Online content moderation: the invisible hand of intermediary service providers in the fight against cyberviolence

"Private actors as judges and enforcers in the technology-driven world"

Conference, University of Luxembourg

4 July 2023

Prof. Dr Vanessa Franssen

Outline

Outline

- Introduction: general presentation of the @ntidote research
- Research objectives and methodology
- Challenges encountered
- First results
- Conclusions

Introduction: general presentation of the @ntidote research

Introduction: general presentation of the @ntidote research (1)

- **@ntidote**
  - 2-year research project funded by Belspo (Belgian Science Policy Office
  - Interuniversity
  - Interdisciplinary
    - Law
    - Communications sciences
    - Criminology (psychology)
    - Anthropology
  - https://www.antidoteproject.be/

Introduction: general presentation of the @ntidote research (2)

- **@ntidote (cont'd)**
  - Two forms of cyberviolence
    - Online hate speech (OHS)
    - Non-consensual distribution of intimate images (NCII)
  - Target population:
    - 'Digital natives' (15-25y)
    - Online service providers (OSPs)

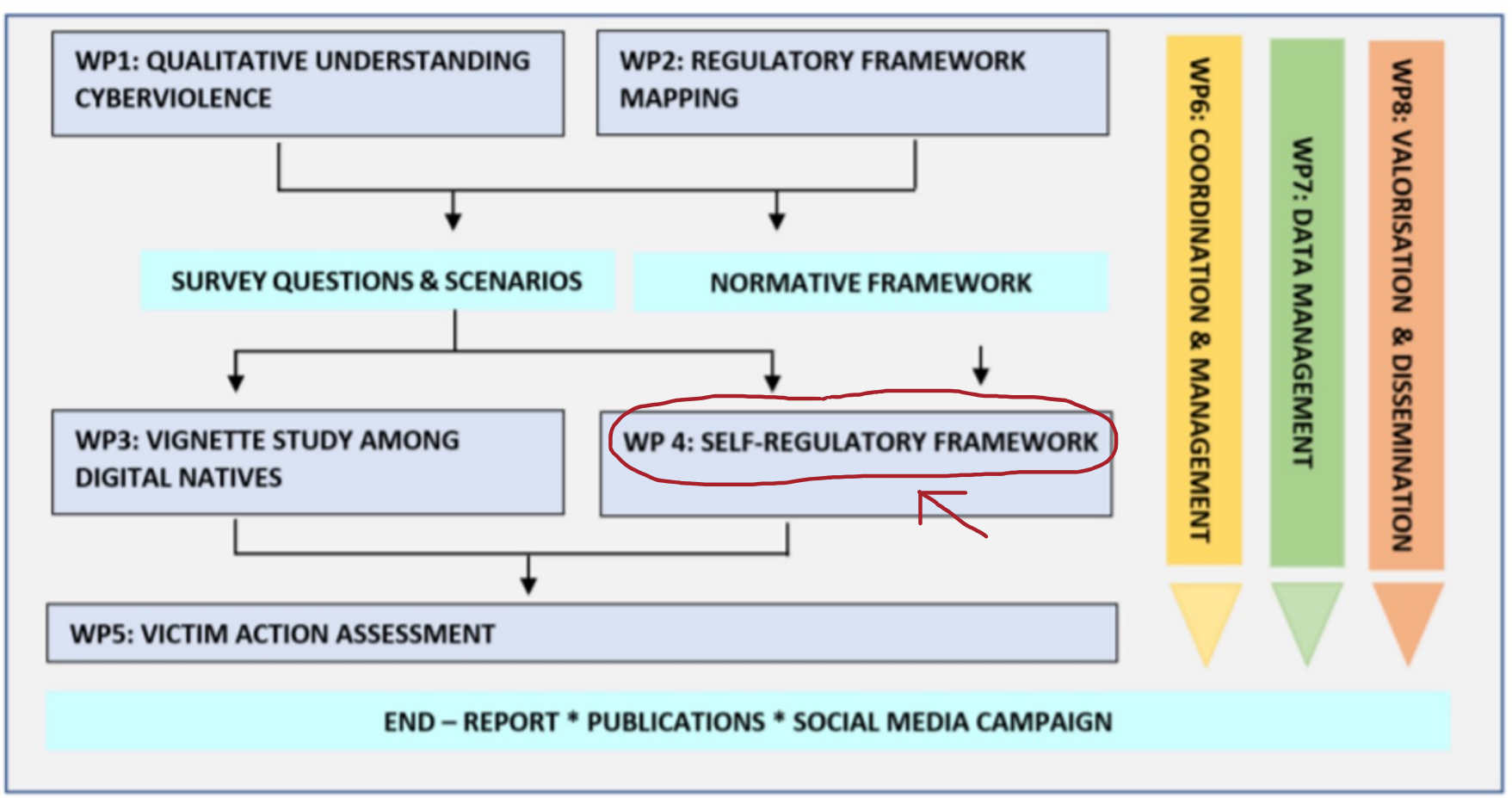
Introduction: general presentation of the @ntidote research (3)

- **@ntidote (cont'd)**
  - Objectives
    - Better understanding phenomena
      - Perception of permissible/harmful behaviour
      - Prevalence
    - How these phenomena are or can be tackled
      - Legally
        - Legislation
        - Case law
      - Self-regulation and soft law
      - Coping strategies of victims

Incl. liability OSPs

Online content moderation

Introduction: general presentation of the @ntidote research (4)



Research objectives and methodology

Research objectives and methodology (1)

- Objectives
  - **Assessment of the self-regulatory framework of selected OSPs**
    - Term ‘online service providers’
    - Self-regulatory framework = community rules, terms of service, guidelines, policies, transparency reports, etc.
  - **Delineation by OSPs of (im)permissible online behaviour**
    - Proactive & reactive -> content moderation, both human and technical
  - **Link with WP2: analysis and implementation of liability of OSPs under EU law**

Research objectives and methodology (2)

- **Methodology**
  - **Analysis of legal framework (link with WP2)**
  - **Roundtable with industry**
    - Companies & associations
    - Future possibilities to tackle online hate speech (OHS) and NCII
    - Role of industry and cooperation with LEAs
    - Technical tools

Research objectives and methodology (3)

- **Methodology (cont'd)**
  - **Literature study (legal, social sciences)**
  - **Survey with moderators/OSPs: questionnaire with scenarios**
    - **Collect data** on the permissibility of behaviours
    - **Map criteria decisive for assessment** of behaviour as permissible
    - **Assessment of technical solutions** to remove and prevent content in the light of the normative framework developed under WP2
  - **Analysis of self-regulatory framework**
    - Selection OSPs
    - "Coding technique" -> analytical grid
      - Eg general information on self-regulatory framework, definition cyberviolence, information on moderators, proactive and reactive content moderation, follow-up + transparency

Challenges encountered

Challenges encountered

- **Delineation** of research
  - BE market → EU market
  - Selection of OSPs: various criteria
  - Definition of phenomena
- **Design** of survey
  - Traditional issues: confidentiality, length, clarity, potential biases, incomplete answers, methodological consistency…
  - Comparability with WP3 (perception of digital natives)
- **Intensive 'recruitment' process**
  - Contacting moderators
    - Various strategies
  - Industry's willingness to cooperate

First results

First results (1)

- **Survey**
  - **Sample limited** (13 moderators + 2 companies)
  - **Confidentiality** = major hurdle

- **Questionnaire**
  - **Profile** of moderators: great diversity
    - Gender: good balance
    - Age
    - Language
    - Qualifications: higher education, various disciplines
      - Due to recruitment process?
    - Recruitment companies/online platforms
    - Current and former moderators
    - >< literature

First results (2)

- **Questionnaire (cont'd)**
  - **Training** seems to raise few issues
    - Sufficient, some gaps
      - Eg 'easy work'
      - Eg trainer had no moderator experience, insufficient to deal with borderline cases, 'in a hurry to start'
    - >< literature
  - **Feedback**
    - Mostly focused on quality(!)
  - **Main objective(s) of moderation?**
    - Various: to protect users against (the most) harmful content, to create a safe online environment, to respect the law, to reply to users flagging illegal, harmful or disturbing content, to respect the platform's policy rules, to safeguard the platform's reputation
    - >< (some) literature

First results (3)

- **Questionnaire (cont'd)**
  - **Moderation process**
    - How does the content you have to moderate end up on your desk?
      - Almost general use of AI tools
      - Flagged by
        - Users (frequent)
        - Non-professional content moderator (in some cases)
        - Law enforcement authorities (LEAs) (in some cases)
    - **Time and volume** to moderate
      - Variety
      - Depends on type of content

First results (4)

- **Questionnaire (cont'd)**
  - **Moderation process (cont'd)**
    - During the moderation process, what happens with the content pending the decision?
      - Remains online, temporarily removed, tagged
      - Depends on the type of content (according to some respondents)
      - Several respondents: prefer not to answer or do not know(!)
    - **Reporting to LEAs** as individual moderator?
      - Several respondents: prefer not to answer

First results (5)

- **Questionnaire (cont'd)**
  - **Challenges?**
    - Quite some diversity!
      - >< literature
    - Tensions between policies and moderators' own perception: seem to be limited
      - Mantra: 'apply the company rules'
    - Time constraints: too limited to take sound decisions
    - Moderators
      - Not enough
      - Rotation
    - Policy rules
      - Application in practice
      - Change daily
    - Language barriers
    - Cultural barriers (limited!)
    - Lack of psychological support (only mentioned by one)

First results (6)

- **Survey - Scenarios**
  - Objective: How do SPs delineate (im)permissible online behaviour?
  - Which scenarios?
    - Borderline cases, based on analysis WP1 (qualitative research) and WP2 (case law)
    - Same as for WP3 (for sake of comparability)
  - But all versions presented
    - Variables
      - OHS: ethnicity, gender
      - NCII: level of nudity, sexual orientation
  - Questions
    - Detection by AI tools? + reaction?
    - Reaction if flagged by user?
    - Reaction if no consent?

First results (7)

- **Scenarios (cont'd)**
  - Some preliminary results
    - Only variable that seems to matter *significantly* is level of nudity
    - Reaction = more than removal
    - User notification (whether victim or another person) matters
    - OHS less easily detected by AI tools than NCII

First results (8)

- **OSPs self-regulatory framework**
  - Again: great variety!
  - First impressions
    - Level of sophistication differs
    - Moderation process differs
    - Level of transparency differs

Conclusions

Conclusions

- Survey gives an interesting insight in moderation process
  - Despite limited sample
  - Quite nuanced
- Further analysis
  - Scenarios
    - Comparison answers moderators with those of OSPs
    - Comparison with WP3
  - OSPs self-regulatory framework
- Basis for future research!

Questions?

vanessa.franssen@uliege.be