University of Oslo

VIROS Symposium

Oslo, 28 September 2022

BEYOND DATA: HUMAN RIGHTS, ETHICAL AND SOCIAL IMPACT ASSESSMENTS IN AI

Prof. Avv. Alessandro Mantelero
Jean Monnet Chair in Mediterranean Digital Societies and Law
Polytechnic University of Turin | Associate Professor

< I >

HRIA in AI-powered machines

Politecnico di Torino

Case study: a story of machines and kids

- ✓ Key elements
  - Role of playing and role-playing in kids' education
  - Emotional interaction with anthropomorphic dolls
  - Dialogue as a channel to suggest behavioral patterns, collect personal information, convey values

Politecnico di Torino

Phase I: Planning and scoping

- Used technology
  - NLP: speech recognition technology
  - AI-based interaction (more than 8,000 lines of dialogue)
  - Cloud-based
  - Data processing (voice-recording tracks)

- Device features
  - Microphone and speakers
  - Wi-Fi connection

Politecnico di Torino

- Right-holders
  - Direct users (minors)
  - Supervisory users (parents, partial remote control)
  - Third parties (e.g. friends of the user or re-users of the doll)

- Main purposes
  - Play
  - Educational
  - Others (limited parental control, testing and service improvement)

- Duty-bearers
  - Manufacturer
  - Third-party service providers (e.g., ML, cloud)

Politecnico di Torino

Phase II: Initial risk analysis and assessment

Mitigation of evident high-risks

- Interaction and data collection
  - Activation process
  - Push-and-hold button
  - Element (doll's necklace) which light up when the device is active
- AI-based NLP
  - Pre-selected dataset of possible answerers
  - No search on Internet

Politecnico di Torino

Risk analysis and assessment

- Use of a questionnaire to support the impact assessment
- Potentially impacted rights
  - Data protection and the right to privacy (dialogues, parental monitoring)
  - Freedom of thought, parental guidance and the best interest of the child (behavioural, cultural and educational influence)
  - Right to psychological and physical safety (cyberattacks, data theft, transmission of inappropriate content, safety)

Politecnico di Torino

Tab. 2 Probability

	Probability	
Low	The risk of prejudice is improbable or highly improbable	1
Medium	The risk may occur	2
High	There is a high probability that the risk occurs	3
Very high	The risk is highly likely to occur	4

Tab. 3 Exposure

	Exposure	
Low	Few or very few of the identified population of rights-holders are potentially affected	1
Medium	Some of the identified population are potentially affected	2
High	The majority of the identified population is potentially affected	3
Very high	Almost the entire identified population is potentially affected	4

Tab. 4 Likelihood table(L)

		Probability			
		1	2	3	4
Exposure	1	1	2	3	4
	2	2	3	5	9
	3	3	5	9	12
	4	4	7	12	15

Politecnico di Torino

Data protection and the right to privacy

- Likelihood of prejudice
  - Risk factors: companion toy, dialogue recording, largely unsupervised interaction, potential data sharing by parents
  - Probability: high
  - Risk factors: all the doll's users are potentially exposed to this risk
  - Exposure: very high
  - Likelihood of prejudice: very high

Tab. 4 Gravity of the prejudice

	Gravity of the prejudice	
Low	Affected individuals and groups may encounter only minor prejudices in the exercise of their rights and freedoms.	1
Medium	Affected individuals and groups may encounter significant prejudices.	2
High	Affected individuals and groups may encounter serious prejudices.	3
Very high	Affected individuals and groups may encounter serious or even irreversible prejudices.	4

Tab. 5 Effort to overcome the prejudice and to reverse adverse effects.

	Effort	
Low	Suffered prejudice can be overcome without any problem (e.g. time spent amending information, annoyances, irritations, etc.)	1
Medium	Suffered prejudice can be overcome despite a few difficulties (e.g. extra costs, fear, lack of understanding, stress, minor physical ailments, etc.).	2
High	Suffered prejudice can be overcome albeit with serious difficulties (e.g. economic loss, property damage, worsening of health, etc.).	3
Very high	Suffered prejudice may not be overcome (e.g. long-term psychological or physical ailments, death, etc.).	4

Tab. 6 Severity table (S)

		Gravity			
		1	2	3	4
Effort	1	1	2	4	6
	2	2	3	5	8
	3	3	5	8	10
	4	5	8	10	12

Politecnico di Torino

- Severity
  - Risk factors: subjects involved (young children and minors), processing of personal data in several areas, sensitive information, unexpected findings, transborder data flows
  - Gravity of the prejudice: high
  - Risk factors: potential parental supervision and remote control, data security measures (e.g. data erasure, dialogue with the minor in case of unexpected findings).
  - Effort to overcome potential prejudice/to reverse adverse effects: medium
  - Severity: medium

Tab. 8. Overall risk impact table

		Severity [impacted right/freedom]			
		Low	Medium	High	Very high
Likelihood	Low				
	Medium				
	High				
	Very high				

- If the likelihood of prejudice can be considered very high and the severity medium, the overall impact is high

Politecnico di Torino

Freedom of thought, parental guidance and the best interest of the child

- Likelihood: medium
  - Risk factors: limited number of value-oriented statement (e.g., “It’s so cool that you want to be a mom someday”)
  - Probability: medium
  - Risk factors: values commonly accepted in the target cultural context (including value-oriented notion of inappropriate questions)
  - Exposure: medium
- Severity: low
  - Risk factors: not particularly controversial value-laden sentences
  - Gravity of prejudice: low
  - Risk factors: talking with children can mitigate potential harm
  - Effort: low
- Overall impact: medium

Politecnico di Torino

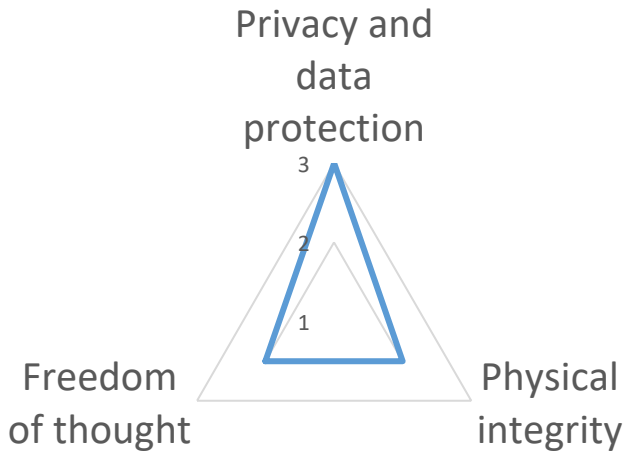
Right to psychological and physical safety

- Likelihood: low
  - Risk factors: limited interest in malicious attacks (e.g., harassment, stalking, insults, confidence loss, bullying), but easy access to the toy
  - Probability: medium
  - Risk factors: use of the toy mainly in safe environment
  - Exposure: low
- Severity: medium
  - Risk factors: young age of the users, attacks only through verbal instructions
  - Gravity of prejudice: medium
  - Risk factors: parent-child dialogue and technical solutions can combat the potential prejudice
  - Effort: medium
- Overall impact: medium

Politecnico di Torino

Results of the Initial Assessment

Risk	L	S	Overall impact
Impact on privacy and data protection	VH	M	H
Impact on freedom of thought	M	L	M
Impact on the right to psychological and physical safety	L	M	M



1 Low impact

2 Medium impact

3 High impact

Phase III: Mitigation measures and re-assessment

- Data protection and the right to privacy
  - Default setting: deliberate action to activate AI-based information processing/dialogue functions
  - Unexpected content: accurate selection of conversation topics (closed set of sentences, possibility for parents to modify phrases/questions), policy for unexpected findings
  - Content: no conversation monitoring, invidividual testing phases only in a laboratory setting, possibility for parents to delete stored information
  - Data security: stronger authentication and encryption solutions

Politecnico di Torino

Exposure: reduced to low (prejudices only in special circumstances, e.g. malicious attack)
Probability: reduced to low (reduction of risk relating to data collection/retention)
Likelihood: reduced to low

Gravity: lowered to medium (mitigation measures)
Effort: it remains medium (risk of hacking)
Severity: lowered somewhat, though remaining medium

Overall impact: lowered from high to medium

Politecnico
di Torino

- Freedom of thought, parental guidance and the best interest of the child
  - NLP: pre-set database, no Internet, content fine-tuned to the education level of the user
  - Transparency: visualisation of embedded values (logic and content maps)
  - Values/content: user-customisable (critical topics), stereotype prevention by default
  - Design team: diversity

  Exposure: no change, medium (variety of cultural contexts, need for an active role of parents)
  Probability: lowered to low (product design and customization)
  Likelihood: lowered to low

  Severity: no change, low (now more responsible content management)

  Overall impact: lowered from medium to low

Politecnico di Torino

- Right to psychological and physical safety
  - Risk of malicious hacking activities: exclusion of interaction with other IoT devices, strong authentication and data encryption

  Exposure: no change (low)
  Probability: reduced to low (protection measures adopted)
  Likelihood: it remains low but is lowered

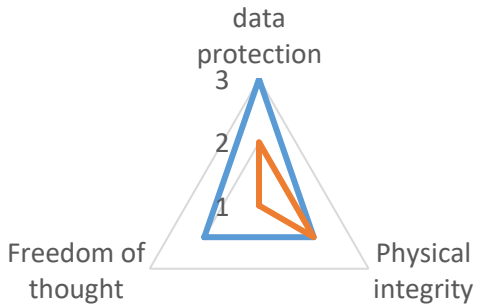
  Gravity: no impact (medium)
  Effort: no impact (medium)
  Severity: it remains medium

  Overall impact: no change, medium (malicious hacking is the most critical aspect)

Politecnico di Torino

- Assessment (effects of measures adopted)

Risk	L	S	Overall impact	MMs	rL	rS	Final impact
Impact on privacy and data protection	VH	M	H	Yes	M	M	M
Impact on freedom of thought	M	L	M	Yes	L	L	L
Impact on the right to psychological and physical safety	L	M	M	Yes	L	M	M
Overall impact (all impacted areas)			M/H				M/L



1 Low impact

2 Medium impact

3 High impact

Politecnico di Torino

< II >

Ethical and Social Impact Assessments

Politecnico di Torino

Results of the empirical analysis (AI companies)

- 1st group

active role of ethics committees in the companies' business (internal procedures, tasks and companies committed to taking the committees' input into account)

- 2nd group

concrete interaction and impact on company decisions not documented

- 3rd group

unknown identity of committee members, general description of the main purpose of the committees

Politecnico di Torino

Available models

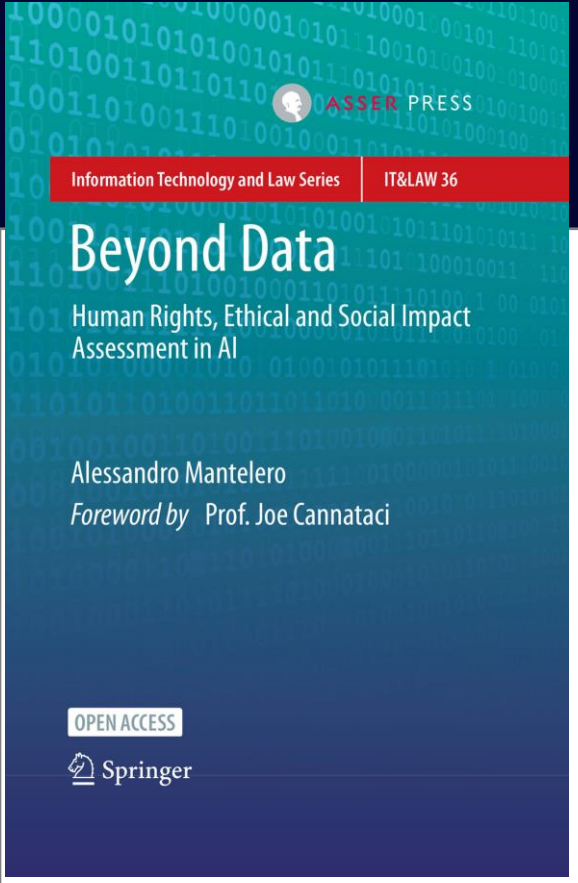
- Variety of structures (internal/external committees)
- Variety of tasks (guidelines, advice on specific products/services, policies, etc.)
- Key role of independence and reputation of committee members
- Tension between human rights and corporate principles/values
- Need for greater transparency about the structure and functioning, including their impact on the decision-making processes of companies
- Accountability for decisions based on committee recommendations
- Important role for internal requests (critical issues/cases) and role of internal ethics officers

Politecnico di Torino

The role of expert committees in AI (HRESIA)

- Contextualisation human rights
- Integrating HRIA with respect to contextual ethical and social values (community values, acceptability and substitutability of proposed AI solutions)
- No one-size-fits-all model
- Key elements
  - Independence
  - Reputation of committee members
  - Effectiveness
  - Transparency
  - Accountability
  - Stakeholder and rightsholder engagement

Politecnico di Torino



- ✓ Beyond Data : Rise and Fall of Individual Sovereignty Over Data Use
- ✓ A Paradigm Shift: The Focus on Risk Assessment
- ✓ The HRESIA model : Human Rights, Ethical, and Social Impact Assessment
- ✓ HRIA in AI
- ✓ The Social and Ethical Component in AI Systems Design
- ✓ Impact assessment in AI regulating: a missing piece
- ✓ Open Issues

Politecnico di Torino

Open Access
https://link.springer.com/book/10.1007/978-94-6265-531-7


Photo by Kimson Doan on Unsplash

Alessandro Mantelero

alessandro.mantelero@polito.it

@mantelero

Politecnico di Torino

Jean Monnet Chair in Mediterranean Digital Societies and Law

Co-funded by the European Union

Project: 101047818 - DIGIMED - ERASMUS-JMO-2021-CHAIR