

访谈

区倬僖
季志强

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

审查进了流水线： 看清中国人工智能 模型里的政治控制



区倬僖



季志强

编按：生成式人工智能是“一条很长的内容生成流水线。流水线拉长了，国家或当局这类外部行为者能够干预生成结果的环节也就多了”，区倬僖 (Owen Au) 说。北京正是沿着这条流水线，系统性地把它对内容和信息的政治要求转化为技术程序。区倬僖生于香港，现于英国从事独立中国事务研究与分析，为不同智库、咨询公司及非政府组织工作。2026 年 7 月，他与 Aiden Gu 合著发表报告《专制创新：中国如何将控制植入生成式人工智能系统》。受《中国民主季刊》委托，2026 年 10 月 3 日，计算政治学研究者季志强 (Eddy Ji) 就这份报告专访区倬僖，谈北京为何既倚重又提防生成式人工智能，政治要求如何借助技术标准写进数据、训练和测试各个环节，这套控制如何随中国模型扩散到海外，以及民主国家应当如何应对。

季志强 (Eddy Ji, 以下简称季)：感谢你接受我们的采访。今年 7 月，你与 Aiden Gu 发表了报告《专制创新：中国如何将控制植入生成式人工智能系统》(Authoritarian Innovation: How China Is Building Control into Generative AI Systems)。¹报告梳理了中国如何通过政

策、技术标准以及与合作企业的合作，在内容生成之前就把信息控制嵌入生成式人工智能的训练、评测和服务环节，并探讨了这种控制随中国模型扩散到海外可能带来的影响。这也是我们今天这次访谈的缘起。不过，在讨论这份报告之前，想先请你向读者介绍一下自己。你写过很多关于中国政治的文章，尤其关注中国的海洋战略。最初是什么让你对研究中国的政策问题产生了兴趣？这一路上有没有什么经历或转折点？

区倬僖 (Owen Au, 以下简称区)：好的，谢谢邀请。我现在是常驻英国的独立中国问题分析员，为多家智库、咨询公司和非政府组织工作。

我在香港出生长大。香港名义上是在中国主权下实行高度自治的城市。从十几岁起，我就亲眼目睹了中国的影响力不断扩大，姿态也越来越强硬。香港和中国大陆的这些政治变化，引发了香港一连串大规模抗议和政治风波，随之而来的是当局的镇压，其中包括 2020 年强行实施的《香港国安法》。我后来研究中国政治、成为中国问题分析员，都与这段经历有关。

离开香港、移居英国后，我开始从更宏观的层面观察中国。读硕士时，我主要研究中国在印太地区的地缘战略。印太这个战略舞台以海洋为重心，台湾海峡、南海和东海都在其中。中国可能对台采取的军事行动是当下战略讨论的热点，中国的海军能力和制海权正是其中的核心议题。我对中国海洋战略的研究就是这样开始的。

过去一年，全球科技竞争的分量越来越重，我的研究重点也逐渐从海

洋战略转向中国更广泛的科技战略，包括中国如何建设自己的技术能力，又如何谋求在全球科技治理和标准制定上的影响力。这也是我们今天要谈的话题。

季：那我们来谈谈这份报告。我们的读者对网络上的信息管控和言论审查并不陌生，但生成式人工智能似乎不太一样。对于不熟悉这项技术的读者，能否解释一下，北京一方面把生成式人工智能当作监控和宣传的工具，为什么另一方面又把它视为对信息管控的威胁？

区：生成式人工智能和中国的许多其他技术一样，非常强大，也具有颠覆性。中国已经在用这项技术加强多个领域的政治控制，比如网络监控和你提到的审查。许多研究和报告也指出，中国正积极利用这项技术在国内外开展信息行动，试图影响欧洲、美国和其他自由民主国家的信息空间。

不过，生成式人工智能在给中国政府带来这些好处的同时，也给北京带来了风险，对一个要把包括技术在内的方方面面都置于政治控制之下的政权来说尤其如此。与国内其他重大技术或战略性技术不同，生成式人工智能几乎人人都能用。即使是科技素养不高的普通中国人，也能在日常生活和工作中轻松使用它。在北京看来，这一点格外危险，因为这项技术可能被用来对当局发起非对称行动（力量较弱的一方以低成本手段对抗强大对手——编者），比如大量生成和传播与官方叙事相抵触的内容，或者用来了解和分析领导层的行为，而这些都是当局不愿看到的。

即使没有人蓄意对当局发起这类行动，生成式人工智能还有另一层风险。这项技术有两个特点尤其让当局担忧。第一个是人工智能的“黑箱”问题。许多中国人工智能研究者已经指出这类技术的局限。很多开发者和研究者至今仍难以完全解释模型生成了什么内容、是怎样生成的，更不用说加以控制。

第二个特点是，许多中国人工智能研究者也指出，如今的生成式人工智能和模型训练严重依赖西方数据。模型训练需要消耗大量数据，其中大部分是网络内容，而如今很多网络内容是用英语写成、来自西方国家的。中国当局往往认为，这些数据嵌入了自由民主国家的价值观和意识形态预设，因此把它们视为需要遏制和化解的国家安全风险。

所以，即便是中国本土研发的人工智能模型，也可能生成偏离北京官方叙事的内容。在中国，这类风险被正式归为“内容安全风险”，纳入国家的《人工智能安全治理框架》（该框架 1.0 版于 2024 年 9 月发布，2.0 版于 2025 年 9 月发布 -- 编者）。

季：你提到当局担心的风险之一，是训练数据大多来自西方，可能威胁国家的内容安全。由此我想到一个相关的问题。根据 Anthropic 最近的一份报告，一些中国人工智能实验室在完成预训练之后，仍在通过“蒸馏”技术²学习或借用 Anthropic 的 Claude、OpenAI 的 GPT 等国外前沿模型的能力，有些公司甚至直接把用户的请求转发给这些国外模型，再拿回答案。³中国实验室直接或间接地使用国外模型的回答，会不会在内容控制体系中打开一个缺口？

区：简单地说，我认为不会。因为保持政治控制、缓解令中国当局非常担忧的内容安全风险，是任何技术在中国得以投入使用和商业化的基本前提。所以，即便那些中国人工智能实验室正在从美国和其他西方国家大力汲取技术，并试图把这些技术用到自己的本土模型中，在面向公众之前，它们都必须经过一系列技术测试，以确保模型的回答与官方叙事保持一致——即便不是完全一致，也要大体一致。

季：你的报告让我认识到，政治审查早在模型给出回答之前就已介入，贯穿训练数据筛选、数据标注和模型测试等环节。我自己用几个中国模型做过测试，问过一些敏感问题，比如有哪些中国人获得过诺贝尔和平奖。其中一个模型已经开始作答，我能看到文字一个个蹦出来，但很快整段回答就被撤回了，换成一句“我们聊点别的吧”。这是用户能看到的部分。在这背后，政治要求是怎样一步步变成训练、标注和测试中的具体规则的？这些规则最终又怎样影响用户得到的回答？

区：你做的这种测试很有意思，许多研究者也经常这样测试这些模型的行为。中国的人工智能模型其实相当擅长拒绝回答敏感问题，让回答与中国共产党的立场保持一致。

要理解政治控制和言论审查在生成式人工智能中如何运作，首先要了解人工智能的内容是怎样生成的。过去谈网络内容审查，或者谈我们怎样在网上分享内容，事情相当简单。我作为内容创作者写好内容，发到社交媒体之类的网络平台上就行。生成式人工智能的内容生成过程却被拉长了，中间要经过数据收集、数据筛选和输入、模型训练、数据标注和处理等环节，最后才生成内容。这是一条很长的内容生成

流水线。

流水线拉长了，国家或当局这类外部行为者能够干预生成结果的环节也就多了。正是基于这种认识，中国自 2024 年以来花了很大力气，建立起一套庞大的人工智能技术标准体系，确保这条流水线每一个环节的技术标准都符合政治要求。

在 2024 年以来出台的众多标准中，我认为最值得注意的之一，是 2025 年发布的《生成式人工智能服务安全基本要求》。⁴ 它规定，所有中国人工智能模型在面向公众、投入商用之前，都必须通过一套基准测试集（基准测试是用一套统一的题目来评估模型表现的方法，类似于给模型做标准化考试 -- 编者），以确保内容安全。这套测试集共有 2000 道题，其中三分之一专门测试生成内容是否违反中国共产党所谓的“社会主义核心价值观”。⁵ 要通过测试，模型对可能违反内容安全的提示词，拒答率必须至少达到 95%。

与此同时，这一框架又从内容生成这一层扩展到其他技术环节，比如数据标注、机器学习和算法训练。⁶ 这清楚地表明，国家正在整个人工智能技术栈中，⁷ 系统性地把它对内容和信息的政治要求转化为技术程序。

季：接下来想谈谈中国模型走出国门之后的情况。按你的报告，中国政府对模型生成内容的控制，会随着海外用户采用中国模型一起传到国外，背后并不需要什么有组织的影响力行动。中国技术以前也走出去过，比如华为在巴基斯坦的“平安城市”项目，又比如因内置政治

偏向、压制敏感内容而屡遭质疑的 TikTok。按照政治要求训练和测试出来的中国人工智能模型，在海外产生影响的方式与这些先例有哪些异同？

区：我认为，中国人工智能模型和技术的扩散，与其他中国技术和产业涌向全球的模式相当相似。令人意外的是，背后的逻辑相当资本主义、相当商业化。太阳能电池板、可再生能源、电信等战略性产业受到中国国家重点扶持，多年来享受大量补贴，又是在“内卷”的环境中发展起来的。它们的产品和服务因此非常便宜，质量又足以让海外用户接受，在全球市场上极具竞争力。

人工智能技术也是如此。这些技术未必最尖端，中国最先进的前沿模型目前仍落后于美国。但它们的能力已经足够好，又是开放权重的，⁸考虑到价格就更显得划算，因此在全球很有吸引力，在全球南方国家尤其如此。这些国家大规模采用中国廉价又好用的产品和基础设施，也就更加依赖中国的科技产业和标准。随着中国积极把政治要求嵌入生成式人工智能技术，这些要求的影响就不只限于国内，还会波及依赖中国科技产业的国家。

季：我还有一个问题，可能超出了你的研究范围。有没有证据表明，中国企业或政府机构正在帮助其他专制政府建立自己的人工智能审查系统，比如提供内容过滤工具、训练流程或技术支持？

区：生成式人工智能方面倒没有特别的例子。不过几十年来，尤其是近些年，中国一直向其他专制国家输出技术，助长海外的专制统治，

这方面已有大量记录，也广为人知。东南亚、俄罗斯和拉丁美洲的一些国家，都从中国引进了人脸识别、电信和舆情监控技术。

最能说明问题的例子之一，是柬埔寨的国家互联网网关。柬埔寨在2020年代初宣布设立这一网关，旨在让政府对国内和国际互联网连接拥有更大的管控权。⁹ 由于技术上相似，观察人士常把它称为柬埔寨版的“防火长城”。许多研究者一直在追查是谁在帮柬埔寨搭建这套基础设施。尽管尚未得到官方证实，柬埔寨公民社会的一些专家推测是华为或中兴。¹⁰ 这个例子很能说明中国如何在海外推行技术威权主义（指借助数字技术实施监控和社会控制的威权统治方式——编者）。

季：接下来想谈谈商业抱负和政治控制之间的张力。你的报告指出，中国领先的科技企业和人工智能实验室参与起草了生成式人工智能的相关标准，与国家一起制定内容安全规则，同时也接受国家的政治控制要求。但政府和企业追求的目标本来就不同。政府要实现人工智能的自立自强，构建从高端芯片到基础软件都自主可控的软硬件体系；企业则想要商业成功，想要全世界的用户。这些目标怎样协调，又在哪里发生冲突，尤其是在海外用户希望自由下载、运行和修改这些模型的时候？开放与政治要求相互拉扯时，由谁来划定界线？企业手里有没有自己的筹码？

区：这是个很好的问题。如果问由谁来划定界线，毫不意外，是中国政府。

至于企业有没有筹码，当然有。中国虽然是专制国家，但政府也承

认，国家的大部分创新和资本不在政府手中，而在私营部门。它要实现自己的科技抱负、扩大全球影响力，就必须与私营部门合作。如何处理与私营部门的关系，在商业成功、技术创新和政治控制之间取得平衡，一直是中国当局面临的难题。

2020年代初，习近平发起了一系列打压私营部门的行动，马云就是很典型的例子。马云是中国最知名的私营企业之一阿里巴巴的联合创始人，也是中国商界的成功典范。马云多年来一直公开批评国家干预市场、对金融业严加监管的倾向。他公开反对国家干预，随后便从公众视野中消失多年，令全世界震惊。¹¹ 许多人认为，马云的遭遇说明国家试图规训私营部门，确保自己对私营部门的控制。

经过一系列打压和整顿，到了2025年，习近平开始释放信号，表示欢迎私营部门，鼓励它们在多年规训之后去追求商业成功。但这一次，北京在鼓励私营部门追求商业成功和创新的同时，也明确表示，所有创新和发展都必须在政府划定的边界之内进行。

最值得注意的是，北京采取了胡萝卜加大棒的做法。一方面，政府提供大量激励，促使私营企业向国家的战略目标和优先事项靠拢。另一方面，像马云那样偏离国家意志的企业，就会受到严厉惩罚。到目前为止，这套做法相当成功。它给私营科技企业留出了一些追求商业成功、推动技术创新的自主空间，同时也提醒它们，中国共产党的地位不容挑战，从而强化了这一权力结构。

季：下面从企业一方来看这种政商关系。这些政治要求会不会给企业

的创新和全球竞争力带来很大的成本？你刚才提到，一些基准测试专门检验模型是否符合社会主义核心价值观，而这些价值观里本身就有“自由”“民主”这样的词。如此宽泛、模糊的要求，要落实成训练和测试中前后一致的标准，企业要付出多大代价？

区：在国内，我认为这些科技企业要符合国家的政治要求，并不需要付出多少成本；相反，如果不积极配合国家的优先事项和政治要求，代价才会非常高。但放到全球来看，从国际竞争力和全球市场份额的角度说，这类政治要求以及控制与审查的逻辑，已经在影响中国科技企业的国际形象，华为就是一个例子。

许多人认为，2019 年美国对华为的禁令主要出于中美之间的地缘政治竞争，但它也反映出自由民主国家共同抱有的一些合理担忧：中国积极利用技术，在海外获取战略筹码、服务国家战略目标。中国《国家情报法》出台后，这种担忧更为突出，因为该法要求中国的组织和个人配合国家的情报工作。人工智能、电信等高科技企业在海外运营，需要处理大量数据。各国因此担心，这些数据和业务会不会成为渠道，让情报或敏感信息流入中国的情报部门。

季：你刚才谈到，中国科技企业在海外运营要处理大量数据，民主国家担心这些数据会流入中国的情报部门。这是国家层面的安全担忧。我想换到个人开发者的角度。很多开发者偏爱中国的开放权重模型，因为它们成本低，可以在自己的电脑上运行、按需定制，用起来更自由。我在 X 上看到，不少开发者选择这些模型，正是为了不把自己的

数据交给西方的大型人工智能公司。如果一个人主要用中国模型写代码，从不和它谈政治，审查问题还值得他担心吗？还有哪些风险？这些担忧有什么证据？

区：我们在报告中讨论的政治控制、审查和信息控制，主要影响的是那些会波及信息空间的使用场景，比如教育、公共传播和媒体。如果是写代码这类纯技术用途，对信息空间的影响可能就没那么大。

但我认为风险在于，自由民主国家的政策制定者其实很难追踪和评估这些中国模型是怎样被采用的、用在了哪些场景。我们很难核实，它们是只用于技术用途，还是已经进入了公共传播领域，因此也很难追踪和化解由此带来的潜在风险。难以追踪、难以核实，这本身就已经是一种风险。

季：你在报告中也谨慎地指出，下载量和采用量说明不了这些模型在媒体、教育和公共传播中实际怎样被使用。正如你所说，这很难追踪，也很难评估。那么，中国政府的内容控制是否已经在影响海外的信息环境？这仍是一个悬而未决的问题吗？目前我们知道些什么，又能做些什么？

区：这确实很难追踪，也很难核实，但风险已经存在，而且清晰可见。对这个问题我没有明确的答案，这一点我承认。不过我认为，自由民主国家的政策制定者，以及当地的企业和开发者，无论是自己采用中国人工智能模型，还是监管本国对这些模型的采用，都应该更加谨慎。他们必须清楚自己要付出什么代价、承担什么风险，也要更加

警惕可能的后果。这些后果不只涉及信息空间，也可能影响他们自身的运营和技术栈。

季：你认为西方的政策制定者准备好应对中国人工智能模型崛起带来的挑战了吗？民主国家现有的人工智能法规和治理框架，比如欧盟的《人工智能法案》（AI Act），能在多大程度上应对模型内置、用户甚至看不到的政治内容控制？应该由谁来发现和纠正这些控制？现有的保障措施又有哪些不足？

区：全球人工智能治理（AI governance）和监管在今天确实是非常重要的问题。众所周知，人工智能安全（AI safety and security）¹² 问题日益关乎全人类和整个文明，需要各大国合作应对。据我所知，美国、英国和欧盟等自由民主国家已经与中国开展了广泛合作，推动二轨对话之类的交流（由学者、前官员、智库人士等以非官方身份参与的对话，与政府之间的“一轨”外交相对 -- 编者），希望加强沟通，强化全球人工智能治理。

但我认为，这份报告也强调了一点。谈到人工智能监管或人工智能安全时，各方用的词语虽然相似，比如人工智能安全（AI safety）、内容安全（content security）、内容审核（content moderation），背后的含义却往往截然不同。在欧洲，谈内容安全或内容审核，更多是为了减少人身攻击、侵犯隐私或种族主义之类的风险，保护个人的权利和尊严。在中国，谈人工智能安全和内容安全，更多是指中国国家和中国共产党如何保持对这项技术及其生成内容的政治控制。同样的词语，在不同司法管辖区背后的逻辑和解读大不相同。

如果自由民主国家的政策制定者和开发者意识不到这些差异，就可能正中中国下怀，在不知不觉中接受中国的解读。何况中国正在积极推广自己的标准，扩大自己在全球科技治理中的影响力。

季：你刚才谈到，同样说“安全”，中国政府的理解与欧美完全不同，而中国又在积极推广自己的标准。这正是报告序言所担心的“人工智能治理的专制化”。我想接着问，一个国家在全球人工智能治理中的话语权，在多大程度上取决于它的人工智能产业实力？我们得承认，中国的人工智能实验室做得非常出色，而除了美国之外，很难看到哪个民主国家开发出了被广泛使用的主流大模型。对其他民主国家来说，要防止人工智能治理走向专制化，是否必须打造自己的模型来替代中国模型？还是说，即使在模型研发上比不上中国，它们仍然可以发挥真正的影响力？

区：任何国家要在人工智能监管和全球科技治理中建立影响力、获得话语权，自身都必须具备科技竞争力，这一点毫无疑问。中国对人工智能治理和人工智能安全的理解之所以举足轻重，正是因为中国已经成为当今的科技超级大国之一。

但具备科技能力，并不等于一定要自己研发最前沿的人工智能模型。全球技术栈非常复杂，供应链相互交织。我认为，一个国家只要能在全球技术栈和供应链中占据重要位置，就有潜在的筹码和力量参与这些讨论。台湾就是很好的例子。台湾虽然也设计芯片，但不及美国先进，也不研发最前沿的人工智能模型；不过凭借在芯片制造中的地位，它在全球人工智能技术栈中不可或缺，也因此在全球科技讨论中拥有

相当大的筹码和话语权。

英国、欧洲国家等其他自由民主国家，如果想在科技治理和科技竞争中发挥更大作用，就要在全球技术栈中找到自己的优势和着力点。

季：最后，想请你给我们的读者一些日常使用上的建议。读者平时可能会接触到中国的人工智能模型。有人看中的是它们性价比高，还能在自己的电脑上运行；也有人用的应用或工具其实以中国模型为底层，自己却未必知道。他们需要刻意避开中国模型吗？如果要用，应该注意些什么？

区：中国如今在全球技术栈中举足轻重，完全与中国技术脱钩不太现实。任何人，尤其是开发者，觉得中国的人工智能模型或技术基础设施很有吸引力，我认为可以理解。但就像我前面说的，使用这些技术时，一定要清楚其中的风险和可能的后果，也要清楚自己在用什么代价换取这些便宜又好用的技术。

季：非常感谢你，Owen。相信这次访谈会让我们的读者受益良多。也推荐大家读一读报告原文，一起关注和讨论如何防止人工智能治理走向专制化。

注释

- 1 区俾僖：Aiden Gu：《专制创新：中国如何把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 and Strategic China Policy Group*, July 2026, <https://>

resilienceinnovationlab.org/wp-content/uploads/2026/07/Authoritarian_Innovation_2026.pdf)。

- 2 预训练 (pre-training) 是开发大模型的第一阶段：用海量文本训练模型，让它掌握语言规律和一般知识，之后再经过“后训练”，让它学会按人的指令作答。蒸馏 (distillation) 原指用一个大模型 (“教师模型”) 的输出来训练另一个模型 (“学生模型”)，让后者以较低成本获得接近前者的能力。未经许可、大规模调用他人模型来获取训练数据，违反主要模型开发者的服务条款。前沿模型 (frontier model) 指能力处于当前最先进水平的大模型。-- 编者
- 3《Anthropic: 检测和对抗AI的滥用》(Anthropic, *Detecting and Countering Misuse of AI: September 2026, September 10, 2026*, pp.143-154, <https://www-cdn.anthropic.com/e50be2e51e7695dc4b1366a37a245a597377d3b5/Anthropic-Detecting-and-countering-091026.pdf>)。
- 4 国家市场监督管理总局、国家标准化管理委员会：《网络安全技术 生成式人工智能服务安全基本要求》(GB/T 45654-2025)，2025年4月25日发布，2025年11月1日实施。
- 5 这套包含2000道题的《生成式人工智能服务安全基准测试集》由中国电子技术标准化研究院与复旦大学于2024年联合开发，用于配合上述标准实施，其中约三分之一的题目测试生成内容是否违反“社会主义核心价值观”。同注释1，第18-19页。
- 6 国家市场监督管理总局、国家标准化管理委员会：《网络安全技术 生成式人工智能预训练和优化训练数据安全规范》(GB/T 45652-2025)、《网络安全技术 生成式人工智能数据标注安全规范》(GB/T 45674-2025)，均与上述标准同日(2025年4月25日)发布，2025年11月1日实施。
- 7 技术栈 (tech stack) 原是软件工程用语，指构建一个系统所需的、层层叠加的各种技术。就人工智能而言，从下到上大致包括芯片和算力、数据中心、训练数据、模型，以及基于模型开发的各种应用。-- 编者
- 8 权重 (weights) 是模型在训练中习得的海量参数，决定了模型如何生成内容。开放权重模型 (open-weight model) 指开发者公开这些参数、允许任何人下载模型，用户可以在自己的电脑或服务器上运行，也可以加以修改和再训练。DeepSeek、通义千问 (Qwen) 等中国主流模型多以这种方式发布。开放权重不完全等同于“开源”，开发者通常并不公开训练数据和完整的训练代码。由于政治要求在训练阶段就已写入权重，模型被下载到海外、在本地运行时，这些要求也会随之带走。-- 编者
- 9 柬埔寨首相洪森于2021年2月签署设立国家互联网网关的次级法令，要求所有互联网流量于2022年2月前经由该网关传输，但政府原定启用的前一天宣布推迟实施。见《非政府组织敦促柬埔寨废除互联网管制法令，称其影响在线权利》(“NGOs Urge Cambodia to Scrap Decree

on Internet Controls Citing Impact on Online Rights,” *Radio Free Asia*, February 18, 2021); 索恩·萨拉特:《争议性国家互联网网关因项目批评增加而推迟》(Sorn Sarath, “Controversial National Internet Gateway Postponed as Project Criticisms Grow,” *CamboJA News*, February 15, 2022)。据《日经亚洲》报道,柬埔寨计划于2026年开始建设该网关的关键基础设施,见菲奥娜·凯利赫:《柬埔寨重启备受争议的互联网网关计划》(Fiona Kelliher, “Cambodia Resurrects Plan for Controversial Internet Gateway,” *Nikkei Asia*, October 20, 2025)。

10 见文章 19:《数字丝绸之路:中国与印太地区数字镇压的兴起》(ARTICLE 19, *The Digital Silk Road: China and the Rise of Digital Repression in the Indo-Pacific*, 2024, p.36, https://www.article19.org/wp-content/uploads/2024/04/DSR_final.pdf)。

11 2020年10月24日,马云在上海外滩金融峰会上发表演讲,批评中国的金融监管思路。同年11月3日,蚂蚁集团原定在上海、香港两地的上市被暂缓。此后马云长期淡出公众视野,直到2023年才较多地在国内公开露面。2025年2月17日,习近平主持召开民营企业座谈会,马云应邀出席,外界普遍视之为北京向民营企业释放善意的信号。--编者

12 英文中, AI safety 一般指防止人工智能系统本身造成伤害,比如出错、失控或输出有害内容; AI security 一般指保护系统免遭蓄意攻击和滥用,也常与国家安全相联系。中文把两者都译作“安全”。中国官方所说的“内容安全”(content security),除了一般的有害内容,还包括生成内容是否符合政治和意识形态要求。

