人工智能时代的意识形态风险及其化解

余 杰

【摘 要】人工智能技术深刻变革了人类的交往方式与信息的传播方式，在网络中展现出了一幅政治与技术深度互嵌、意识与万物泛在互联、真实与虚假杂糅互构、人类与机器双向互驯的媒介景观，并使意识形态风险呈现出隐蔽化、全域化、复杂化和动态化的特点。人工智能时代，意识形态面临着多重风险。为此，应以智能平权防范意识形态排斥风险，以共识凝聚规避意识形态撕裂风险，以群己共律应对意识形态操纵风险，以价值重塑消弭意识形态解构风险，从而实现对人工智能时代意识形态风险的化解。

【关 键 词】人工智能；意识形态安全；意识形态风险；风险治理

【作者简介】余杰，武汉大学马克思主义学院（湖北 武汉 430072）。

【原文出处】《思想理论教育》（沪），2022.12.84～89

【基金项目】国家社科基金青年项目“新时代中国共产党意识形态话语权的群众基础研究”（项目批准号：19CKS043）。

意识形态工作是党与国家的一项极端重要的工作，它“关乎旗帜、关乎道路、关乎国家政治安全”。[1]人工智能在给意识形态工作带来诸多机遇的同时，也造成了严峻的挑战，并使意识形态风险呈现出更为错综复杂的样貌。防范和化解人工智能时代的意识形态风险，成为意识形态工作的当务之急。基于此，深入揭示人工智能时代意识形态风险的表征与样态，探究人工智能时代意识形态风险的因应之道，对于抵御人工智能时代的意识形态风险，维护我国意识形态安全，具有重要的意义。

一、人工智能时代意识形态风险的表征

风险作为一种不确定的因素，具有潜在损益的可能，其“本质并不在于它正在发生，而在于它可能会发生”。[2]人工智能技术犹如一把“双刃剑”，既打开了生活的无限可能空间，又增大了应对意识形态风险的难度，使意识形态风险更趋隐蔽化、全域化、复杂化和动态化。

1.政治与技术的深度互嵌：意识形态风险隐蔽化

人工智能技术“政治无涉”的假象背后，政治与技术实则紧密勾连。一方面，政治镶嵌于人工智能技术研发过程中。人工智能技术是政治价值的承载，技术研发者为人工智能技术所选择的数据、所编写的代码，都是技术研发者的政治观念与价值取向的反映。这种观念源头式地自觉或不自觉植入与算法黑箱式的不透明运行，使技术研发者政治观念的渗透更加难以为人们所觉察。另一方面，政治镌刻于人工智能技术运行的赋权与推荐之中。人工智能时代，一些西方媒体利用人工智能技术查封发布支持中国言论的社交账号，限制发布支持中国言论账号的社交流量，提升反华言论推荐权重，将明面上的“言论自由”绘制成暗地里的“民主谎言”。这种基于人工智能技术的有针对性地信息推送，能够于潜移默化中形塑人们的政治认知，改变人们的政治倾向，让人们在难以觉察的同时疏于应对，对我国意识形态安全造成严重威胁。

2.意识与万物的泛在互联：意识形态风险全域化

人工智能时代，意识形态风险的发生地与发酵池从互联网延伸至物联网，人与物都可能成为意识

形态风险的传播源与扩散器，这使意识形态风险由人扩展到物，呈现出全域化的特征。随着移动通信技术的飞速发展，基于泛在网络的万物互联即将成为现实。通过泛在网络，人与人、人与物、物与物都被连接起来。“在未来信息生产系统的各个环节，参与主体将不仅是人，机器及万物都可能成为信息的采集者，而机器也可以完成信息的智能化加工。”[3]由“万众皆媒”到“万物皆媒”的演进，将有效提高信息采集与加工的效率，使信息变得更为易得与易传。但当越来越多的现实物品接入互联网，被现实物品包围的人类也易将自身置于越来越多的不可控的风险之中。当前，物联网网络安全防护仍处于起步阶段，安全防护能力较弱，许多智能家居设备尚不具备防火墙等安全防护功能。这使物联网相较于传统无线网而言，更易受到黑客的威胁与攻击。一旦物联网被入侵者破坏，人们身边的物品就将成为入侵者的工具，就存在发生信息泄漏、信息篡改与被价值渗透的可能，构成个人信息安全与国家意识形态安全的隐患。

3.真实与虚假的杂糅互构：意识形态风险复杂化

真实与虚假，是人们对事物是否与客观事实相符合的判断。但人工智能时代，一些事物真中有伪，伪中含真，真伪混杂，伪真共生，事物呈现出真实与虚假杂糅互构的状态。以“深度伪造”为例，“深度伪造”是一项基于人工智能的人体图像合成技术。“深度伪造”借助机器学习，能够替换一个视频中人物的面容、声音、口型、表情与动作，轻易实现视频“换脸”，快速合成肉眼无法分辨的假视频。“深度伪造”技术的诞生与普及，瓦解了公众对视觉文本的信任，使曾经“有图有真相”的时代一去而不复返，真相变得更加难以确证，社会信任危机日益凸显。而“深度伪造”技术在合成政治人物虚假视频方面的应用则易动摇公众的政治信任，煽动网民的非理性情绪，为网络舆情治理带来挑战。除此之外，“社交机器人”的存在亦严重扰乱了社交生态，它早已不仅是传播媒介，更成为传播中的对话者。“社交机器人”可以模仿人类在社交媒体上进行发布、评论、转发与关注等行为，并以此来操纵舆论与影响人类认知，使公众的思想为机器人及其背后之人所操控。人工智能时代，眼见不一定为实，构建事实的基础材料在逐渐失去效力；说者不一定为人，人际互动与人机互动在逐渐失去界限。这种真伪难辨的境况与别有用心的势力故意以假乱真的行径，令意识形态风险走向复杂化。

4.人类与机器的双向互驯：意识形态风险动态化

人工智能时代，是人类驯化机器，还是机器驯化人类?这是一个需要人类以一种高度警醒的态度来面对的问题。辩证而言，人类与机器的驯化关系从来都并非是单向的，而是双向交互作用的。尤其是人工智能技术赋予了机器以自主性与成长性，使人类与机器的互驯成为一个循环交替作用与即时交互作用共在的动态过程。人类对人工智能的驯化体现为人类将观念融入数据并植入算法。而人工智能则以其算法反过来影响人类观念，实现了人工智能对人类的驯化。深度学习方法的出现，使人工智能拥有了随着自主学习而不断改进自身的能力。人类与人工智能的每一次互动，都会使人工智能习得人类的观念和行为，亦使人类的观念与行为受到人工智能的影响。当人类意识到喂食人工智能的数据与编写的人工智能的算法将对人类观念与行为造成影响时，人类便以其观念与行为试图修正数据与算法，而修正后的数据与算法将再次影响人类的观念与行为，周而复始，使意识形态风险在不断的交互作用中呈现出演进发展的动态化特征。

二、人工智能时代意识形态风险的样态

面对社会智能化的阵痛，技术向善是当前所有有识之士的共同追求。但技术有时成为社会的“负催化剂”，意识形态的排斥、撕裂、操纵与解构成为人工智能时代意识形态风险的可能样态。

1.“智能鸿沟—数据偏见—算法歧视”的意识形态排斥风险

人工智能时代的意识形态排斥风险主要表现在三个维度：一是智能鸿沟。“智能鸿沟”是人工智能技术在应用过程中所导致的社会分化现象，是“数字鸿沟”在人工智能时代的新表现。一方面，“智能鸿沟”横亘于人与人之间。对于难以接触到人工智能技

术，更不知如何使用人工智能技术维护自身权益、赋能自身发展的人而言，他们更少地被数据化和代码化，甚至成为人工智能技术“看不见的人”，在人工智能时代被动缺席与噤声。另一方面，“智能鸿沟”还横亘于国与国之间。对于人工智能技术研发与应用能力弱的国家而言，面对来自“智能强国”的“信息殖民”与“价值入侵”，它们束手无策、无力抵御。二是数据偏见。数据是人工智能运行的基石，数据偏见既来源于“智能鸿沟”导致的数据代表性不足，亦来源于渗透人类偏见的对数据的不当标注。由于人工智能技术具有自我强化的倾向，一旦存在偏见的数据被用于机器学习，算法运行的结果也将带有偏见，人工智能将继承人类社会的原始偏见并不断循环扩散，使偏见被进一步复制与强化。三是算法歧视。算法是人工智能运行的灵魂，但看似客观中立的算法有时却饱含歧视，并且这种歧视与排斥将在算法隐蔽的循环运行中得到不断加强和固化，对社会发展产生消极影响。

2.“认知窄化—价值分化—群体极化”的意识形态撕裂风险

人工智能时代的意识形态撕裂风险摆脱了各种势力角力的狰狞，在技术的掩护下蒙上了温和的面纱，却不改其本质，遵循“认知窄化—价值分化—群体极化”的内在逻辑，层层嵌套，逐层深入，对价值共识凝聚工作提出了严峻挑战。首先，认知窄化是意识形态撕裂风险的序曲。人工智能时代的智能推送是一种“量身定制”的个性化推送，亦是一种“投其所好”的迎合性推送，一切不符合用户认知与偏好的观点都会被算法筛选后过滤，呈现在用户面前的都是符合用户认知与偏好的内容。这种成瘾机制下对用户刻意的迎合与讨好，容易造成“信息偏食”，使用户被困于大量同质化信息编织而成的“信息茧房”之中，难以接触到异质性观点，认知逐渐趋于偏狭与窄化。其次，价值分化是意识形态撕裂风险的主旋律。如果从个体的角度看，个体是被困于“信息茧房”之中，那么从个体与个体之间的角度看，“信息茧房”就成为一堵无形的价值观念的隔离墙，使不同的个体相互隔绝，弱化了个体与个体之间的交流，使每个个体都成为信息海洋中的一座“孤岛”，在信息区隔与观念封闭中人们彼此疏离，价值逐渐分化。最后，群体极化是意识形态撕裂风险的高潮。虽然个体成为“孤岛”，但无论是基于血缘、地缘、学缘、业缘、趣缘等因素，个体仍归属于一定群体，就如同“孤岛”并非处于真空中，周围仍有其他临近的“孤岛”。从群体与群体之间的角度看，人工智能时代价值分化后的群体内部聚合着高度同质化的用户，社群内部的一致性强而多样性弱，这就导致群体内部的自我纠偏能力匮乏，易使窄化后的片面认知不断强化。

3.“全景监狱—计算宣传—数字霸权”的意识形态操纵风险

在人工智能时代，人们的学习、工作、生活的便捷程度得到极大提高，但意识形态的操纵也在人工智能技术的辅助下进一步升级。第一，在“全景监狱”操纵之下人们沦为“数据囚徒”。人工智能时代，私人领域在逐步消失，每个人都宛若“数字囚徒”，戴上了数字痕迹的镣铐。第二，在“计算宣传”操纵之下人们沦为“算法靶子”。“计算宣传”是人工智能时代技术、资本与政治三者合谋的产物，“全景监狱”的存在为“计算宣传”提供了数据原料，提高了“计算宣传”的准确度。基于庞大的数据与智能的算法，“计算宣传”能够根据用户画像精准投放内容，对人们施加针对性的影响，进行有目的的引导，甚至利用“深度伪造”与“社交机器人”等手段编造谎言，以此来达到操纵公众舆论的目的。一些西方国家“通过高效隐蔽和灵活多变的信息分发对我国进行意识形态渗透和价值分化，规训受众价值，使其对主流价值产生厌恶和偏见，进而在潜移默化中削弱社会主义核心价值观的凝聚力和引领力”。[4]在此情境之下，人们沦为算法的“靶子”，自以为所拥有的自由不过是“被操纵的自由”。第三，在“数字霸权”操纵之下人们沦为“数字劳工”。“数字霸权”是人工智能时代意识形态操纵的本质彰显。“全景监狱”与“计算宣传”的出现本质上源于“数字霸权”的诞生，“全景监狱”与“计算宣传”的运行亦强化了“数字霸权”。一方面，“数字霸权”表现为侵害隐私与民主的政治控制，高度集中的权力缔造了“超级公司”与“超级国家”以及原子

化的个体。另一方面,“数字霸权”则表现为对数据等资源的垄断与对个体劳动的剥削。在“数字利维坦”面前,个体的数据成为资本贩卖的商品,人们沦为资本免费的劳工。

4.“调控失位—公共失序—人本失落”的意识形态解构风险

人工智能时代主流意识形态面临着调控失位、公共失序与人本失落的三重解构风险,三者彼此关联并层层叠加,对传统意识形态工作造成了极大的冲击。其一,调控失位。随着人工智能技术的快速发展与对社会影响的日益加深,治理权力的来源与格局发生深刻变革。“人工智能时代,得数据者得天下,控算法者控天下。相比财力雄厚、职能单一、目标明确的科技巨头公司,政府在数据采集、算法研发、人才储备、资金投入、技术创新等方面有所滞后。”[5]在人工智能时代,政府与企业之间的技术势差使权力结构的中心发生转移,政府不得不面对调控能力被削弱的境况。此外,“算法黑箱”所导致的责任主体模糊,亦使政府难以归责与问责,客观上增加了政府调控的难度。其二,公共失序。公共失序具体表现为:智能算法推荐下公共把关的失守,流量至上逻辑下公共道德的失范,娱乐话题充斥下公共议题的失焦,用户偏好导向下公共信息的失衡与信息茧房束缚下公共视野的失明。在公共失序的媒体环境中,价值逻辑被资本逻辑侵蚀,有意义的信息被有意思的信息取代,娱乐荒诞、低俗色情、血腥暴力、虚假猎奇的信息挤占屏幕,人们在自我的小天地中“孤芳自赏”“自娱自乐”,对公共议题“视而不见”或“固执己见”。其三,人本失落。调控失位与公共失序导致了社会主流价值的消解,使“以人为本”转向“以数据为本”或“以算法为本”。一方面,人本失落表现为人的主体地位的动摇。数据化是人工智能时代的一个重要趋势,但当人们习惯于以数据衡量一切、一切皆可量化时,人们就已经为数据所挟持,眼中只有冷冰冰的数据,而看不到活生生的人。当人被数据化、标签化,人的意义就交由数据度量。另一方面,人本失落表现为人的自主能力的削弱。人工智能时代,人们依赖算法,导致人的自主性被消解,人们受制、服从于算法,成为算法的俘虏。当工具理性盛行而价值理性式微,当机器变得越来越像人而人变得越来越像机器,人工智能就成为外在于人的异己的力量。

三、人工智能时代意识形态风险的化解

面对人工智能时代意识形态的风险,推进智能平权、强化共识凝聚、注重群己共律与重塑核心价值,是人工智能时代意识形态风险可能的化解路径。

1.以智能平权防范意识形态排斥风险

智能平权是减少数据偏见、破解算法歧视、弥合智能鸿沟的关键举措,以智能平权防范意识形态排斥风险,应从人工智能的技术研发、设备接入与能力培训三个维度入手。首先,聚焦人工智能的技术研发,要着眼于以下几方面:一是在研发团队方面,要努力做到团队成员所代表群体利益的多样化和学科背景的多元化,推动技术与人文的统一。二是在研发数据方面,由于数据正义是实现算法正义的前提,因而要努力做到数据采集的全面化、数据存储的加密化与数据标注的审查化,增强人工智能数据的代表性、安全性与正当性。三是在研发算法方面,由于算法正义是实现智能正义的关键,因而要努力做到算法决策的可解释化、算法过程的可溯源化与算法后果的可追责化,从而打开“算法黑箱”,使算法歧视在公开透明之下无处遁形。其次,拓展人工智能的设备接入,要注重倾斜“智能穷人”。当人工智能设备已经成为人们生活的必需品时,政府应使贫困人口拥有接触人工智能设备的可能,增强“智能穷人”在人工智能时代的参与度与能见度。再次,加强人工智能的能力培训,要侧重思维的转变与能力的提升。要通过培训,使“智能穷人”认识到人工智能的重要性并学会灵活运用人工智能,实现“智能脱贫”。此外,面对横亘于国与国之间的智能鸿沟,应增强网络空间命运共同体意识,加强人工智能意识形态风险的国际协同治理,携手防范人工智能时代的地缘政治风险。

2.以共识凝聚规避意识形态撕裂风险

在人工智能时代,促进社会共识凝聚,规避意识形态撕裂风险,要多措并举,方能走出“认知窄化—

价值分化—群体极化”的泥沼。第一,通过双管齐下突破认知窄化。认知窄化现象的形成既受制于智能算法,又受控于用户本身。因而,突破认知窄化,也需从这两方面入手。一方面,对于智能算法而言,在信息推送过程中要兼顾多样性。促进个性化固然是智能算法的效率密码,但兼顾多样性也是智能算法的责任担当。智能算法在研发之初就应植入“破茧”基因,延展算法模型的信息选择范围,丰富算法模型的信息挑选维度,并有意识地为用户推送与其认知不符但对凝聚共识起着重要作用的异质性观点,让“信息偏食”者“营养均衡”。另一方面,对于用户本身而言,在信息接收过程中要掌握主动。被动接收信息是智能算法推荐下用户的自然状态,而主动调试信息则是智能算法推荐下用户的应然选择。在现有算法模型下,用户可通过反向利用算法,主动跳出舒适区,通过有目的的信息检索,设计算法推荐内容,拓宽信息推荐维度,自主装修自己的“信息茧房”,让算法为己所用。第二,通过分众传播弥合价值分化。正所谓“解铃还须系铃人”,技术发展带来了风险,但其自身亦蕴藏着化解风险的钥匙,人工智能时代高精准度与高智能化的分众传播技术就是那把弥合价值分化的钥匙。人工智能时代的分众传播技术可以基于用户画像对群体进行科学分类,从而根据不同群体的特点精准施策,有针对性地策划、推送具有思想引导与价值引领作用的信息,通过量体裁衣、因材施教、对症下药,收获事半功倍的效果。除此之外,分众传播不仅可以用来消除人工智能技术带来的价值分化问题,更可以利用人工智能技术精准消除歧视等价值观念问题,为营造和谐社会氛围、凝聚社会共识提供助力。第三,通过疏导对话预防群体极化。从群体外部来看,预防群体极化,要加强对敏感议题、公众情绪、社会舆情的监测与预警,健全突发舆情处置机制,拓宽公众诉求反馈渠道,增加公众政治参与途径,通过对话协商形成理性共识。从群体内部来看,预防群体极化,要积极培育不同群体中的优质“意见领袖”,通过“意见领袖”加强对群体中非理性言论的引导与纠偏,以及对合理诉求的吸纳与上报。同时,对于群体中别有用心的谣言散布者与恶意煽动者,政府要及时惩处,防止他们成为“群体极化”的幕后推手。

3.以群己共律应对意识形态操纵风险

人工智能时代的意识形态操纵风险,是在多元主体的交互作用中生成的。因此,应对人工智能时代的意识形态操纵风险,应从不同主体入手,通过主体协同、群己共律,形成应对人工智能时代意识形态操纵风险的合力。首先,人工智能时代立法者要以“善法”推动“善治”。一方面,立法者要运用法律手段加强个人信息保护。尽管人工智能的发展需要庞大的数据作后盾,但立法者应处理好隐私保护与技术发展的矛盾,把握好二者的平衡。具体而言,企业获取用户数据应以用户的知情同意为前提,以对用户数据的分级保护与脱敏处理为基础,在合理限度内收集用户数据,禁止违法收集用户数据与滥用公民个人信息。另一方面,立法者要运用法律手段加强智能操纵治理。法律手段是规制智能操纵的最有力手段。可通过法律明确利用人工智能技术生成虚假新闻信息与恶意操纵网络舆论属于犯罪行为,以此来达到落实主体责任、震慑不法之徒的目的。其次,运营者要以“良心”铸就“良芯”。人工智能“良芯”的铸就,离不开运营者对良心的恪守。以良心铸就“良芯”,应从以下几方面着力:第一,加强行业自律,制定人工智能行业的自律公约,在行业内部建立自我净化机制,推动人工智能行业整体向上向善发展。第二,保障用户权利,向用户主动公开算法推荐相关规则与用户个人画像标签,为用户提供便捷的关闭算法推荐服务的选项与删改用户个人画像标签的权利,并自觉接受社会监督。第三,研发保护技术,通过研发信息加密技术、匿名处理技术与源头追溯技术,防止恶意收集信息,破解问责难题;同时,通过研发“深度伪造”识别技术与“社交机器人”检测技术,打击人工智能技术的不当应用,守护社会信任体系,维护国家意识形态安全。最后,使用者要以“自为”超越“自在”。从“自在”到“自为”的转变,根源于人工智能时代使用者的主体意识觉醒,具体表现在以下几方面:一是数据主体意识。使用者应意识到自身是作为数据主体的存在,从而在使用人工智能

的过程中，注意数据安全与隐私保护，谨慎为人工智能提供数据收集权限。二是媒介主体意识。使用者应意识到自身是作为媒介主体的存在，从而有意识地提升自身的媒介素养，增强媒介信息鉴别能力与媒介价值判断能力，深入了解“计算宣传”的手法与特征，练就辨识操纵的火眼金睛，从而真正享受不被操纵的自由。三是劳动主体意识。使用者应意识到自身是作为劳动主体的存在，认识到一般情况下作为“数字劳工”的“无酬劳动”的本质，从而合理使用人工智能。

4.以价值重塑消弭意识形态解构风险

综观人工智能时代调控失位、公共失序与人本失落的意识形态解构风险，价值重塑是消弭意识形态解构风险的有效路径。实现人工智能时代的价值重塑，应坚持三条重要原则：第一，坚持政府主导与企业驱动相互补的原则。人工智能作为引领未来发展的战略性技术，也是未来全球科技竞争的焦点。政府应加大对人工智能技术的研发投入，增强对人工智能技术的研发力度，建立政府主导、企业驱动的人工智能发展格局。首先，政府应致力于推动政府内部数据的共享，加速应用平台与政府数据的对接，实现政府对大数据的监管，构建由政府主导的大数据系统。其次，政府应积极整合各方力量投入算法研发，组建由政府信息部门主管，企业、高校和科研院所参与的算法研发机构，打破技术对资本的依附，确保算法的民生导向与公共属性。最后，政府应强化算法问责，督促企业提升算法透明度，并积极探索事前评估、事中监管、事后追责的全过程算法监管模式。第二，坚持主流价值与智能算法相融通的原则。助力主流价值与智能算法相融通，关键在于理解三个表达式：一是“算法推荐+人工推荐”。算法推荐与人工推荐在内容分发上有着各自的优势与短板，将算法推荐与人工推荐结合起来，有利于既发挥人工智能的推送效率优势，又发挥人的价值纠偏优势，在人机互补的基础上更好地净化内容生态。二是“算法推荐×主流价值”。主流价值与智能算法的融通不是简单叠加，而是深度融合。对于智能算法的研发而言，应在研发之初就将主流价值融入算法推荐的核心技术之中，为算法推荐植入主流价值的基因，提高主流价值权值在算法优先级中的比重，使智能算法成为传播主流价值的利器。三是“算法推荐÷不良信息”。智能算法不仅要成为主流价值的助推器，还要成为不良信息的过滤器。一方面，智能算法应通过对不良信息的智能识别、剔除与追踪，对发布不良信息的账号主体进行处罚，从根源上减少不良信息的传播；另一方面，智能算法不得将不良信息关键词记入用户画像标签，以使用户免受不良信息的二次侵蚀。第三，坚持以人为本与以物为用相统一的原则。人工智能的发展有助于将人从机械枯燥的重复劳动中解放出来，进一步释放人的创造力。但人工智能不是万能的神话，有着自身的应用范围与使用条件。唯有明晰这一点，人方能摆脱对人工智能的过度依赖，发挥主体作用，以价值理性驾驭工具理性，在以人为本与以物为用的统一中，使人工智能成为促进人自由而全面发展的有力手段。

参考文献：

[1]习近平关于总体国家安全观论述摘编[M].北京：中央文献出版社，2018：111.

[2][英]芭芭拉·亚当，乌尔里希·贝克，约斯特·房·龙.风险社会及其超越：社会理论的关键议题[M].译者：赵延东，马缨等.北京：北京出版社，2005：3.

[3]彭兰.新媒体用户研究：节点化、媒介化、赛博格化的人[M].北京：中国人民大学出版社，2020：10.

[4]张林.算法推荐时代凝聚价值共识的现实难题与策略选择[J].思想理论教育，2021(1).

[5]庞金友.人工智能与未来政治的可能样态[J].探索，2020(6).