

分析文章

代码不会独自飞行

一份基于公开来源的分析论文，区分人工智能的使用、产出、独立验证与运行能力，并把也门北部案例放回区域历史，同时不越过归属判断与证据的边界。

文章日期

2026年9月22日

作者

Safwan AlModhayan

关键词: 人工智能影响评估 · 网络安全 · 多智能体系统 · 也门 · 威胁归属 · 人工智能治理

本文是作者阿拉伯语公开来源分析的简体中文版本，证据截至2026年9月22日；它不是独立数字取证调查，不包含新的武器性能实验，也未经同行评议。

摘要

本文从 Anthropic 披露的一起事件出发，考察“人工智能提升敏感安全环境中的技术能力”这一主张。该公司称，也门北部一个集群把 Claude Code 用于武器研发相关的软件工作。文章提出，应把公共叙事中经常混为一谈的四个层级分开：观察到使用、产生了输出、完成独立验证、得到可接受结果。现有来源能证明软件活动、一次似乎失败的测试以及此后的失败分析，却不能证明用户就是胡塞武装、已经部署可运行系统，或人工智能产生了多大因果增量。^{[1][2]}

文章将该事件置于本地区网络攻击、导弹和无人机能力，以及能源和航运受影响的历史中，同时让每项事实停留在其来源能够支持的范围内。随后讨论生产率测量、验证成本、多智能体复核的局限、模型提供商的可见范围，并提出沙特机构可采用的治理要求：最小权限、独立测试、决策记录，以及按每项可接受结果测量完整成本。这里给出的是分析结论，并非作者实施新实验所得的结果。

方法与证据边界

本文采用公开来源的分析性叙述综述。只要条件允许，优先使用第一手或距离事件较近的材料，包括企业对其服务内部活动的报告、联合国和政府文件、受影响机构的公告及研究论文；对于缺少完整原始记录的回应和进展，则使用可靠新闻报道。证据截止日期为2026年9月22日。即使某些信息在正式发布时已经出现，只要晚于这一日期，就不进入本文结论。

这不是系统综述：本文没有执行数据库检索方案，也没有对研究质量作定量评分。作者无法接触 Anthropic 原始日志、该集群的文件、独立测试数据或现场。因此，服务提供商的陈述只被视为它对平台内所见活动的直接证据，而非平台外表现的独立证明。政府及交战方声明同样被视为需要权衡的归属叙述，不会自动成为中立事实。

分析使用了主张记录：每项主张都对应来源、证据类型、观察边界和归属状态。可能转化为武器研发或目标选择实用指导的细节被排除，技术空白也没有用推测填补。来源相互冲突时，文章保留冲突，不用强于来源的措辞替它裁决。

历史案例按分析功能选择，用来区分破坏与间谍、指控与定罪、实体影响与主观感受、支持与直接指挥。这份清单并不穷尽本地区所有事件。最终表述也重新对照来源边界，确保旧事件不会被拿来填补新案例中缺失的证据。

1. 在比喻接管叙事之前，还原原始事件

Anthropic 在2026年9月发布的报告中称，也门北部一个集群把 Claude Code 用于多个武器研发项目中的部分软件工作。公司表示，用户把写作、检索和复核分配给多个实例，进行了一次似乎失败的发射测试，之后又回到服务中分析原因。公司还称已经封禁相关账户，并有证据表明用户制作了一套可在本地运行的模拟工具。^[1]

报告最重要的限制，是公司没有证据表明行为者成功部署了可运行系统。这并不让活动变得无关紧要：按来源说法，它已经推进到实体测试，并试图从失败中学习。但这一限制阻止我们把观察结果写成武器已经就绪或精度已经改善。公开材料没有告诉我们软件产出是否正确、没有工具时行为者会做到什么，也没有证明封号后本地工具仍然有用。^[1]

Anthropic 也没有称用户属于胡塞武装。它只说集群位于也门北部；美联社则根据地理与背景推断其与胡塞武装有关，并刊登了该组织政治办公室成员 Hizam al-Assad 的回应。相关方的回应对于理解叙事争议是必要材料，却既不能独立验证技术数据，也不能把数据推翻。方法上的结论应是：公司报告确认了地点；组织身份未确认；运行成功未得到证明。^[2]

2. 区域历史并非始于2026年

胡塞运动产生于也门本地冲突，早于其2014年控制萨那；2015年，以沙特为首的联军开始干预，以支持得到国际承认的也门政府。不能把这段历史缩减成一种新的技术关系，也不能删去其中的外部因素。联合国专家组报告呈现了有关外部物质支持、技术援助和训练的证据与评估，认为这些因素推动了胡塞武装能力发展；伊朗则一直否认向其提供武器。^{[12] [16]}

支持、知识转移、协调和直接行动指挥，不是同一件事的不同名称。来源可能证明存在部件或技术相似性，却不证明某次行动的决策在哪里作出；也可能证明训练或技术援助，却不证明支持方控制此后的每一次使用。若把“伊朗与胡塞武装”当成一个不分时期、任务和关系类型的变量，因果分析就会失真。

针对沙特及红海的发射和袭击记录证明，相关实体能力在新案例出现前已经存在。联合国材料记录了大量导弹、无人机活动，以及针对民用和经济设施的袭击；红海行动后来又把影响扩大到国际航运。但这些记录不能证明 Anthropic 所称集群在组织上直接延续自任何既有单位。历史让假设值得追问，却不会替新事件补上来源没有公布的证据。^{[9] [10] [11]}

3. 从擦除磁盘到窃取知识

2012年8月，沙特阿美遭遇“沙蒙”攻击，数万台数字设备受影响；2016年，更新后的攻击浪潮再次针对沙特机构。来源区分了沙特调查公布的内容，以及情报评估或安全企业对伊朗相关行为者的归属判断。机构数字系统遭到大规模破坏是确定的，但这不等于证明工业生产停摆；归属强度也必须和作出判断的来源一起说明，不能写成一项笼统的司法结论。^[18]

2017年，Mandiant 评估 APT33 为伊朗政府工作，并记录其对航空、能源和石化行业的兴趣，目标中包括与沙特有关的机构。2022年，美国机构把 MuddyWater 描述为伊朗情报与安全部的下属组成部分。两起案例的来源、事实和置信度不同，不会因为政治背景接近就变成同一起事件。^{[19] [20]}

2026年8月18日，美国司法部公布一份扩大的替代起诉书，起诉 Mabna Institute 的17名成员，其中9人已在2018年受到起诉。起诉书指控他们代表伊朗实体实施大规模行动，窃取学术数据与知识产权。在法院作出裁判前，这些内容仍是司法指控；不过，该案确实说明研究知识本身可以成为战略目标。^[7]

这段记录解释了为什么模型提供商会进入威胁情报视野。传统问题是：谁进入了网络，窃取了什么，又破坏了什么？现在还要再问：用户试图借助平台建立何种能力？问题之间的相似性不能成为无证据合并行为者或案例的理由。归属判断必须把具体事件、具体影响和具体来源连接起来，而不是挑选政治记忆中最醒目的名字。

4. 当风险离开屏幕

2019年9月14日，沙特阿美宣布，布盖格和胡赖斯遇袭导致每日570万桶产量暂停。袭击发生及其运行影响由受害机构公告确认；谁宣称负责、设备从何而来、从哪里发动、政治决策由谁作出，则需要不同证据回答的不同问题。^[21]

2022年1月17日，胡塞武装宣称负责的一次阿布扎比袭击造成3名平民死亡。此后，红海袭击扩大了经济影响。国际货币基金组织估计，2024年前两个月经苏伊士运河的贸易量较2023年同期下降约一半，船舶改道，交付时间延长。这一数字描述的是特定时期，不能充当2026年的实时指标；但它说明，安全事件如何从袭击地点传导到保险、库存与供应链。^{[22] [23]}

2026年9月，也门战事与流离失所加剧，世界粮食计划署扩大了应对。该机构和路透社的材料都没有把升级局势与 Claude 案例相连，编辑时也不能这样处理。纳入人道背景，是为了说明能力评估不能止于一件技术产品的表现；社会后果还会出现在粮食、交通和服务连续性上。^{[4] [5]}

从系统分析看，能力不是单一部件，而是一条依赖链：信息、软件、硬件、集成、测试、运行和后勤。公共分析无需描述目标选择或传感器整合的技术。只要承认任一环节的薄弱都可能限制结果，软件工作改善也不证明整条链已经改善，论点就已经成立。

5. 先定义增量，再讨论测量

说人工智能“提高了能力”，必须先定义结果。它可能是任务时间减少、专家工时下降、质量提高、尝试次数增加，也可能是可行项目清单扩大。这些指标并非同义。产出可以增加而可靠性下降；写作可以缩短而成本转移到复核；技术环节可以成功而整个项目继续受测试与验收约束。

用因果推断的语言说，我们想知道有工具与没有工具时结果之差。但历史只展示一条路径，不会让同一事件重演两次。看到用户请求帮助，不能告诉我们没有帮助时需要多久；看到测试失败，也不能说明工具让其更接近还是更远离成功。因此，也门北部案例无法提供一个数字化的能力提升系数。

仍可提出一个较窄的假设：模型也许能降低认知工作的摩擦，让检索、写作、比较和草稿生成更快或更便宜。如果这一点成立，战略影响可能是尝试次数增加，而非每次尝试都必然成功。它是可检验的假设，不是从一份报告中直接抽取的结论。

2026年9月9日，Google Threat Intelligence Group 描述了部分对手从简单提示转向智能体 workflows 和相互连接的自动化。该报告支持一种更广泛的工作组织趋势，却没有验证也门案例，也没有证明多智能体结构优于人类工作或单一工具。^[3]

6. 当秒表与使用者的感受相反

METR 提供了一个理解主观感受与实际测量差距的有用例子。在2025年7月发表的一项随机实验中，16名资深开发者在熟悉的代码库里完成246项任务。在该研究条件下，获得人工智能工具的开发者的时间反而多花了19%，尽管他们在实验后仍认为工具让自己更快。这一结果不能证明人工智能会拖慢所有程序员；它只适用于特定使用者、任务、工具与时期。^[26]

2026年2月，METR 在更新中表示，智能体普及改变了愿意参加研究的人，也改变了他们提交何种任务，削弱了旧设计估计增量大小的能力。机构认为，2025年末的工具可能加快了部分工作，却指出选择偏差和并行时间测量问题使数据不足以确定幅度。严谨的研究者不会给旧结论永久居留权，也不会把不确定的新进展包装成相反叙事的胜利。^[27]

在后来的备忘录中，METR 又区分了工具对旧任务的影响、工具出现后人们选择的新任务，以及对总体价值的影响。技术可能改变我们愿意尝试的事项清单，而不只是加快原来的清单。这一区分对安全案例尤其重要：过去不值得投入的尝试可能因为工具而变得可行，同时每项结果的质量仍未改善。^[28]

这里的教训关乎方法。问卷衡量的是信念，计时记录衡量特定时段，质量测试衡量产出属性，运行结果衡量真实环境中的价值。任何一个指标都不该冒充全部指标。

7. 验证成本，以及不会阅读演示文稿的现实

考虑一个民用示例：软件草稿生成需要两小时，复核、测试和集成需要八小时。即使前两小时完全消失，项目时间也没有消失。这是说明性算例，不是实验结果；它揭示了一种常见错觉：最显眼的环节加速了，真正的约束却仍在测试、批准和运行。

敏感工作里有两个不同问题：系统是否按规格构建？规格是否代表现实问题？软件可能通过执行无误的测试，而测试依赖的假设并不完整。模拟也可能在自身世界中产出一致结果，却被现实中一个从未纳入的条件推翻。因此，“模拟成功”这句话若不说明范围与排除项，信息仍然不足。

理查德·费曼在“挑战者”号事故调查报告附录中写道，现实必须优先于公共关系，因为自然无法被愚弄。这一观点的价值，在于迫使我们追问：哪里存在能够反驳自身叙事的测试？如果同一系统给出解决方案、解释其正确性，再为自己签发合格证，我们建立的是一个语言上自洽的闭环，而非独立验证。^[29]

因此，本文提出四级证据阶梯：观察到使用、产生了输出、完成独立验证、得到可接受结果。跨越每一级都需要新证据。就来源所示，也门北部案例达到了使用与产出阶段，并出现一次似乎失败的测试；它没有达到得到证实的运行验收。

8. 多个智能体不等于多个独立错误源

把工作分给一个写作智能体、一个检索智能体和一个复核智能体，可以带来好处。但独立性不来自窗口数量。若它们共享同一批数据、同一成功标准或同一个错误假设，一致意见不会修复假设；复核者若在设计测试前就看见答案，其角色也可能退化成润色，而不是挑战结论。

一篇题为《多智能体大语言模型系统为何失败？》的 arXiv 预印本分析了7种框架中的1642条执行记录，归纳出14种失败模式，涉及系统设计、智能体协调不良、任务验证与终止。专家先从初始样本建立分类，再用一种经人工标注校验的自动工具把标签扩展到完整数据集；并非所有记录都由专家逐条手工标注。该研究没有分析也门集群，却足以反驳“多个智能体自然等于可靠复核”的跳跃。^[25]

真正的复核需要证据和任务上的差异。复核者应从清晰规格出发，能够接触作者没有生成的第一手来源或测试，并有权宣布证据不足。其价值可以用发布前发现的缺陷衡量，而不是用对话中新增了多少条评论衡量。

可靠性工程不能只问系统是否完成任务，还要问它是否识别出应该停止的状态，是否保存决策轨迹，是否清楚传达不确定性。这里的独立性是设计与测试的属性，不是工作流程图上的角色名称。

9. 产出会留下，平台只看得见自己的窗口

Anthropic 表示，用户在账户被封前已经制作出本地模拟工具。工具存在不证明它正确或有用，却说明终止访问不会自动收回产出。机构终止外部订阅后，仍可能保留在工具协助下产生的文件、代码或决策。因此，评估干预应问三个问题：什么访问被切断？哪些产出已知被导出？有什么证据说明它们后来被使用？^[1]

另一方面，模型提供商日志提供了其他安全工具不具备的视野。提供商可能看见问题解决过程的一部分、请求如何变化、工具如何调用。这足以让其报告成为重要来源，却不能让它知道文件下载后发生了什么，或经过其服务的环境里发生了什么。

Anthropic 报告还说明，所选案例并不代表常见使用，而是因醒目或新颖被纳入。公开案例数量因此不能用于估计流行率。披露增多可能反映活动增多、检测改善、公开政策变化，或三者并存。缺少分母，就不能把数量换算为比例。^[1]

新闻报道数量增加也不会自动解决这一问题。五十篇文章重复同一份披露，可能补充解释和采访，却不会变成五十名关于原始主张的独立证人。证据是否独立，要按具体主张判断，而不是按媒体标识数量判断。

10. 以完整价值为尺度的沙特治理

沙特既有直接理由关注人工智能滥用，也有直接理由用它改善生产与韧性。因此，治理不应成为禁用的同义词。需要设计的是一种可证明价值、可控制错误，并在工具失效或变化时维持服务的使用方式。

2026年7月5日，沙特国家网络安全局就人工智能网络安全指南启动公开征求意见，范围包括治理、防御、韧性和第三方，并延伸至生成式和智能体式系统。来源证明咨询及其议题存在，却不证明最终且具有约束力的文本已在当日发布。^[8]

第一项执行原则是最小权限：智能体只获得任务所需权限，并区分读取与修改、测试与生产、建议与执行。流畅表达或“自主智能体”这一名称都不构成访问权。任何走向高影响动作的转换，都需要明确条件、留有记录的人类复核，或与风险相称的技术控制。^[30]

第二项原则是结果记录。复核一项有人工智能参与的决策时，应能查明工具版本、相关信息来源、所作修改、执行的测试和批准使用的人。这并不要求保存员工的所有话语，而是按专业审计所需保留适量材料，并控制访问与保存期限。

第三项原则，是按每项通过独立测试的可接受结果计算成本。成本包括任务准备、工具运行、复核、返工和后续缺陷处理；结果则必须对应明确质量，而不是文件已完成或工单已关闭。用这一尺度，才能区分活动增加与价值增加。

11. 本文的限制与不能支持的结论

核心事件依赖 Anthropic 的披露，现有来源中没有对账户日志或实体测试的独立复核。本文无法确认该集群的组织身份、测量代码质量，也不知道封号后产出去向。反事实同样无法估计：没有 Claude 时，用户会在多长时间内完成什么。

区域来源并不均质。联合国报告、政府与军方公告、武装组织声明、企业报告和新闻报道，各有不同的观察窗口、利益关系和证据标准。本文汇集它们是为了建立背景，而不是把它们熔成一份统一记录。部分袭击和归属仍存在相互冲突的说法，文章没有强行裁决。

本文是公开来源的叙述综述，不是系统综述或统计荟萃分析。案例选择可能让知名或容易公开的事件获得更大权重。模型提供商报告也受检测能力和披露政策影响，且没有给出估计流行率所需的分母。

本文不评估具体武器系统，也不说明如何研发此类系统。设计、目标选择与干扰相关细节被概括处理，因为它们并非论证所需，还可能把分析变成操作材料。结论讨论的是认知工作、治理和证据边界，不是武器工程。

公开记录还可能存在发表偏差：企业只展示自己发现且选择公开的案例，受害者未必披露全部影响，成功与失败也不会以相同比率出现。来源没有提供可用来计算流行率或比较提供商的共同分母。因此，本文不会仅凭报告数量增加就推出定量趋势。

12. 一份民用研究议程，以及仍然存在的问题

第一条研究路径是既有经验的作用。可以在安全的民用编程任务中比较经验不同的群体，随机分配工具，对质量进行盲评，并测量完整耗时。问题是：工具会让善于发现其错误的人获益更多，还是确实缩小差距？研究设计必须允许两种结果出现。

第二条路径是验证负担是否转移。任务准备、草稿生成、复核、测试和返工应分别记录，再测量验收前后发现的缺陷。工具也许同时加快生产和复核；若数据如此，“验证成本”假设就应修改，不能只因概念优雅便加以保护。

第三条路径是自动复核的独立性。可以比较两类复核者：一类先看作者答案，另一类直接从规格和原始证据开始；随后测量各自发现的缺陷和共同漏过的问题。变量不应只有智能体数量，还应包括证据差异、反对权与停止标准。

第四条路径是结果在阿拉伯语和本地工作环境中的有效性。研究不预设表现更差或更好，而应使用机构实际采用的语言、领域和文档进行测试。目的不是制作模型营销榜单，而是了解语言和语境变化时，质量、成本与疏漏风险在哪里发生变化。

第五条路径可以直接研究治理。让一组团队采用最小权限、决策记录和独立测试，另一组在相近的安全民用任务中不采用这些控制；比较延迟出现的错误、恢复时间和审计成本，而不只比较满意度。这样，治理会成为可测量的假设，而不是挂在墙上的漂亮原则清单。

Claude 案例把两个问题同时摆在面前：一边是袭击和工具演变得到记录的区域历史，另一边是证据不足以证明军事成功影响的新事件。历史让问题变得紧迫，却不能填补来源空白。战略分析的任务，是说明什么已经改变、什么尚不确定是否改变，以及哪一种测试能够区分两者。

风险无需夸大，也足以值得关注；信心写得再漂亮，也不会自动变成保护。代码在屏幕上无论多么令人信服，都不会独自飞行。它与结果之间还有硬件、数据、测试和责任。研究与治理的共同价值，就位于这段距离之中。

从使用到影响的证据阶梯

1	2	3	4
观察到使用	产生了输出	完成独立验证	得到可接受结果

不能跨级推断：每一次前进都需要新的标准，以及独立于产出方的证据。

参考文献

1. 检测并遏制人工智能滥用：2026年9月
Anthropic
2. Anthropic称，也门胡塞武装控制区用户曾尝试借助人工智能开发先进武器
Associated Press, 2026年9月11日
3. GTIG人工智能威胁追踪：从提示到自主——对抗性人工智能的演变
Google Threat Intelligence Group, 2026年9月9日
4. 也门战事导致数千人流离失所，世界粮食计划署扩大响应
World Food Programme, 2026年9月17日
5. 官员称饥饿风险上升，世界粮食计划署扩大也门响应
Reuters, 2026年9月21日
7. 17名伊朗人被控代表伊斯兰革命卫队及其他伊朗实体实施大规模网络窃取
U.S. Department of Justice, 2026年8月18日
8. 国家网络安全局就人工智能网络安全指南启动公开征求意见
National Cybersecurity Authority, 2026年7月5日
9. 也门问题专家小组最终报告
United Nations Security Council
10. 联合国也门国家工作队2021年度报告
United Nations in Yemen
11. 红海弹道导弹袭击引发亚洲对防御系统的关注
Reuters, 2024年2月21日

12. 也门问题专家小组最终报告
United Nations Security Council, 2024年10月11日
16. 也门问题专家小组最终报告
United Nations Security Council, 2025年10月17日
18. 沙蒙2: Disttrack擦除程序卷土重来
Unit 42, 2016年11月30日
19. APT33: 对伊朗网络间谍活动的观察
Mandiant, 2017年9月20日
20. 伊朗政府支持的行为者对全球政府与商业网络实施网络行动
CISA, 2022年2月24日
21. 布盖格与胡赖斯事件
Saudi Aramco, 2019年9月14日
22. 安全理事会关于阿拉伯联合酋长国阿布扎比恐怖袭击的新闻声明
United Nations Security Council, 2022年1月21日
23. 红海袭击扰乱全球贸易
International Monetary Fund, 2024年3月7日
25. 多智能体大语言模型系统为何失败?
arXiv, 2025年3月17日
26. 衡量2025年初人工智能对资深开源开发者生产率的影响
METR, 2025年7月10日
27. 我们正在调整开发者生产率实验设计
METR, 2026年2月24日
28. 任务替代与增量
METR, 2026年5月8日
29. “挑战者”号航天飞机事故总统委员会报告, 附录F
NASA
30. 计算机系统与信息保护
Proceedings of the IEEE