圆桌

柯珍雅
萧强
亚历克斯·格莱德斯坦
季志强
刘爽

中国民主季刊
第 4 卷　第 4 期
2026 年 10 月
CC-BY-NC-ND
自由、免费转载分发，但须注明作者与出处，并不得修改及商用

人工智能时代的民主机遇

编按：人工智能对专制政权的影响是学者、观察人士高度关注的议题。专制政权积极运用人工智能技术巩固统治，例如利用人脸识别、智能监控等技术强化社会控制，[1]并借助大语言模型开展宣传、舆论引导和意识形态灌输。[2]因而，学者和观察家相继提出了“数字专制主义”（digital authoritarianism）和“数据中心专制主义”（data-centric authoritarianism）等概念，[3]以概括人工智能时代专制统治的新特征。但最近的研究开始揭示另一面：专制政权驾驭人工智能有其内在局限，而这项技术同样可以为民主力量所用。[4]

人工智能在为专制政权提供新的统治工具的同时，是否也给其带来了重要的冲击与挑战，为反对派、公民社会或民主力量创造了新的机会？从目前的发展趋势来看，这种冲击和挑战主要体现在哪些方面？

《中国民主季刊》邀请了柯珍雅、萧强、亚历克斯·格莱德斯坦、季志强和刘爽五位学者和专业人士，对此进行了讨论。



柯珍雅（Genia Kostka，柏林自由大学中国政治学教授）：人工智能正在改变政治：它扩大了政府处理信息、自动作出决策和干预社会的能力。相关讨论大多沿用一种人们熟悉的政体二分法：按照这种看法，民主政府运用人工智能来改善公共服务，专制政府则用它来监控公民、压制异见。这种区分抓住了一些重要差别，但已越来越不够用。民主政府和专制政府使用着许多相同的技术，既用来改进行政管理，也用来进行社会控制。真正关键的问题在于：政府开发的是哪些人工智能应用的组合，这些应用投入使用的范围有多广，以及国家是

否具备足够的监管能力和行政能力，来建立有效的保障机制、降低风险，并对公共和私营行为者进行问责。

政府通常借助四类相互交叠的人工智能能力。识别系统用于辨认人脸、车辆、语音、物体或行为。分析与预测系统对个案进行分类、发现异常，并预测风险或公共需求。语言模型处理文件、回答问题、翻译材料并生成文本。推荐与自动化系统确定个案处理的优先次序、分配资源，有时还直接执行决策。这些能力可以为公共行政提供支持，但也可以帮助政府追踪个人、管控边境、预判集体行动以及审查信息。

上述四种由人工智能赋能的国家能力，并非民主政体或专制政体所独有。从这些能力在美国和中国的应用可以看出，即便身处不同的制度环境、受到不同的约束，相似的能力也会在不同的政治体制中出现。在美国，国税局利用人工智能处理文件、识别可能存在的税务违规行为，并回答纳税人的问题。在中国，政府同样在各项公共服务中运用人工智能。地方政府设置了全天候的自助政务服务终端，杭州的“城市大脑”分析交通状况，深圳的法院则使用专门的语言模型来归纳证据、辅助法官审判。地方当局还在探索利用大数据和大语言模型识别腐败风险。北京的 12345 热线是又一个例子，说明数字技术可以通过便利民众反映诉求、提高行政回应性来增强国家能力。市民通过这条热线反映问题、申请服务，热线背后的数字基础设施则对他们的诉求进行分类，并分派给各个部门。它建立了一条由国家管理的渠道，用于收集民众意见、实现回应型治理，不过民众的参与仍集中于具体的服务诉求，而非政治诉求。[5] 这些例子说明，中国的地方政府正在利用人工智能及相关数字技术提升行政能力，并在各个政策领域扩大国家

的数字触达范围。[6] 由此可见，在这两个国家，政府都在把人工智能投入使用，以增强国家能力、提高行政效率，并改善公共服务的供给和对公民的回应，尽管实际成效参差不齐，也难以评估。

两国也都在使用人工智能赋能的监控，尽管其规模、政治目的和制度约束存在重大差异。在中国，人脸识别系统帮助当局识别个人、监控集体活动。政治动荡会促使地方政府增加对人脸识别技术的采购，从而把国家对监控的需求与人工智能企业的发展联系在一起。[7] 政府机构及其承包商还使用自动化工具扫描网上讨论、识别敏感内容。机器学习系统使平台能够大规模过滤政治言论。政治控制也影响着大语言模型生成的信息：与其他地方开发的模型相比，源自中国的模型拒绝回答政治敏感问题，或者给出更简短或更不准确回答的可能性要高得多；而训练数据中若包含国家控制的媒体内容，可能使模型的输出更偏向政府。[8] 在民主国家，基于人工智能的监控基础设施同样在扩张。在美国，地方警察局向 Flock Safety 公司购买了自动车牌识别摄像头，用途往往是一般的犯罪预防。联网访问使这些信息的流向远远超出了原本的本地用途。据报道，警方曾为联邦机构进行了数千次与移民相关的检索，一项试点计划还让国土安全部获得了直接访问权限。[9] 这个案例表明，一项为正当行政目的而采用的技术，可能变成一个范围大得多的监控体系的一部分。

传统的区分 —— 民主国家主要用人工智能来扩大公民参与、改善公共服务，而专制政权则用它来监控和控制——过于简单化了。在不同的政治体制中，同一政策领域的机构往往出于相似的目标采用人工智能，例如提高效率、改进决策、发现风险以及管理人口。此外，人工智能

系统的用途可以灵活转换：为提供服务而设计的数据基础设施，也可以便利对人口的监控；为安全目的而开发的技术，也能为行政规划提供信息。与此同时，依赖私营部门开发的人工智能系统，可能使权力向大型科技公司转移：政府所依赖的数据、基础设施和技术专长，正越来越多地掌握在这些公司手中。数据集中于少数企业，进一步加剧了这种权力不对称。最后，人工智能技术及其相关风险演变得如此之快，监管者往往难以跟上，从而在民主体制和专制体制中都造成了治理缺口。

在部署和治理人工智能时面临相似的挑战，并不意味着政体差异已经消失。各国在以下几个方面各不相同：服务型应用与强制型应用之间的平衡，人工智能系统的规模和整合程度，以及受影响的公民能够获得的法律保护和救济途径。每个国家都会形成一套独特的人工智能使用组合。塑造这一组合的因素有四个。政治优先事项和对威胁的认知，影响着政府把人工智能能力导向服务还是控制。行政能力决定了人工智能是停留在小规模试验阶段，还是在各部门之间整合运用。法律、法院、监督机构和申诉机制，影响着政府可以使用哪些应用，以及公民能否对有害的决定提出质疑。最后，国家与企业的关系，影响着政府获取哪些技术，以及由谁为这些技术造成的后果承担责任。

中国的各类社会信用体系本身虽然不是单一的人工智能系统，但它们的发展为人工智能治理提供了一个有益的启示：公众是否接受数据驱动的国家权力，可能取决于公民如何理解这种权力的运作方式及其后果。我在 2019 年发表的一项研究发现，公众对社会信用体系的认可度很高，其中较富裕、受教育程度较高和居住在城市的受访者支持尤为强烈。[10] 这种支持并不一定意味着人们在知情的前提下偏好监控。相

反，许多受访者似乎把社会信用视为弥补现有制度漏洞和监管缺口的一种手段。后来的一项实验研究表明，向受访者说明社会信用体系可能具有的压制作用，会显著降低他们的支持度；而强调它对维护社会秩序的贡献，则并未提高支持度。[11] 因此，数据驱动治理的合法性，可能部分建立在一种信息落差之上：这些系统对外的呈现方式，与公民对其潜在用途的理解之间存在落差。

缩小这种信息落差，并不会自动使人工智能治理或专制政治走向民主化。人工智能既能扩大人们获取知识的渠道，也同样容易制造错误信息、助长个性化的操纵。但它也能帮助公民获取和解读政治信息、记录滥权行为，并把官方说法与实际做法加以对照。如果人工智能系统如何收集数据、如何给人分类、如何共享信息变得透明，就有可能改变公众的态度，并为质疑与争论打开空间。

因此，人工智能的民主潜力，与其说在于技术本身，不如说在于围绕它的制度和社会实践。公民需要知道政府何时在使用人工智能，了解自动化系统如何影响自己，能够对有害的决定提出质疑，并围绕共同承受的后果组织起来。无论在哪种政体之下，核心的政治博弈都将围绕这样一个问题展开：谁能够看清、质疑并约束人工智能赋能下不断扩张的政府权力。



萧强（《中国数字时代》创办人、主编）：过去十年，我主要关注人工智能如何强化中国的数字专制：人脸识别、智能监控、算法驱动的舆论引导。这幅图景今天依然成立。但生成式大模型带来一个新问题：当

信息控制不再只是拦截文字，而必须进入模型本身，它会付出什么代价？这种代价，又能否被体制之外的力量转化为新的信息能力？

中国的网络审查长期建立在“拦截”的逻辑上：关键词过滤、删帖、封号。大模型却是生成式的。同一个问题有无数种问法，也有无数种回答。控制因此不能只停留在输出端，而必须进入训练、对齐、检索和生成机制，规定哪些关联可以形成，哪些事实不能抵达某些结论。

我把这种代价称为一种“智能税”。

大模型的能力来自海量语言、知识和经验之间的关系。历史、政治和社会事实并不是彼此隔绝的数据库。模型可以被要求在敏感问题上拒绝回答，但当限制不断深入，就可能出现某种认识上的断裂：同一事件在不同语言中的回答不一致，历史叙述在某个节点突然中止，已经承认的事实却不能继续进入因果推理。这并不意味着中国的大模型必然整体变笨。更精细的生成式人工智能同时改变了我们“看见”信息控制的方式。

专制的信息控制不仅决定人能不能说什么，也决定什么能留下，哪些事件彼此隔绝，哪些经验无法积累成共同记忆。一条被删掉的帖子只是一个瞬间；但如果十万条经过验证的删除记录与时间、事件、政策和官方叙事连接起来，长期的控制模式就可能显现出来。

当然，这里仍然只能说“可能”。模型能帮助我们识别真实结构，也可能把偶然拼接成貌似完整的故事。因此，大模型降低的并不是“获

得真相”的门槛，而是处理、比较和连接大量信息的门槛。

我认为，民主力量新的机会恰恰出现在这里。

国家与公民社会之间最根本的不对称之一，就是组织能力、计算能力和传播能力的不对称。生成式人工智能第一次显著降低了其中一部分能力的门槛。

第一个杠杆是记忆。海外中文媒体和公民社会几十年来保存的被删报道、帖子、文件和个人经验，可以从供人阅读的档案转化为机器可检索、可比对的知识资源。但档案必须保留出处、时间、版本和证据标准。否则进入机器的既可能是记忆，也可能是污染。

第二个杠杆是关联。大模型能帮助我们把不同时间、不同平台上的审查、宣传和社会事件放进同一分析框架。但关联不等于因果，模式也可能是模型制造的。正因为机器更善于寻找模式，人的核查和证据边界反而比过去更重要。

第三个杠杆是规模。小型媒体和公民组织今天可以借助大模型，完成过去需要更大团队才能承担的监测、整理和跨语言传播。但国家当然也获得了同样的能力：大规模生成宣传、测试叙事、制造噪音，甚至淹没公共空间。规模本身并不天然属于民主一方。

大模型带来的不是自动兑现的“民主技术窗口”。重要的是，我们能否把暂时降低的信息处理成本，转化为长期存在的知识基础设施：保

存原始材料，建立共同标准，抵抗信息污染，使机器的判断能回到证据，也使未来的人仍然知道，一件事情为什么被认为是真的。

裂缝不会自动变成杠杆。

真正的问题是：在裂缝仍然存在的时候，我们能否建立一些即使裂缝重新闭合，也不会随之消失的东西。



亚历克斯·格莱德斯坦（Alex Gladstein，“人权基金会”首席战略官）：公民抵抗运动通常由资源有限的小型团体发起和推动，他们往往悄然运作，难以获得视频制作、调查、研究或战略规划等领域顶尖人才的协助。人工智能代理（AI agents）正在改变这一局面。如今，一名普通活动人士就能调动过去唯有资金充裕的团队才具备的能力——创意制作、复杂研究、资助申请、筹款、大规模传播，以及把过去需要数日才能完成的工作实现自动化。由于专制政权本就掌握庞大资源，这一变化的影响是不对称的：对一个资源匮乏的组织而言，其意义远远大于对一个本就应有尽有的安全机构。我认为，人工智能代理如今已是人权捍卫者手中最重要的新工具之一。

具体到中国议题，情况可分为截然不同的两个层面。在中国境外——流亡社群、台湾与香港的活动人士、藏人流亡政府、海外维吾尔人——这一效果几乎可以肯定会显现。中国人权工作长期资金匮乏；我很难相信，除去人道主义工作不计，全球每年投入中国公民自由组织的资金总额会超过一千万美元。大多数组织的预算都在一两百万美元以

下，依靠一两笔来自官僚化大型基金会的拨款维持运作，仅是申请和汇报流程本身就要耗费大量时间。人工智能代理对这类组织而言正是直接的能力倍增器——而这类组织，正是当下大部分中国人权工作的落脚之处，因为在中国境内从事这项工作实在太过危险。而对于愿意掌握这些工具的活动人士——无论是运行本地 AI 模型，还是接入由保密计算（confidential computing）保障的推理服务——同样的赋能效果，正越来越多地惠及从事台湾、香港议题的人，以及为海外藏人和维吾尔人工作的人，不论他们各自的人权工作眼下还能在何处、以何种程度展开。

至于中国境内的情况，我确实不知道——我并非中国问题专家。但中共自己似乎相信，失控的人工智能可能带来重大冲击：几天前，中共通过其喉舌媒体表态，称人工智能构成的头号威胁，是对其政治秩序的冲击。[12] 这一点我深以为然——因为开放权重（open-weight）模型是任何人都能下载的。中国境内的大多数人不会主动去找这些模型，他们会继续像其他人一样，使用被打包、被审查过的访问渠道。但本地模型正变得越来越小、越来越便宜，也越来越难以被限制。一旦权重开放、模型装进了你的设备，你就可以对它进行定制——如果它被审查过，你也可以解除审查、进行“越狱”处理，想怎么改就怎么改。对于愿意走到这一步的人来说，这将是真正的变局，也是过去十年里，国家对数字空间近乎全面的控制所面临的第一个真正的技术挑战。

关于信任问题，还有一点更技术性、但值得一提。可信执行环境（Trusted Execution Environment，简称 TEE）的原理，[13] 是让数

据从你发出请求的那一刻起，一路保持加密状态，直到进入 GPU 芯片内部一块受硬件保护的“飞地”（enclave，芯片内部一块与外界隔离的安全区域）才会被解密、进行计算。也就是说，只要你用来跟模型对话的那套程序运行在一台安全的服务器上，“飞地”之外的任何人——包括托管这台服务器的云服务商——都看不到多少东西。哪怕是国家级的监控者，能截获到的也只是元数据（比如你什么时候发了请求、数据量多大，而不是你问了什么、模型答了什么），这和加密通讯软件 Signal 给用户的保障类似。但这解决的只是“你信不信任这家云服务商”的问题，跟“国家能不能在网络层面监控你的流量”完全是两个问题：后者根本不需要任何公司配合，是在网络基础设施这一层（比如互联网服务商、防火墙）直接进行的。这一块我们还需要更多研究和思考，目前我自己也没有完整的答案。

让我觉得真正奇怪的一点是：如果你今天想要这种 TEE 级别的保障，你几乎肯定要用一个中国模型。目前，只有中国的实验室在发布可在保密计算环境中投入使用的前沿级开放权重模型；而美国的前沿实验室则完全没有发布前沿级的开放权重模型，因为他们更在意的是抢占 AI 市场的垄断地位。现实情况是，大多数想要这种信任保障的人，最终都是在美国的云基础设施上运行中国的开放权重模型。这意味着，人权维权者最终要依靠中共自己的技术栈，来组织对抗中共本身——这是一个奇怪而令人不安的处境。我希望看到美国和欧洲的实验室在这个领域展开竞争，而不是拱手让出这块阵地。

需要说明的是，这一切并不意味着这些模型本身有什么危险。一个模型说到底就是一个从网上下载下来的文件；你自己定制它、自己运行

它，完全不需要接触任何中国的网络。真正奇怪的，不是使用这些模型安不安全，而是欧美的公司竟然没能在这个领域占据领先地位。

我还想指出一个研究空白: 本地运行的开放权重模型，在中国境内究竟能不能真正被封禁、或在硬件与软件层面被有效管控——也就是说，普通人手上的设备软硬件，能不能被管控到足以确保没有人在使用本地模型的程度——目前似乎没有人认真研究过这个问题。我确实不知道答案。我猜测应该做不到，但也不敢肯定；这会是你们的读者值得深入研究的、影响更为深远的问题之一，尽管对海外流亡群体而言，这个问题相对没那么重要——不论答案如何，本地化、注重隐私保护的 AI 工具都会大幅提升他们的能力。

最后我想强调的是：想不靠人工智能去对抗一个已经用人工智能武装起来的政权，这是不可能的。这就好比拿着打字机，去挑战一个已经电脑化的独裁者。中国政府的监控已经近乎天罗地网，而且持续了差不多十年，这场仗有多难打，大家心里都清楚，几乎不可能赢。但现在，我们手里多了一样近乎魔法的新技术，它能让一个人、一名普通的活动人士，能力大幅提升——相当于给每个人配了一整支后援队伍。不管你对 AI 抱着什么态度，这已经不是选择题了，我们只能拥抱它，不然就只会被甩在后面。



季志强（政治哲学博士、机器学习工程师、计算政治学研究者）：人工智能可能给民主化及民主所依赖的平等带来新的挑战。前沿人工智能模型的开发集中在少数公司手中，专制政府也在加强对这些工具的政治

管控。例如，中国在 2023 年 7 月公布的《生成式人工智能服务管理暂行办法》，就对面向境内公众的生成式服务提出了明确的政治内容要求。[14] 然而，开发能力的集中与政府的管控，并不意味着普通人无法借助这些工具扩展自身的能力。人工智能有助于言论自由的实现，在获取信息与表达传播这两个重要维度上都降低了门槛。

获取信息的自由，需要有实际寻找、理解和判断信息的能力。过去，核实一条消息往往要花很大力气查找出处、比较说法，碰到陌生领域还得补习相关知识。如今，人工智能可以帮助寻找线索、整理分歧、解释概念，人再循着线索核对原始材料。大模型学到的知识也不必然局限于某个政府允许传播的内容：敏感问题上的拒答，并不意味着相关知识不存在。当然，模型也会出错，甚至复述宣传或编造信息；它的价值在于辅助查证，而不是成为新的真理权威。跨语言翻译又使更多人能够接触外文报道、研究和不同社会的经验。有条件接触和比较不同来源的信息，普通人才更有可能突破政府试图强加于民众的单一叙事，形成独立判断。

表达与传播的能力同样在扩展。时政自媒体、政治活动人士、研究者和记者，已经在用人工智能制作海报和节目封面，为视频添加字幕、翻译内容。这些工作过去往往需要设计、后期和翻译人员协助，如今个人或小团队就能承担其中不少环节。2024 年 12 月，民主运动活动人士、组织者、社会科学家和技术专家在哈佛肯尼迪学院阿什民主治理与创新中心（Ash Center for Democratic Governance and Innovation）参加工作坊，讨论人工智能如何支持民主运动。其报告介绍了生成式工具帮助资源有限的团体制作和传播内容的潜力。[15] 人工智能降低了

内容制作与传播的门槛，让这些个人与小团队更容易把想表达的意见制作成能够分享、传播的作品，使他们的声音有更多机会触达听众、进入公共讨论。

这种能力还可以用来建立人与人的联系。不精通外语的人，也能借助人工智能与其他国家的行动者通信、交换经验，把本地发生的事情讲给更广泛的公众。民主理论家埃莱娜·兰德莫尔（Hélène Landemore）曾讨论人工智能在多语言参与者之间充当即时译者的可能。[16] 翻译与制作门槛的降低，也方便活动者、研究者和普通公民分享材料、共同开展工作。对于资源分散的民主力量，这种交流和协作的便利值得重视。

更深一层的机会，是借助人工智能理解监控与审查技术，探索意想不到的应对策略。专制政府能够利用大数据、机器学习和人工智能扩大监控、识别潜在异议并强化管控，但这些系统并非没有边界。美国《民主季刊》的一篇文章就建议，民主力量应当公开、反复地解释人工智能的能力与局限，为技术“去魅”。[17] 了解系统如何运作，不仅有助于辨认真实风险，也可能揭示新的应对空间。

举一个与政治无关、却颇有启发的例子。据 TechRadar 报道，一位音频工程师分析了苹果 2026 年全球开发者大会的演讲视频，发现音频中有些频段被删去了。这样处理让观众仍能听懂“Hey Siri”，设备却不再把它当作唤醒指令，从而减少误唤醒。[18] 这样的办法很难仅凭日常经验想到，只有了解系统处理信号的方式，才可能发现其中的空间。突破信息封锁、应对言论审查，也可能需要类似的技术想象力。过去，摸索这些路径需要相当的专业积累；现在，普通人可以借助人

工智能阅读技术资料，在持续提问和讨论中形成假设，再制作工具、快速试验，逐步摸清系统的能力边界，探索哪些限制可能绕开、哪些办法行不通。机会在于，更多人能够参与原本难以着手的技术探索。

这些能力不会自动转化为民主进展。不同地区的工具接入条件并不相同，便利不保证安全，联系也不能替代信任与组织。人工智能对民主的总体影响尚难判断，但这不妨碍辨认已经出现的机会：让本来就关心公共事务的人，更有条件获取信息、表达意见、相互合作并探索应对压制的办法。能力扩展不是民主的充分条件，却可以成为普通人参与民主实践的重要资源。



刘爽（独立时事评论人、AI 写作尝试者）：人工智能正在成为专制统治的新基建，这已无需赘言。令人玩味的是北京对这套系统的技术规训：按照官方标准，面对政治敏感问题，模型的拒答率不得低于 95%；但一旦触及“制度、信仰、形象、文化、英烈”等宏大叙事，拒答率又被卡死在 5% 以下——它不仅必须开口，还必须答得“政治合格”。

这种既要“绝对封口”又要“积极背书”的要求，像一把精密的尺子，把专制权力的内在焦虑强行焊进了模型的权重里。控制越是严丝合缝，体制自身的荒谬与漏洞，就越深地投射在代码之中。在机器的这种撕扯中，我看到了专制技术极权的三重内在悖论，以及在裂缝之间，民间社会尚存的腾挪空间。

1、机器深知它不可言说之物　一个模型要完美回避或篡改“六四”，

其前提恰恰是：它必须先精准地识别出“六四”。换言之，遗忘的先决条件是全知。

研究者在测试阿里的 Qwen3 时发现，它在涉法轮功或八九民运的问题上常给出伪造的回答，但偶尔又会吐露真相——这暴露出那些被勒令压制的记忆，其实早已沉淀在它的知识底座里。 2025 年初，DeepSeek 上演过更戏剧性的一幕：当被问及言论自由在中国是否正当时，它先在屏幕上敲出“在中国，主要威胁是国家本身”，句子还没打完，便像触电般整段撤回，尴尬地改口：“我们还是改聊数学、编程和逻辑题吧。”

过去的审查是在看不见的数据机房里悄悄完成的；而今天，大模型的审查却像一场被迫在用户眼皮底下发生的应激反应。更精巧的控制往往不是单纯的静默，而是一种被精心矫饰的谎言。但正因如此，我们对模型的每一次追问、试探与比对，实际上都在给这个政权的恐惧绘制一份精准的心电图。

2、若要机器聪明，就不能让它太听话 大模型的灵性来自对混乱、真实且庞杂的人类数据的吸收，而政治高压的第一步，就是把这些真实的数据污染、删改或碾碎。

学者 Eddie Yang 的实验揭示了一个致命的规律：政治压制越深，训练集里的有效信息就越匮乏，审查模型的真实表现就越钝化——尤其在突发危机时刻。一个最残酷的讽刺是：如果不给 AI 额外灌输特定的敏感词，大模型根本无法理解为什么街头有人“仅仅举起一张白纸”

就构成了一场抗议。网信办在整治专项中痛斥模型“审核过滤能力不足……纠偏能力不足”，实际上是在责怪物理规律。

一个既想在全球科技竞争中夺魁，又想把国民大脑装进套子里的政权，注定会陷入这种精神分裂：它想要一个能干的头脑，却只允许它长出顺从的神经。

3、当“全知”的神话被技术刺破 所有基于统计与预测的控制系统，本质上都在走钢丝：阈值放宽则漏网，收紧则殃及无辜。正如两位美国学者指出的直白真相：“只要是预测，就一定会犯错。”

过去，专制统治依赖一种“老大哥可能在暗处注视你”的神秘威慑。可一旦这种威慑全盘委托给算法，神话就极易褪色。2022 年郑州将一千多名储户的健康码粗暴变红，瞬间撕掉了“科学防疫”的遮羞布；同年冬天，在监控探头密度极高的各大城市，白纸运动依然连轴爆发。人们很快意识到：这套庞大的机器不仅有盲区，而且常常露出权力滥用时的愚蠢与狼狈。当神秘感被代码的漏洞打破，弥漫在社会上空的敬畏与顺从，便开始松动。

4、在数字裂隙中，重构民间社会 面对权力的巨兽，民间社会或许资源匮乏，但在技术的天平上，AI 绝非仅仅属于统治者。

在流亡的俄罗斯独立媒体中，已有超过九成将 AI 深度融入日常工作，用来处理海量的泄露文档、音频转写和多语种传播。“自由之家”的“中国异见监测”团队，正是依靠机器学习工具，每天在网信办的删

帖大网落下前，抢救式地抓取散落在社交媒体上的抗议火星，至今已记录了逾万起反抗事件。

专制机器越是依赖自动化，它在算法世界里留下的脚印就越深。今年年初 OpenAI 曝光的案例极具象征意义：一名中国执法人员竟然把 ChatGPT 当作办公备忘录，把针对海外异见人士的整套打压流程原原本本地“喂”给了大模型，反倒给外界留下了一份铁证。

当然，我们必须保持冷峻的现实感：按照现行法规，国内的 AI 服务商有义务对用户的“违规提问”留痕、上报。用国产机器人去试探敏感话题，无异于在警局门口大声密谋。在墙内的交互界面里，谨慎依然是第一法则。

5、结语 技术从来不会自动开出民主的花朵。它更像是一个无情的放大器：原本有纠错能力的制度，会被它赋予更强的修复力；而原本就扭曲失衡的系统，只会被它加速推向逻辑的死角。

但有一件事是明确的，也是值得所有不甘沉默的人去坚守的：未来的大模型，终究是从人类今天留下的全部字符中汲取养分的。审查者穷尽国力，要让机器遗忘；可机器生来就是要读取全部的真实。我们今天每多记录下一条被删的帖子，每在历史的暗夜里擦亮一个被抹杀的名字，就是在为明天的智能系统埋下一颗真相的种子。

要让它在未来开口，我们今天就不能放弃说出真相。

注释

1 萧强：《通往数字不自由之路：习近平的监控国家》（Xiao Qiang, "The Road to Digital Unfreedom: President Xi's Surveillance State," *Journal of Democracy*, Vol. 30, No. 1, January 2019, pp. 53-67.）

2 许成钢、王天成：《中国与民主资本主义体系：第四次产业革命中的制度竞争》，《中国民主季刊》2026 年第 1 期；区倬僖、Aiden Gu：《专制创新：中国如何将控制植入生成式人工智能系统》（Owen Au and Aiden Gu, *Authoritarian Innovation: How China Is Building Control into Generative AI Systems*, Resilience Innovation Lab, July 2026, https://resilienceinnovationlab.org/wp-content/uploads/2026/07/Authoritarian_Innovation_2026.pdf.）

3 《萧强谈中国的数字专制主义模式》（*"Xiao Qiang On China's Model Of Digital Authoritarianism,"* Hoover Institution, February 11, 2021, https://www.hoover.org/research/xiao-qiang-chinas-model-digital-authoritarianism-episode-2102.）；瓦伦丁·韦伯：《数据中心专制主义：中国前沿技术的发展可能如何使压制全球化》（Valentin Weber, *Data-Centric Authoritarianism: How China's Development of Frontier Technologies Could Globalize Repression*, National Endowment for Democracy, February 2025, https://www.ned.org/wp-content/uploads/2025/02/NED_FORUM-China-Emerging-Technologies-Report.pdf.）

4 L. 杰森·阿纳斯塔索普洛斯、连杰（音译）：《专制人工智能的局限》[L. Jason Anastasopoulos and Jie (Jason) Lian, "The Limits of Authoritarian AI," *Journal of Democracy*, Vol. 37, No. 2, April 2026, pp. 5-17, https://www.journalofdemocracy.org/articles/the-limits-of-authoritarian-ai/.]

5 萨布丽娜·哈比希 - 索比加拉、郑欢（音译）、弗兰齐斯卡·普吕默：《北京数字化市民诉求系统的时间治理与问责成本》（Sabrina Habich-Sobiegalla, Huan Zheng, and Franziska Plümmer, "Temporal Governance and Accountability Costs of Beijing's Digital Citizen Request System," *Regulation & Governance*, Vol. 20, No. 1, 2026, pp. 18-34, https://doi.org/10.1111/rego.12612.）

6 柯珍雅、安东·博格斯：《中国国家不断扩展的数字触达：中国的数字治理》（Genia Kostka and Anton Bogs, "The Expanding Digital Reach of the Chinese State: Digital Governance in China," *Regulation & Governance*, Vol. 20, No. 1, 2026, pp. 3-17, https://onlinelibrary.wiley.com/doi/10.1111/rego.70078.）

7 马丁·贝拉哈、安德鲁·考、大卫·Y·杨和诺姆·尤赫特曼：《人工智能专制统治》（Martin Beraja, Andrew Kao, David Y. Yang, and Noam Yuchtman, "AI-tocracy," *Quarterly Journal of Economics*, Vol. 138, No. 3, 2023, pp. 1349-1402.）

8 詹妮弗·潘、徐旭（音译）：《源自中国的大语言模型中的政治审查》（Jennifer Pan and Xu Xu, "Political Censorship in Large Language Models Originating from China," *PNAS Nexus*, Vol. 5, No. 2, 2026, pgag013, https://academic.oup.com/pnasnexus/article/5/2/pgag013/8487339.）；汉娜韦特等：《国家媒体控制影响大语言模型》（Hannah Waight, Eddie Yang, Yin Yuan, Solomon Messing, Margaret E. Roberts, Brandon M. Stewart, and Joshua A. Tucker, "State Media Control Influences Large Language Models," *Nature*, Vol. 655, No. 8123, 2026, pp. 685-693, https://doi.org/10.1038/s41586-026-10506-7.）

9 艾奥·多兹：《揭秘数据被提供给美国移民与海关执法局的人工智能警务科技公司》（Io Dodds, "Inside the AI Police Tech Firm Whose Data Is Being Fed to ICE," *The Independent*, January 19, 2026, https://www.independent.co.uk/news/world/americas/ice-ai-cameras-surveillance-flock-safety-b2903365.html.）

10 柯珍雅：《中国的社会信用体系与公众舆论：解释高认可度》（Genia Kostka, "China's Social Credit Systems and Public Opinion: Explaining High Levels of Approval," *New Media & Society*, Vol. 21, No. 7, 2019, pp. 1565-1593, https://doi.org/10.1177/1461444819826402.）

11 徐旭（音译）、柯珍雅、曹勋（音译）：《信息控制与中国公众对社会信用体系的支持》（Xu Xu, Genia Kostka, and Xun Cao, "Information Control and Public Support for Social Credit Systems in China," *Journal of Politics*, Vol. 84, No. 4, 2022, pp. 2230-2245, https://doi.org/10.1086/718358.）

12 陈一新：《全面筑牢人工智能安全屏障，推动人工智能健康有序发展》，《中国网信》2026 年第 8 期，https://mp.weixin.qq.com/s/jsc97cKVYOcHuI_WisvOxw。

13 可信执行环境（TEE）是芯片厂商设计的一种硬件安全技术，能在计算过程中把数据和运行环境隔离保护起来，即使云服务商想看，也看不到里面的内容——编者。

14 国家互联网信息办公室等七部门：《生成式人工智能服务管理暂行办法》，2023 年 7 月 13 日公布于中国网信网，第 2、4 条，https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm。

15 艾瑞卡·切诺维斯编：《人工智能如何支持民主运动：研究与实践工作坊总结报告》[Erica Chenoweth, ed., *How AI Can Support Democracy Movements: Summary Report of a Research and Practice Workshop*, Occasional Papers Series (Ash Center for Democratic Governance and Innovation, Harvard Kennedy School, February 2025), https://ash.harvard.edu/wp-content/uploads/2025/02/How-AI-Can-Support-Democracy-Movements-Final.pdf.]

16 埃莱娜·兰德莫尔：《人工智能能否将审议带给大众？》[Hélène

Landemore,"Can Artificial Intelligence Bring Deliberation to the Masses?"in *Conversations in Philosophy, Law, and Politics*, ed. Ruth Chang and Amia Srinivasan (Oxford University Press, 2024), p. 62, https://doi.org/10.1093/oso/9780198864523.003.0003.]

17 同注释 4。

18 亚历克斯·布莱克：《一位音频工程师解开苹果 Siri 之谜：频谱图揭示苹果如何在 2026 年全球开发者大会期间避免数百万台 iPhone 和 HomePod 被误唤醒》(Alex Blake,"An Apple Siri Mystery Has Been Solved by an Audio Engineer — Spectrogram Reveals How Apple Stopped Voice Assistant Setting Off Millions of iPhones and HomePods during Its WWDC 2026 Event," *TechRadar*, June 24, 2026, https://www.techradar.com/ai-platforms-assistants/an-apple-siri-mystery-has-been-solved-by-an-audio-engineer-spectrogram-reveals-how-apple-stopped-voice-assistant-setting-off-millions-of-iphones-and-homepods-during-its-wwdc-2026-event.)