一项欧洲科研项目研究如何监管人工智能创造的虚假信息

马德里卡洛斯三世大学（UC3M）参与了一项名为SOLARIS的欧洲研究项目，分析通过人工智能（AI）创建的多媒体内容所涉及的政治风险，从法律领域提出创新的监管方法，从而打击利用该技术创建的虚假新闻和虚假信息。

生成对抗网络（GANs）是一类能够创建近乎真实多媒体内容（音频和视频）的人工智能模型。该技术所面临的主要挑战与所所谓的深度伪造（deep fakes）有关——即所生成的虚假图像或视频能以极高的精确度模拟真实事件从而达到以假乱真的效果。如最近出现的教皇方济各穿着巴黎世家大衣的图片，后来被证明是虚假的。“由于已被用于传播虚假新闻和信息，这项技术构成了紧迫的政治威胁，并对民主治理和监管提出了严峻挑战。提高GANs的问责制、透明度和可靠性已迫在眉睫。”SOLARIS 项目的研究人员之一、马德里卡洛斯三世大学（UC3M）国家公共法系教师，Antonio Estella de Noriega表示。

这个在欧洲研发与创新（I+D+i）框架下的联合团队将用两种方式应对挑战：一方面，通过分析与这些技术相关的政治风险来防止对欧盟民主体制可能产生的负面影响。为此，须建立新的监管机制来检测和减轻深度造假带来的风险。另一方面，团队将评估GANs所创造的重振公民对民主参与的机会。

为了实现这一目标，团队将在SOLARIS项目框架下进行三个案例研究：一、通过实验室条件下的对照实验，了解 GANs 感知可靠性的心理方面；二、通过模拟GANs在社交网络上传播威胁性内容，检测风险并设计缓解策略；三、通过共同创建基于正确价值观的GANs内容，提高对全球关键民主问题（如气候变化、性别维度或人类移民）的认知。

“GANs也是提升民主意识，扩大积极、包容性公民身份的机会。”同时任职于UC3M 欧洲经济治理法让·蒙内特（Jean Monnet）机构特设教授的Antonio Estella de Noriega 表示并补充：“在这方面，GANs同样可以用于做好事，并在新闻、历史和法律等领域发挥积极作用。”

UC3M进行的研究是项目的监管部分，其中包括对该领域的法律法规提出建议。 “最大的挑战恰恰是变化发生的速度极快。 本质上来说，法律只能对现实的某种冻结图像起作用。一般来说，现实的发展速度不如人工智能。”他解释并补充说明：“我们今天能监管的内容可能在六到九个月后因为失去时效性变得毫无价值。”

SOLARIS项目（通过基于价值观的生成对抗网络加强民主参与）(Strengthening demOcratic engagement through vaLue-bAsed geneRative adversarIal networkS) 由欧盟的欧洲研究和创新地平线计划 （GA 101094665）提供近 300 万欧元的研究经费。该项目在2023年至2026年间进行，由阿姆斯特丹大学协调，参与项目的有来自阿尔巴尼亚、德国、保加利亚、斯洛文尼亚、西班牙、意大利、荷兰和英国的十几个学术机构和私营企业。

更多信息：

SOLARIS项目网页： https://projects.illc.uva.nl/solaris/

视频：https://youtu.be/JLKsGXdoe78