全球AI深伪威胁与信息可信度指数（先行版）

Global AI Deepfake Vulnerability & Information Credibility Index (Pilot Edition)

编号	TJRI-WP-2026-002
编制单位	天机研究院（International Communication Organization, ICO）
法理框架	联合国全球数字契约（GDC）｜UNESCO媒体与信息素养联盟
发布日期	2026年6月
版本	V1.2 先行版

核心发现：

· 五大场景中，"社交关系与紧急求助"场景DFVI最高（78分，Ⅳ级·高危）

· 十国韧性评估中，欧盟和中国领先但仍有缺口，全球南方国家韧性严重不足

· 全球深伪治理呈现"法规先行、执行滞后"的结构性特征

2026年，人工智能生成内容已从技术新奇品变为全球性安全威胁。Europol预测，高达90%的在线内容可能包含AI生成素材；世界经济论坛将"错误信息和虚假信息"连续列为全球五大风险；语音深伪欺诈同比增长680%，金融领域增长达1,300%。 然而，全球尚无一个指数产品能回答最核心的问题：在特定场景下，普通人和机构有多大可能被深伪内容欺骗？社会有多大的能力抵御？ 本指数提出DFVI（Deepfake Vulnerability Index，深伪易感度指数），以"威胁度×脆弱度−韧性度"三维模型，量化评估五大高风险场景的深伪易感程度，并对比十个代表性国家/地区的治理韧性。 核心发现： 五大场景中，"社交关系与紧急求助"场景DFVI最高（78分，Ⅳ级·高危），普通公众在此场景中几乎无法自保 十国韧性评估中，欧盟和中国领先但仍有缺口，全球南方国家韧性严重不足 全球深伪治理呈现"法规先行、执行滞后"的结构性特征

一、背景与紧迫性

1.1 一个正在崩塌的信任基石

2026年4月，世界经济论坛发布《2026年全球风险报告》，将"错误信息和虚假信息"与地缘经济对抗、武装冲突、极端天气、社会极化并列为全球五大短期风险。这是该风险连续第二年位列前二，且AI技术滥用与虚假信息的相关系数达0.87——两者已形成相互强化的恶性循环。 更严峻的是，Europol预测到2026年，高达90%的在线内容可能包含AI生成素材。当"真实"不再是默认选项，而是需要被证明的例外，数字社会的信任基石就在崩塌。 这不是危言耸听。一组数据足以说明： | 指标 | 数值 | 来源 | | --- | --- | --- | | 语音深伪欺诈同比增长 | +680% | Pindrop 2025语音安全报告 | | 金融领域深伪增长 | +1,300% | 同上 | | 保险业合成语音攻击增长 | +475% | 同上 | | 克隆语音所需音频 | 仅30秒（2022年需30分钟） | Deloitte分析 | | 暗网深伪工具起价 | $20 | 行业报告汇编 | | 制作Biden深伪电话成本 | $1 | Bright Defense | | 人类识别高质量深伪的准确率 | 仅0.1%能全部正确 | iProov 2025 | | CNN检测器真实场景精度下降 | 15%以上 | 2026深伪检测状态报告 |

中国本土感知： 中国公安部和国家反诈中心数据显示，2025年AI换脸和拟声诈骗案件立案数同比激增，涉案金额数以亿计。国家反诈中心已将"AI冒充熟人诈骗"列为重点预警类型，多省市反诈中心发布专项提示。随着《AI生成合成内容标识办法》2025年9月施行，中国成为全球最早对AI生成内容实施强制标识的国家，但法规落地与犯罪迭代之间的时间差，仍是当下最紧迫的防护缺口。

1.2 为什么需要这个指数

当前全球深伪治理领域存在一个关键缺口：没有人从"普通人有多大可能被骗"的视角做系统性评估。 | 现有产品 | 做了什么 | 没做什么 | | --- | --- | --- | | WEF全球风险报告 | 宏观风险排序，专家感知调查 | 无场景级量化，无国家排名 | | MITRE ATLAS | 攻击技术分类（170+技术） | 仅攻击者视角，无防御/韧性评估 | | GDI（全球虚假信息指数） | 新闻网站虚假信息风险评级 | 不涉及AI深伪，不评估场景易感度 | | Pindrop/Sumsub行业报告 | 欺诈统计数据 | 行业数据碎片化，无可比框架 | | C2PA/ITU AMAS | 技术标准与政策框架 | 标准≠执行，不评估实际防护效果 | | 学术数据集（DFDC等） | 检测算法基准 | 实验室数据，不反映真实威胁态势 |

DFVI的差异化定位： 全球第一个以"深伪易感度"为核心、整合"威胁-脆弱-韧性"三维的公共指数产品。不是技术报告，不是政策追踪，而是回答一个每个人都在问、却没人系统性回答的问题—— "我有多容易被骗？"

1.3 法理基础

本指数的编制和发布依托以下国际法理框架： 联合国全球数字契约（GDC）：ICO 系 GDC 背书机构，本指数围绕 GDC 数字信任与安全原则展开评估 UNESCO媒体与信息素养联盟：CRNA/CDNA/ICO 为 MIL 联盟三个独立成员，信息可信度评估是 MIL 方法论的延伸 IGF互联网治理论坛：CRNA 运营 IGF 远程中心，数字治理与数字可信属于其核心议题

二、指数模型：DFVI

2.1 设计哲学

DFVI不是衡量"深伪技术有多强"——那是技术问题，大厂在做。 DFVI衡量的是：在特定场景下，普通人和机构有多大可能被深伪内容欺骗，以及社会有多大的能力识别和抵御它。 这一定位决定了DFVI的独特价值：它不服务于技术迭代，而服务于公共决策——帮助政策制定者识别最脆弱的场景，帮助公众理解风险等级，帮助国际组织评估治理缺口。

2.2 三维指标体系

DFVI由三个维度构成： DFVI = αT + βV − γR T（Threat，威胁度）：深伪"矛"有多锋利 V（Vulnerability，脆弱度）："盾"有多薄 R（Resilience，韧性度）："城墙"有多厚 先行版采用等权设定（α=β=γ=1/3），后续版本将基于熵权法和专家打分法动态校准。 **DFVI越高，说明该场景越容易被深伪攻破**。

2.3 指标定义与评分标准

第一维：威胁度（T）——深伪"矛"有多锋利 | 指标 | 含义 | 1分 | 3分 | 5分 | | --- | --- | --- | --- | --- | | T1 生成门槛 | 生成该场景深伪内容所需的技术水平与成本 | 专业团队+高成本 | 需一定提示词工程或开源本地部署门槛，但单次生成成本极低 | 零基础个人可免费批量生成 | | T2 逼真度 | 当前最先进深伪在该场景下的欺骗能力 | 肉眼可辨 | 需仔细观察才能发现 | 专业鉴伪师也难以区分 | | T3 扩散速率 | 深伪内容的传播速度与覆盖面 | 小众传播 | 中等范围跨平台 | 跨平台病毒式扩散 | | T4 攻击多样性 | 已出现的深伪攻击类型数量 | 单一类型 | 2-3种类型 | 多模态跨渠道组合攻击 |

第二维：脆弱度（V）——"盾"有多薄 | 指标 | 含义 | 1分 | 3分 | 5分 | | --- | --- | --- | --- | --- | | V1 人群识别力 | 普通用户对该场景深伪的识别成功率 | 多数能识别 | 部分能识别 | 几乎无法识别 | | V2 信任惯性 | 目标人群是否习惯性信任该信息渠道 | 普遍质疑 | 选择性信任 | 几乎无条件信任 | | V3 损失放大 | 欺骗成功后的损失规模 | 个人级小额 | 企业级中等 | 社会系统性风险 | | V4 高危暴露 | 老年人、未成年人等脆弱群体的暴露比例 | 低频偶发 | 中等频率 | 高频日常暴露 |

第三维：韧性度（R）——"城墙"有多厚 | 指标 | 含义 | 1分 | 3分 | 5分 | | --- | --- | --- | --- | --- | | R1 法规完备度 | 深伪相关法规覆盖程度 | 无法规 | 有框架但执法弱 | 专项立法+执法机制完善 | | R2 检测覆盖率 | AI鉴伪能力在主流平台的部署程度 | 无部署 | 部分部署 | 全链路实时检测+自动处置 | | R3 标准采纳度 | C2PA等内容来源认证标准的行业采纳率 | 无采纳 | 少数平台自愿采纳 | 主流平台强制采纳 | | R4 公众素养 | 媒体信息素养教育的普及程度 | 无系统教育 | 有项目但未普及 | 纳入国民教育体系 |

2.4 评分方法

每个场景的各指标由研究团队基于以下数据综合评估： A级数据：官方统计、法律文本、国际组织报告（最高权重） B级数据：学术论文、检测基准、行业白皮书 C级数据：平台透明度报告、安全厂商报告 D级数

据：新闻报道、专家评估（需交叉验证） 各维度得分为其子指标的算术平均（1-5分），综合得分按公式计算后归一化为0-100分。

2.5 归一化与DFVI计算

各维度T、V、R的子指标评分范围为1-5分，维度得分为子指标的算术平均值。 先行版采用简化公式进行归一化： DFVI =（T + V − R）/ 10 × 100 其中 T、V、R 为各维度的加权平均分（先行版等权 α=β=γ=1/3），分母 10 为理论极差（T_max + V_max − R_min = 5 + 5 − 0 = 10，R 最低取 0 作为理论下界）。

敏感性分析： 为检验等权假设的稳健性，我们模拟了 α、β、γ 各浮动 ±0.1（总和保持为 1）的情景。结果表明：在合理权重浮动范围内，五大场景的 DFVI 排名保持不变（社交 > 政治 ≥ 金融 ≥ 新闻 > 娱乐），但绝对分值波动可达 6–8 分。正式版（GDTI）将通过熵权法与专家打分法确定最优权重，并启用完整归一化公式（T+V−R+3）/ 12 × 100 以消除边界压缩效应。

2.6 风险等级

DFVI区间	等级	含义
0-20	Ⅰ级·安全	深伪威胁可控，韧性充足
21-40	Ⅱ级·关注	存在特定风险点，需加强防护
41-60	Ⅲ级·警告	威胁与防御严重失衡
61-80	Ⅳ级·高危	公众几乎无法自保，深伪攻击高发
81-100	Ⅴ级·失守	信息真实性系统性崩塌

三、五大场景评估

场景一：政治新闻与选举信息

场景描述： 深伪政治人物发言视频、伪造选举公告、AI生成假新闻、国家主导的AI信息战 场景定位： 本场景侧重于地缘政治对抗与民主选举合法性——深伪对政治体系的攻击不仅在于传播虚假信息，更在于对选举合法性和制度信任的根本性侵蚀。政治深伪的损失放大（V3=5）不在个人，而在社会系统性灾难：民主根基动摇、社会撕裂、选举结果可能被改写。 威胁度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | T1 生成门槛 | 4 | 开源换脸工具泛滥，政治人物公开影像素材充足，生成门槛极低 | | T2 逼真度 | 4 | 2024超级选举年已出现多起高仿真政治深伪，部分连专业团队亦需技术手段鉴别 | | T3 扩散速率 | 5 | 政治内容天然具备争议性，社交媒体算法放大，跨平台病毒式传播 | | T4 攻击多样性 | 4 | 换脸视频、克隆语音、AI生成文本、伪造文件组合攻击 |

T维度均分：4.25 脆弱度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | V1 人群识别力 | 4 | iProov研究显示仅0.1%能正确识别全部深伪；政治内容情绪化，理性识别力进一步下降 | | V2 信任惯性 | 4 | 选民对"权威来源"的信任惯性显著，深伪利用党派确认偏误 | | V3 损失放大 | 5 | 民主制度根基受损，社会撕裂，可能影响选举结果——这是所有场景中唯一可能改写国家治理走向的损失等级 | | V4 高危暴露 | 3 | 全民级暴露，但选举周期性降低了日常暴露频率 |

V维度均分：4.00 韧性度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | R1 法规完备度 | 2 | 多数国家无选举深伪专项立法；中国标识办法覆盖面广但执法刚起步；欧盟AI Act 2026年8月方全面适用；美国联邦层面空白 | | R2 检测覆盖率 | 2 | 主流社交平台有一定检测能力，但政治内容审核涉及言论自由争议，平台多持审慎态度 | | R3 标准采纳度 | 1 | C2PA在新闻机构中采纳率极低，绝大多数政治内容无来源认证 | | R4 公众素养 | 2 | 仅芬兰等少数国家将媒体素养纳入K-12；全球多数国家缺乏系统教育 |

R维度均分：1.75 **DFVI计算** T=4.25, V=4.00, R=1.75 DFVI =（T + V − R）/ 10 × 100 =（4.25 + 4.00 − 1.75）/ 10 × 100 = 6.50/10 × 100 = 65 **风险等级：IV级·高危 典型案例** 2024年美国大选深伪潮： 选举期间大量AI生成的政治内容在社交媒体传播，包括伪造的候选人发言视频和克隆语音电话。FCC裁定AI生成语音电话须披露，但裁定仅覆盖自动拨号电话，不涵盖社交媒体内容。美国联邦层面至今无统一深伪立法，各州标准不一，跨州执法困难。 印度2024年大选： 多个政党的社交媒体账号使用深伪技术制作对手的虚假视频。在印度22种官方语言、数百种方言的多语种环境中，事实查核严重滞后于深伪传播速度。

场景二：金融高管发言与投资决策

场景描述： 伪造CEO指令视频/语音、AI生成虚假财报分析、克隆语音授权转账、深伪视频会议诈骗 威胁度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | T1 生成门槛 | 4 | 克隆高管语音仅需30秒公开演讲音频，视频换脸工具成本低于$50 | | T2 逼真度 | 5 | 金融高管公开影像丰富，生成素材充足；2024年已出现实时视频会议深伪 | | T3 扩散速率 | 3 | 多为定向攻击，非大规模传播，但精准度极高 | | T4 攻击多样性 | 5 | 语音克隆+视频换脸+伪造邮件+伪造文件+视频会议伪造，多渠道组合 |

T维度均分：4.25 脆弱度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | V1 人群识别力 | 4 | 员工对"老板指令"服从惯性极强；某Fortune 500测试中12名财务分析师11人上当 | | V2 信任惯性 | 5 | 企业科层制度下，下级对

上级指令的信任近乎本能 | | V3 损失放大 | 5 | 单次成功攻击平均损失$450,000（Pindrop）；Deloitte预测2027年美国GenAI欺诈损失达$400亿 | | V4 高危暴露 | 3 | 非全员暴露，但财务、行政等关键岗位高频暴露 |

V维度均分：4.25 韧性度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | R1 法规完备度 | 2 | 金融深伪诈骗可适用现有欺诈法律，但无深伪专项条款 | | R2 检测覆盖率 | 3 | 大型金融机构已部署行为生物识别、语音鉴伪等系统；但中小企业覆盖率极低 | | R3 标准采纳度 | 2 | C2PA在金融领域几乎无采纳 | | R4 公众素养 | 2 | 企业反钓鱼培训有覆盖，但针对深伪的专项培训极少 |

R维度均分：2.25 **DFVI计算** T=4.25, V=4.25, R=2.25 DFVI =（4.25 + 4.25 − 2.25）/ 10 × 100 = 6.25/10 × 100 = 63 **风险等级：IV级·高危 典型案例** 2024年香港跨国公司深伪诈骗案： 诈骗者使用深伪技术伪造跨国公司CFO及其他高管的视频会议，诱使财务人员分15次转账2亿港元。这是目前已知金额最大的深伪诈骗案件之一，关键特征是攻击者利用实时视频会议的"权威场景"降低了受害者的警觉。 2026年Fortune 500企业测试： 安全团队伪造CEO语音指令财务团队电汇$25万。12名财务分析师中11人打开附件，仅1人因"声音略带压缩感"而报告——这不是可规模化的防御能力。 **防范要点** 企业不能再依赖"视频见单"式的传统审批流程。面对深伪攻击，必须引入多因子身份验证：建议采用基于区块链的数字身份公钥验证，或建立线下多渠道多签审批（Multi-Sig）机制——任何涉及资金转账的音视频指令，必须经过至少一条独立于即时通讯的渠道进行二次确认。

场景三：社交关系与紧急求助

场景描述： 克隆亲友声音/面容实施诈骗、"紧急借钱"类深伪攻击、伪造求救信息 威胁度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | T1 生成门槛 | 5 | 克隆语音仅需30秒社交平台音频；换脸仅需数张照片；暗网工具$20起 | | T2 逼真度 | 4 | 熟人场景下逼真度已足够欺骗受害者；实时语音克隆技术成熟 | | T3 扩散速率 | 4 | 通过即时通讯和社交平台点对点传播，精准且快速 | | T4 攻击多样性 | 4 | 语音克隆+换脸视频+伪造聊天记录+伪造短信 |

T维度均分：4.25 脆弱度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | V1 人群识别力 | 5 | 熟人关系下信任惯性极强；"紧急"场景压缩了理性判断时间；iProov研究显示人类检测率接近0% | | V2 信任惯性 | 5 | 熟人关系是信任惯性的最强场景——"父母不会骗我"→"父母的声音不会假" | | V3 损失放大 | 3 | 多为个人级损失，但中老年群体毕生积蓄可能一次清零 | | V4 高危暴露 | 5 | 中老年群体高频暴露于即时通讯，数字素养最薄弱 |

V维度均分：4.50 韧性度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | R1 法规完备度 | 1 | 个人间深伪诈骗可适用反诈骗法律，但事前防护几乎为零 | | R2 检测覆盖率 | 1 | 即时通讯平台端到端加密，平台无法扫描内容；个人终端无鉴伪能力 | | R3 标准采纳度 | 1 | 消费级通讯应用几乎无内容来源认证 | | R4 公众素养 | 1 | 中老年群体AI素养极低，大多数不知道深伪攻击的存在 |

R维度均分：1.00 **DFVI计算** T=4.25, V=4.50, R=1.00 DFVI =（4.25 + 4.50 − 1.00）/ 10 × 100 = 7.75/10 × 100 = 78 **风险等级：IV级·高危** 【特别警示】这是五大场景中DFVI最高的场景——深伪时代最致命的软肋不是政治或金融，而是人与人的信任。当韧性度仅为1.00时，社交关系场景的易感性已接近"失守"边界。端到端加密的通讯架构、个人终端鉴伪能力的缺失、中老年群体AI素养的真空，构成了一个几乎无防护的高危场景。 **典型案例** "妈妈，我出事了"： 2024-2025年，多国报告了克隆子女声音拨打父母电话的诈骗案

件。攻击者只需从社交平台获取30秒语音样本，即可克隆声音拨打电话，以"出车祸""被绑架"等理由要求紧急转账。中老年受害者在"紧急"和"亲情"的双重压力下，几乎不可能理性判断真伪。

场景四：娱乐流媒体与网络身份

场景描述： AI换脸直播、伪造明星内容、虚拟人冒充真人、深度伪造色情内容 威胁度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | T1 生成门槛 | 5 | 换脸App免费可用；虚拟人直播工具开箱即用 | | T2 逼真度 | 3 | 娱乐场景对逼真度要求相对较低，部分伪造仍有瑕疵 | | T3 扩散速率 | 4 | 娱乐内容天然具备传播力，粉丝社群放大效应显著 | | T4 攻击多样性 | 3 | 以换脸和虚拟人为主，类型相对集中 |

T维度均分：3.75 脆弱度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | V1 人群识别力 | 3 | 粉丝群体对偶像的信任降低识别力，但娱乐内容非关键决策 | | V2 信任惯性 | 3 | 娱乐内容的信任惯性中等，但对特定粉丝群体影响显著 | | V3 损失放大 | 3 | 名誉侵权、色情深伪（97%受害者为女性）损失严重，但一般不涉及社会系统性风险 | | V4 高危暴露 | 4 | 未成年人高度暴露于流媒体平台，辨别力最弱 |

V维度均分：3.25 韧性度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | R1 法规完备度 | 3 | 部分国家有深伪色情禁令（加州AB 602、韩国等），但覆盖不全面 | | R2 检测覆盖率 | 3 | 主流平台有AI内容标记机制，但执行不一致 | | R3 标准采纳度 | 2 | 部分直播平台试点数字水印，但未普及 | | R4 公众素养 | 2 | 青少年数字原住民有一定识别力，但深伪色情受害者（多为女性）缺乏保护意识 |

R维度均分：2.50 **DFVI计算** T=3.75, V=3.25, R=2.50 DFVI =（3.75 + 3.25 − 2.50）/ 10 × 100 = 4.50/10 × 100 = 45 **风险等级：III级·警告 典型案例** 深伪色情泛滥： Deeptrace报告显示，97%的深伪视频为非自愿色情内容，受害者几乎全为女性。尽管部分国家已立法禁止，但制作成本低（$20起）、跨境传播难以追诉，实际保护效果有限。

场景五：新闻媒体与事实查核

场景描述： AI生成虚假新闻稿件、伪造采访视频、批量生成虚假评论、媒体信任度系统性下降 场景定位：本场景侧重于日常新闻生态的污染与"信息冷漠症"（Reality Apathy） ——当假新闻泛滥到一定程度，公众不再区分真假，开始怀疑一切媒体内容。与场景一（政治深伪的损失放大属性）不同，新闻媒体场景的核心威胁在于"扩散速率"（T3=5）：AI可批量生成、多语言分发、24小时不间断，将新闻生态变成真假难辨的"迷雾"。 威胁度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | T1 生成门槛 | 4 | LLM生成新闻文本零门槛，伪造采访视频技术成熟 | | T2 逼真度 | 4 | AI生成新闻文本在语言流畅度上已与人类写作无异；伪造视频在新闻场景下欺骗力强 | | T3 扩散速率 | 5 | 新闻内容天然具备传播力；AI可批量生成、多语言分发、24小时不间断——这是深伪的"自动化军工厂" | | T4 攻击多样性 | 4 | 文本+图片+视频+评论+社交账号，全链条伪造 |

T维度均分：4.25 脆弱度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | V1 人群识别力 | 4 | 公众对"新闻"格式的信任惯性降低但未消失；AI生成文本的14.2%事实错误率难以肉眼发现 | | V2 信任惯性 | 4 | "可信度危机"（reality apathy）正在蔓延——公众开始怀疑一切媒体内容，连真新闻也不再信任 | | V3 损失放大 | 4 | 媒体信任是民主社会的基石，系统性崩塌影响深远 | | V4 高危暴露 | 4 | 全民日常暴露 |

V维度均分：4.00 韧性度评估 | 指标 | 评分 | 依据 | | --- | --- | --- | | R1 法规完备度 | 2 | 中国标识办法要求AI生成新闻内容须标识，但执法覆盖有限；欧盟AI Act对新闻内容标注有要求但尚未全面适用 | | R2 检测覆盖率 | 2 | 事实查核机构（如ICIF）存在但覆盖面不足；全球事实查核多聚焦主流平台，视频和图片平台监测

不足 | | R3 标准采纳度 | 2 | C2PA在少数新闻机构试点，但绝大多数媒体未采纳 | | R4 公众素养 | 2 | UNESCO MIL框架在部分国家推广，但全球覆盖率极低 |

R维度均分：2.00 **DFVI计算** T=4.25, V=4.00, R=2.00 DFVI =（4.25 + 4.00 − 2.00）/ 10 × 100 = 6.25/10 × 100 = 63 **风险等级：Ⅳ级·高危 典型案例** AI批量生成假新闻： 2024-2025年，多个案例显示AI被用于批量生成虚假新闻文章。某公众号运营者利用AI生成"30万居民撤离上海"等耸动内容，引发公众恐慌，最终被行政拘留。这类案例暴露了AI生成内容在新闻领域的双面性：技术降低了好新闻的生产门槛，也同时将假新闻的制造能力武器化。 事实查核的困境： 学术研究显示，47%的事实查核聚焦国内事务，53%涉及跨境内容。但事实查核机构主要关注Facebook和X等主流平台，对YouTube、Instagram、TikTok等视频和图片平台的监测严重不足——而深伪内容恰恰在这些平台最为泛滥。 五大场景DFVI总览 | 排名 | 场景 | DFVI | 风险等级 | 最致命短板 | | --- | --- | --- | --- | --- | | 1 | 社交关系与紧急求助 | 78 | Ⅳ级·高危 | 韧性极低（1.00），个人通讯无任何防护 | | 2 | 政治新闻与选举信息 | 65 | Ⅳ级·高危 | 损失极大（V3=5），选举合法性面临系统性侵蚀 | | 3 | 金融高管发言与投资决策 | 63 | Ⅳ级·高危 | 损失极大（V3=5），单次攻击可致巨额损失 | | 4 | 新闻媒体与事实查核 | 63 | Ⅳ级·高危 | 扩散极快（T3=5），AI可批量多语种生成 | | 5 | 娱乐流媒体与网络身份 | 45 | Ⅲ级·警告 | 相对可控，但深伪色情（97%针对女性）是突出痛点 |

关键洞察： 四个场景进入高危等级，仅娱乐场景处于警告级别。这不是局部问题，而是全面性的信任危机。其中社交关系场景的韧性得分仅为1.00（满分5），是整个深伪防御体系最大的盲区。场景一（政治）与场景五（新闻）DFVI同为高危但本质不同：前者是"信任被武器化"的灾难性后果，后者是"信任被稀释"的慢性中毒。

四、十国/地区深伪治理韧性对比

4.1 评估框架

本部分评估十个代表性国家/地区在深伪治理韧性（R维度）上的表现，使用2.3节定义的四项韧性指标（法规完备度、检测覆盖率、标准采纳度、公众素养），评分1-5分。

4.2 国别评估

🇨🇳 中国

指标	评分	依据
R1 法规完备度	4	《AI生成合成内容标识办法》2025年9月1日施行，配套国标GB45438—2025，三维责任体系（服务提供者+分发平台+用户），全球最严格的显式+隐式双标识制度
R2 检测覆盖率	3	头部平台已部署AI鉴伪，但中小平台覆盖不足；跨平台一致性待验证
R3 标准采纳度	2	国内标准推进中，与国际标准（C2PA）互操作性待建立
R4 公众素养	2	部分城市试点媒体素养教育，但未纳入国民教育体系

韧性均分：2.75 ｜ 韧性等级：中 **优势： 全球最早实施AI内容强制标识的国家，法规框架完备度高。国内隐式水印（数字水印）技术已实现产业化部署，随着中国企业大规模出海，该技术有望随供应链外溢，形成事实上的全球技术标准雏形。 短板： 标识办法对境外平台执行效力有限；隐式标识跨平台互操作性未验证；公众素养教育滞后于法规建设**

🇪🇺 欧盟

指标	评分	依据
R1 法规完备度	4	AI Act第50条深伪标识义务，违规最高罚全球营业额6%；2026年8月全面适用
R2 检测覆盖率	3	DSA要求超大型平台透明度报告；AI4TRUST等EU项目推进监测
R3 标准采纳度	3	C2PA在欧洲部分媒体机构试点；欧盟积极推动数字身份框架
R4 公众素养	3	多国推行媒体素养教育（芬兰为全球标杆），但成员国间差异大

韧性均分：3.25 ｜ 韧性等级：中高 **优势： 最严格的处罚机制（营业额6%）；芬兰等国公众素养全球领先；C2PA在欧洲部分媒体机构试点推进中，欧盟积极推动数字身份框架（EUDI Wallet） 短板： AI Act 2026年8月方全面适用，目前处于执法真空期；27国执法协调成本高；部分成员国（如匈牙利、罗马尼亚）的检测技术部署和公众素养远落后于北欧标杆**

🇰🇷 韩国

指标	评分	依据
R1 法规完备度	4	《AI基本法》2026年1月生效，亚洲首部综合性AI立法；深伪色情犯罪专项立法严苛
R2 检测覆盖率	3	科技企业检测能力较强，但深伪色情犯罪仍高发
R3 标准采纳度	2	国际标准采纳起步中
R4 公众素养	2	深伪事件频发提高了公众警觉，但系统教育不足

韧性均分：2.75 | 韧性等级：中 **优势： 立法速度快，深伪色情问题推动立法动力强 短板： AI基本法实施细则待完善；深伪色情犯罪规模仍远超执法能力**

🇸🇬 新加坡

指标	评分	依据
R1 法规完备度	2	Model AI Governance Framework完全自愿；POFMA可用于深伪内容移除但非专项
R2 检测覆盖率	3	政府数字能力较强，AI Verify Foundation提供开源测试工具
R3 标准采纳度	2	自愿采纳模式，企业参与度尚可
R4 公众素养	3	教育体系较完善，数字素养基础好

韧性均分：2.50 | 韧性等级：中 **优势： 小国响应快，AI Verify工具开源可复用 短板： 无强制力，完全依赖企业自律**

🇬🇧 英国

指标	评分	依据
R1 法规完备度	3	Online Safety Act覆盖部分深伪内容，但非专项立法
R2 检测覆盖率	3	平台问责机制建立中
R3 标准采纳度	2	C2PA在BBC等机构试点
R4 公众素养	3	媒体素养教育有一定基础

韧性均分：2.75 | 韧性等级：中 **优势： Online Safety Act建立了较完整的平台问责框架；Ofcom作为监管机构执行力较强；BBC等主流媒体在内容来源认证方面走在前列；媒体素养教育有较好基础 短板： 非专项深伪立法，监管覆盖面有限；C2PA等标准仅在少数机构试点；脱欧后与欧盟数字治理框架的协调成本增加**

🇺🇸 美国

指标	评分	依据
R1 法规完备度	2	联邦层面无统一深伪立法；FCC仅裁定AI语音电话须披露；各州立法碎片化（加州、德州等各有侧重）
R2 检测覆盖率	3	技术能力全球最强（Intel FakeCatcher、Resemble AI、Sensity等），但平台审核受言论自由争议制约
R3 标准采纳度	2	C2PA由Adobe/Microsoft等美国企业发起，但平台采纳仍属自愿
R4 公众素养	2	公众对深伪威胁认知两极分化，系统教育缺失

韧性均分：2.25 | 韧性等级：中低 **优势： 深伪检测技术全球领先，已形成可观的产业化能力：Intel FakeCatcher实现96%实时检测准确率并部署于多家新闻机构；Meta主办的Deepfake Detection Challenge（DFDC）汇集全球逾2,000支团队参与，推动了检测算法的系统性进步；Resemble AI和Sensity提供商业化语音/视频鉴伪SaaS服务；C2PA内容来源认证标准由Adobe和Microsoft联合发起，已获BBC、Reuters、AP等国际媒体机构试点采纳。美国在"技术供给侧"的领先优势是客观事实，但技术能力向公共防护的转化面临制度性障碍 短板： 由于美国独特的联邦体制与宪法第一修正案（言论自由）的深度博弈，其治理呈现"技术工具高度市场化、公共政策显著碎片化"的结构性特征。联邦层面缺乏统一深伪立法，各州标准不一导致跨州执法困难，技术优势难以通过公共政策渠道转化为全民防护。这并非技术能力不足，而是制度设计使然——言论自由的传统保护机制在深伪时代面临前所未有的张力**

🇯🇵 日本

指标	评分	依据
R1 法规完备度	2	《AI推进法》2025年5月通过，无禁止性条款、无强制评估、无罚款，仅依赖"声誉制裁"
R2 检测覆盖率	2	企业自愿部署
R3 标准采纳度	1	几乎无采纳
R4 公众素养	3	教育水平高，但AI专项素养不足

韧性均分：2.00 | 韧性等级：低 **优势： 社会自律传统 短板： 软法模式对深伪威胁约束力严重不足**

🇧🇷 巴西

指标	评分	依据
R1 法规完备度	2	TSE（最高选举法院）对选举深伪有禁令，但非综合性立法
R2 检测覆盖率	1	平台审核能力弱，WhatsApp端到端加密无法扫描
R3 标准采纳度	1	无采纳
R4 公众素养	1	公众媒体素养低，深伪内容在社交媒体广泛传播

韧性均分：1.25 ｜ 韧性等级：极低 **短板：社交媒体深伪重灾区，WhatsApp加密通信无检测能力，公众素养极低**

🇮🇳 印度

指标	评分	依据
R1 法规完备度	1	无深伪专项立法，仅有IT Act一般条款
R2 检测覆盖率	1	数字基础设施覆盖不足，农村地区几乎无检测能力
R3 标准采纳度	1	无采纳
R4 公众素养	1	22种官方语言环境下，事实查核多语种能力严重不足

韧性均分：1.00 ｜ 韧性等级：极低 **短板：多语种环境是深伪治理的最大挑战；政党自身使用深伪进一步侵蚀治理基础**

🇳🇬 尼日利亚

指标	评分	依据
R1 法规完备度	1	无AI/深伪专项法规
R2 检测覆盖率	1	基本数字基础设施尚不完善
R3 标准采纳度	1	无采纳
R4 公众素养	1	AI素养几乎为零

韧性均分：1.00 ｜ 韧性等级：极低 **短板：代表全球南方数字基础设施与治理能力缺口的典型**

4.3 十国韧性排名

排名	国家/地区	韧性均分	等级	关键特征
1	🇪🇺 欧盟	3.25	中高	最严处罚机制+芬兰素养标杆，但执法真空期
2	🇨🇳 中国	2.75	中	全球最早强制标识，数字水印技术有出海标准输出潜力
2	🇰🇷 韩国	2.75	中	立法最快，深伪色情驱动立法动力
2	🇬🇧 英国	2.75	中	平台问责推进中
5	🇸🇬 新加坡	2.50	中	自愿模式，小国响应快但无强制力
6	🇺🇸 美国	2.25	中低	技术供给侧全球领先，但联邦体制与言论自由的制度张力使政策显著碎片化
7	🇯🇵 日本	2.00	低	软法模式约束力严重不足
8	🇧🇷 巴西	1.25	极低	社交媒体深伪重灾区

排名	国家/地区	韧性均分	等级	关键特征
9	🇮🇳 印度	1.00	极低	多语种治理噩梦
9	🇳🇬 尼日利亚	1.00	极低	基础设施缺口

关键洞察： 无一国韧性达到"高"等级——即使排名第一的欧盟，韧性均分也仅为3.25/5，说明全球深伪治理尚处于起步阶段 全球南方韧性断层——巴西、印度、尼日利亚的韧性均分仅为1.00-1.25，与发达国家差距悬殊。需特别说明：本报告对全球南方国家的评估受限于公开数据的严重不足，实际深伪风险可能被低估而非高估——在缺乏系统性监测和统计能力的地区，"看不见的风险"往往比"看得见的风险"更大 法规先行、执行滞后——中国和欧盟法规框架完备度高，但执法覆盖和标准采纳仍处于早期 美国悖论——深伪检测技术全球领先（Intel FakeCatcher 96%实时检测准确率、DFDC全球检测挑战赛、C2PA标准由Adobe/Microsoft发起并获BBC/Reuters/AP采纳），但联邦体制与宪法第一修正案的深度博弈，使技术优势难以通过公共政策渠道转化为全民防护

五、核心发现与行动建议

5.1 六大核心发现

发现一：社交信任是深伪时代最致命的软肋

社交关系场景DFVI高达78分，是五大场景中风险最高的。这不是因为深伪技术在这里最强，而是因为防御几乎为零——即时通讯端到端加密使平台无法检测，个人终端无鉴伪能力，中老年群体AI素养极低。深伪最致命的攻击面不是政治或金融，而是人与人的信任。

发现二：全球深伪治理处于"法规先行、执行滞后"阶段

中国和欧盟已建立最严格的法规框架，但标识办法和AI Act的实际执法覆盖仍处于早期。法规文本≠治理实效——从"纸面合规"到"实质防护"仍有巨大距离。

发现三：全球南方韧性严重断层

巴西、印度、尼日利亚的韧性均分仅为1.00-1.25，与发达国家差距悬殊。深伪威胁没有国界，但治理能力有。全球南方在深伪治理中的"赤字"，既是风险也是机会——国际组织（UNESCO、ITU）的MIL框架和标准体系在此有最大的落地空间。

发现四：检测技术的"实验室-实战鸿沟"正在扩大

CNN检测器在真实场景中精度下降15%+，对抗性扰动可将精度降至接近随机。实验室AUC 99.6%的数据在实战中并不代表真实防护能力。当前检测技术与生成技术的差距约为24个月，且在扩大。

发现五：公众素养是全球治理的最短木板

除芬兰等极少数国家外，全球绝大多数国家缺乏系统的媒体信息素养教育。深伪检测不可能完全依赖技术——当技术检测滞后于生成技术时，公众的批判性思维和验证习惯是最后一道防线，而这道防线几乎不存在。

发现六：深伪威胁正从"单点攻击"向"组合攻击"演进

2026年的深伪攻击已不再是单一的换脸视频或克隆语音，而是语音+视频+伪造文件+伪造邮件+伪造聊天记录的多模态组合攻击。金融场景的攻击多样性评分已达5分（最高），这意味着防御必须从"检测单一伪造"升级为"验证信息全链路"。

5.2 行动建议

对政策制定者 填补社交场景防护空白——当前法规聚焦内容标识和平台责任，但个人通讯场景几乎无法规覆盖。建议：在反诈骗法律中增设深伪专项条款，要求即时通讯平台提供"通话验证"功能 推动C2PA等标准强制采纳——自愿采纳速度太慢。建议：在新闻、金融、选举等高风险领域强制要求内容来源认证 建立全

球南方深伪治理能力建设机制——UNESCO MIL框架应向全球南方倾斜资源 **对平台企业** 从"检测伪造"升级为"验证真实" ——被动检测永远滞后于生成技术，应转向主动的内容来源认证（C2PA）和数字水印 为个人通讯场景提供验证工具——开发"通话真伪验证"功能，让用户在接听可疑来电时可一键验证对方身份 **对公众** 建立"交叉验证"习惯——收到涉及金钱或紧急事务的音视频消息时，务必通过第二条独立渠道（如回拨对方已知号码、当面确认）进行核实 为家中老人设置"安全词"——约定一个只有家人知道的暗语，深伪无法获取【延时验证法】凡是涉及资金转账的音视频请求，即使画面实时、声音熟悉，也建议强制执行以下任一验证动作： 要求对方做出AI难以实时渲染的动作（如在脸前挥动手掌观察面部边缘是否撕裂，或侧头观察面部轮廓是否变形） 挂断后由自己"原路回拨"对方已知电话号码 在2026年，AI只需3秒语音即可模拟任何人。"确认一下"这个简单动作，可能比任何技术检测都更有效。

六、方法论延伸展望：从数字可信到实物可信

6.1 方法论的普适性

DFVI的核心方法论——"去伪存真"——本质上是对信息真实性与可信度的量化评估。这一框架在逻辑上可延伸至实体消费品领域： | 数字深伪 | 实物造假 | | | --- | --- | --- | | 威胁 | AI生成以假乱真的数字内容 | 工业化生产以假乱真的仿冒产品 | | 脆弱 | 公众无法辨识AI生成内容 | 消费者无法辨识正品与仿品 | | 韧性 | 法规标识+技术检测+素养教育 | 溯源标准+品质认证+消费者教育 | | 核心问题 | "这个内容是真的吗？" | "这个产品是真的吗？" |

从方法论角度看，DFVI所构建的"威胁-脆弱-韧性"三维分析框架具有跨领域适用性。数字世界的信息可信度评估与物理世界的产品可信度评估，在底层逻辑上具有同构性。

6.2 潜在研究方向：消费品溯源验证

以沉香品类为例，该市场存在假货泛滥、标准缺失、消费者认知模糊等典型问题，是检验DFVI方法论向实物领域延伸的合适场景。研究团队已完成相关品类的文献考证与化学指标分析，未来可探索"区块链数字存证+现代波谱化学分析"双重验证在消费品可信度评估中的应用可能性，将DFVI的方法论从数字可信延伸至实物可信领域。

七、方法论说明与局限性

7.1 数据来源

本报告数据主要来源于以下公开信息： 国际组织报告：WEF全球风险报告、UNESCO AI伦理建议书、ITU AMAS标准、OECD AI政策观测站 政府法规文本：中国《AI生成合成内容标识办法》及配套国标、欧盟AI Act、韩国《AI基本法》等 学术数据集：FaceForensics++、DFDC、Celeb-DF、DeepfakeBench 行业报告：Pindrop、Sumsub、Resemble AI、Deloitte、Bright Defense 平台透明度报告：Meta、TikTok、EU DSA数据库 新闻报道：仅采用具名来源，不采用匿名来源

7.2 局限性

数据覆盖不均：全球南方国家的深伪案例和治理数据严重不足，本报告对巴西、印度、尼日利亚的评估可能低估了实际风险——在缺乏系统性监测和统计能力的地区，"看不见的风险"往往比"看得见的风险"更大 权重暂定等权：α=β=γ=1/3为初始设定，后续版本将基于专家打分法和熵权法动态校准。本版已增加敏感性分析（见2.5节），在±0.1权重浮动范围内排名保持稳健，但绝对分值变化可达6-8分 场景有限：先行版仅覆盖5个场景，医疗、教育、军事等重要场景留待正式版扩展 半定量方法：部分指标当前缺乏系统性量化数据，先行版采用基于公开数据的专家评估方法。本版已在附录C中披露各指标的核心评分锚点和数据来源，完整文献列表将在正式版方法论附录中披露 国家韧性评估聚焦制度层面：未深入评估各国的技术检测能力差异（这需要更详细的行业数据），但本版已对美国技术优势的落地案例做了补充 时效性：深伪技术发展极快，部分数据可能在发布后6个月内需要更新 归一化公式：先行版采用简化公式（T+V−R）/ 10 × 100，正式版将切换至完整归一化公式（T+V−R+3）/ 12 × 100，届时分数量体会有所上移

7.3 更新计划

节点	时间	内容
先行版	2026年6月	本报告
季度更新	2026年9月	更新DFVI数据，扩展场景和国家覆盖，补完敏感性分析和方法论附录
正式版（GDTI）	2026年Q3	整合为GDTI全球数字信任指数旗舰产品，引入动态权重，启用完整归一化公式

附录A：术语表

术语	英文	定义
DFVI	Deepfake Vulnerability Index	深伪易感度指数，衡量特定场景被深伪攻破的可能性
深伪	Deepfake	利用深度学习技术生成以假乱真的图像、音频、视频
GDC	Global Digital Compact	联合国全球数字契约
MIL	Media and Information Literacy	媒体与信息素养
C2PA	Coalition for Content Provenance and Authenticity	内容来源与真实性联盟
AUC	Area Under Curve	模型检测准确度指标
EER	Equal Error Rate	等错误率，鉴伪系统核心指标
GDCC	Global Digital Credibility Certification	全球数字可信认证
GDTI	Global Digital Trust Index	全球数字信任指数

附录B：数据源索引

类别	名称	来源
威胁统计	Pindrop 2025语音安全报告	pindrop.com
欺诈报告	Sumsub身份欺诈报告	sumsub.com
深伪统计	Bright Defense深伪统计汇编	brightdefense.com
风险评估	WEF 2026全球风险报告	weforum.org
攻击模型	MITRE ATLAS	atlas.mitre.org
监管追踪	中国标识办法	gov.cn
监管追踪	欧盟AI Act	artificialintelligenceact.eu
数据集	FaceForensics++	github.com/ondyari/FaceForensics
数据集	DFDC	ai.facebook.com/datasets/dfdc
标准	UNESCO AI伦理	unesco.org
标准	ITU AMAS	iso.org
标准	ISO/IEC 42001	iso.org
产业联盟	C2PA	c2pa.org

附录C：各指标评分依据与数据来源

本附录披露各维度子指标的核心打分依据及对应数据来源，提升评分透明度。 威胁度（T）指标评分依据 | 指标 | 评分锚点 | 核心数据来源 | | --- | --- | --- | | T1 生成门槛 | 1分：专业团队+高成本（2020年前）；3分：提示词工程/开源部署+低成本（2024-25现状）；5分：零基础免费批量生成（娱乐场景已实现） | Deloitte分析（克隆语音30秒→30分钟缩减）；Bright Defense（$1 Biden深伪电话）；暗网工具$20起 | | T2 逼真度 | 1分：肉眼可辨；3分：需仔细观察；5分：专业鉴伪师也难区分（金融实时视频会议已达此水平） | iProov 2025（人类识别率0.1%）；2026深伪检测状态报告（CNN真实场景精度下降15%+） | | T3 扩散速率 | 1分：小众传播；3分：中等跨平台；5分：跨平台病毒式（政治/新闻场景：算法放大+多语种AI生成） | WEF 2026风险报告（错误信息与AI滥用相关系数0.87）；Europol 90%在线内容预测 | | T4 攻击多样性 | 1分：单一类型；3分：2-3种；5分：多模态跨渠道组合（金融场景：语音+视频+邮件+文件+会议伪造） | Pindrop 2025（金融深伪+1,300%）；香港2亿港元案（实时视频会议伪造） |

脆弱度（V）指标评分依据 | 指标 | 评分锚点 | 核心数据来源 | | --- | --- | --- | | V1 人群识别力 | 1分：多数能识别；3分：部分能识别；5分：几乎无法识别（社交紧急场景：信任惯性+时间压力） | iProov 2025（仅0.1%全部正确）；Fortune 500测试（12人中11人上当） | | V2 信任惯性 | 1分：普遍质疑；3分：选择性信任；5分：几乎无条件信任（社交：亲情信任；金融：科层服从） | 社会心理学文献（确认偏误、权威服从）；中国反诈中心（"AI冒充熟人"重点预警） | | V3 损失放大 | 1分：个人小额；3分：企业中等；5分：社会系统性风险（政治：民主根基；金融：单次$45万均值） | Pindrop（单次攻击平均$450,000）；Deloitte（2027年美国GenAI欺诈$400亿预测） | | V4 高危暴露 | 1分：低频偶发；3分：中等频率；5分：高频日常暴露（中老年即时通讯；青少年流媒体） | 中国国家反诈中心数据；UNESCO MIL素养调查 |

韧性度（R）指标评分依据 | 指标 | 评分锚点 | 核心数据来源 | | --- | --- | --- | | R1 法规完备度 | 1分：无法规；3分：有框架但执法弱；5分：专项立法+执法机制完善 | 中国标识办法+GB45438（gov.cn）；欧盟AI Act第50条；韩国AI基本法；美国FCC裁定文本 | | R2 检测覆盖率 | 1分：无部署；3分：部分部署；5分：全链路实时检测+自动处置 | 各平台透明度报告（Meta、TikTok）；Intel FakeCatcher部署数据；AI4TRUST项目文档 | | R3 标准采纳度 | 1分：无采纳；3分：少数平台自愿采纳；5分：主流平台强制采纳 | C2PA官网成员列表；ITU AMAS标准文档；中国隐式水印产业化部署报告 | | R4 公众素养 | 1分：无系统教育；3分：有项目但未普及；5分：纳入国民教育体系 | UNESCO MIL全球调查；芬兰媒体素养教育政策文献；OECD AI政策观测站 |

说明： 以上为各指标的核心评分锚点和主要数据来源，非穷尽清单。部分指标（如V2信任惯性、V4高危暴露）涉及社会心理学和行为学研究，评分依据为学术文献与官方预警的综合判断，具体文献列表将在正式版方法论附录中完整披露。 编制单位： 天机研究院（International Communication Organization, ICO） 法理框架： 联合国全球数字契约（GDC） | UNESCO媒体与信息素养联盟 免责声明： 本报告基于公开数据和研究团队专业评估编制，仅供参考，不构成任何投资、法律或政策建议。报告中的DFVI评分反映评估时点的最佳判断，可能随数据更新而调整。