Tackling deepfakes in European policy

- **Rathenau Institute (The Netherlands),**
- **Institute for Technology Assessment and Systems Analysis (Germany),**
- **Fraunhofer Institute for Systems and Innovation Research (Germany),**
- **Technology Centre of the Academy of Sciences (Czech Republic)**

We define *Deepfakes* as ***manipulated or synthetic audio or visual media that seem authentic, which feature people that appear to say or do something they never said or did, produced using artificial intelligence techniques, including machine learning and deep learning***.

TA Study Outline

- Societal context
- **Benefits, risks and impacts**
- Current regulatory landscape
- **Regulatory gaps**
- **Policy options**

Benefits

- Audio graphic productions
- Human-machine interactions
- Video conferencing
- Satire
- Creative expression
- Medical (research) applications

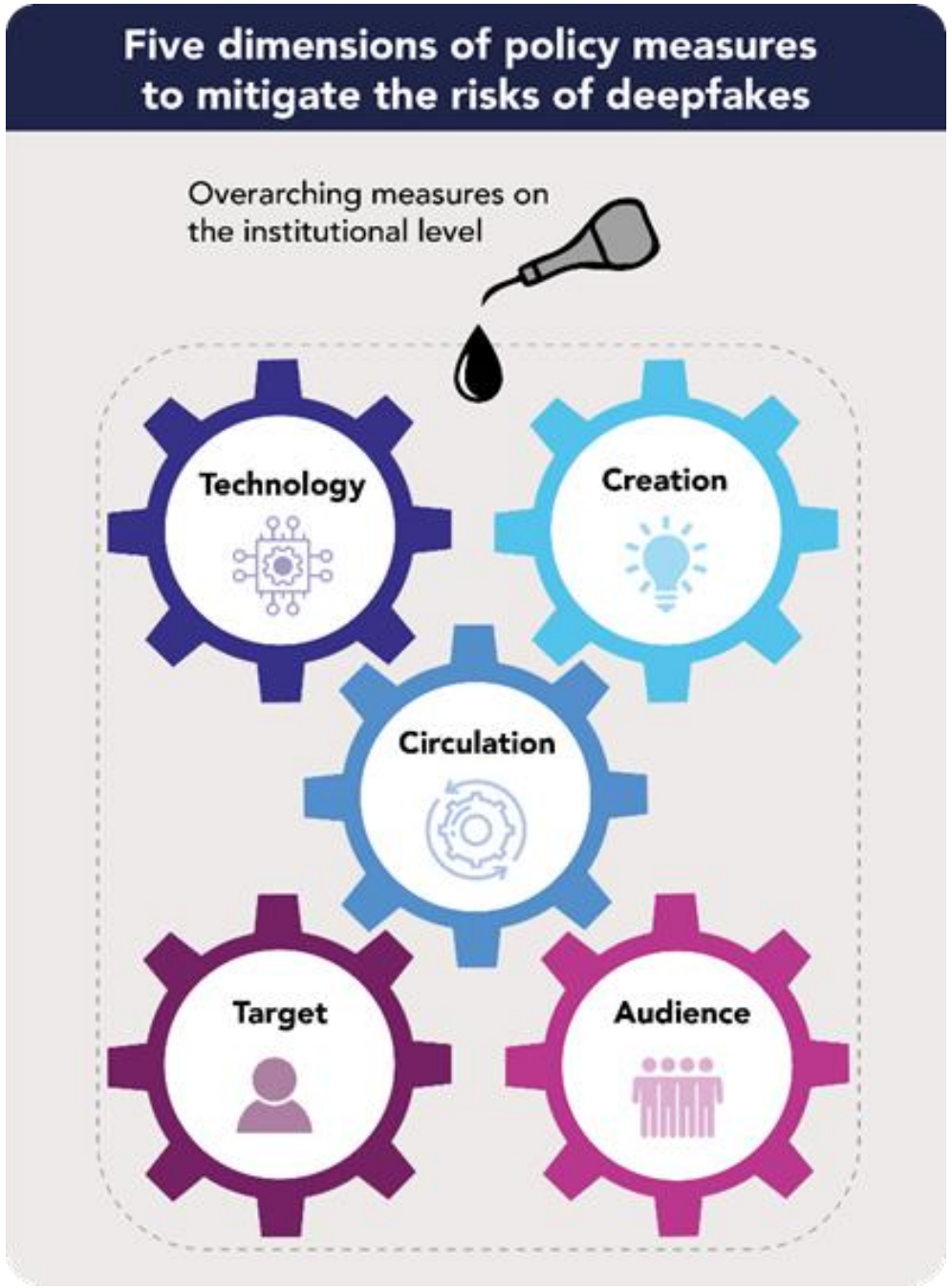
Overview of different categories of **risks** associated with deepfakes

Psychological harm	Financial harm	Societal harm
• (S)extortion • Defamation • Intimidation • Bullying • Undermining trust	• Extortion • Identity theft • Fraud (e.g. insurance/payment) • Stock-price manipulation • Brand damage • Reputational damage	• News Media manipulation • Damage to economic stability • Damage to the justice system • Damage to the scientific system • Erosion of trust • Damage to democracy • Manipulation of elections • Damage to international relations • Damage to national security

Cascading *impacts* of deepfakes

Regulatory gaps

- Victims may be protected on paper, but need support to claim their rights in practice
- Multiple actors involved in the lifecycle of a deepfake:
  - Victim
  - Perpetrator
  - Audience
  - Technology provider
  - Platform users
  - Platforms
- Competing rights and obligations
- Perpetrators often act anonymously
- Platforms play a pivotal role


Five dimensions of policy measures to mitigate the risks of deepfakes
Overarching measures on the institutional level
Technology
Creation
Circulation
Target
Audience

Options for the AI Framework

Technology dimension

- Clarify which AI practices shall be prohibited under the AI Framework
- Create legal obligations for deepfake technology providers
- **Regulate deepfake technology as high risk**
- Place limits on the spread of deepfake detection technology

Creation dimension

- Clarify the guidelines for the manner of labelling
- Limit the exceptions for the deepfake labelling requirement
- Ban certain applications

Options for the DSA

Circulation dimension

- Oblige platforms to have deepfake detection systems in place
- Oblige platforms to have systems in place to detect authenticity
- Establish labelling and take-down procedures
- Oblige platforms to have an appeal procedure in place
- Limit the decision-making authority of platforms to unilaterally decide on the legality and harmfulness of content
- Increase transparency
- Slow down the speed of circulation

In summary

TA study impact:

- information on state-of-art
- awareness about problem outreach
- analysis of current regulations
- development of policy options
- establishment of new regulation