消除战争一样，对网络中的信息战，也不能只寄希望于通过国际合作进行治理。国际法可以在一定程度上规范战争，但这也并不表示国家不需要建立起能够避免战争或是赢得战争以保卫自身安全的制度。《塔林手册》(Tallinn Manual)这样的国际规则，也已经尝试对网络战争进行一定的规范，基本立场是现有国际法规范完全可以适用于网络战争，无需创制新的国际法规范以管辖网络行为。〔78〕 而机器人参与的信息战不是指向具体的物理或数据目标，尚未被规定在《塔林手册 2.0 版》的“网络攻击”定义之中，表明以此为代表的国际法所规制的网络武装冲突尚未纳入这一范畴。〔79〕 在这种背景下，对我国而言，尤其需要探讨如何以总体国家安全观为指导，通过加强国内的规制，保证网络空间的此类信息安全。

针对机器人言论的干扰，有学者提出了“算法素养”概念，寄希望于通过公众的批判性思维来对机器人言论进行区分识别，从而避免受到误导。〔80〕 这方面确实也是我国公众的薄弱所在，根据调查显示，中美两国网民群体对社交机器人的认知有显著差异，中国网民认为社交机器人存在的比例显著低于美国，并且更多持有正面认识，而非美国网民的负面认识。〔81〕 因此，如果从这一思路出发，规制的重点也就是加强对公众的教育。同时，个人也可以从言论自由权利出发对平台未能对机器人进行有效规制进行投诉。〔82〕 对那些较为明显可以发现的社交机器人，对其存在的认知的确有助于公众更好的理解社会舆论环境和群体分化。但这种情况相对较少，在道高一尺魔高一丈的机器人技术发展过程中，个人即使有这样的意识，也很难保证在参与网络交流的过程中总能辨别出机器人言论。因此，如果要通过算法素养解决机器人言论的问题，要求个人在网络中始终保持更高程度的注意，对

〔75〕 参见罗昕，见前注〔65〕，第 27 页。

〔76〕 参见罗昕，见前注〔65〕，第 31－32 页。

〔77〕 See Henry Farrell and Charles L. Glaser, “How Effects, Saliencies, and Norms Should Influence U.S. Cyberwar Doctrine”, in Herbert Lin and Amy Zegart (ed.), *Bytes, Bombs, and Spies: The Strategic Dimensions of Offensive Cyber Operations*, Washington, D. C.: Brookings Institution Press, 2018, p. 70.

〔78〕 参见陈颀：“网络安全、网络战争与国际法——从《塔林手册》切入”，《政治与法律》2014 年第 7 期，第 147－160 页。

〔79〕 参见张艳：“对《塔林手册 2.0 版》‘网络攻击’定义的解读”，《武大国际法评论》2019 年第 3 期，第 124－138 页。

〔80〕 参见罗昕，见前注〔65〕，第 34 页。

〔81〕 参见张洪忠等：“中美特定网民群体看待社交机器人的差异——基于技术接受视角的比较分析”，《西南民族大学学报(人文社会科学版)》2021 年第 5 期，第 161－163 页。

〔82〕 See Marechal, supra note 72, p. 5029.

个人成本过高。不愿意承担这种成本的个人,就难以在网上保持足够的警惕。这样的思路虽然有意义,但局限性也非常明显。

就"算法素养"这一概念而言,与其将识别机器人的成本配置给个人,配置给平台会更有效率。平台可以运用技术手段,结合人工审查,鉴别社交机器人账号,并视情况进行标签标注、限制发言频次、禁言乃至封号等措施。因此,计算机科学的有关研究大都致力于探讨更好检测社交机器人的技术手段。如果以技术对抗技术,让平台加强监管显然比起个人识别能更好遏制社交机器人的负面影响。但由于流量经济已成为平台的基本商业模式,而社交机器人的使用创造了流量,带来了数据资源,决定了平台的竞争力,因此平台对社交机器人的运用实际上持欢迎态度,缺乏足够的激励去加强监管。而且平台的股权结构通常也较为复杂,涉及到的境外资本也对此更加缺乏激励。而对监管平台的机构而言,难以证明平台有技术能力识别机器人却不去运用,因此难以提出要求。

既然要求个人或平台加强对机器人的识别都存在这样的局限,公权力的规制就更为重要。规制不能离开对社交机器人的幕后控制者的追踪,以技术手段"刺破机器人的面纱",通过直接作用于幕后控制者的惩罚措施对其进行约束。尤其是明显具有政治目的的政治机器人的使用,从总体国家安全观指引出发,应当将其视为一种信息战攻击,采取与此相对应的制裁手段。当公权力对社交机器人进行识别和追踪时,掌握数据的平台应当配合。

但对社交机器人的发现和追踪,本身都存在着技术上的制约,以这种手段作为对社交机器人的防御,还应当考虑互联网"易攻难守"的架构对网络安全形成的挑战,如果能够使之变得"易守难攻"则可以一定程度上改变安全态势。〔83〕 例如,加强对算法推荐的监管,虽然不直接针对社交机器人,但是机器人作用的发挥则离不开算法推荐,只有通过算法推荐才更容易被普通用户注意到。中央网信办制定的《互联网信息服务算法推荐管理规定》,其中一系列规定其实就已经指向了这一方面。加强对个人数据的保护,也能够在这一方面体现出其作用。对社交机器人而言,更有效率发挥影响力需要有针对性的算法推荐,尤其是针对本文所讨论的"意见",因为并不涉及客观信息的虚构,而是主观观点的表达,更需要进入有针对性的特定用户的社交网络中才能充分发挥作用,而这就需要同个人数据结合在一起。从这一点出发,也提出了强化对用户个人数据的保护的要求,对个人数据的保护并不完全出于个人权利,而是具有国家安全价值。从增加人类用户与社交机器人之间的隔离带这样的视角出发,需要突出保护的个人敏感信息和认证更紧密地结合在一起,不是一种本质性的"敏感",而是因为能够同个人认证相联系而变得敏感。〔84〕 由于境外的社交机器人活动有着更为突出的政治目的,尤其要避免的是个人数据被泄露于境外用于分析。数据在国家之间的跨境流动,也就成为一个突出的国家安全问题,需要将数据跨境流动作为一个国家主权规制的领域,按照主权国家的国家安全标准进行

〔83〕 参见左亦鲁:"国家安全视域下的网络安全——从攻守平衡的角度切入",《华东政法大学学报》2018年第1期,第148－157页。

〔84〕 参见胡凌:"功能视角下个人信息的公共性及其实现",《法制与社会发展》2021年第5期,第183页。

分类规制和审查。〔85〕此外,通过跨平台 cookie,能够结合用户在不同社交媒体的行为数据进行更精确的用户画像,从而使社交机器人能够对不同个人进行更具有针对性的信息操控。〔86〕因此,加强对跨平台 cookie 的限制也有利于强化对社交机器人的规制。

从总体国家安全观出发,还有一个重要视角,在于统筹网络空间与物理空间,降低网络空间言论操纵的现实影响力。在中国语境中,言论自由的重点不是美国式的"思想市场",而是更强调保障"言路畅通",以通畅的信息渠道实现决策者的"兼听则明",尤其是要避免中间层次的官僚过滤掉渠道中的信息。这实际上要求公共部门提升自身的算法素养,强化国家安全意识。相对个人用户而言,去识别机器人言论的成本更应当配置给公权力机构。这需要在一定程度上改变公权力机构尤其是政府的"舆情"观念。如果说在美国言论自由的教条构成了一种制约的话,中国语境中则更多要考虑重视舆论监督的教条。〔87〕过去一段时间中,由于公权力机构对网络的理解尚不够充分,对舆情的重视往往基于少数社交媒体平台,以流量和热度作为需要回应的标准,并将社交媒体意见作为公众意见形成回应对策。从总体国家安全观的视野出发,考虑到社交机器人特别是政治机器人的言论,就需要更审慎地进行舆情分析,结合技术和人工手段识别过滤机器人言论,将线上线下的社会舆论进行统筹考虑,以形成正确认知。尤其要避免基层部门各自为政,在缺少技术能力应对时,及时由更高层级的机关介入处理。在涉及公共领域议程设定时,要更审慎对待网络舆情,注重群众路线的传统。同时,对传统媒体的监管,要格外重视其基于社交平台热点问题进行报道,避免传统媒体将流量等同于热点并以其公信力为网络言论背书。

当然,在中国语境下讨论强化国家对社交机器人的规制,也不能忽略我国网络本身所具有的特性。在接入互联网时,由于 GFW 防火墙的存在,已经使用包括 DPI(deep packet inspection,深度包检测)技术在内的手段,对数据流量采取了控制和过滤。在这样的背景下,是否还应当主张国家的强有力干预?但值得注意的是,在讨论中国接入互联网时,不应理解为与一个抽象的"世界"接轨,更不能想象接入到的互联网是一个与主权国家无关的自生自发秩序,而是要看到互联网本身就是由冷战期间美国国防部建立的阿帕网(Arpanet)转型而来,美国政府通过其对根域名的最终权威显示了其主权。〔88〕因此,面对美国主导的网络,采取相应的防御性对策,也是维护国家主权的必要措施。基于人民主权自身建立在代表性基础上的正当性,需要回答的不是主权能否限制信息自由,而是何为契合特定国家人民需求的"限制边界"。〔89〕至于这种防御手段强

〔85〕 参见石静霞、张舵:"跨境数据流动规制的国家安全问题",《广西社会科学》2018 年第 8 期,第 128—133 页。

〔86〕 关于基于 cookie 的用户画像与个性化推荐的挑战,以及对其如何设计规制框架,参见丁晓东:"用户画像、个性化推荐与个人信息保护",《环球法律评论》2019 年第 5 期,第 82—96 页。

〔87〕 对传媒监督在当代中国存在的一些认识误区及其影响,参见陈柏峰:《传媒监督的法治》,法律出版社 2018 年版,第 25—31 页。

〔88〕 参见刘晗:"域名系统、网络主权与互联网治理:历史反思及其当代启示",《中外法学》2016 年第 2 期,第 518—535 页。

〔89〕 参见张新宝、许可:"网络空间主权的治理模式及其制度构建",《中国社会科学》2016 年第 8 期,第 144 页。

度应该如何设计，其对国内公民权利产生的影响应当如何平衡，并进而追问主权国家对言论的干预在政治哲学意义上的正当性，则是另一个主题之下需要回应的问题。但无论如何，在主权国家仍然作为国际秩序的基本组成单元、互联网并非超越国家主权而存在的虚拟空间的情境下，如果认识到机器人可以成为主权国家使用的信息战力量，就应当从国家安全的视角进行关注。

六、结　语

社交机器人并非独立存在的现象，而是同各类人工智能技术结合，立足于社交媒体的全网络覆盖和流量经济的基本市场结构发挥作用。而在目前的法学研究中，谈起“机器人”更多还是实体化的想象，却忽略了这种更真实的存在。然而，较之于讨论机器人的民事主体地位或是刑事责任能力，从国家安全的视角思考社交机器人对言论自由的影响，则是更重要的问题。不同于关注机器人传播虚假事实的研究，本文更进一步讨论了机器人言论中包含的不涉及客观真伪判断的主观意见所产生的影响，并指出从国家安全的视角来看，对这种形式上符合言论自由的言论应当加以规制。

在社交媒体深刻嵌入公众生活的背景下，社交机器人兴起，通过自动化程序掌控账号，模仿人类用户的行为模式，以发言、转发、点赞、评论、添加好友、发布状态等一系列行动共同构成“言论”。人工智能技术的发展，使社交机器人能够更好隐藏自己的身份，突破平台的过滤，也不容易被普通用户识别，从而使其言论在社交媒体中大量传播，形成人类用户与社交机器人共存的言论生态。

基于这种共存的背景，某些信息被机器人更多传播，某些信息则相应被淡化了，导致社交媒体中的社会流瀑、群体极化等效应更进一步得到强化。社交机器人对信息交流更强有力的影响能力，使人类用户被误导，因此使社交媒体成为一个严重不对等的思想市场而非自由市场。社交媒体的这种变化，通过作用于人们的心理和预期进一步产生对线下社会的影响，包括经济、文化、社会观念，而更重要的则是政治影响。机器人言论的政治影响，导致社交机器人发展出了政治机器人这一特定形态，其基于政治目的进入社交媒体，通过言论来形成政治影响。政治机器人通过干扰社会舆论，影响公共决策的议程安排，反映错误的社会群体分化，干扰竞争性选举，制造社会矛盾乃至群体对立。由于网络的全球性，这也成为一个国际政治问题，成为国家间信息战的一种形式。对我国而言，国际政治问题更为突出，对社交机器人的重视，关键点不是对言论自由和民主的损害，而是对国家安全与主权的挑战。而对此采取的规制手段，其正当性依据和目标首先也立足于国家安全视角。

因此，对社交机器人的规制，需要从总体国家安全观的视角来加以思考。从这一视角出发，不能只依赖于全球协同治理、平台自治和用户个人的算法素养来遏制社交机器人的负面影响，而是需要公权力更多发挥作用。公权力机构要通过技术手段追踪社交机器人幕后的控制者加以规制，也要加强对个人数据跨境流动的保护和对跨平台 cookie 的规制来改变信息战攻守态势，同时要求自身的算法素养提升，能够运用总体国家安全观来全面认识网络言论，增强对舆情的认知能力，结合对传统媒体的规制来统筹线上与线下治理。

总而言之，本文强调的是观察视角的更新。对言论自由的理解，需要动态地放在社会历史语境中，关注言论作为一种信息如何在社会中发挥作用，因此又如何受到信息的生产与传递的技术革命的影响。从人类的信息处理能力出发，不同的信息数量级也要求不同的制度。形成于农业社会的人类认知习惯，难以应对信息社会中的信息过载，这就为各种操控言论并进而影响社会的手段提供了前提。经典的言论自由理论，正如布兰代斯大法官(Louis Dembitz Brandeis)所指出的那样，认为可以通过更多言论而不是沉默来解决言论中的危险，但由于当下公众面对的信息的数量和速度都和此前的社会有了根本差异，因此已经很难以信息流动的多样性来解决信息扭曲，思想自由市场的隐喻在现代信息技术的环境中已经失效。〔90〕 如果仍然从言论自由的传统观念来理解被机器人重塑的言论生态，那么就无法有效回应新的挑战，可能会面临一系列负面的社会与政治后果，甚至可能是对国家安全的严重威胁。正视社交机器人的出现，重视机器人言论对言论自由的挑战，并且从总体国家安全观思考对策，才能更好应对人工智能影响下的互联网，整体上实现数字社会的治理。

Abstract: The development of Internet and artificial intelligence technology and the flow economy have led to the rise of social bots in social media. The rise of social bots may not only spread fake news, but also have a significant impact on the expression of opinions that do not involve "true / false" issues. The speech of social bots will interfere with the understanding of opinions on the Internet and in reality, and may have social consequences acting on the physical space, thus incurring the challenge to the freedom of speech. This challenge is more prominently reflected in the speech of political bots trying to influence politics, which has become a national security issue due to the "information war" in international politics. Therefore, in the era of artificial intelligence, speech of social bots in the social media should be a matter of concern and regulated under the guidance of the holistic approach to national security, based on which the theoretical understanding of freedom of speech should be updated.

Key Words: Social Media; Algorithms; Social Bots; Political Bots; Freedom of Speech

(责任编辑：章永乐)

〔90〕 See Lin, supra note 62, p. 39.