European Parliament

Tackling deepfakes in European policy

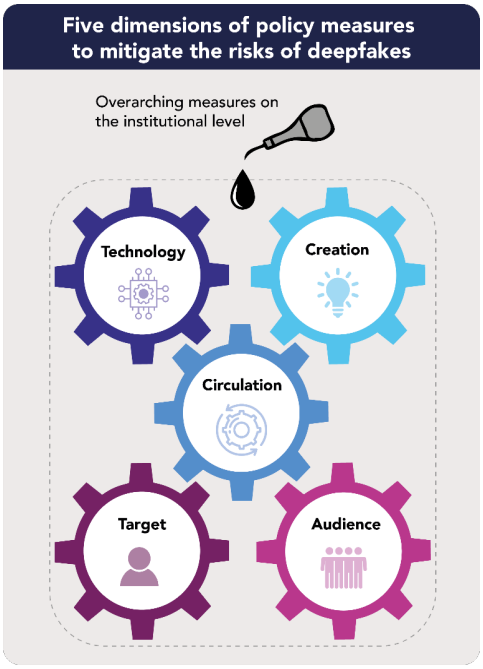
Novel artificial intelligence (AI) and other contemporary digital advances have given rise to a new generation of manipulated media known as deepfakes. Their emergence is associated with a wide range of psychological, financial and societal impacts occurring at individual, group and societal levels. The Panel for the Future of Science and Technology (STOA) requested a study to examine the technical, societal and regulatory context of deepfakes and to develop and assess a range of policy options, focusing in particular upon the proposed AI (AIA) and digital services acts (DSA), as well as the General Data Protection Regulation (GDPR). This briefing summarises the policy options developed in the study. They are organised into five dimensions – technology, creation, circulation, target and audience – and are complemented by some overarching institutional measures.

Technology dimension

The technology dimension concerns the underlying technologies and tools that are used to generate deepfakes, and the actors that develop deepfake production systems.

Clarify which AI practices should be prohibited under the AIA: The proposed AIA mentions four types of prohibited AI practices that could relate to certain applications of deepfake technology. However, the formulation of these sections is open to interpretation. Some deepfake applications appear to fulfil the criteria of high-risk applications, such as enabling deceptive manipulation of reality, inciting violence or causing violent social unrest.

Create legal obligations for deepfake technology providers: As proposed, the AIA would not oblige technology providers to label deepfake content, so the responsibility for labelling deepfakes currently lies with deepfake creators. The AIA could be extended to oblige the producers of deepfake creation tools to incorporate labelling features.



Source: Rathenau Instituut

Regulate deepfake technology as high risk: Deepfake applications of AI could be defined as high risk by including them in annex III of the proposed AIA. This may be justified by risks to fundamental rights and safety, a criterion that is used to determine whether AI systems are high risk. Doing so would place explicit legal requirements on the providers of deepfake technologies, including risk-assessment, documentation, human oversight and ensuring high-quality datasets.

Limit the spread of deepfake detection technology: While detection technology is crucial in halting the circulation of malicious deepfakes, knowledge of how they work can help deepfake producers to circumvent detection. Limiting the diffusion of the latest detection tools could give those that possess them an advantage in the 'cat-and-mouse game' between deepfake production and detection. However, limiting detection technology to too narrow a group of actors could also restrict others from legitimate use.

Invest in the development of AI systems that restrict deepfake attacks: While technology solutions cannot address all deepfake risks, mechanisms such as Horizon Europe could be mobilised to invest in the development of AI systems that prevent, slow, or complicate deepfake attacks.

Invest in education and raise awareness amongst IT professionals: Familiarity with the impacts of deepfakes (and other AI applications) could become a standard part of the curriculum for information technology professionals, in particular AI researchers and developers. This may also provide an opportunity to equip them with a greater understanding and appreciation of the ethical and societal impacts of their work, as well as the legal standards and obligations in place.

Creation dimension

While the technology dimension concerns the production of deepfake generation systems, the creation dimension concerns those that actually use such systems to produce deepfakes. Those that do so for malicious purposes may actively evade identification and enforcement efforts.

Clarify the guidelines for labelling: While standardised labels may help audiences to identify deepfakes, the proposed AIA does not state what information should be provided in the labels, or how it should be presented.

Limit the exceptions for the deepfake labelling requirement: The proposed AIA places a labelling obligation on users of deepfake technology. However, it also creates exemptions when deepfakes are used for law enforcement, in arts, sciences, and where the use 'is needed for freedom of expression'. Liberal interpretation of these exceptions may allow many deepfakes to remain un-labelled.

Ban certain applications: Transparency obligations alone may be insufficient to address the severe negative impacts of specific applications of deepfakes such as non-consensual deepfake pornography or political disinformation campaigns. While an outright ban may be disproportionate, certain applications could be prohibited, as seen in some jurisdictions including the United States of America, the Netherlands and the United Kingdom. Given the possible strong manipulative effect of deepfakes in the context of political advertising and communications, a complete moratorium on such applications could be considered. However, any such bans should be sensitive to potential impacts upon freedom of expression.

Diplomatic actions and international agreements: The use of disinformation and deepfakes by foreign states, intelligence agencies and other actors contributes to increasing geopolitical tension. While some regional agreements are in place, there are no binding global agreements to deal with information conflicts and the spreading of disinformation. Intensified diplomatic actions and international cooperation could help to prevent and de-escalate such conflicts, and economic sanctions could be considered when malicious deepfakes are traced back to specific state actors.

Lift some degree of anonymity for using online platforms: Anonymity serves as protection for activists and whistle-blowers, but can also provide cover for malicious users. Users of online platforms in China need to register with their identity (ID). If some degree of platform anonymity is considered essential, more nuanced approaches could require users to identify themselves before uploading certain types of content, but not when using platforms in other ways.

Invest in knowledge and technology transfer to developing countries: The negative impacts of deepfakes may be stronger in developing countries. Embedding deepfake knowledge and technology transfer into foreign and development policies could help improve these countries' resilience.

Circulation dimension

Policy options in the circulation dimension are particularly relevant in the context of the proposed DSA, which provides opportunities to limit the dissemination and circulation of deepfakes and, in doing so, to reduce the scale and the severity of their impact.

Detecting deepfakes and authenticity: Platforms and other intermediaries could be obliged to embed deepfake detection software and enforce labelling, or to detect the authenticity of users to counteract amplification in the dissemination of deepfakes and disinformation.

Establish labelling and take-down procedures: Platforms could be obliged to label content detected as a deepfake and to remove it when notified by a victim or trusted flagger. This could be done transparently, under human oversight, and with proper notification and appeal procedures. A distinction could be made between reporting by any person and reporting by persons directly affected.

Limit platforms' decision-making authority decide unilaterally on the legality and harmfulness of content: Independent oversight of content moderation decisions could limit the influence of platforms on freedom of expression and the quality of social communication and dialogue.

Increase transparency: To support monitoring activities, the DSA's reporting obligations could be extended to include deepfake detection systems, their results and any subsequent decisions.

Slow the speed of circulation: While freedom of speech is a fundamental right, freedom of reach is not. Platforms could be obliged to slow the circulation of deepfakes by limiting the number of users in groups, the speed and dynamics of sharing patterns, and the possibilities for micro-targeting.

Target dimension

Malicious deepfakes can have severe impacts on targeted individuals, and these may be more profound and long-lasting than many traditional patterns of crime.

Institutionalise support for victims of deepfakes: National advisory bodies could provide accessible judicial support to help victims ensure take-downs, identify perpetrators, launch civil or criminal proceedings, and access psychological support. They could also contribute to the long-term monitoring of deepfakes and their impacts.

Strengthen the capacity of data protection authorities (DPAs) to respond to the use of personal data for deepfakes: Since deepfakes tend to make use of personal data, DPAs could be equipped with specific resources to respond to the challenges they raise.

Provide guidelines on GDPR in the context of deepfakes: DPAs could develop guidelines on how the GDPR framework applies to deepfakes, including the circumstances in which a data protection impact assessment is required and how freedom of expression should be interpreted in this context.

Extend the list of special categories of personal data with voice and facial data: The GDPR could be extended to include voice and facial data, to specify the circumstances under which their use is permitted and clarify how freedom of expression should be interpreted in the context of deepfakes.

Develop a unified approach for the proper use of personality rights. Personality rights are comprised of many different laws including various rights of publicity, privacy and dignity. The 'right to the protection of one's image' could be developed and clarified in light of deepfake developments.

Protect personal data of deceased persons. Deepfakes can present deceased persons in misleading ways without their consent. A 'data codicil' could be introduced to help people control how their data and image is used after their death.

Address authentication and verification procedures for court evidence: Various types of digital evidence, such as electronic seals, time stamps and electronic signatures, have been established as admissible as evidence in legal proceedings. Guidelines could be provided to help address authentication and verification issues and support courts when dealing with digital evidence of questionable authenticity.

Audience dimension

Audience response is a key factor in the extent to which deepfakes can transcend the individual level and have wider group or societal impacts.

Establish authentication systems: In parallel to labelling measures, authentication systems could help recipients of messages to verify their authenticity. These could require raw video data, digital watermarks or information to support traceability.

Invest in media literacy and technological citizenship: Awareness and literacy of deepfake technologies could increase the resilience of citizens, organisations and institutions against the risks of deepfakes. These could target different profiles, such as young children, professionals, journalists and social media users.

Invest in a pluralistic media landscape and high quality journalism: The European democracy action plan recognised a pluralistic media landscape as a prerequisite for access to truthful information, and to counter disinformation. Support for journalism and media pluralism at European and national levels could help maintain this.

Institutional and organisational measures

Overarching options for institutional and organisational action could support and complement measures in all five dimensions discussed above.

Systematise and institutionalise the collection of information with regards to deepfakes: Systemic collection and analysis of data about the development, detection, circulation and impact of deepfakes could inform the further development of policies and standards, enable institutional control of deepfake creation, and may even transform deepfake creation culture. This option corresponds with the European democracy action plan, the European action plan against disinformation and the European Digital Media Observatory that is currently being formed. The European Union Agency for Cybersecurity (ENISA) and European Data Protection Board (EDPB) could also play a role in this.

Protecting organisations against deepfake fraud: Organisations could be supported to perform risk assessments for reputational or financial harm caused by malicious deepfakes, to prepare staff and establish appropriate strategies and procedures.

Identify weaknesses and share best practices: Assessments of national regulations in the context of deepfakes could reveal weaknesses to be addressed, as well as best-practices to be shared. An EU-wide comparative study could be promoted within the framework of Horizon Europe.

This document is based on the STOA study 'Tackling deepfakes in European policy' (PE 690.039) published in July 2021. The study was written by Mariëtte van Huijstee, Pieter van Boheemen and Djurre Das (Rathenau Institute), Linda Nierling and Jutta Jahnel (Institute for Technology Assessment and Systems Analysis), Murat Karaboga (Fraunhofer Institute for Systems and Innovation Research) and Martin Fatun (Technology Centre ASCR), with the assistance of Linda Kool (Rathenau Institute) and Joost Gerritsen (Legal Beetle). It was requested by the Panel for the Future of Science and Technology (STOA) and managed by the Scientific Foresight Unit, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament. STOA administrator responsible: Philip Boucher.

DISCLAIMER AND COPYRIGHT

stoa@ep.europa.eu (contact)
http://www.europarl.europa.eu/stoa/ (STOA website)
www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)