公民抵抗

L. 杰森·阿纳斯塔索普洛斯、连杰 文

洪范 译

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

L·杰森·阿纳斯塔索普洛斯（L. Jason Anastasopoulos），是佐治亚大学公共行政与政策系的副教授，同时担任《公共行政评论》（Public Administration Review）的副主编

连杰（Jie Jason Lian），是哈佛大学“非暴力行动实验室”（Nonviolent Action Lab）的博士后研究员，以及哈佛大学“阿什民主治理与创新中心”（Ash Center for Democratic Governance and Innovation）的访问学者

译者洪范为中国大陆专业人士

专制主义人工智能的局限

人工智能的出现被认为给专制统治提供了一种异常强大的控制工具。L. 杰森·阿纳斯塔索普洛斯（L. Jason Anastasopoulos）、连杰（Jie Jason Lian）在这篇文章中提出，专制政权在运用人工智能时也存在内在的局限性。人工智能系统会迫使统治者面对一个无法回避的“校准困境”：降低判断阈值，可以减少漏报，抓住更多潜在挑战者，但也会增加误报，把更多无辜者卷入其中，从而引起反弹。这种结构性的不稳定会产生所谓的“阈值抽风”，即在不断收紧与突然放松之间循环摆动，中国的情况就是一个例子。民主力量可以通过消解对算法权力的神秘化、建立保护性规范，以及通过策略性抵抗挑战“全景监狱式监控”的假象，来利用这些脆弱性。

本文原载于美国《民主季刊》2026 年第 2 期（*Journal of Democracy*, Volume 37, Number 2, April 2026, pp. 5-17）。

人工智能常被称为“神奇利器”（silver bullet），它能帮助专制统治者以更低的成本、更快的速度和更精准的方式实施镇压。这种说法未必有错，至少在短期内如此。数字监控能够压制言论、识别反对者，并帮助政权扩大强权统治。但关于人工智能政治力量的最大论断——即实现预测性、先发制人的大规模控制——却面临着一个约束性的限制。

任何处理概率的系统都必须决定：将一个人、一种行为或一条信息标记为“风险”的阈值应设得高还是低。降低阈值（threshold），政权可以减少漏报（false negatives），抓住更多潜在挑战者，但也会增加误报（false positives），把更多无辜者卷入其中，从而引起反弹。提高“风险”的阈值，政权可以减少殃及无辜的镇压（collateral

repression），维护自身合法性。但这样也会漏掉真正的威胁，产生对手可以利用的盲区。简而言之，预测必然伴随错误，至少在某种程度上如此。唯一的选择在于，政权更愿意容忍哪一种错误。

我们把这个问题称为“独裁者的校准困境”（calibration dilemma）。这不是一个会随着下一次更新、更多数据或更好的工程技术而消失的“漏洞”。这是预测系统应用于政治现实时的一个结构性特征——尤其是在可信威胁稀少、公民不断适应监控、官僚和党的忠诚分子又有动机操纵指标的情况下。

在中国，这一困境尤为突出，因为监管与审查已经把“校准”变成了一项官僚事务。自 2022 年 3 月起，中国国家互联网信息办公室（网信办）要求那些推荐系统具有影响“舆论”或“社会动员”能力的主要平台，向国家算法备案系统提交关键信息和自我评估，并在系统发生变化时更新备案。[1] 2022 年 8 月，网信办公布了阿里巴巴、腾讯、字节跳动等公司的首批备案信息。针对“深度合成媒体”（deep-synthesis media）（如 AI 生成的视频和语音）以及生成式 AI 服务的类似规定，还增加了身份验证、内容治理要求，以及对政治敏感功能的人工审核。[2]

这样一来，政府的职责不仅限于观察结果，还包括监督人工智能的设置：重大更新可能会触发安全审查和行政审查。表面上看，这像是一种集中控制。但在实践中，这揭示了更深层的问题——即围绕政权能容忍多少错误，以及惩罚过多无辜者与漏掉过多反对者哪个更安全，始终存在着一场持续的博弈。

这场博弈牵涉到中国中央政权、地方官僚和公民。对地方官员的考核，依据的是中央下达的维稳目标。这些官员有强烈动机去杜绝“群体性事件”和上访，操纵合规指标。[3] 公民则通过戴口罩、使用暗语、从公开发帖转向私下聊天等方式来适应。如果管控范围过广而导致反弹，北京就会介入以重新确立控制权。它会发布新指示、集中权力，并展开危机公关（damage-control efforts）。其结果就是我们所说的“阈值抽风”（threshold whiplash）：围绕政治敏感时刻收紧，随后是突然放松、清理和宣传修复，如此循环。

2022 年末中国“清零”政策的迅速撤销，便是一个生动的例子。在近三年时间里，地方官员以维护公共卫生为由，严格执行严苛的封控及其他管控措施。2022 年 11 月，反对这些严苛限制的抗议活动在全国范围内爆发。在这些“白纸”示威活动中，人们举起空白标语牌，不写下任何被禁止的内容而巧妙地表达异议。这使得中共政权难以在不付出高昂代价的情况下加以镇压。

在经济和行政压力日益加剧的背景下，中国于 12 月 7 日发布了自疫情爆发以来最全面的放宽限制措施。[4] 此后，官方宣传的基调从“人民战争”转向了安抚民众和倡导自我防护，高层官员公开强调新冠最新变异毒株的严重性已“减弱”。[5] 然而，即便政策放宽，当局仍对部分参与抗议活动的人士实施拘留或施加压力，并压制有关这些抗议活动的公开讨论或纪念活动。简而言之，当局在行政层面放松管制的同时，却在政治层面收紧了控制。[6]

这种反复波动暴露了监控性政权的一个关键局限。人工智能无法解决

专制者的困境，因为这些系统的训练数据源自充满恐惧的民众，其所学习到的信号只能是隐晦的、被审查过的，且经过策略性操纵的。而人工智能所能做的，是通过把这种权衡取舍转移到模型、阈值、备案和安全审查之中，从而将不确定性变成官僚程序，在此过程中，这种不确定性变得无法回避、具有政治后果，且越来越明显。

独裁者的校准困境

这种不确定性的转移，迫使这个政权必须去面对统计误差带来的具体运作成本。尽管人工智能许诺精准，但 AI 赋能的镇压仍受制于预测错误，而这些错误会带来确定的政治惩罚。一方面是误报（I 类错误），系统将无辜者卷入其中：被人脸识别误认的邻居、被过滤器标记的无害帖子，或是被误判为可疑的正常活动。这些绝非单纯的统计上的烦扰。它们会累积民怨，使安全部门被噪音淹没，并让公民意识到，服从并不能保证安全。另一方面是漏报（II 类错误），即系统漏掉了真正的威胁。随着算法被调整以减少噪音，真正的挑战者会漏网，而（反抗）网络也学会了以模型无法察觉的方式进行协调。在这种情况下，即使反对派活动已转入地下，人工智能却会给专制政权带来一种令人心安却虚假的稳定感。

大规模自动化的三个特点使这一两难困境更加尖锐。第一是自主性权衡（autonomy trade-off）。赋予算法广泛的自主权，虽能最大限度地提高速度和覆盖范围，但也放大了哪怕是微小错误率所带来的后果。一个在全国范围内运行的模型，可能会产生数千次错误警报，从而将微小的统计误差变成重大的政治负担。而通过要求人工核查来限

制算法的自主性，虽然可以避免部分过度反应，但也会造成官僚程序上的瓶颈。警报不断堆积，分析人员不停地进行甄别，而当初让人工智能具有吸引力的速度和规模优势，也开始逐渐消失。

人工智能第二个出问题的特点在于算法的僵化（algorithmic rigidity）。模型是基于历史数据训练的，并假定昨天的模式能够预测明天的威胁。但在监控下，“真实情况”始终在变化。公民通过发明新暗语、改变路线、更换平台或将讨论转移到私密渠道等方式，来进行调整适应。结果就是模型迅速过时：模型上个月学到的东西，对今天的行为已难以提供有效指导。这迫使政权不得不进行持续的重新训练和重新校准，而这种灵活性正是大型监控官僚机构难以维持的。

放大这些风险的是第三个特点，即“稀有事件问题”（rare-events problem）。在任何社会中，活跃的异见者都是稀少的，仅占社会中极小的一部分。当算法试图搜寻他们时，即便是高度精准的系统，产生的误报数量也会远多于真实检测结果。如果一个监控模型的准确率达到 99%，但扫描的人口中每一万名公民中只有一人是真正的威胁，那么被标记出来的绝大多数人都是无辜的。其结果是一种危险的反馈循环（feedback loop）：安全机构被虚假“警报”淹没，因殃及无辜的镇压而导致政权的合法性遭到侵蚀，而真正的威胁则可能被埋没在海量的数据噪音之中。

这种两难的校准困境，根源于国家统治术（statecraft）长期以来的一种追求——使社会“变得可识别、可治理”的追求。现代国家长期以来一直力图把复杂性压缩为可以被统计、比较和治理的类别。19 世

纪，阿道夫·凯特勒（Adolphe Quetelet）曾承诺建立一种“社会物理学”（social physics），通过统计学将混乱的人口纳入行政管理。[7] 詹姆斯·斯科特（James C. Scott）后来警告说，当这种对社会清晰可控的野心与“高度现代主义”（high modernism）合流时，国家就会强加一套简化的社会和政治生活图景，这不可避免地会与地方性知识（local knowledge）和人类行为发生冲突。[8]

数字专制主义（digital authoritarianism），正是这种高度现代主义野心在算法时代的全面升级，其特征在于持续的量化评估、实时分类和自动化干预。然而，正如斯科特所指出的，地图终究不是疆域（the map is not the territory）。一个政权越是挖空心思刺探公民的“忠诚度”与“风险值”，并据此采取行动，这个政权就越会激发出相应的策略性调适。公民学会了如何避免被系统标记，而官员则修饰（tweak）统计数字以取悦上级。监控国家可能在量化数据变得更多的同时，实际掌握的信息却变得更少，用真实的认知换取了一种掌控一切的表象。在中国近期应对新冠疫情的官方措施中，这种权衡取舍体现得最为明显，也最为反复无常。

阈值抽风

中国以不同寻常的力度追求着斯科特所描述的高度现代主义野心，将算法工具嵌入日常治理和国内安全之中。在疫情“清零”时期使用的“健康码”系统，展示了基于阈值的管控在实践中是如何运作的。为了大规模管理公共卫生，当局将复杂的流行病学现实转化为简单的颜色编码状态，通常为绿色、黄色或红色。状态的变化可能

决定一个人能否出行、进入公共场所，乃至能否离开家。实际上，数字阈值（cutoff）已成为强制措施的触发点。而由于这一阈值过于僵化，系统极易引发剧烈波动。

尽管有阿里巴巴等主要平台的强大技术支持，“健康码”体系在社会层面仍显脆弱。两大失误引发了民众的不满。其一，高灵敏度的筛查产生了误报。人们在没有明确解释或可行申诉渠道的情况下，被标记为“红码”并遭隔离。[9] 其二，不透明性为政治操纵留下了空间，进而威胁到系统的合法性。最明显的例子发生在 2022 年河南村镇银行危机期间。[10] 据报道，面对腐败丑闻曝光后的银行挤兑，当地官员通过远程将储户的健康码转为红色，阻止他们外出抗议。这一事件暴露出，该系统并非中立的安全仲裁者，而是一个可为政治便利随意操控的开关。

随着时间的推移，人工智能的刚性控制没有带来稳定，反而走向了反面。任意限制和未解决的不满情绪不断积累，非但没有悄无声息地压制异议，反而激起了公众的愤怒，最后演变为公开抗议。2022 年 10 月，北京四通桥上出现的横幅标语，直接抨击监控国家和“独裁者”。次月，多座城市爆发了“白纸”抗议。除了对封控措施的愤怒之外，人们还表达了政治诉求。原本为识别个体而调校的监控体系，在面对由共同不满驱动的群体抗议时，却难以加以遏制。面对这种系统性失灵，中央政权采取了一种典型的“急转弯”策略（whiplash maneuver）。北京方面突然放松了关键限制措施，扭转了官方宣传口径，并着手进行“记忆抹除”（memory suppression）和对部分参与者的惩罚。

对抗监控的低技术手段，还包括香港抗议者使用雨伞、口罩和激光笔来干扰摄像头。在俄罗斯，抵抗既有技术形式也有法律形式。一些活动人士使用“迷彩”化妆来迷惑系统，另一些人则在被算法监控系统标记后寻求法律救济。[11] 活动人士尼古拉·格鲁欣（Nikolay Glukhin）在 2019 年的一次个人示威中被莫斯科地铁的监控系统识别，此后他将案件诉至欧洲人权法院（European Court of Human Rights, ECHR）。2023 年，他赢得了一项具有里程碑意义的裁决：缺乏严格法律保障的大规模人脸识别，构成对隐私权的侵犯。[12] 尽管俄罗斯已于 2022 年 9 月撤销了对欧洲人权法院的承认，但由于该案处理的是在撤销承认日期之前提出的投诉，因此仍具有可诉性（并为成员国确立了有效先例）。

在伊朗，伊斯兰政权针对 2022 年爆发的“妇女、生命、自由”（Women, Life, Freedom）抗议运动，安装了由人工智能驱动的街头监控摄像头，用于识别违反头巾着装规定的女性，这个规定要求女性完全遮盖头发和身体。早在 2026 年 1 月大规模起义爆发之前，这项技术就未能镇压公民抗命运动，因为女性仍在公共场所持续主张自己的权利。[13] 这些监控摄像头也未能阻止数百万人走上街头，在伊朗全部 31 个省份掀起自 1979 年以来规模最大的抗议浪潮。这个政权得以存续，并非依靠技术控制，而是通过极端暴力手段：实弹射击、外国什叶派民兵，以及可能超过三万人的死亡。在每一种情况下，人工智能都促使抵抗运动进行适应，却不能将其平息。

需要明确的是，在某些情况下，人工智能确实会强化专制统治。它能在受控的环境中识别已知个体，用可预测的线索来过滤内容，并对线索进行分级筛选，从而简化人工审核流程。就预测性、面向大规模人

口的警务而言，如果政府已锁定嫌疑人，人工智能系统能够确认其身份；但在缺乏对特定个体的可疑根据的情况下，若要求其从数百万普通公民中筛选并标记“高风险”人群，AI 系统的可靠性就要低得多。识别一个已有存档影像的人，与对那些可能毫无“前科”的个人做出大规模自动化判断，二者之间的差异不是程度上的，而是性质上的。这种差异，也正是当局在应对“校准困境”时最难处理的领域。

为何即使有了更好的 AI，这一困境仍会存在

这些失灵仅仅是成长的阵痛吗？未来的政权，若拥有更强大的算力和更复杂的模型，最终能解决校准问题吗？答案是否定的。准确率的提升所受到的限制，并非来自计算能力，而是复杂社会中预测本身的性质。即便是“更好的人工智能”，也无法消除这种权衡或取舍（tradeoff）；它只能改变权衡发生的位置。有三个因素确保了这一两难困境将持续存在。

第一，异议的稀有性意味着，从整体人口层面来看，大量误报在所难免。预测性警务系统无异于在草堆里找针。当目标行为的基础比率——即潜在发生率——极低时，即便是高度精准的模型，标记出的无辜者数量也会远多于真正的威胁。这种现象有时被称为“误报悖论”（false-positive paradox）。在政治语境中，误报会带来沉重代价，表现为殃及无辜的镇压及其引发的反弹。一个系统扫描了一百万人，其中只有十人是真正的威胁，如果准确率为 99%，虽能捕获大多数真正的威胁，但也会错误地将数千名无辜者标记为“威胁风险”。当局随后必须决定，是忽略这些警报（这意味着政府在花钱支付一个它并不使用

的系统），还是追查这些警报，从而耗尽稀缺资源，同时让公众产生敌对情绪。再强的算力也无法改变这种稀有事件的算术规律。

第二，应用环境并非一成不变。公民会做出调整适应。机器学习模型是基于“未来将与过去相似”这一假设构建的。它们从历史数据中学习，但在政治领域，这些“数据”会发生反作用。这正是“古德哈特定律”（Goodhart's Law）的一种体现，该定律由英国经济学家查尔斯·古德哈特（Charles Goodhart）在半个世纪前提出。他的结论是：当一项指标（measure）成为目标时，它就不再是一项好的指标了。[14] 就本文而言，这意味着一旦当局开始依赖某些特定信号，如关键词、行动轨迹或社交关系，人们便开始学会避免发出这些信号。他们使用暗语，将手机留在家中，或回到非数字化的交流方式（analog communication）。

政府或许仍能突破这些应对措施，但它们确实增加了监控的难度。在措施与反制的迭代对抗中，侦测与规避成为一场博弈。当局做出调整，人们便想法躲避或绕过；随后当局再次调整，人们又作出回应，如此循环往复。监控模型总是被“概念漂移”（concept drift）甩在身后：它试图预测的行为，已经因预测本身而发生了改变。

让这个困境持续存在的第三个也是最后一个因素，是专制主义的政治经济体系对输入数据（inputs）的扭曲。人工智能常被宣传为“客观真相的机器”（objective-truth machine），但在独裁体制下，数据本身就是政治资本。基层官员深知，他们必须拿出能向中央证明自己称职的数字。如果最高领导要求“稳定”，地方当局便可能压制警报

或少报事件。同样，上层对“警惕”的要求，可能会促使地方官员夸大威胁指标，以此展示忠诚和工作力度。向政府出售“黑箱”（black-box，意思是内部运作原理不透明——译者）监控技术的私营供应商，可能会进一步加剧这种扭曲，他们夸大其可靠性，淡化实际存在的不确定性边界，而这本来需要由人工判断。拥有一套自动化系统，并不能防止“垃圾进，垃圾出”（garbage in, garbage out）的问题，甚至可能使问题更加严重，因为它会让垃圾般的输出看起来像是硬性、权威的数据。看到这些数据，领导者会忍不住将“测量结果”误当成“真知灼见”，并自以为无所不知，而实际上远非如此。

几十年来困扰着专制政权的“信息认知难题”（knowledge problem），并没有被人工智能解决，只是被掩盖了。通过在策略性数据和不断变化的社会情境之上叠加概率推断，这些系统只能为本质上仍充满不确定性的现实，披上一层精确性的外衣。打造一个“完美”警察国家的梦想，不仅在政治上代价高昂，在数学上也根本无法实现。

虚张声势的“全景监狱”

如果这些制约确实存在，为什么人工智能看起来依然如此强大？诚然，人工智能是控制民众的一个有力工具，但这并不妨碍我们指出，出于感知和政治上的双重原因，它的能力被夸大了。首先，重要的是区分人工智能“实际能做什么”和“人们相信它能做什么”。十八世纪英国哲学家杰里米·边沁（Jeremy Bentham）曾提出“全景监狱”（panopticon）的构想，在这种圆形建筑中，一名看守可以随时观察任何囚犯，这个构想提示我们，持续的监视本身并不那么重要，重要

的是“人们相信自己随时可能被监视”这一信念。

当政府和科技公司将人工智能描绘成无所不知——拥有远超人类能力的“智能”——它们助长了一种寒蝉效应：公民们的自我审查和服从程度，远超出政权的实际监控能力所覆盖的范围，这并非因为他们正在被监视，而是因为他们认为自己可能被监视。技术乐观主义者承诺，人工智能将带来增长和互联互通的“数字红利”（digital dividend）。且不论这一承诺是否属实，但不可否认的是，专制政权已经从人工智能中赚到了一种“独裁者红利”（dictator’s dividend）——即由恐惧驱动的服从，其程度远超这些政权实际监控民众的能力。“算法”所带来的模糊而无处不在的威胁，促使公民进行自我审查，甚至无需执行一行代码，就为政权完成了它要做的工作。

但这种红利是以隐性债务为代价换来的。全景监视的威胁所产生的稳定是脆弱的，因为它依赖于经济学家蒂穆尔·库兰（Timur Kuran）所说的“偏好伪装”（preference falsification）。[15] 公民表面假装忠诚，私下却心怀不满，由此形成的对服从的评估，能够使统治者误判真实的不满程度。政权越是宣扬其无所不知，一旦这种无所不知被揭穿为片面、可被操纵或彻头彻尾的虚假，其付出的政治代价就越高。

这里同样存在一个动态的目标。随着异议人士和当局不断调整策略，对民众的控制也会随之变化。起初，政权往往占据上风。先进的监控系统成本高昂，需要高度的集中和协调，而这正是政府的强项。但随着时间推移，公民会了解这些系统的运作方式，制定应对措施，并更广泛地获得属于自己的数字工具。

这正是“全景监狱”的虚张声势可能崩塌之处。当公民目睹系统失灵——无论是河南“红色健康码”的随意性，还是德黑兰基于摄像头执法所显露的局限——这种神秘感便烟消云散。这个“全知之眼”（all-seeing eye）被揭穿为一台易犯错的机器，它可能被口罩蒙骗，也可能被腐败地方官员所操纵。关于“算法厌恶”（algorithm aversion）的研究表明，一旦人们观察到算法犯错，就可能不那么愿意听从它，从而加速这种虚张声势的瓦解。[16] 一旦公众普遍认识到，国家监控系统可能会因“噪声”挤占“信号”而陷入瘫痪，且其覆盖范围存在漏洞，恐惧便会迅速转化为愤怒。到那时，政权就只剩下一个在技术上不堪重负、在心理上丧失威慑力的监控机器。

民主派人士能做些什么

人工智能确实会重塑镇压方式，但它不会造就一个天衣无缝、无所不见的警察国家。这项技术改变了政权与其对手较量的场域。专制政府或许会在某些情境下占得优势，但亲民主力量同样可以调整并予以反击。以下三种策略值得关注。

第一，要公开并反复地对人工智能进行“祛魅”（demystify）。人工智能的心理威慑力往往超过其技术能力。清晰、易懂地解释这些系统能做什么、不能做什么，可以戳破政权刻意营造的“全景监狱”式氛围，削弱“算法无所不知”带来的寒蝉效应（chilling effects）。同样重要的是，民主国家和公民社会组织应当记录并公开专制政权所用人工智能系统的种种失灵：中国河南省任意标记红色健康码、错误的身份识别，以及其他可见的故障。揭穿“帷幕背后的巫师”（wizard

behind the curtain），揭示这台机器往往只是在猜测、存在偏见，或者干脆就是坏掉的，这能够减少恐惧，鼓励人们采取切实可行的应对措施。这类有针对性的叙事，通过证明 AI 赋能的控制既非天衣无缝、也非无懈可击，能鼓舞民主人士的士气。

专制政权从人工智能中获得了一种“独裁者红利”——即由恐惧驱动的服从，远超出政权监控民众的实际能力

第二，建立民主标准和防御能力。民主国家以及与之结盟的公民社会组织，应在政治敏感领域投入使用人工智能时，对许可的数据获取、透明度和问责制，推动制定更明确的规范，并在可行的情况下辅以监测、审计和检查。他们还应划出“红线”，以界定何时构成对人工智能的滥用。其中包括：缺乏保障措施的大规模生物识别、政治黑名单，以及旨在通过法律和制度渠道提高违规行为声誉成本的强制性评分或“社会信用”系统。若能诉诸法院，战略性诉讼或许会有所帮助（战略性诉讼是有意识地选择和设计的诉讼案件，以推动更广泛的法律、政策或制度变化——译者）。这方面的裁决可以对人脸识别和自动身份识别施加严格的保障措施，并通过将其重新定性为任意迫害，从而削弱算法治理的“科学”外衣。

鼓励技术民主化的举措，同样能够带来实质性的收益。这包括资助相关研究，探索如何使人工智能和其他安全的公民技术实现民主化，例如端到端加密、规避审查的工具（circumvention tools）和隐私保护协议。在太多地方，压迫性政权垄断了先进技术能力。这种局面应当改变。探索如何利用包括人工智能在内的先进技术，来促进而非压制民主，有助于公民社会对抗压迫、增强组织能力，并发挥这些工具所

蕴含的潜力。[17]

最后，AI 赋能的镇压行为应当被纳入现有人权执法的范围。现有的国际人权机制，往往依赖“可观察到的侵犯行为”（observable violations）来触发国际压力。然而，人工智能如今却被用于“预防性”镇压——在潜在挑战者尚未发展到需要对其采取明显镇压措施的程度之前，就对其进行识别、威慑或使其丧失行动能力。人工智能这种“事前”（before the fact）的特性，给传统的人权机制带来了挑战。权利侵犯和镇压的界定方式亟需重新审视，人工智能在压制异见中所扮演的角色，也需要进行主动审查。国际社会需要进行持续的、制度化的努力，来监测并报告 AI 赋能的胁迫行为。由数字监控和画像所助长的侵犯行为，可以被“点名羞辱”（named and shamed），并施加相应的制裁。关注先进微芯片的出口流向，也是一个好主意。

高度现代主义的幻梦

凭借其强大的能力，人工智能监控向独裁者们许下了高度现代主义的终极幻梦：一个可以像模拟程序一样被建模、被预测、被管理的社会。人工智能驱动的数字专制主义前景，在全球范围引发了民主人士的警觉。然而，正如中国、俄罗斯和伊朗的经验所表明的，这种对全人口规模进行预测性控制的梦想，在某种程度上不过是妄想。通过用僵化的数字阈值取代人类的判断，这些政权用灵活性换取了脆弱性。它们或许在识别个体目标方面变得更加得心应手，却在社会大变局爆发之前丧失了感知其动向的能力。

AI 赋能的镇压，其局限性既源于预测本身的统计特性，也源于经典的“独裁者困境”（dictator's dilemma）。由于模型具有概率性，而真正的威胁是稀少的、具有适应性，且往往被策略性地隐匿起来，因此并不存在用于预防性镇压的“完美”算法。不可避免的错误，尤其是误报，会带来政治代价，表现为资源的错配、民怨的累积，以及合法性的流失。

与此同时，通过强化官僚们钻制度空子（game the system）的动机，人工智能还可能会加剧“独裁者困境”。当地方官员因“不稳定”而受到惩罚时，他们就有强烈的动机在信息上报中央之前压制坏消息、夸大业绩、操控指标。于是，即使混乱加剧，中央领导层收到的指标却仍显示一切井然有序。一个政权越是依赖量化的风险评分来统治，就越接近于使用一张与实际情况不再相符的地图。

这正是自动化警察国家的核心脆弱性。它虽然能营造出精确的表象，却会削弱统治者需要的真实反馈。民主国家看似混乱——充满争论、抗议和显而易见的摩擦——但这种混乱往往能发挥信息的作用。它能暴露错误、反映不满，并开辟纠正的渠道。相比之下，专制人工智能则致力于压制噪音，转而追求那些可以被过滤、修饰乃至误读的校准信号。其结果看似井井有条，却使统治者陷入危险的隔绝状态，并容易过度自信。

对于民主的捍卫者而言，他们的任务不是因 AI 的能力而陷入停顿，而是要理解这些工具在具体情境中是如何运作的，并基于这种理解制定有效的应对措施

尽管在 AI 赋能的镇压之下，民主抵

抗仍存在空间，但要把人工智能固有的弱点转化为切实的民主成果，依然是一项重大挑战。一个核心障碍在于能力失衡——在压制性环境中，政权凭借其资源优势，往往对异见人士和公民活动者占据上风。就目前而言，最佳的策略应对措施仍是上文所述的那些：对人工智能的能力范围进行祛魅，推动制定可执行的关于 AI 政治应用的标准和规范，投资于公民（行动）安全技术，以及加强对 AI 赋能的胁迫行为的跨国监督。

这算是个不大的好消息。要了解坏消息，不妨看看以下三个令人担忧的趋势，它们可能会让 AI 赋能的专制主义更难以遏制。第一，人工智能的开发和投入使用，既未形成政治共识，也缺乏健全的治理和法律标准。作为人工智能领先的两大强国，中国和美国主要从相互竞争的国家安全视角来对待这项技术。两国均未将建立一个限制人工智能用于胁迫或镇压的全球框架作为优先事项。在欧洲，2024 年的《人工智能法案》（欧盟通过的一项在所有成员国均具有法律效力的法规），是一项值得注意的举措，旨在全面规范人工智能的开发和应用。即便如此，要实现有意义的约束，仍需要政府、企业与公民社会之间更广泛的跨国合作与共识。

第二，人工智能开发的领域依然是一个“小圈子”，大量权力集中在少数主体手中。在硬件方面，仅有少数几家公司——其中最突出的是英伟达（NVIDIA）——主导着先进芯片及其配套软件栈的生产。在模型方面，同样只有极少数公司生产并掌控着尖端的人工智能技术。只要这种高度集中的局面依然存在，推动人工智能民主化的努力就将举步维艰。例如，亲民主的公民社会组织，若想缩小独裁者目前享有的

技术优势，就需要获得更多资源。

第三，当前人工智能领域中“开源与闭源”的格局，给民主的支持者带来了一个现实的困境。许多广泛可用的开源模型，是由位于专制体制国家的企业开发的，而美国企业往往通过闭源平台提供性能最强的系统。这并不是说中国公司的开源模型——例如 DeepSeek 或通义千问（Qwen）——不能被用于民主目的。但当研发过程不透明——尤其是涉及所使用的训练数据、微调方式和内置的防护机制——审计这些系统中的偏见、隐藏限制或其他漏洞就会变得十分困难。而由科技巨头掌控的闭源模型，则会带来不同的风险。建立在这些平台之上的工具，容易受到价格或访问权限突然变化的影响。此外，迫使活动人士将数据发送至远程的企业服务器，也会让这些在压制性政权监视之下开展活动的人士陷入危险。

步入人工智能时代，既不意味着一场灾难，也不能保证民主的复兴。正如以往历次技术突破一样，人工智能在镇压方面的运用，将被既有的政治制度和权力关系所塑造，即使它也会带来新的挑战和机遇。对于民主的捍卫者而言，他们的任务不是因 AI 的能力而陷入停顿，而是要理解这些工具在具体情境中是如何运作的，并基于这种理解制定有效的应对措施，以保护公民空间、减少恐惧，并保持民主抗争的可能性始终开放。

注释

1 罗吉尔·克雷默斯、格雷厄姆韦伯斯特、海伦托纳:《译文〈互联网信息服务算法推荐管理规定〉——自 2022 年 3 月 1 日起施行》[Rogier Creemers, Graham Webster, and Helen Toner,"Translation: Internet Information Service Algorithmic Recommendation Management Provisions—Effective March 1, 2022," *DigiChina* (Stanford University), 10 January 2022, https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022.]

2 乔希·叶与布伦达·吴:《中国监管机构称阿里巴巴、腾讯已提交应用算法详情》[Josh Ye and Brenda Goh, "China Regulator Says Alibaba, Tencent Have Submitted App Algorithm Details," *Reuters*, 15 August 2022, *www.reuters.com/technology/china-regulator-says-alibaba-tencent-have-submitted-app-algorithm-details-2022-08-12*; "Provisions on the Administration of Deep Synthesis Internet Information Services," *China Law Translate*, 11 December 2022, www.chinalawtranslate.com/en/deep-synthesis.]

3 王裕华、明克胜:《中国维稳体系的崛起》[Yuhua Wang and Carl Minzner, "The Rise of the Chinese Security State," *China Quarterly* 222 (June 2015): 339–59.]

4 伯纳德·奥尔:《中国严格的"清零"政策开始松动》[Bernard Orr, "China's Rigid Zero-COVID-19 Policy Starts to Thaw," *Reuters*, 7 December 2022, www.reuters.com/world/china/chinas-rigid-zero-covid-19-policy-starts-thaw-2022-12-07.]

5 吴瑞恩:《在防疫政策转向之际,中国对民众的宣传口径从"战争"转为"自我防护"》[Ryan Woo, "In COVID U-Turn, China's Message to the People Shifts from War to Self-Care," *Reuters*, 14 December 2022, www.reuters.com/world/china/covid-u-turn-chinas-message-people-shifts-war-self-care-2022-12-14.]

6 裴敏欣:《清零政策的骤然终结:一项调查》[Minxin Pei, "The Sudden End of Zero-Covid: An Investigation," *China Leadership Monitor* 75 (Spring 2023), www.prcleader.org/post/the-sudden-end-of-zero-covid-an-investigation.]

7 凯文·唐纳利:《阿道夫·凯特勒、社会物理学与"平均人"的科学,1796—1874》[Kevin Donnelly, *Adolphe Quetelet, Social Physics and the Average Men of Science*, 1796–1874 (London: Routledge, 2015).]

8 詹姆斯·C. 斯科特:《国家的视角:那些试图改善人类状况的项目是如何失败的》[James C. Scott, *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed* (New Haven: Yale University

Press, 2020).]

9 陆柯言：《健康码背后隐忧：用户因中国移动错误定位被隔离》[Lu Keyan, "Hidden Worries Behind the Health Code: Users Quarantined Due to China Mobile Positioning Errors"[in Chinese], *Jiemian News*, 12 March 2020, https://finance.sina.cn/2020-03-13/detail-iimxyqvz9892557.d.html?vt=4.]

10 文森特·倪：《中国因银行账户被冻结引发的抗议演变为暴力冲突》[Vincent Ni, "Protest in China over Frozen Bank Accounts Ends in Violence," Guardian, 11 July 2022, www.theguardian.com/world/2022/jul/11/china-violent-clashes-at-protest-over-frozen-rural-bank-accounts.]

11《莫斯科活动人士以面部彩绘抗议大范围人脸识别》["Moscow Activists Protest Widespread Facial Recognition with Face Paint," *Moscow Times*, 7 February 2020, www.themoscowtimes.com/2020/02/07/moscow-activists-protest-widespread-facial-recognition-with-face-paint-a69205.]

12 约马尔特·约尔多舍夫：《格鲁欣诉俄罗斯案：欧洲人权法院迈入 AI 监控时代的第一步》[Jomart Joldoshev, "Glukhin v. Russia: The European Court of Human Rights' First Step into the Age of AI Surveillance," *Federal Bar Association*, 23 September 2025, www.fedbar.org/blog/glukhin-v-russia-the-european-court-of-human-rights-first-step-into-the-age-of-ai-surveillance.]

13 艾米丽·布劳特:《马赫萨·阿米尼去世三年后，伊朗对女性发动的高科技战争遭到抵抗》[Emily Blout,"Resisting Iran's High-Tech War on Women Three Years After Mahsa Amini's Death,"*Stimson Center,* 16 September 2025, www.stimson.org/2025/resisting-irans-high-tech-war-on-women-mahsa-amini.]

14 古德哈特对该定律的正式表述如下："每当政府试图依赖此前观察到的某种统计规律来实施控制时，这种规律性便会崩溃。"而该定律更为精炼、也更为著名的表述——"当一项指标变成目标时，它就不再是一个好指标"——则出自玛丽莲·斯特拉森。[Charles A.E. Goodhart,"Problems of Monetary Management: The U.K. Experience," *Papers in Monetary Economics*, vol. 1 (Sydney: Reserve Bank of Australia, 1975). Marilyn Strathern, "'Improving Ratings': *Audit in the British University System," European Review 5 (October 1997): 308*, https://archive.org/details/ImprovingRatingsAuditInTheBritishUniversitySystem/mode/2up.]

15 蒂穆尔·库兰：《私下的真相，公开的谎言：偏好伪装的社会后果》。[Timur Kuran, *Private Truths, Public Lies: The Social Consequences of Preference Falsification* (Cambridge: Harvard University Press, 1995).]

16 伯克利·J. 迪特沃斯特、约瑟夫·P. 西蒙斯、凯德·梅西：《算法厌恶：人们在目睹算法出错后，会错误地回避使用算法》[Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey, "Algorithm Aversion:

People Erroneously Avoid Algorithms After Seeing Them Err," *Journal of Experimental Psychology*: General 144 (February 2015): 114.]

17 埃莉卡·切诺维斯编:《人工智能如何支持民主运动: 一项研究与实践工作坊的总结报告》[Erica Chenoweth, ed.,"How AI Can Support Democracy Movements: Summary Report of a Research and Practice Workshop," *Ash Center for Democratic Governance and Innovation*, Harvard Kennedy School, 2025.]