Chapter 2
Human Rights Impact Assessment and AI

Contents

Abstract The recent turn in the debate on AI regulation from ethics to law, the wide application of AI and the new challenges it poses in a variety of fields of human activities are urging legislators to find a paradigm of reference to assess the impacts of AI and to guide its development. This cannot only be done at a general level, on the basis of guiding principles and provisions, but the paradigm must be embedded into the development and deployment of each application. To this end, this chapter suggests a model for human rights impact assessment (HRIA) as part of the broader HRESIA model. This is a response to the lack of a formal methodology to facilitate an ex-ante approach based on a human-oriented design of AI. The result is a tool that can be easily used by entities involved in AI development from the outset in the design of new AI solutions and can follow the product/service throughout its lifecycle, providing specific, measurable and comparable evidence on potential impacts, their probability, extension, and severity, and facilitating comparison between possible alternative options.

A. Mantelero, *Beyond Data*, Information Technology and Law Series 36,
https://doi.org/10.1007/978-94-6265-531-7_2

Keywords Data ethics · Democracy · Human rights by design · Human Rights Impact Assessment · Participation · Precautionary principle · Smart cities · Smart toys · Transparency

2.1 Introduction

The debate that has characterised the last few years on data and Artificial Intelligence (AI) has been marked by an emphasis on the ethical dimension of the use of data (data ethics)[1] and by a focus on potential bias and risk of discrimination.[2]

While data processing regulation has been focused for decades on the law, including the interplay between data use and human rights, this debate on data-intensive AI systems has rapidly changed its trajectory, from law to ethics.[3] This is evident not only in the literature,[4] but also in the political and institutional discourse.[5] In this regard, an important turning point was the European Data Protection Supervisor (EDPS) initiative on digital ethics[6] which led to the creation of the Ethics Advisory Group.[7]

As regards the debate on data ethics, it is interesting to consider two different and chronologically consecutive stages: the academic debate and the institutional initiatives. These contributions to the debate are different and have given voice to different underlying interests.

The academic debate on the ethics of machines is part of the broader and older reflection on ethics and technology. It is rooted in known and framed theoretical models, mainly in the philosophical domain, and has a methodological maturity. In contrast, the institutional initiatives are more recent, have a non-academic nature and aim at moving the regulatory debate forward, including ethics in the sphere of data protection. The main reason for this emphasis on ethics in recent years has been the growing concern in society about the use of data and new data-intensive applications, from Big Data[8] to AI.

[1] Floridi et al. 2018; Mittelstadt et al. 2016.

[2] Wachter et al. 2021; Algorithm Watch 2020; Myers West et al. 2019, p. 33; Zuiderveen Borgesius 2020; Mann and Matzner 2019.

[3] Raab 2020, para 3; Bennett and Raab 2018.

[4] E.g. Floridi and Taddeo 2016.

[5] In the context of the legal debate on computer law, at the beginning of the last decade only a few authors focused on the ethical impact of IT, e.g. Wright 2010. Although the reflection on ethics and technology is not new in itself, it has become deeper in the field of data use where new technology development in the information society has shown its impact on society. See also Verbeek 2011; Spiekermann 2016; Bohn et al. 2005, pp. 19–29.

[6] European Data Protection Supervisor 2015b.

[7] European Data Protection Supervisor 2015a.

[8] Council of Europe, Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108) 2017.

Although similar paths are known in other fields, the shift from the theoretical analysis to the political arena represents a major change. The political attention to these issues has necessarily reduced the level of analysis, ethics being seen as an issue to be flagged rather than developing a full-blown strategy for ethically-oriented solutions. In a nutshell, the message of regulatory bodies to the technology environment was this: law is no longer enough, you should also consider ethics.

This remarkable step forward in considering the challenges of new paradigms had the implicit limitation of a more general and basic ethical framework, compared to the academic debate. In some cases, only general references to the need to consider ethical issues has been added to AI strategy documents, leaving the task of further investigation to the recipients of these documents. At other times, as in the case of the EDPS, a more ambitious goal of providing ethical guidance was pursued.

Methodologically, the latter goal has often been achieved by delegating the definition of guidelines to committees of experts, including some forms of wider consultation. As in the tradition of expert committees, a key element of this process is the selection of experts.

These committees were not only composed of ethicists or legal scholars but had a different or broader composition defined by the appointing bodies.[9] Their heterogeneous nature made them more similar to multi-stakeholder groups.

Another important element of these groups advising policymakers concerns their internal procedures: the actual amount of time given to their members to deliberate, the internal distribution of assigned tasks (in larger groups this might involve several sub-committees with segmentation of the analysis and interaction between sub-groups), and the selection of the rapporteurs. These are all elements that have an influence in framing the discussion and its results.

All these considerations clearly show the differences between the initial academic debate on ethics and the same debate as framed in the context of institutional initiatives. Moreover, this difference concerns not only structure and procedures, but also outcomes. The documents produced by the experts appointed by policymakers are often minimalist in terms of theoretical framework and focus mainly on the policy message concerning the relevance of the ethical dimension.

The variety of the ethical approaches, the lack of clear indications on the frame of reference or the reasons for preferring a certain ethical framework make it difficult to understand the key choices on the proposed ethical guidelines.[10] Moreover, the local perspective of the authors of these documents, in line with the context-dependent nature of ethical values, undermines the ambition to provide global standards or, where certain values are claimed to have general relevance, may betray a risk of ethical colonialism.

[9] This is the case, for example, of the Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission, which brought together 52 experts, the majority (27) from industry and the rest from academia (15, including 3 with a legal background and 3 with an ethical background), civil society (6) and governmental or EU bodies (4). See also Access Now 2019; Veale 2020.

[10] Ienca and Vayena 2020.

These shortcomings that characterise a purely ethical discourse on AI regulation – which are analysed in more detail in Chap. 3 – lead us to turn our gaze towards more well-established and commonly accepted frameworks such as that provided by human rights, the implementation of which in the field of AI is discussed in the following sections.

2.2 A Legal Approach to AI-Related Risks

In considering the impact of AI on human rights, the dominant approach in many documents is mainly centred on listing the rights and freedoms potentially impacted[11] rather than operationalising this potential impact and proposing assessment models.

However, case-specific assessment is more effective in terms of risk prevention and mitigation than using risk presumptions based on an abstract classification of high-risk sectors or high-risk uses/purposes, where sectors, uses and purposes are very broad categories which include different kind of applications – some of them continuously evolving – with a variety of potential impacts on rights and freedoms that cannot be clustered *ex ante* on the basis of risk thresholds, but require a case-by-case impact assessment.[12]

Similarly, the adoption of a centralised technology assessment carried out by national ad hoc supervisory authorities[13] can provide useful guidelines for technology development and can be used to fix red lines[14] but must necessarily be complemented by a case-specific assessment of the impact of each application developed.

For these reasons, a case specific impact assessment remains the main tool to ensure accountability and the safeguarding of individual and collective rights and freedoms. In this regard, a solution to the problem could easily be drawn from the human rights impact assessment models already adopted in several fields.

However, these models are usually designed for different contexts than those of AI applications.[15] The latter are not necessarily large-scale projects involving entire

[11] Raso et al. 2018; Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission 2019; Council of Europe, Committee of Ministers 2020; Council of Europe 2018.

[12] Chapter 4, Sect. 4.3.2.

[13] European Parliament 2020, Article 14.2 ("the risk assessment of artificial intelligence, robotics and related technologies, including software, algorithms and data used or produced by such technologies, shall be carried out, in accordance with the objective criteria provided for in paragraph 1 of this Article and in the exhaustive and cumulative list set out in the Annex to this Regulation, by the national supervisory authorities referred to in Article 18 under the coordination of the Commission and/or any other relevant institutions, bodies, offices and agencies of the Union that may be designated for this purpose in the context of their cooperation") and Chap. 4, Sect. 4.3.2.

[14] On the debate on the adoption of specific red lines regarding the use of AI in the field of facial recognition, European Digital Rights (EDRi) 2021. See also Chap. 4.

[15] See below fn 40 and fn 137.

regions with multiple social impacts. Although there are important data-intensive projects in the field of smart cities, regional services (e.g. smart mobility) or global services (e.g. online content moderation provided by big players in social media), the AI operating context for the coming years will be more fragmented and distributed in nature, given the business environment in many countries, often dominated by SMEs, and the variety of communities interested in setting-up AI-based projects. The growing number of data scientists and the decreasing cost of hardware and software solutions, as well as their delivery as a service, will facilitate this scenario characterised by many projects with a limited scale, but involving thousands of people in data-intensive experiments.

For such projects, the traditional HRIA models are too articulated and oversized, which is why it is important to provide a more tailored model of impact assessment, at the same time avoiding mere theoretical abstractions based on generic decontextualised notions of human rights.

Against this background, it is worth briefly considering the role played by impact assessment tools with respect to the precautionary principle as an alternative way of dealing with the consequences of AI.

As in the case of potential technology-related risks, there are two different legal approaches to the challenges of AI: the precautionary approach and the risk assessment. These approaches are alternative, but not incompatible. Indeed, complex technologies with a plurality of different impacts might be better addressed though a mix of these two remedies.[16]

As risk theory states, their alternative nature is related to the notion of uncertainty.[17] Where a new application of technology might produce potential serious risks for individuals and society, which cannot be accurately calculated or quantified in advance, a precautionary approach should be taken.[18] In this case, the uncertainty associated with applications of a given technology makes it impossible

[16] Council of Europe, Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108) 2017, Section IV, paras 1 and 2, where the precautionary approach is coordinated with an impact assessment that also includes ethical and social issues.

[17] On the distinction between the precautionary approach and the precautionary principle, Peel 2004 ("One way of conceptualising what might be meant by precaution as an approach […] is to say that it authorises or permits regulators to take precautionary measures in certain circumstances, without dictating a particular response in all cases. Rather than a principle creating an obligation to act to address potential harm whenever scientific uncertainty arises, an approach could give regulators greater flexibility to respond").

[18] Commission of the European Communities 2000, pp. 8–16; Hansson 2020. Only few contributions in law literature take into account the application of the precautionary approach in the field of data protection, Costa 2012 and Gonçalves 2017; Pieters 2011, p. 455 ("generalised to information technology, it can serve as a trigger for government to at last consider the social implications of IT developments. Whereas the traditional precautionary principle targets environmental sustainability, information precaution would target social sustainability"). On the precautionary approach in data protection, Narayanan et al. 2016; Raab and Wright 2012, p. 364; Lynskey 2015, p. 83; Raab 2004, p. 15.

to conduct a concrete risk assessment, which requires specific knowledge of the extent of the negative consequences, albeit in specific classes of risks.[19]

Where the potential consequences of AI cannot be fully envisaged, as in the case of the ongoing debate on facial recognitions and its applications, a proper impact assessment is impossible, but the potentially high impact on society justifies specific precautionary measures (e.g., a ban or restriction on the use of AI-based facial recognition technologies).[20] This does not mean limiting innovation, but investigating more closely its potentially adverse consequences and guiding the innovation process and research,[21] including the mitigation measures (e.g. containment strategies, licensing, standards, labelling, liability rules, and compensation schemes).

On the other hand, where the level of uncertainty is not so high, the risk-assessment process is a valuable tool in tackling the risks stemming from technology applications. According to the general theory on the risk-based approach, the process consists of four separate stages: (1) identification of risks, (2) analysis of the potential impact of these risks, (3) selection and adoption of the measures to prevent or mitigate the risks, (4) periodic review of the effectiveness of these measures.[22] Furthermore, to enable subsequent monitoring of the effective level of compliance, duty bearers should document both the risk assessment and the measures adopted.

Since neither the precautionary principle nor the risk assessment are an empty list but rather focus on specific rights and freedoms to be safeguarded, they can be seen as two tools for developing a human rights-centred technology. While the uncertainty of some technology solutions will lead to the application of the precautionary principle, a better awareness and management of related risk will enable a proper assessment.

However, the relationship between risk assessment and the precautionary principle is rather complicated and cannot be reduced to a strict alterative. Indeed, when a precautionary approach suggests that a technology should not be used in a certain social context, this does not necessary entail halting its development. On the contrary, where there is no incompatibility with human rights[23] the technology can be developed further to reach a sufficient level of maturity that shows awareness of the related risks and the effective solutions.

This means that, in these cases, human rights can play an additional role in guiding development such that, once it reaches a level of awareness of the potential consequences that exclude uncertainty, will be subject to risk assessment.

[19] Tosun 2013; Aven 2011; Stirling and Gee 2002.

[20] European Parliament – Committee on Civil Liberties, Justice and Home Affairs 2020, paras 14, 15 and 20; Council of Europe, Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108) 2021, para 1.1. See Chap. 4.

[21] Commission of the European Communities 2000, p. 4 (“measures based on the precautionary principle should be maintained so long as scientific information is incomplete or inconclusive, and the risk is still considered too high to be imposed on society”).

[22] Koivisto and Douglas 2015.

[23] Article 5, European Commission, Proposal for Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending legislative acts, COM(2021) 206 final, Brussels, 21 April 2021.

Under this reasoning, two different scenarios are possible. One in which the precautionary principle becomes an outright ban on a specific use of technology and the other in which it restricts the adoption of certain technologies but not their further development. In the latter case, a precautionary approach and a risk assessment are two different phases of the same approach rather than an alternative response.

2.3 Human Rights Impact Assessment of AI in the HRESIA Model

Having defined the importance of a human rights-oriented approach in AI design and use, and the role that impact assessment procedure can play in this respect,[24] it is worth noting that traditional Human Rights Impact Assessment (HRIA) models are often territory-based considering the impact of business activities in a given local area and community, whereas in the case of AI applications this link with a territorial context may be less significant.

There are two different scenarios: cases characterised by use of AI in territorial contexts with a high-impact on social dynamics (e.g. smart cities plans, regional smart mobility plans, predictive crime programmes) and those where AI solutions have a more limited impact as they are embedded in globally distributed products/services (e.g. AI virtual assistants, autonomous cars, recruiting AI-based software, etc.) and do not focus on a given socio-territorial community. While in the first case the context is very close to the traditional HRIA cases, where large-scale projects affect whole communities and the potential impacts cover a wide range of human rights, the second case is characterised by a more limited social impact, often focusing more on individuals rather than on society at large.[25] This difference has a direct effect on the structure and complexity of the model, as well as the tool employed.

Criteria such as the AAAQ framework,[26] for example, or issues concerning property and lands, can be used in assessing a smart city plan, but are unnecessary or disproportionate in the case of an AI-based recruitment software. Similarly, a large-scale mobility plan may require a significant monitoring of needs through interviews of rightsholders and stakeholders, while in the case of an AI-based personal IoT device this phase can be much reduced.

In both these scenarios, the two most relevant novelties introduced by the HRESIA with regard to its HRIA module concern the *ex ante* nature of the assessment carried and the greater focus on quantifiable risk thresholds.

Regarding the former, the *ex ante* approach is required by the guiding role that HRESIA aims to play in project design and development, as opposed to the *ex post*

[24] See also Chap. 1.

[25] This does not mean that the collective dimension does not plays an important role and should be adequately considered in the assessment process, Mantelero 2016.

[26] The Danish Institute for Human Rights 2014.

evaluation centred on corrective policies that often characterises traditional HRIA.[27] Moreover, here, the pervasive and varied nature of data-intensive AI systems and their components leads to a reflection on the challenges that large-scale AI poses with respect to multi-factor scenarios.[28]

Concerning the focus on risk thresholds, this is in line with the requirements emerging in the regulatory debate on AI[29] where the definition of different risk levels is crucial in acceptability of AI products/services and has a direct impact on the obligations of AI manufacturers, providers and users. A quantitative dimension of assessment, in terms of ranges of risks, is therefore needed both for AI deign guidance and legal compliance.

Notwithstanding these important differences influencing the assessment methodology, the main building blocks of the model described here – planning and scoping, data collection (including rightsholder and stakeholder consultation) and analysis – remain the same as those used in HRIA and are examined in detail in the following sub-sections.

2.3.1 Planning and Scoping

The first stage deals with definition of the HRIA target, identifying the main features of the product/service and the context in which it will be placed, in line with the context-dependent nature of the HRIA. Three are the main areas to consider at this stage:

- description and analysis of the type of product/service, including data flows and data processing purposes
- the human rights context (contextualisation on the basis of local jurisprudence and laws)
- identification of rightsholders and stakeholders.

The Table 2.1 provides a non-exhaustive list of potential questions for HRIA planning and scoping.[30] The extent and content of these questions will depend on the specific nature of the product/service and the scale and complexity of its development and deployment.[31] This list is therefore likely to be further supplemented with project-specific questions.[32]

[27] World Bank and Nordic Trust Fund 2013, pp. 8–9.

[28] See Sect. 2.4.2.

[29] See Chap. 4.

[30] Regarding the structure and nature of the questions, Selbst forthcoming, pp. 33–35 and 69–70, who points out how open-end questions are better than top-down questions ("With open-ended questions, you do not need to anticipate the particular problems that might come up, and the answers to them emerge naturally. With top-down questions, no matter how thoughtful they are, the picture will be coarse and general").

[31] E.g. The Danish Institute for Human Rights 2020b, g.

[32] For similar questionnaires, e.g., The Danish Institute for Human Rights 2020a, pp. 30–39.

Table 2.1 Planning and scoping.

Description and analysis of the type of product/service, including related data flows and data processing purposes	– What are the main features of the product/service? – In which countries will the product/service be offered? – Identification of rights-holders: who are the target-users of the product/service? – What types of data are collected (personal, non-personal, special categories)? – What are the main purposes of data processing? – Identification of the duty-bearers: which subjects are involved in data management and what is their role in data processing?
Human rights context (contextualisation based on local jurisprudence and laws)	– Which human rights are potentially affected by the product/service? – Which international/regional legal instruments have been implemented at an operational level? – Which are the most relevant courts or authoritative bodies dealing with human rights issues in the given context? – What are the relevant decisions and provisions in the field of human rights?
Controls in place	– What policies and procedures are in place to assess the potential impact on human rights, including rightsholder and stakeholder engagement? – Has an impact assessment been carried out, developed and implemented in relation to specific issues or some features of the product/service (e.g. use of biometrics)?
Rightsholder and stakeholder engagement	– Which are the main groups or communities potentially affected by the service/product, including its development? – What other stakeholders should be involved, in addition to affected community and groups, (e.g. civil society and international organisations, experts, industry associations, journalists)? – Are there any other duty-bearers to be involved, apart from the product/service developer and users[33] (e.g. national authorities, governmental agencies)?

(continued)

[33] On the distinction between AI system users and end users, see Chap. 1, Sect. 1.3.

Table 2.1 (continued)

	– Were business partners, including suppliers (e.g. subcontractors in AI systems and datasets) involved in the assessment process? – Has the developer conducted an assessment of its supply chain to identify whether the activities of suppliers/contractors involved in product/service development might contribute to adverse human rights impacts? Has the developer promoted human rights standards or audits to ensure respect for human rights among suppliers? – Do the product/service developers publicly communicate the potential impacts on human rights of the service/product? – Does the developer provide training on human rights standards for relevant management and procurement staff?

Source The author

2.3.2 Data Collection and the Risk Analysis Methodology

While the first stage is mainly desk research, the second focuses on gathering relevant empirical evidence to assess the product/service's impact on human rights and freedoms. In traditional HRIA this usually involves extensive fieldwork. But in the case of AI applications, data collection and analysis is restricted to large-scale projects such as those developed in the context of smart cities, where different services are developed and integrated. For the remaining cases, given the limited and targeted nature of each application, data collection is largely related to the product/service's features and feedback from stakeholders.

Based on the information gathered in the previous stage (description and analysis of the type of product/service, human rights context, controls in place, and stakeholder engagement), we can proceed to a contextual assessment of the impact of AI use on human rights, to understand which rights and freedoms may be affected, how this may occur, and which potential mitigation measures may be taken.

Since in most cases the assessment is not based on measurable variables, the impact on rights and freedoms is necessarily the result of expert evaluation,[34] where expert opinion relies on knowledge of case law, the literature, and the legal framework. This means that it is not possible to provide precise measurement of the expected impacts but only an assessment in terms of range of risk (i.e. low, medium, high, or very high).

[34] E.g. Scheinin and Molbæk-Steensig 2021.

The benchmark for this assessment is therefore the jurisprudence of the courts and independent bodies (e.g. data protection authorities, equality bodies) that deal with human rights in their decisions. Different rights and freedoms may be relevant depending on the specific nature of the given application.

Examination of any potentially adverse impact should begin with a general overview followed by a more granular analysis where the impact is envisaged.[35] In line with normal risk assessment procedures, three key factors must be considered: risk identification, likelihood (L), and severity (S). As regards the first, the focus on human rights and freedoms already defines the potentially affected categories and the case specific analysis identifies those concretely affected, depending on the technologies used and their purposes. Since this is a rights-based model, risk concerns the prejudice to rights and freedoms, in terms of unlawful limitations and restrictions, regardless of material damage.

The expected impact of the identified risks is assessed by considering both the likelihood and the severity of the expected consequences, using a four-step scale (low, medium, high, very high) to avoid any risk of average positioning.

Likelihood is the combination of two elements: the probability of adverse consequences and the exposure. The former concerns the probability that adverse consequences of a certain risk might occur (Table 2.2) and the latter the potential number of people at risk (Table 2.3). In considering the potential impact on human rights, it is important not only to consider the probability of the impact, but also its extension in terms of potentially affected people.

Both these variables must be assessed on a contextual basis, considering the nature and features of the product and service, the application scenario, previous similar cases and applications, and any measures taken to prevent adverse consequences. Here, the engagement of relevant shareholders can help to better understand and contextualise these aspects, alongside the expertise of those carrying out the impact assessment.

These two variables are combined in the combinatorial Table 2.4 using a cardinal scale to estimate the overall likelihood level (L). This table can be further

[35] For an analytical description of the main components of impact analysis, based on the experience in the field of data protection, Janssen 2020, which uses four benchmarks covering the traditional areas of risk analysis in the law (impacted rights, risks at design stages and during operation, balancing risks and interests, control and agency over data processing). As for the risk assessment, the model proposed by the author does not provide a methodology to combine the different elements of impact assessment or to estimate the overall impact. Moreover, the model is used for an ex post comparative analysis, rather than for iterative design-based product/service development, as does the model we present here. In this sense, by providing two fictitious basic cases, Janssen tests her model though a comparative analysis (one case against the other) and without a clear analysis of the different risk components, in terms of individual impact and probability, with regard to each potentially affected right or freedom (e.g. "given that the monitor sensor captures every noise in its vicinity in situation (1), it probably has a high impact on a number of privacy rights, including that of intimacy of the home, communication privacy and chilling effects on the freedom of speech of (other) dwellers in the home"), and without a clear description of the assessment of their cumulative effect and overall impact. With a focus on the GDPR, Kaminski and Malgieri 2020. See also Reisman et al. 2018.

Table 2.2 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

Source The author

Table 2.3 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

Source The author

Table 2.4 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

Likelihood	
Low	1
Medium	2
High	3
Very high	4

Source The author

modified on the basis of the context-specific nature of assessed AI systems and feedback received from experts, rightsholders and stakeholders.

The severity of the expected consequences (S) is estimated by considering the nature of potential prejudice in the exercise of rights and freedoms and their consequences. This is done by taking into account the gravity of the prejudice (gravity), and the effort to overcome it and to reverse adverse effects (effort) (Tables 2.5 and 2.6).

As in the case of likelihood, these two variables are combined in a table (Table 2.7) using a cardinal scale to estimate the severity level (S).

A Table 2.8 for the overall assessment charts both variables – likelihood (L) and severity (S) of the expected consequences – against each envisaged risk to rights and freedoms (R1, R2, … Rn).

Table 2.5 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

Source The author

Table 2.6 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

Source The author

Table 2.7 Severity table (S)

		Gravity					Severity	
		1	2	3	4		Low	1
Effort	1	1	2	4	6		Medium	2
	2	2	3	5	8		High	3
	3	3	5	8	10		Very high	4
	4	5	8	10	12			

Source The author

The overall impact for each examined risk, taking into consideration the L and S values, is determined using a further table (Table 2.9). The colours represent the overall impact, which is very high in the dark grey sector, high in the grey sector, medium in the lighter grey sector and is low in the light grey sector.

Table 2.8 Table of envisaged risks

	L	S	Overall impact
R1			
R2			
…			
Rn			

Source The author

Table 2.9 Overall risk impact table

		Severity [impacted right/freedom]			
		Low	Medium	High	Very high
Likelihood	Low	░	▒	▓	▓
	Medium	▒	▒	▓	▓
	High	▒	▓	▓	█
	Very high	▒	▓	█	█

Source The author

Once the potentially adverse impact has been assessed for each of the rights and freedoms considered, a radial graph is charted to represent the overall impact on them. This graph is then used to decide the priority of intervention in altering the characteristics of the product/service to reduce the expected adverse impacts. See Fig. 2.1.[36]

To reduce the envisaged impacts, factors that can exclude the risk from a legal perspective (EFs) – such as the mandatory nature of certain impacting features or the prevalence of competing interests recognised by law – and those that can reduce the risk by means of appropriate mitigation measures (MMs) should be considered.

After the first adoption of the appropriate measures to mitigate the risk, further rounds of assessment can be conducted according to the level of residual risk and its acceptability, enriching the initial table with new columns (Table 2.10).

The first two new columns show any risk excluding factors (EFs) and mitigation measures (MMs), while the following two columns show the residual likelihood (rL) and severity (rS) of the expected consequences, after accounting for excluding and mitigation factors. The last column gives the final overall impact, using rL and rS values and the overall impact table (Table 2.9); this result can also be represented in a new radial graph. Note that it is also possible to estimate the total overall impact, as an average of the impacts on all the areas analysed. But this necessarily treats all the different impacted areas (i.e. rights and freedoms) as having the same importance and is therefore a somewhat imprecise synthesis.[37]

[36] This approach is also in line with the adoption of the Agile methodology in software development.

[37] See also Chap. 4, Sect. 4.3.2.

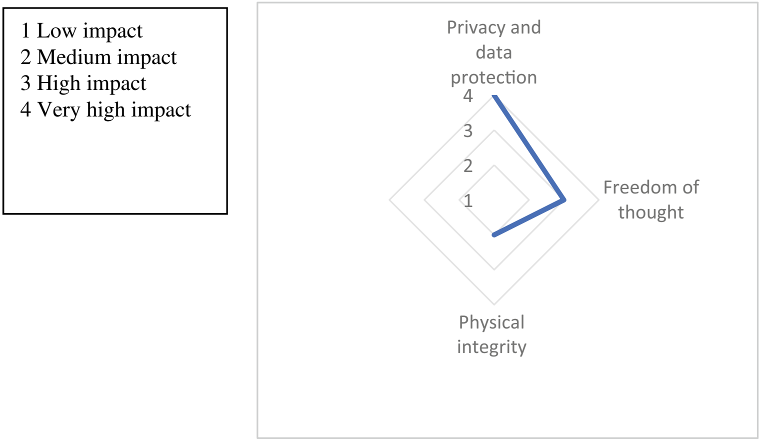

Fig. 2.1 Radial graph (impact) example. *Source* The author

Table 2.10 Comparative risk impact analysis table (before/after mitigation measures and excluding factors)

	L	S	Overall impact	EFs	MMs	rL	rS	Final Impact
R1								
R2								
…								
Rn								

Source The author

In terms of actual effects on operations, the radial graph is therefore the best tool to represent the outcome of the HRIA, showing graphically the changes after introducing mitigation measures. However, an estimation of overall impact could also be made in future since several legislative proposals on AI refer to an overall impact of each AI-based solution,[38] using a single risk scale covering all potential consequences.

[38] Data Ethics Commission 2019, p. 18. See Chap. 4.

2.4 The Implementation of the Model

The next two sub-sections examine two possible applications of the proposed model, with two different scales of data use. The first case, an Internet-connected doll equipped with AI, shows how the impact of AI is not limited to adverse effects on discrimination, but has a wider range of consequences (privacy and data protection, education, freedom of thought and diversity, etc.), given the innovative nature of the application and its interaction with humans.

This highlights the way in which AI does not merely concern data and data quality but more broadly the transformation of human-machine interaction by data-intensive systems. This is even more evident in the case of the smart cities, where the interaction is replicated on large scale affecting a whole variety of human behaviours by individuals, groups and communities.

The first case study (an AI-powered doll) shows in detail how the HRIA methodology can be applied in a real-life scenario. In the second case (a smart city project) we do not repeat the exercise for all the various data-intensive components, because a full HRIA would require extensive information collection, rightsholder and stakeholder engagement, and supply-chain analysis,[39] which go beyond the scope of this chapter.[40] But above all, the purpose of this second case study is different: to shed light on the dynamics of the HRIA in multi-factor scenarios where many different AI systems are combined.

Indeed, a smart city environment is not a single device, but encompasses a variety of technical solutions based on data and algorithms. The cumulative effect of integrating many layers results in a whole system that is greater and more complicated than the sum of its parts.

This explains why the assessment of potential risks to human rights and freedoms cannot be limited to a fragmented case-by-case analysis of each application. Rather, it requires an integrated approach that looks at the whole system and the interaction among its various components, which may have a wider impact than each component taken separately.

Scale and complexity, plus the dominant role of one or a few actors, can produce a cumulative effect which may entail multiple and increased impacts on rights and freedoms, requiring an additional integrated HRIA to give an overall assessment of the large-scale project and its impacts.

[39] Crawford and Joler 2018.

[40] A proper HRIA would require a multidisciplinary team working locally for a significant period of time. For example, the human rights impact assessment of the Bisha Mine in Eritrea, which started in July 2013, issued its final HRIA report in February 2014, followed by an auditing procedure in 2015. See LKL International Consulting Inc. 2014; LKL International Consulting Inc. 2015. See also Abrahams and Wyss 2010.

2.4.1 A Case Study on Consumer Devices Equipped with AI

Hello Barbie was an interactive doll produced by Mattel for the English-speaking market, equipped with speech recognition systems and AI-based learning features, operating as an IoT device. The doll was able to interact with users but did not interact with other IoT devices.[41]

The design goal was to provide a two-way conversation between the doll and the children playing with it, including capabilities that make the doll able to learn from this interaction, e.g. tailoring responses to the child’s play history and remembering past conversations to suggest new games and topics.[42] The doll is no longer marketed by Mattel due to several concerns about system and device security.[43]

This section discusses the hypothetical case, imagining how the proposed assessment model[44] could have been used by manufactures and developers and the results that might have been achieved.

2.4.1.1 Planning and Scoping

Starting with the questions listed in Table 2.1 above and information on the case examined, the planning and scoping phase would summarise the key product characteristics as follows:

(a) A connected toy with four main features: (i) programmed with more than 8,000 lines of dialogue[45] hosted in the cloud, enabling the doll to talk with the user about “friends, school, dreams and fashion”;[46] (ii) speech recognition technology[47] activated by a push-and-hold button on the doll's belt buckle; (iii) equipped with a microphone, speaker and two tri-colour LEOs embedded

[41] Mattel, ‘Hello Barbie FAQ’ Version 2 (2015). http://hellobarbiefaq.mattel.com/faq/. Accessed 12 November 2020.

[42] Hello Barbie FAQ (fn 41).

[43] Shasha et al. 2019 (with regard to Hello Barbie, see Appendix A, para A.3).

[44] On the safeguard of human rights and the use of HRIA in the business context, United Nations 2011 (“The State duty to protect is a standard of conduct. Therefore, States are not per se responsible for human rights abuse by private actors. However, States may breach their international human rights law obligations where such abuse can be attributed to them, or where they fail to take appropriate steps to prevent, investigate, punish and redress private actors’ abuse”) and more specifically Principles 13, 18 and 19.

[45] The comprehensive list of all the lines Hello Barbie says as of 17 November 2015 is available at http://hellobarbiefaq.mattel.com/wp-content/uploads/2015/11/hellobarbie-lines-v2.pdf. Accessed 28 November 2020.

[46] Hello Barbie FAQ (fn 41). Cloud service was provided by ToyTalk, see the following footnote.

[47] This technology and services were provided by ToyTalk, a Mattel partner.

in the doll's necklace, which light up when the device is active; (iv) a Wi-Fi connection to provide for two-way conversation.[48]

(b) The target-user is an English-speaking child (minor). Theoretically the product could be marketed worldwide in many countries, but the language barrier represents a limitation.

(c) The right-holders can be divided into three categories: direct users (minors), supervisory users (parents, who have partial remote control over the doll and the doll/user interaction) and third parties (e.g. friends of the direct user or re-users of the doll).

(d) Regarding data processing, the doll collects and stores voice-recording tracks based on dialogues between the doll and the user; this information may include personal data[49] and sensitive information.[50]

(e) The main purpose of the data processing and AI is to create human–robot interaction (HRI) by using machine learning (ML) to build on the dialogue between the doll and its young users. There are also additional purposes: (i) educational; (ii) parental control and surveillance[51] (parents can listen, store

[48] Hello Barbie FAQ (fn 41).

[49] Hello Barbie FAQ (fn 41) ("Q: Can Hello Barbie say a child's name? No. Hello Barbie does not ask for a child's name and is not scripted to respond with a child's name, so she will not be able to recite a child's name back to them"). But Leta Jones 2016, p. 245 who reports this reply in the dialogue with the doll: "Barbie: Sometimes I get a little nervous when I tell people my middle name. But I'm really glad I told you! What's your middle name?".

[50] Hello Barbie FAQ (fn 41) ("Although Hello Barbie was designed not to ask questions which are intended to elicit answers that might contain personal information, we cannot control whether a child volunteers such information without prompting. Parents who are concerned about this can monitor their child's use of Hello Barbie, and parents have the power to review and delete any conversation their child has with Hello Barbie, whether the conversations contain personal information or not. If we become aware of any such personal information captured in recordings, it is our policy to delete such information, and we contractually require our Service Providers to do the same. This personal information is not used for any purpose").

[51] Hello Barbie FAQ (fn 41) ("Hello Barbie only requires a parent's email address to set up an account. This is necessary so that parents can give permission to activate the speech recognition technology in the doll. Other information, such as a daughter's birthday, can be provided to help personalize the experience but are not required"). See also fn 52.

and re-use recorded conversations);[52] (iii) direct advertising to parents;[53] (iv) testing and service improvement.[54]

(f) The chief duty-bearer is the producer, but in connected toys other partners – such as ToyTalk in the Hello Barbie case – may be involved in the provision of ML, cloud and marketing services.

Another important set of data to be collected at this stage concerns the potential interplay with human rights and the reference framework, including main international/regional legal instruments, relevant courts or other authoritative bodies, and relevant decisions and provisions.

As regards the rights potentially affected, depending on the product's features and purposes, data protection and the right to privacy are the most relevant due to the possible content of the dialogue between the doll and the user, and the parental monitoring. Here the legal framework is represented by a variety of regulations at different levels. Compliance with the US COPPA[55] and the EU GDPR[56] can cover large parts of the potential market of this product and international guiding Principles[57] can facilitate the adoption of global policies and solutions.

[52] Hello Barbie FAQ (fn 41) ("Hello Barbie recording and storing conversations girls have with the doll? Yes. Hello Barbie has conversations with girls, and these conversations are recorded. These audio recordings are used to understand what is being said to Hello Barbie so she can respond appropriately and also to improve speech recognition for children and to make the service better. These conversations are stored securely on ToyTalk's server infrastructure and parents have the power to listen to, share, and/or delete stored recordings any time").

[53] Hello Barbie FAQ (fn 41) ("Q. Are conversations used to market to children? No. The conversations captured by Hello Barbie will not be used to contact children or advertise to them." This was confirmed by the analysis carried out by Shasha et al. 2019. Regarding the advertising directs to parents, this is the answer provided in the FAQ: "Q: Your Privacy Policy says that you will use personal information to provide consumers with news and information about events, activities, promotions, special offers, etc. That sounds like consumers could be bombarded with marketing messages. Can parents elect not to receive those communications? Yes. Opting out of receiving promotional emails will be an option during the set up process and you can opt out at any time by following the instruction in those emails. Note that marketing messages will not be conveyed via the doll itself").

[54] Hello Barbie FAQ (fn 41) ("Conversations between Hello Barbie and consumers are not monitored in real time, and no person routinely reviews those conversations. Upon occasion a human may review certain conversations, such as in order to test, improve, or change the technology used in Hello Barbie, or due to support requests from parents. If in connection with such a review we come across a conversation that raises concern about the safety of a child or others, we will cooperate with law enforcement agencies and legal processes as required to do so or as we deem appropriate on a case-by-case basis").

[55] Federal Trade Commission 2017; Haber 2019.

[56] Information Commissioner's Office 2020.

[57] E.g. Council of Europe, Convention 108+. See also Council of Europe 2018, para 36 ("With respect to connected or smart devices, including those incorporated in toys and clothes, States should take particular care to ensure that data-protection principles, rules and rights are also respected when such products are directed principally at children or are likely to be regularly used by or in physical proximity to children"); Mantelero 2021.

Moreover, in relation to data processing and individual freedom of choice, the potential effects of marketing strategies can also be considered as forms of freedom of expression[58] and freedom to conduct a business.

Given the broad interaction between the doll and the user and the behavioural, cultural and educational influence that the doll may have on young users,[59] further concerns relate to freedom of thought and diversity.[60]

In the event of cyberattack and data theft or transmission of inappropriate content to the user through the doll, safety issues also arise and may impact on the right to psychological and physical safety and health.

With the potentially global distribution of the toy, the possible impacts need to be further contextualised within each relevant legal framework, taking into consideration local case law and that of regional supranational bodies like the European Court of Human rights. In this regard, it is necessary during the scoping phase to identify the significant provisions and decisions in the countries/regions where the product is distributed.

The last aspect to be considered in planning and scoping HRIA concerns the identification and engagement of potential stakeholders. In the case of connected toys, the most important stakeholders are likely to be parents' associations, educational bodies, professional associations (e.g. psychologists and educators), child, consumer and data protection supervisory bodies, as well as trade associations. Stakeholders may also include the suppliers involved in product/service development. In the latter case, the HRIA must also assess the activities by these suppliers and may benefit from an auditing procedure[61] or the adoption of standards.

The following sections describe an iterative assessment process, starting from the basic idea of the connected AI-equipped toy with its pre-set functionality and moving on to a further assessment considering additional measures to mitigate unaddressed, or only partially addressed, concerns.

[58] Universal Declaration of Human Rights, Article 19, and International Covenant on Civil and Political Rights, Article 19(2). See also International Covenant on Civil and Political Rights, Human Rights Committee 2011, para 11; UNICEF 2012, principle 6 (Use marketing and advertising that respect and support children's rights).

[59] Mertala 2020 ("As Hello Barbie is able to speak, the child no longer performs the role through the doll, but in relation to the doll. This changes the nature of the performative element from dominantly transitive to dominantly performative, in which the child occupies and embodies a role in relation to the toy"). See also the following statement included in the list of all the lines Hello Barbie says as of 17 November 2015 (fn 45) "It's so cool that you want to be a mom someday".

[60] Hello Barbie FAQ (fn 41) ("The doll's conversation tree has been designed to re-direct inappropriate conversations. For example, Hello Barbie will not repeat curse words. Instead, she will respond by asking a new question"). However, besides the example given, there is no clear description of what is considered appropriate or not, and this category (appropriateness) is significantly influenced by the cultural component and potentially also by corporate ethics that may create forms of censorship or oriented behavior and thinking in the young user. Even when the FAQs refer to "school age appropriate content" ("All comments made by Hello Barbie are scripted with school age appropriate content"), they implicitly refer to a benchmark dependent the educational standards of developed economies.

[61] But see European Commission 2020, pp. 73–74.

2.4.1.2 Initial Risk Analysis and Assessment

The basic idea of the toy is an interactive doll, equipped with speech recognition and learning features, operating as an IoT device. The main component is a human-robot voice interaction feature based on AI and enabled by Internet connection and cloud services.

The rights potentially impacted are data protection and privacy, freedom of thought and diversity, and psychological and physical safety and health.[62]

Data Protection and the Right to Privacy

While these are two distinct rights, for the purpose of this case study we considered them together.[63] Given the main product features, the impact analysis is based on following questions:[64]

- Does the device collect personal information? If yes, what kind of data is collected, and what are the main features of data processing? Can the data be shared with other entities/persons?
- Can the connected toy intrude into the users' private sphere?
- Can the connected toy be used for monitoring and surveillance purposes? If yes, is this monitoring continuous or can the user stop it?
- Do users belong to vulnerable categories (e.g. minors, elderly people, parents, etc.)?
- Are third parties involved in the data processing?
- Are transborder data flows part of the processing operations?

Taking into account the product's nature, features and settings (i.e. companion toy, dialogue recording, personal information collection, potential data sharing by parents) the likelihood of prejudice can be considered very high (Table 2.4). The extent and largely unsupervised nature of the dialogue between the doll and the user, as well as the extent of data collection and retention make the probability high (Table 2.2). In addition, given its default features and settings, the exposure is very high (Table 2.3) since all the doll's users are potentially exposed to this risk.

Regarding risk severity, the gravity of the prejudice (Table 2.5) is high, given the subjects involved (young children and minors), the processing of personal data in several main areas, including sensitive information,[65] and the extent of data collection. In addition, unexpected findings may emerge in the dialogue between the

[62] Keymolen and Van der Hof 2019 ("Smart toys come in different forms but they have one thing in common. The development of these toys is not just a feature of ongoing technological developments; their emergence also reflects an increasing commercialisation of children's everyday lives").

[63] UN Convention on the Rights of the Child, Article 16; European Convention on Human Rights, Article 8.

[64] For a more extensive list of guiding questions, see e.g. UNICEF 2018.

[65] Pre-recorded sentences containing references to, for instance, religion and ethical groups. See the full list of all lines for Hello Barbie (fn 45) (e.g. "Sorry, I didn't catch that. Was that a yes or a no to talking about Kwanzaa?").

user and the doll, as the harmless topics prevalent in the AI-processed sentences can lead young users to provide personal and sensitive information. Furthermore, the data processing also involves third parties and transborder data flows, which add other potential risks.

The effort to overcome potential prejudice or to reverse adverse effects (Table 2.6) can be considered as medium, due to the potential parental supervision and remote control, the nature of the doll's pre-selected answers and the adoption of standard data security measures that help to overcome suffered prejudice with a few difficulties (e.g. data erasure, dialogue with the minor in case of unexpected findings). Combining high gravity and medium effort, the resulting severity (Table 2.7) is medium.

If the likelihood of prejudice can be considered very high and the severity medium, the overall impact according to Table 2.9 is high.

Freedom of Thought, Parental Guidance and the Best Interest of the Child
Based on the main features of the product, the following questions can be used for this analysis:

– Is the device able to transmit content to the user?
– Which kind of relationships is the device able to create with the user?
– Does the device share any value-oriented messages with the user?

  - If yes, what kind of values are communicated?
  - Are these values customisable by users (including parents) or on the basis of user interaction? If so, what range of alterative value sets is provided?
  - Are these values the result of work by a design team characterised by diversity?

Here the case study reveals the critical impact of AI on HRI owing to the potential content imparted through the device. This is even more critical in the context of toys where the interactive nature of AI-powered dolls changes the traditional interaction into a relational experience.[66]

In the model considered (Hello Barbie), AI creates a dialogue with the young user by selecting the most appropriate sentence from the more than 8,000 lines of dialogue available in its database. On the one hand, this enables the AI to express opinions which may also include value-laden messages, as in this sentence: "It's so cool that you want to be a mom someday".[67] On the other, some value-based considerations are needed to address educational issues concerning "inappropriate questions"[68] where the problem is not the AI reaction (Hello Barbie responds "by asking a new question"[69]), as previously, but the notion of appropriateness, which necessarily involves a value-oriented content classification by the AI system.

[66] See Mertala 2020.

[67] See fn 45. On gender stereotypes in smart toys, see Norwegian Consumer Council 2016.

[68] See fn 60.

[69] Hello Barbie FAQ (fn 41).

As these value-laden features of AI are inevitably defined during the design process, the composition of the design team, its awareness of cultural diversity and pluralism are key elements that impact on freedom of thought, in terms of default values proposed and the availability of alternative settings. In addition, the decision to provide only one option or several user-customisable options in the case of value-oriented content is another aspect of the design phase that can limit parents' freedom to ensure the moral and religious education of their children in accordance with their own beliefs.

This aspect highlights the paradigm shift brought by AI to freedom of thought and the related parental guidance in supporting the exercise by children of their rights.[70] This is even more evident when comparing AI-equipped toys with traditional educational products, such as books, serious games etc., whose contents can be examined in advance by parents.[71]

The AI-equipped doll is different. It delivers messages to young users, which may include educational content and information, but no parent will read all the 8,000 lines the doll can use or ask to have access to the logic used to match them with children's statements.

As AI-based devices interact autonomously with children and convey their own cultural values,[72] this impacts on the rights and duties of parents to provide, in a manner consistent with the evolving capacities of the child, appropriate direction and guidance in the child's freedom of thought, including aspects concerning cultural diversity.

In terms of risk assessment, the probability (Table 2.2) is medium, considering the limited number of sentences involving a value-oriented statement, and the exposure (Table 2.3) is medium, due to their alignment with values commonly accepted in many cultural contexts. The likelihood is therefore medium (Table 2.4).

Taking into account the nature of the product and its main features (i.e. some value-laden sentences used in dialogue with the young user),[73] the gravity of prejudice (Table 2.5) can be considered low in the case in question, as the value-laden sentences concern cultural questions that are not particularly controversial. The effort (Table 2.6) can also be considered low, as talking with children can mitigate potential harm. Combining these two values, the severity is therefore low (Table 2.7).

Note that this assessment would be completely altered if the dialogue content were not pre-selected but generated by AI on the basis of information resulting from

[70] UN Convention on the Rights of the Child, Articles 5, 14, and 18. See also See UNICEF 2018, p. 9; Murdoch 2012, p. 13.

[71] UN Convention on the Rights of the Child, Articles 17(e) and 18.

[72] E.g. Norwegian Consumer Council 2016 referring to the connected doll Cayla ("Norwegian version of the apps has banned the Norwegian words for "homosexual", "bisexual", "lesbian", "atheism", and "LGBT" […]" "Other censored words include 'menstruation', 'scientology-member', 'violence', 'abortion', 'religion', and 'incest'").

[73] Steeves 2020.

web searches,[74] where the potential risk would be much higher.[75] Similarly, the inclusion in the pre-recorded database of a greater number of value-laden sentences would directly increase the risk.

Considering the likelihood as medium and the severity of the prejudice as low, the overall impact (Table 2.9) is medium.

Right to Psychological and Physical Safety

Connected toys may raise concerns about a range of psychological and physical harms deriving from their use, including access to data and remote control of the toy.[76] Based on the main features of the product examined, the following questions can be used for this analysis:

– Can the device put psychological or physical safety at risk?
– Does the device have adequate data security and cybersecurity measures in place?
– Can third parties perpetrate malicious attacks that pose a risk to the psychological or physical safety of the user?

As regards the probability, considering the third-party origin of the prejudices and the limited interest in malicious attacks (no business interest, distributed and generic target), but also how easy it is to hack the toy, the probability (Table 2.2) of an adverse impact is medium. Exposure (Table 2.3) is low, given the prevalent use of the device in a supposedly safe environment, such as schools and home, where malicious access and control of the doll is difficult and adult monitoring is more frequent. The likelihood (Table 2.4) is therefore low.

Taking into account the nature of the product examined, the young age of the user, and the potential safety and security risks,[77] the gravity of prejudice (Table 2.5) can be considered medium. This is because malicious attacks can only be carried out by speech, and no images are collected. Nor can the toy – given its size and characteristics – directly cause physical harm to the user. The effort (Table 2.6) can be considered medium since parent-child dialogue and technical solutions can combat the potential prejudice. The severity (Table 2.7) is therefore medium.

Considering the likelihood as low and the severity of the prejudice as medium, the overall impact is medium (Table 2.9).

[74] In the case examined, the content provided by means of the doll was handcrafted by the writing team at Mattel and ToyTalk, not derived from open web search. Hello Barbie FAQ (fn 41).

[75] E.g., Neff and Nagy 2016.

[76] E.g. de Paula Albuquerque et al. 2020, whose authors refer to harassment, stalking, grooming, sexual abuse, exploitation, paedophilia and other types of violence blackmail, insults, confidence loss, trust loss and bullying; Shasha et al. 2019. See also Federal Bureau of Investigation 2017.

[77] See fn 41.

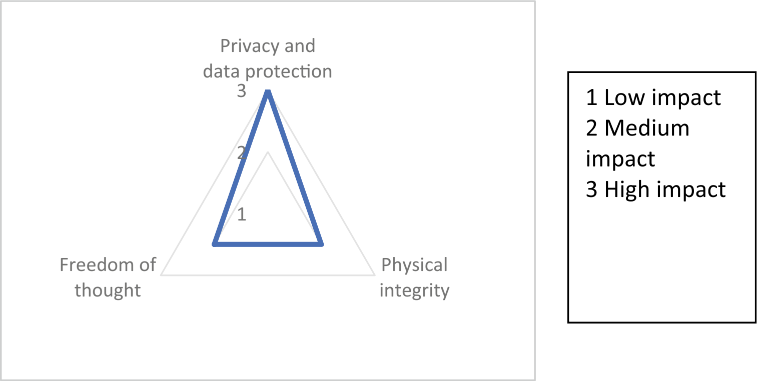

Fig. 2.2 Radial graph (impact) of the examined case. *Source* The author

2.4.1.3 Results of the Initial Assessment

The following table (Table 2.11) shows the results of the assessment carried out on the initial idea of the connected AI-equipped doll described above:

Based on this table, we can plot a radial graph representing the overall impact on all the affected rights and freedoms. The graph (Fig. 2.2) shows the priority of mitigating potentially adverse impacts on privacy and data protection, followed by risks related to physical integrity and freedom of thought.

This outcome is confirmed by the history of the actual product, where the biggest concerns of parents and the main reasons for its withdrawal related to personal data and hacking.[78]

Table 2.11 Table of envisaged risks for the examined case (L: low, M: medium; H: high; VH: very high)

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

Source The author

[78] Gibbs 2015.

2.4.1.4 Mitigation Measures and Re-assessment

Following the iterative assessment, we can imagine that after this initial evaluation of the general idea, further measures are introduced to mitigate the potential risks found. At this stage, the potential rightsholders and stakeholders (users, parents associations, educational bodies, data protection authorities etc.) can make a valuable contribution to better defining the risks and how to tackle them.

While the role of the rightsholders and stakeholders cannot be directly assessed in this analysis, we can assume that their participation would have shown great concern for risks relating to communications privacy and security. This conclusion is supported by the available documentation on the reactions of parents and supervisory authorities in the Hello Barbie case.[79]

After the first assessment and given the evidence on the requests of rightsholders and stakeholders, the following mitigation measures and by-design solutions could have been adopted with respect to the initial prototype.

(A) Data protection and the right to privacy

Firstly, the product must comply with the data protection regulation of the countries in which it is distributed.[80] Given the product's design, we cannot exclude the processing of personal data. The limited number of sentences provided for use by AI, as in the case of Hello Barbie, does not exclude the provision of unexpected content by the user, including personal information.[81]

Risk mitigation should therefore focus on the topics of conversation between the doll and the young user, and the safeguards in processing information collected from the user.

As regards the first aspect, an effective way to limit the potential risks would be to use a closed set of sentences, excluding phrases and questions that might induce the user to disclose personal information, and making it possible to modify these phrases and questions by the owner of the toy.[82]

[79] E.g. BEUC 2016; Neil 2015; McReynolds et al. 2017.

[80] In this regard Hello Barbie was certified as compliant with the US COPPA, see 'Hello Barbie FAQ' (fn 41).

[81] Hello Barbie FAQ (fn 41) ("we cannot control whether a child volunteers such information without prompting").

[82] In this case, the conditions are largely present, although there is evidence of minor issues. E.g. Hello Barbie FAQ (fn 41) ("Hello Barbie does not ask for a child's name and is not scripted to respond with a child's name, so she will not be able to recite a child's name back to them"), but see the interaction reported in Leta Jones 2016, p. 245 ("Barbie: Sometimes I get a little nervous when I tell people my middle name. But I'm really glad I told you! What's your middle name?! !"). Hello Barbie FAQ (fn 41) also points out the privacy-oriented design of the product with regard to dialogue content: "Although Hello Barbie was designed not to ask questions which are intended to elicit answers that might contain personal information".

Regarding the processing of personal data, the doll's AI-based information processing functions should be deactivated by default, giving the parents control over its activation.[83] In addition, to reduce the risk of constant monitoring, deliberate action by the child should be required to activate the doll's AI-equipped dialogue functions.[84] This would also help to make users more aware of their interaction with the system and related privacy issues.[85]

Ex post remedies can also be adopted, such as speech detection to remove personal information in recorded data.[86]

Conversations are not monitored, except to support requests from parents. To reduce the impact on the right to privacy and data protection, human review of conversations – to test, improve, or change the technology used – should be avoided, even if specific policies for unexpected findings have been adopted.[87] Individual testing phases or experiments can be carried out in a laboratory setting or on the basis of user requests (e.g. unexpected reactions and dialogues). This more restrictive approach helps to reduce the impact with respect to the initial design.

Further issues, regarding the information processing architecture and its compliance with data protection principles, concern data storage. This should be minimised and parents given the opportunity to delete stored information.[88]

With regard to the use of collected data, while access to, and sharing of, this information by parents[89] are not per se against the interest of the child, caution should be exercised in using this information for marketing purposes. Given the early age of the users and the potentially large amount of information they may

[83] Hello Barbie FAQ (fn 41) ("Hello Barbie only requires a parent's email address to set up an account. This is necessary so that parents can give permission to activate the speech recognition technology in the doll. Other information, such as a daughter's birthday, can be provided to help personalize the experience but are not required […] If we discover that, in violation of our terms of service, an account was created by a child, we will terminate the account and delete all data and recordings associated with it.").

[84] In the Hello Barbie case, the doll was not always on but it was activated by pressing the belt buckle.

[85] In the examined case this was also emphasized because the two tri-colour LEOs embedded in the doll's necklace lighted up to indicate she was active.

[86] Hello Barbie FAQ (fn 41) ("If we become aware of any such personal information captured in recordings, it is our policy to delete such information, and we contractually require our Service Providers to do the same. This personal information is not used for any purpose").

[87] See fn 50.

[88] Hello Barbie FAQ (fn 41) ("Parents who are concerned about this can monitor their child's use of Hello Barbie, and parents have the power to review and delete any conversation their child has with Hello Barbie, whether the conversations contain personal information or not"). Considering the young age of the user this seems not to be a disproportionate monitoring with regard to their activities and right to privacy. This does not exclude a socio-ethical relevance of this behaviour, see e.g. Leta Jones and Meurer 2016 ("the passive nature of Barbie's recording capabilities could prove perhaps more devastating to a child who may have placed an implicit trust in the doll. In order to determine the extent of the parent's involvement in their child's recordings, we extended our analysis to include the adult oversight capabilities").

[89] See above fn 52.

provide in their conversation with the doll, plus the lack of active and continuous parental control, the best solution would be not to use child-doll conversations for marketing.[90]

The complexity of data processing activities in the interaction between a child and an AI-equipped doll inevitably affects the form and content of the privacy policies and the options offered to users, as provided by many existing legislations.

A suitable notice and consent mechanism, clear and accessible and legally compliant, is therefore required,[91] but meeting this obligation is not so simple in the case in question. The nature of the connected toy and the absence of any interface limits awareness of the policies and distances them from direct interaction with the device. This accentuates the perception of the notice and consent mechanism as a mere formality to be completed to access the product.

The last crucial area concerns data security. This entails a negative impact that goes beyond personal data protection and, as such, is also analysed below under impact on the right to psychological and physical safety.

As the AI-based services are hosted by the service provider, data security issues concern both device-service communications and malicious attacks to the server and the device. Encrypted communications, secure communication solutions, and system security requirements for data hosted and processed on the server can minimise potential risks, as in the case study, which also considered access to data when the doll's user changes.[92]

None of these measures prevent the risks of hacking to the device or the local Wi-Fi connection, which are higher when the doll is used outdoors.[93] This was the chief weakness noted in the case in question and in IoT devices more generally. They are often designed with poor inherent data security and cybersecure features for cost reasons. To reduce this risk, stronger authentication and encryption solutions have been proposed in the literature.[94]

Taking into account the initial impact assessment plus all the measures described above, the exposure is reduced to low, since users are thus exposed to potential prejudices only in special circumstances, primarily malicious attack. Probability also becomes low, as the proposed measures mitigate the risks relating to dialogue

[90] This was the option adopted in the Hello Barbie case, see fn 53. But Steeves 2020 on the sentences used by Hello Barbie to indirectly reinforce the brand identity and encourage the child to adopt that identity for his/her own.

[91] In the case examined, one of the main weakness claimed with regard to Hello Barbie concerned the privacy policies adopted, the interplay between the different entities involved in data processing, and the design of these policies and access to them, which were considered cumbersome. Leta Jones and Meurer 2016.

[92] Hello Barbie FAQ (fn 41) ("Conversations and other information are not stored on the doll itself, but rather in the associated parent account. So, if other users are using a different Wi-Fi network and using their own account, Hello Barbie would not remember anything from the prior conversations. New users would need to set up their own account to enable conversations with Barbie").

[93] Leta Jones 2016, p. 244.

[94] See also below under (C).

between doll and user, data collection and retention. Likelihood (Table 2.4) is therefore reduced to low.

Regarding severity of prejudice, gravity can be lowered to at least medium by effect of the mitigation measures, but effort remains medium, given the potential risk of hacking. Severity is therefore lowered somewhat (from 5 to 3 in Table 2.7), though remaining medium.

If the severity and the likelihood are medium in Table 2.9, the overall impact is lowered from high to medium.

(B) Impact on freedom of thought

As described in Sect. 2.4.1.2, the impact on freedom of thought is related to the values conveyed by the doll in dialogue with the user. Here the main issue concerns the nature of the messages addressed to the user, their sources and their interplay with the rights and duties of parents to provide appropriate direction and guidance in the child's exercise of freedom of thought, including issues of cultural diversity.

A system based on Natural Language Processing allows AI various degrees of autonomy in identifying the best response or sentence in the human-machine interaction. Given the issues considered here (the nature of the values shared by the doll with its young user) the two main options are to use a closed set of possible sentences or search for potential answers in a large database, such as the Internet. A variety of solutions can also be found between these two extremes.

Since the main problem is content control, the preferable option is the first, and this was indeed the solution adopted in the Hello Barbie case.[95] Content can thus be fine-tuned to the education level of the user, given the age range of the children.[96] This reduces the risk of unexpected and inadequate content and, where full lines of dialogue are available (this was the case with Hello Barbie), parents are able to get an idea of the content offered to their children.

Some residual risks remain however, due to intentional or unintentional cultural models or values, including the difference between appropriate and inappropriate content.[97] This is due to the special relationship the toy generates[98] and the only limited mitigation provided by transparency on pre-recorded lines of dialogue.

To address these issues, concerning both freedom of thought and diversity, the AI system should embed a certain degree of flexibility (user-customizable content) and avoid stereotyping by default. To achieve this, the team working on pre-recorded sentences and dialogues should be characterised by diversity, adopting a by-design approach and bearing in mind the target user of the product.[99]

[95] Hello Barbie FAQ (fn 41).

[96] Hello Barbie FAQ (fn 41) ("All comments made by Hello Barbie are scripted with school age appropriate content").

[97] See fn 60.

[98] See fn 59.

[99] On the different attitude in pre-recorded sentences with regard to different religious topics, see Steeves 2020.

Moreover, taking into account the parents' point of view, mere transparency, i.e. access to the whole body of sentences used by the doll, is not enough. As is demonstrated extensively in the field of data protection, information on processing is often disregarded by the user and it is hard to imagine parents reading 8,000 lines of dialogue before buying a doll.

To increase transparency and user awareness, therefore, forms of visualisation of these values through logic and content maps could be useful to easily represent the content used. In addition, it would be important to give parents the opportunity to partially shape the AI reactions, customising the values and content, providing other options relating to the most critical areas in terms of education and freedom of thought.

With regard to the effects of these measures, they mitigate both the potentially adverse consequences of initial product design and the lack of parental supervision of content, minimising the probability of an adverse impact on freedom of thought. The probability (Table 2.2) is therefore lowered to low.

Given the wide distribution of the product, the potential variety of cultural contexts and the need for an active role of parents to minimise the risk, the exposure remains medium, although the number of affected individuals is expected to decrease (Table 2.3).

If the probability is low and the exposure is medium, the likelihood (Table 2.4) is lowered to low after the adoption of the suggested mitigation measures and design solutions.

The gravity of prejudice and the effort were originally low and the additional measures described can further reduce gravity through a more responsible management of content which might support potentially conflicting cultural models or values. Severity therefore remains low.

Considering both likelihood and severity as low, the overall impact (Table 2.9) is reduced from medium to low, compared with the original design model.

(C) Impact on the right to psychological and physical safety

The potential impact in this area is mainly related to malicious hacking activities[100] that might allow third parties to take control of the doll and use it to cause, psychological and physical harm to the user.[101] This was one of the most widely debated issues in the Hello Barbie case and one of the main reasons that led Mattel to stop producing this toy.[102] Possible mitigation measures are the exclusion of

[100] Gibbs 2015.

[101] Chang et al. 2019 ("For example, the attackers can spread content through the audio system, which is adverse for children's growth through the built-in audio in the smart toys").

[102] See also Shasha et al. 2019.

Table 2.12 Comparative risk impact analysis table (examined case)

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

Source The author

interaction with other IoT devices,[103] strong authentication and data encryption.[104]

As regards likelihood, considering the protection measures adopted and the low interest of third parties in this type of individual and context-specific malicious attack, the probability is low (Table 2.2). Although the suggested measures do not affect the exposure, this remains low due to the limited circumstances in which a malicious attack can be carried out (Table 2.3). The likelihood therefore remains low but is lowered (from 2 to 1 in Table 2.4).

Regarding severity, the proposed measures do not impact on the gravity of the prejudice (Table 2.5), or the effort (Table 2.6) which remain medium. Severity therefore remains medium (Table 2.7).

Since the final values of neither likelihood nor severity change, overall impact remains medium (Table 2.9), with malicious hacking being the most critical aspect of the product in terms of risk mitigation.

The Table 2.12 shows the assessment of the different impacts, comparing the results before and after the adoption of mitigation measures.

In the case in question, there is no Table 2.10 EF column since there are no factors that could exclude risk, such as certain mandatory impacting features or overriding competing interests recognised by law.

The radial graph in this Fig. 2.3 shows the concrete effect of the assessment (the blue line represents the initial impacts and the orange the impacts after adoption of the measures described above). It should be noted that the reduction of potential impact is limited as the Hello Barbie product already included several options and measures to mitigate adverse effects on rights and freedoms (pre-recorded sentences, no Internet access, data encryption, parental access to stored data, etc.). The effect would have been greater starting from a general AI-equipped doll using Natural Language Processing interacting with children, without mitigation measures.

[103] Doll's speech content was hand crafted by the writing team at Mattel and ToyTalk, not derived from open web search. See 'Hello Barbie FAQ' (fn 41).

[104] Demetzou et al. 2018; Gonçalves de Carvalho and Medeiros Eler 2018.

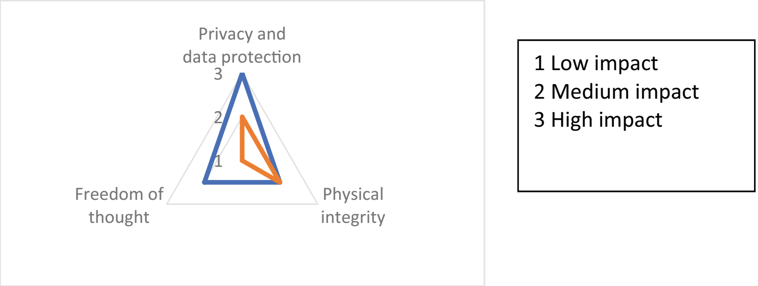

Fig. 2.3 Final radial graph of the examined case. *Source* The author. [Blue line: original impact. Orange line: final impact after adoption of mitigation measures and design solutions]

In this regard, the HRIA model proposed is in line with a human rights-by design approach, where the design team is asked to consider human rights impact from the earliest product design stages, discarding those options that have an obvious negative impact on human rights. With this approach, there is no HRIA 0 where the proposed product is completely open to the riskiest scenarios (e.g. a connected doll equipped with unsupervised AI that uses all available web sources to dialogue with young users, with unencrypted doll-user communication sent to a central datacentre where information is stored without a time limit and used for further purposes, including marketing communications direct to doll users).

In human rights-oriented design, HRIA thus becomes a tool to test, refine and improve adopted options that already entail a risk-aware approach. In this way, HRIA is a tool for testing and improving human rights-oriented design strategies.

2.4.2 A Large-Scale Case Study: Smart City Government

Large-scale projects using data-intensive AI applications are characterised by a variety of potentially impacted areas concerning individual and groups. This produces a more complex and multi-factor scenario which cannot be fully assessed by the mere aggregation of the results of HRIAs conducted for each component of these projects.

An example is provided by data-driven smart cites, where the overall effect of an integrated model including different layers affecting a variety of human activities means that the cumulative impact is greater than the sum of the impacts of each application.

In such cases, a HRIA for AI systems also needs to consider the cumulative effect of data use and the AI strategies adopted, as already happens in HRIA practice with large-scale scenario cases. This is all the more important in the field of AI where large-scale projects often feature a unique or dominant technology partner who benefits from a general overview of all the different processing activities ('platformisation'[105]).

The Sidewalk project in Toronto is an example of this 'platformisation' effect and a case study in the consequent impacts on rights and freedoms. This concluded smart city project was widely debated[106] and raised several human rights-related issues common to other data-intensive projects.

The case concerned a requalification project for the Quayside, a large urban area on Toronto's waterfront largely owned by Toronto Waterfront Revitalization Corporation. Based on an agreement between the City of Toronto and Toronto Waterfront,[107] in 2017, through a competitive Request for Proposals, Waterfront Toronto hired Sidewalk Labs (a subsidiary of Alphabet Inc.) to develop a proposal for this area.[108]

This proposal – the Master Innovation and Development Plan or MIDP[109] – outlined a vision for the Quayside site and suggested data-driven innovative solutions across the following areas: mobility and transportation; building forms and construction techniques; core infrastructure development and operations; social service delivery; environmental efficiency and carbon neutrality; climate mitigation strategies; optimisation of open space; data-driven decision making; governance and citizen participation; and regulatory and policy innovation.[110]

[105] Goodman and Powles 2019.

[106] Carr and Hesse 2020b; Flynn and Valverde 2019.

[107] The Waterfront Revitalization Corporation (which was renamed Waterfront Toronto) was a partnered not-for-profit corporation, created in 2003 by the City of Toronto, Province of Ontario and the Government of Canada (see also Province's Toronto Waterfront Revitalization Corporation Act) to oversee and deliver revitalization of Toronto's waterfront; further information are available at https://www.toronto.ca/city-government/accountability-operations-customer-service/city-administration/city-managers-office/agencies-corporations/corporations/waterfront-toronto/. Accessed 30 December 2020. See also Toronto Waterfront Revitalization: Memorandum of Understanding between the City of Toronto, City of Toronto Economic Development Corporation and Toronto Waterfront Revitalization Corporation. https://www.toronto.ca/legdocs/2006/agendas/council/cc060131/pof1rpt/cl027.pdf. Accessed 30 December 2020; City of Toronto, Executive Committee 2018a.

[108] Waterfront Toronto and Sidewalk Labs entered into a partnership Framework Agreement on October 16, 2017. The Framework Agreement was a confidential legal document, see City of Toronto, Executive Committee 2018a. A summary of this agreement is available in City of Toronto, Executive Committee 2018b, Comments, para 2 and Attachment 2.

[109] Sidewalk Labs was charged with providing Waterfront Toronto with a MIDP for evaluation, including public and stakeholder consultation. Following the adoption of the MIDP by the Waterfront Toronto's Board of Directors, the City of Toronto was to complete an additional assessment programme focused on feasibility and legal compliance, including public consultation. See City of Toronto, Deputy City Manager, Infrastructure and Development 2019.

[110] City of Toronto, Executive Committee 2018a.

This long list of topics shows how this data-intensive project went beyond mere urban requalification to embrace goals that are part of the traditional duties of a local administration, pursuing public interest purposes[111] with potential impacts on a variety of rights and freedoms.

The Sidewalk case[112] suggests several takeaways for the HRIA model. First, an integrated model, which combines the HRIAs of the different technologies and processes adopted within a multi-factor scenario, is essential to properly address the overall impact, including a variety of socio-technical solutions and impacted areas.

Second, the criticism surrounding civic participation in the Sidewalk project reveals how the effective engagement of relevant rightsholders and stakeholders is central from the earliest stages of proposal design. Giving voice to potentially affected groups mitigates the risk of the development of top-down and merely technology driven solutions, which have a higher risk of rejection and negative impact.

Third, the complexity and extent of large-scale integrated HRIA for multi-factor scenarios require a methodological approach that cannot be limited to an internal self-assessment but demand an independent third-party assessment by a multidisciplinary team of experts, as in traditional HRIA practice.

These elements suggest three key principles for large-scale HRIA: independence, transparency, and inclusivity. Independence requires third-party assessors with no legal or material relationship with the entities involved in the projects, including any potential stakeholders.

Transparency concerns both the assessment procedure, facilitating rightsholder and stakeholder participation, and the public availability of the assessment outcome,[113] using easily understandable language. In this sense, transparency is linked to inclusivity, which concerns the engagement of all the different rightsholders and stakeholders impacted by the activities examined (Table 2.13).

[111] Wylie 2020; Goodman and Powles 2019.

[112] For a more extensive discussion of this case: Scassa 2020; Morgan and Webb 2020; Artyushina 2020; Flynn and Valverde 2019; Peel and Tretter 2019; Carr and Hesse 2020a; Goodman and Powles 2019.

[113] Mantelero 2016, p. 766, fn 94 ("It is possible to provide business-sensitive information in a separate annex to the impact assessment report, which is not publicly available, or publish a short version of the report without the sensitive content").

Table 2.13 Multi-factor scenario HRIA: main stages and tasks

Main stage	Sub-section	Main tasks
I. Planning and scoping	A. Preliminary analysis	– Collection of information on the project, parties involved (including supply-chain), rightsholders, potential stakeholders, and territorial target area (country, region)[114] – Human rights reference framework: review of applicable binding and non-binding instruments, gap analysis
	B. Scoping	– Identification of main issues related to human rights to be examined – Drafting of a questionnaire for HRIA interviews and main indicators
II. Risk analysis and assessment	A. Fieldwork	– Interviews with rightsholders and internal/external project stakeholders,[115] interviews with experts, case studies on particular groups and individuals, and data collection[116] – Understanding of contextual issues (political, economic, regulatory, and social)
	B. Analysis and assessment	– Data verification and validation, comparing and combining fieldwork results and desk analysis – Further interviews and analysis, if necessary – Impact analysis for each project branch and impacted rights and freedoms – Integrated impact assessment report[117]
III. Mitigation and further implementation	A. Mitigation	– Recommendations – Prioritisation of mitigation goals
	B. Further implementation	– Post-assessment monitoring – Grievance mechanisms – Ongoing rightsholder and stakeholder engagement

Source The author

An additional important contribution of the integrated HRIA is its ability to shed light on issues that do not emerge in assessing single components of large-scale AI systems, as the cumulative effect of such projects is key. Here, the human rights layer opens up to a broader perspective which includes the impact of socio-technical solutions on democratic participation and decisions.

The Urban Data Trust created by Sidewalk and its role in the Toronto project is an example in this sense. The Urban Data Trust was tasked with establishing "a set

[114] The Danish Institute for Human Rights 2020c, pp. 13–18.

[115] Various interview techniques can be used in the assessment, such as focus groups, women-only group interviews, one-on-one interviews (key persons) and interviews with external stakeholders.

[116] Taking into account the circumstances, e.g. vulnerable groups, data could be collected anonymously through written submissions.

[117] The Danish Institute for Human Rights 2020e.

of RDU [Responsible Data Use] Guidelines that would apply to all entities seeking to collect or use urban data" and with implementing and managing "a four-step process for approving the responsible collection and use of urban data" and any entity that wishes to collect or use urban data in the district "would have to comply with UDT [Urban Data Trust] requirements, in addition to applicable Canadian privacy laws".[118]

This important oversight body was to be created by an agreement between Waterfront Toronto and Sidewalk Lab[119] and composed of a board of five members (a data governance, privacy, or intellectual property expert; a community representative; a public-sector representative; an academic representative; and a Canadian business industry representative) acting as a sort of internal review board and supported by a Chief Data Officer who, under the direction of the board, was to carry out crucial activities concerning data use.[120] In addition, the Urban Data Trust would have to enter into contracts with all entities authorised to collect or use urban data[121] in the district, and these data sharing agreements could also "potentially provide the entity with the right to enter onto property and remove sensors and other recording devices if breaches are identified".[122]

Although this model was later abandoned, due to the concerns raised by this solution,[123] it shows the intention to create an additional layer of data governance, different from both the individual dimension of information self-determination and the collective dimension of public interest managed by public bodies, within a

[118] Side Walk Labs 2019, vol. 2, p. 419 and vol. 3, p. 69. On the interplay the role of the Urban Data Trust in setting requirements for data processing and the legal framework into force in Canada and in Toronto, Scassa 2020.

[119] Scassa 2020, p. 55 ("in proposing the UDT, Sidewalk Labs chose a governance model developed unilaterally, and not as part of a collective process involving data stakeholders").

[120] Side Walk Labs 2019, vol. 2, p. 421 ("the Chief Data Officer would be responsible for developing the charter for the Urban Data Trust; promulgating RDU Guidelines that apply to all parties proposing to collect urban data, and that respect existing privacy laws and guidelines but also seek to apply additional guidelines for addressing the unique aspects of urban data […]; structuring oversight and review processes; determining how the entity would be staffed, operated, and funded; developing initial agreements that would govern the use and sharing of urban data; and coordinating with privacy regulators and other key stakeholders, as necessary").

[121] The notion of urban data is a novel category proposed by Sidewalk, referring to "both personal information and information that is not connected to a particular individual […] it is collected in a physical space in the city and may be associated with practical challenges in obtaining meaningful consent […] Urban data would be broader than the definition of personal information and include personal, non-personal, aggregate, or de-identified data […] collected and used in physical or community spaces where meaningful consent prior to collection and use is hard, if not impossible, to obtain", Side Walk Labs 2019, vol. 2, p. 416. But see, for critical comments on this category and its use, Scassa 2020, pp. 51–54; Goodman and Powles 2019, p. 473.

[122] Side Walk Labs 2019, vol. 2, pp. 420–422.

[123] Open Letter from Waterfront Toronto Board Chair, 31 October 2019. https://waterfrontoronto.ca/nbe/wcm/connect/waterfront/waterfront_content_library/waterfront+home/news+room/news+archive/news/2019/october/open+letter+from+waterfront+toronto+board+chair+-+october+31%2C+2019. Accessed 8 March 2021.

process of centralisation and privatisation of data governance regarding information generated within a community.[124]

In this sense, the overall impact of AI applications in urban spaces and their coordination by a dominant player providing technological infrastructure raise important questions about the cumulative effect on potentially impacted rights, and even more concerning democracy and the socio-political dimension of the urban landscape,[125] particularly in terms of the division of public and private responsibilities on matters of collective interest.

This privatisation of the democratic decision process, based on the 'platformisation' of the city, directly concerns the use of data, but is no longer just about data protection. In socio-technical contexts, data governance is about human rights in general, insofar as the use of data by different AI applications raises issues about a variety of potentially adverse effects on different rights and freedoms.[126] If data becomes a means of managing and governing society, its use necessarily has an impact on all the rights and freedoms of individuals and society. This impact is further exacerbated by the empowerment enabled by AI technologies (e.g. the use of facial recognition to replace traditional video-surveillance tools).

For these reasons, cumulative management of different data-intensive systems impacting on the social environment cannot be left to private service providers or an ad hoc associative structure, but should remain within the context of public law, centred on democratic participation in decision-making processes affecting general and public interest.[127]

Large-scale data-intensive AI projects therefore suggest using the HRIA not only to assess the overall impact of all the various AI applications used, but also to go beyond the safeguarding of human rights and freedoms. The results of this assessment therefore become a starting point for a broader analysis and planning of democratic participation in the decision-making process on the use of AI, including democratic oversight on its application.[128]

In line with the approach adopted by international human rights organisations, the human rights dimension should combine with the democratic dimension and the rule of law in guiding the development and deployment of AI projects from their earliest stages.[129]

[124] Artyushina 2020.

[125] Carr and Hesse 2020a; Powell 2021.

[126] E.g. Raso et al. 2018.

[127] The right to participate in public affairs (Covenant, Article 25) is based on a broad concept of public affairs, which includes public debate and dialogue between citizens and their representatives, with close links to freedom of expression, assembly and association. See UN Human Rights Committee (HRC) 1996. See also UN Committee on Economic, Social and Cultural Rights (CESCR) 1981, para 5.

[128] Mantelero 2020, pp. 82–88.

[129] See the Council of Europe's proposal discussed in Chap. 4.

The findings of the HRIA will therefore also contribute to addressing the so-called 'Question Zero' about the desirability of using AI solutions in socio-technical systems. This concerns democratic participation and the freedom of individuals, which are even more important in the case of technological solutions in an urban context, where people often have no real opportunity to opt out due to the solutions being deeply embedded in the structure of the city and its essential services.

A key issue then for the democratic use of AI concerns architecture design and its impact on rights and freedoms. The active role of technology in co-shaping human experiences[130] necessarily leads us to focus on the values underlying the technological infrastructure and how these values are transposed into society through technology.[131] The technology infrastructure cannot be viewed as neutral, but as the result of both the values, intentionally or unintentionally, embedded in the devices/services and the role of mediation played by the different technologies and their applications.[132]

These considerations on the power of designers – which are widely discussed in the debate on technology design[133] – are accentuated in the context of smart cities and in many large-scale AI systems. Here, the key role of service providers and the 'platformisation' of these environments[134] shed light on the part these providers play with respect to the overall impact of the AI systems they manage.

In this scenario, the HRIA can play an important role in assessing values and supporting a human rights-oriented design that also pays attention to participatory processes and democratic deliberation governing large-scale AI systems. This can facilitate the concrete development of a truly trustworthy AI, in which trust is based on respect for human rights, democracy and the rule of law.

[130] Manders-Huits and van den Hoven 2009, pp. 55–56.

[131] Ihde 1990.

[132] Latour and Venn 2002.

[133] Winner 1980; Winner 1983, p. 105 ("let us recognize that every technology of significance to us implies a set of political commitments that one can identify if one looks carefully enough. To state it more directly, what appear to be merely instrumental choices are better seen as choices about the form of the society we continually build, choices about the kinds of people we want to be"); Verbeek 2011, pp. 109, 129, and 164–165 ("Accompanying technological developments requires engagement with designers and users, identifying points of application fir moral reflection, and anticipating the social impact of technologies-in-design [...] In order to develop responsible forms of use and design, we need to equip users ad designer with frameworks and methods to anticipate, assess, ad design the mediating role of technologies in people's lives and in the ways we organize society").

[134] Council of Europe, Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108) 2019; Council of Europe, Committee of Ministers 2020.

2.5 Summary

The recent turn in the debate on AI regulation from ethics to law, the wide application of AI and the new challenges it poses in a variety of fields of human activities are urging legislators to find a paradigm of reference to assess the impacts of AI and to guide its development. This cannot only be done at a general level, on the basis of guiding principles and provisions, but the paradigm must be embedded into the development and deployment of each application.

With a view to providing a global approach in this field, human rights and fundamental freedoms can offer this reference paradigm for a truly human-centred AI. However, this growing interest in a human rights-focused approach needs to be turned into effective tools that can guide AI developers and key AI users, such as municipalities, governments, and private companies.

To bridge this gap with regard to the potential role of human rights in addressing and mitigating AI-related risks, this chapter has suggested a model for human rights impact assessment (HRIA) as part of the broader HRESIA model. This is a response to the lack of a formal methodology to facilitate an ex-ante approach based on a human-oriented design of product/service development.

The proposed HRIA model for AI has been developed in line with the existing practices in human rights impacts assessment, but in a way that better responds to the specific nature of AI applications, in terms of scale, impacted rights and freedoms, prior assessment of production design, and assessment of risk levels, as required by several proposals on AI regulation.[135]

The result is a tool that can be easily used by entities involved in AI development from the outset' in the design of new AI solutions, and can follow the product/service throughout its lifecycle. This assessment model provides specific, measurable and comparable evidence on potential impacts, their probability, extension, and severity, facilitating comparison between alternative design options and an iterative approach to AI design, based on risk assessment and mitigation.

In this sense, the proposed human rights module of the HRESIA is no longer just an assessment tool but a human rights management tool, providing clear evidence for a human rights-oriented development of AI products and services and their risk management.

In addition, a more transparent and easy-to-understand impact assessment model facilitates a participatory approach to AI development by rightsholders and potential stakeholders, giving them clear and structured information about possible options and the effects of changes in AI design, and contributing to the development of the ethical and social components of the HRESIA.[136]

[135] See Chap. 4.

[136] See Chap. 3.

Finally, the proposed model can also be used by supervisory authorities and auditing bodies to monitor risk management in relation to the impact of data use on individual rights and freedoms.

Based on these results, several conclusions can be drawn. The first general one is that conducting a HRIA should be seen not as a burden or a mere obligation, but as an opportunity. Given the nature of AI products/services and their features and scale, the proposed assessment model can significantly help companies and other entities to develop effective human-centric AI in challenging contexts.

The model can also contribute to a more formal and standardised assessment of AI solutions, facilitating the decision between different possible approaches. Although HRIA has already been adopted in several contexts, large-scale projects are often assessed without using a formal evaluation of risk likelihood and severity.[137] Traditional HRIA reports often describe the risks found and their potential impact, but with no quantitative assessment, providing recommendations without grading the level of impact, leaving duty bearers to define a proper action plan.

This approach to HRIA is in line with voluntary and policy-based HRIA practice in the business sector. However, once HRIA becomes a legal tool – as suggested by the European Commission and the Council of Europe[138] –, it is no longer merely a source of recommendations for better business policy. Future AI regulation will most likely bring specific legal obligations and sanctions for non-compliance in relation to risk assessment and management, as well as specific risk thresholds (e.g. high risk).

Analysis of potential impact will therefore become an element of regulatory compliance, with mandatory adoption of appropriate mitigation measures, and barriers in the event of high risk. A model that enables a graduation of risk can therefore facilitate compliance and reduce risks by preventing high-risk AI applications from being placed on the market.

With large-scale projects, such as smart cities, assessing each technological component using the proposed model and mitigating adverse effects is not sufficient. A more general overall analysis must be conducted in addition. Only an integrated assessment can consider the cumulative effect of a socio-technical system[139] by measuring its broader impacts, including the consequences in terms of democratic participation and decision-making processes.

This integrated assessment, based on broader fieldwork, citizen engagement, and a co-design process, can evaluate the overall impact of an entire AI-based environment, in a way that is closer to traditional HRIA models.

In both cases, figures such as the human rights officer and tools like a HRIA management plan, containing action plans with timelines, responsibilities and indicators, can facilitate these processes,[140] including the possibility of extending them to the supply chain and all potentially affected groups of people.

[137] E.g. The Danish Institute for Human Rights 2020f. But also see Salcito and Wielga 2015.

[138] See Chap. 4.

[139] Selbst et al. 2019.

[140] Abrahams and Wyss 2010.

Finally, the proposed model for the human rights component of the HRESIA model, with its more formalised assessment, can facilitate the accountability and monitoring of AI products and services during their lifecycle,[141] enabling changes in their impacts to be monitored through periodic reviews, audits, and progress reports on the implementation of the measures taken. It also makes it possible to incorporate more precise human rights indicators in internal reports and plans and make assessment results available to rightsholders and stakeholders clearly and understandably, facilitating their cooperation in a human rights-oriented approach to AI.

References

Abrahams D, Wyss Y (2010) Guide to Human Rights Impact Assessment and Management (HRIAM). International Business Leaders Forum, International Finance Corporation and UN Global Compact, Washington.

Access Now (2019) Laying down the Law on AI: Ethics Done, Now the EU Must Focus on Human Rights. https://www.accessnow.org/laying-down-the-law-on-ai-ethics-done-now-the-eu-must-focus-on-human-rights/. Accessed 7 April 2021.

Algorithm Watch (2020) Automating Society report 2020. https://automatingsociety.algorithmwatch.org/wp-content/uploads/2020/12/Automating-Society-Report-2020.pdf. Accessed 23 January 2021.

Artyushina A (2020) Is civic data governance the key to democratic smart cities? The role of the urban data trust in Sidewalk Toronto. 55 Telematics and Informatics, DOI: https://doi.org/10.1016/j.tele.2020.101456.

Aven T (2011) On Different Types of Uncertainties in the Context of the Precautionary Principle. Risk Analysis 31(10): 1515-1525.

Bennett CJ, Raab CD (2018) Revisiting the Governance of Privacy: Contemporary Policy Instruments in Global Perspective. Regulation & Governance 14(3): 447-464.

BEUC (2016) Connected Toys Do Not Meet Consumer Protection Standard. Letter to Mr Giovanni Buttarelli, European Data Protection Supervisor. https://www.beuc.eu/publications/beuc-x-2016-136_mgo_letter_to_giovanni_buttarelli_-_edps_-_connected_toys.pdf. Accessed 12 November 2020.

Bohn J, Coroamă V, Langheinrich M, Mattern F, Rohs M (2005) Social, Economic, and Ethical Implications of Ambient Intelligence and Ubiquitous Computing. In: Weber W, Rabaey JM, Aarts E (eds) Ambient Intelligence. Springer, Berlin, pp 5-29.

Carr C, Hesse M (2020a) Sidewalk Labs closed down – whither Google's smart city. Regions. https://regions.regionalstudies.org/ezine/article/sidewalk-labs-closed-down-whither-googles-smart-city/. Accessed 28 December 2020a.

Carr C, Hesse M (2020b) When Alphabet Inc. Plans Toronto's Waterfront: New Post-Political Modes of Urban Governance. Urban Planning 5:69-83.

Chang V, Li Z, Ramachandran M (2019) A Review on Ethical Issues for Smart Connected Toys in the Context of Big Data. In: Firouzi F, Estrada E, Mendez Munoz V, Chang V (eds) COMPLEXIS 2019 - Proceedings of the 4th International Conference on Complexity, Future Information Systems and Risk. SciTePress, Setúbal, pp 149–156.

[141] The Danish Institute for Human Rights 2020d, pp. 25–33.

City of Toronto, Deputy City Manager, Infrastructure and Development (2019) Report for action. EX6.1. https://www.toronto.ca/legdocs/mmis/2019/ex/bgrd/backgroundfile-133867.pdf. Accessed 30 December 2020.

City of Toronto, Executive Committee (2018a) Executive Committee consideration on January 24. 2018a.EX30. 9. http://app.toronto.ca/tmmis/viewAgendaItemHistory.do?item=2018a.EX30.9. Accessed 30 December 2020.

City of Toronto, Executive Committee (2018b) Executive Committee consideration on January 24, 2018b, 2018b.EX30. 9. Report and Attachments 1 and 2 from the Deputy City Manager, Cluster B on Sidewalk Toronto. https://www.toronto.ca/legdocs/mmis/2018b/ex/bgrd/backgroundfile-110745.pdf. Accessed 31 December 2020.

Commission of the European Communities (2000) Communication from the Commission on the precautionary principle, COM(2000) 1 final.

Costa L (2012) Privacy and the precautionary principle 28(1) Computer Law & Security Review 14–24.

Council of Europe (2018) Algorithms and Human Rights. Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications. Strasbourg.

Council of Europe, Committee of Ministers (2018) Recommendation CM/Rec(2018)7. Guidelines to Respect, Protect and Fulfil the Rights of the Child in the Digital Environment.

Council of Europe, Committee of Ministers (2020) Recommendation CM/Rec(2020)1 on the human rights impacts of algorithmic systems.

Council of Europe, Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108) (2017) Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data, T-PD(2017)01.

Council of Europe, Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108) (2019) Guidelines on Artificial Intelligence and Data Protection, T-PD(2019)01.

Council of Europe, Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108) (2020) Guidelines on Facial Recognition, T-PD(2020)03rev4.

Crawford K, Joler V (2018) Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources. AI Now Institute and Share Lab, New York. http://www.anatomyof.ai. Accessed 27 December 2019.

Data Ethics Commission (2019) Opinion of the Data Ethics Commission. https://www.bmj.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2. Accessed 7 June 2020.

de Paula Albuquerque O, Fantinato M, Kelner J, de Albuquerque Wheler AP (2020) Privacy in smart toys: Risks and proposed solutions. 39 Electronic Commerce Research and Applications, DOI: https://doi.org/10.1016/j.elerap.2019.100922.

Demetzou K, Böck L, Hanteer O (2018) Smart Bears don't talk to strangers: analysing privacy concerns and technical solutions in smart toys for children. In: IET Conference Proceedings. The Institution of Engineering & Technology, Stevenage, DOI: https://doi.org/10.1049/cp.2018.0005.

European Commission (2020) Study on Due Diligence Requirements through the Supply Chain: Final Report. Publications Office of the European Union.

European Commission (2021) Proposal for Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending legislative acts, COM(2021) 206 final, Brussels.

European Data Protection Supervisor (2015a) Decision of 3 December 2015a establishing an external advisory group on the ethical dimensions of data protection ('the Ethics Advisory Group') 2016/C 33/01 OJEU.

European Data Protection Supervisor (2015b) Opinion 4/2015b. Towards a new digital ethics: Data, dignity and technology.

European Digital Rights (EDRi) (2021) Civil Society Calls for AI Red Lines in the European Union's Artificial Intelligence Proposal. https://edri.org/our-work/civil-society-call-for-ai-red-lines-in-the-european-unions-artificial-intelligence-proposal/. Accessed 15 March 2021.

European Parliament (2020) Framework of ethical aspects of artificial intelligence, robotics and related Technologies European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL)), P9_TA-PROV(2020)0275.

European Parliament - Committee on Civil Liberties, Justice and Home Affairs (2020) Opinion of the Committee on Civil Liberties, Justice and Home Affairs for the Committee on Legal Affairs on artificial intelligence: questions of interpretation and application of international law in so far as the EU is affected in the areas of civil and military uses and of state authority outside the scope of criminal justice. 2020/2013(INI).

European Union Agency for Fundamental Rights and Council of Europe (2018) Handbook on European Data Protection Law. http://fra.europa.eu/en/publication/2018/handbook-european-data-protection-law. Accessed 25 May 2018.

Federal Bureau of Investigation (2017) Consumer Notice: Internet-Connected Toys Could Present Privacy and Contact Concerns for Children' Alert Number I-071717 (Revised)-PSA. https://www.ic3.gov/Media/Y2017/PSA170717. Accessed 15 December 2020.

Federal Trade Commission (2017) Enforcement Policy Statement Regarding the Applicability of the COPPA Rule to the Collection and Use of Voice Recordings. https://www.ftc.gov/public-statements/2017/10/federal-trade-commission-enforcement-policy-statement-regarding. Accessed 28 November 2020.

Floridi L et al. (2018) AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds & Machine 28:689–707.

Floridi L, Mariarosaria T (2016) What is data ethics? Phil. Trans. R. Soc. A. 374(2083), doi: https://doi.org/10.1098/rsta.2016.0360.

Flynn A, Valverde M (2019) Where The Sidewalk Ends: The Governance Of Waterfront Toronto's Sidewalk Labs Deal. Windsor Yearbook of Access to Justice 36:263–283.

Gibbs S (2015) Hackers can hijack Wi-Fi Hello Barbie to spy on your children, The Guardian, 26 November 2015. https://www.theguardian.com/technology/2015/nov/26/hackers-can-hijack-wi-fi-hello-barbie-to-spy-on-your-children. Accessed 12 November 2020.

Gonçalves ME (2017) The EU data protection reform and the challenges of big data: remaining uncertainties and ways forward. Inform. Comm. Tech. Law 26(2):90-115.

Gonçalves de Carvalho L, Medeiros Eler M (2018) Security Tests for Smart Toys. In: Proceedings of the 20th International Conference on Enterprise Information Systems 111–120. http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=https://doi.org/10.5220/0006776101110120. Accessed 23 December 2020.

Goodman E, Powles J (2019) Urbanism Under Google: Lessons from Sidewalk Toronto. Fordham Law Review 88:457–498.

Haber E (2019) Toying with Privacy: Regulating the Internet of Toys. Ohio State Law Journal 80:399.

Hansson SO (2020) How Extreme Is the Precautionary Principle? NanoEthics 14:245–257.

Ienca M, Vayena E (2020) AI Ethics Guidelines: European and Global Perspectives. In: Council of Europe. Towards regulation of AI systems. Global perspectives on the development of a legal framework on Artificial Intelligence systems based on the Council of Europe's standards on human rights, democracy and the rule of law. DGI (2020)16, pp 38–60.

Ihde D (1990) Technology and the Lifeworld: from garden to earth. Indiana University Press, Bloomington.

Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission (2019) Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed 15 April 2019.

Information Commissioner's Office (2020) Age appropriate design code. https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/age-appropriate-design-a-code-of-practice-for-online-services/. Accessed 20 February 2021.

International Covenant on Civil and Political Rights, Human Rights Committee (2011) General Comment no. 34. CCPR/C/GC/34.

Janssen HL (2020) An approach for a fundamental rights impact assessment to automated decision-making. International Data Privacy Law 10(1):76–106.

Kaminski ME, Malgieri G (2021) Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations. International Data Privacy Law 11(2):125–144.

Keymolen E, Van der Hof S (2019) Can I still trust you, my dear doll? A philosophical and legal exploration of smart toys and trust. Journal of Cyber Policy 4(2):143-159.

Koivisto R, Douglas D (2015) Principles and Approaches in Ethics Assessment. Ethics and Risk. Annex 1.h Ethical Assessment of Research and Innovation: A Comparative Analysis of Practices and Institutions in the EU and selected other countries. Project Stakeholders Acting Together on the Ethical Impact Assessment of Research and Innovation – SATORI. Deliverable 1.1. http://satoriproject.eu/work_packages/comparative-analysis-of-ethics-assessment-practices/. Accessed 15 February 2017.

Latour B, Venn C (2002) Morality and Technology: The End of the Means. Theory, Culture and Society 19(5-6):247-260.

Leta Jones M (2016) Your New Best Frenemy: Hello Barbie and Privacy Without Screens. Engaging Science, Technology, and Society 2:242-246.

Leta Jones M, Meurer K (2016) Can (and Should) Hello Barbie Keep a Secret? IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS), doi: https://doi.org/10.1109/ETHICS.2016.7560047.

LKL International Consulting Inc. (2014) Human Rights Impact Assessment of the Bisha Mine in Eritrea. https://media.business-humanrights.org/media/documents/files/documents/Nevsun_HRIA_Full_Report__April_2014_.pdf. Accessed 26 October 2020.

Lynskey O (2015) The Foundations of EU Data Protection Law. Oxford University Press, Oxford.

MacNaughton G, Hunt P (2011) A Human Rights-based Approach to Social Impact Assessment. In: Vanclay F, Esteves AM (eds) New Directions in Social Impact Assessment: Conceptual and Methodological Advances. Edward Elgar, Cheltenham, doi:https://doi.org/10.4337/9781781001196.00034.

Manders-Huits N, van den Hoven J (2009) The Need for a Value-Sensitive Design of Communication Infrastructures. In: Sollie P, Düwell M (eds) Evaluating New Technologies. Methodological Problems for the Ethical Assessment of Technology Developments. Springer, Dordrecht, pp 51–60.

Mann M, Matzner T (2019) Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. Big Data & Society 6(2), doi: https://doi.org/10.1177/2053951719895805.

Mantelero A (2016) Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection. Computer Law & Security Review 32 (2):238-255.

Mantelero A (2020) Analysis of international legally binding instruments. In Council of Europe. Towards regulation of AI systems. Global perspectives on the development of a legal framework on Artificial Intelligence systems based on the Council of Europe's standards on human rights, democracy and the rule of law. DGI (2020)16, pp 61–119.

Mantelero A (2021) The future of data protection: Gold standard vs. global standard. Computer Law & Security Review 40, doi: https://doi.org/10.1016/j.clsr.2020.105500.

McReynolds E, Hubbard S, Lau T, Saraf A, Cakmak M, Roesner F (2017) Toys That Listen: A Study of Parents, Children, and Internet-Connected Toys. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (ACM 2017). https://doi.org/10.1145/3025453.3025735. Accessed 12 November 2020.

Mertala P (2020) How Connectivity Affects Otherwise Traditional Toys? A Functional Analysis of Hello Barbie. Int. J. Child. Comput. Interact. 25, doi: https://doi.org/10.1016/j.ijcci.2020.100186.

Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: Mapping the debate. Big Data & Society. 3(2), doi: https://doi.org/10.1177/2053951716679679.

Morgan K, Webb B (2020) Googling the City: In Search of the Public Interest on Toronto's 'Smart' Waterfront. Urban Planning 5:84–95.

Murdoch J (2012) Protecting the Right to Freedom of Thought, Conscience and Religion under the European Convention on Human Rights. Council of Europe.

Myers West S, Whittaker M, Crawford K (2019) Discriminating Systems. https://ainowinstitute.org/discriminatingsystems.pdf. Accessed 13 June 2020.

Narayanan A, Huey J, Felten EW (2016) A Precautionary Approach to Big Data Privacy. In: Gutwirth S, Leenes R, De Hert P (eds) Data Protection on the Move. Springer, Dordrecht, pp 357-385.

Neff G, Nagy P (2016) Automation, Algorithms, and Politics| Talking to Bots: Symbiotic Agency and the Case of Tay. International Journal of Communication 10:4915–4931.

Neil M (2015) Moms Sue Mattel, Saying "Hello Barbie" Doll Violates Privacy. ABA Journal, December 9. https://www.abajournal.com/news/article/hello_barbie_violates_privacy_of_doll_owners_playmates_moms_say_in_lawsuit. Accessed 20 March 2021.

Norwegian Consumer Council (2016) #Toyfail An analysis of consumer and privacy issues in three internet-connected toys. https://fil.forbrukerradet.no/wp-content/uploads/2016/12/toyfail-report-desember2016.pdf. Accessed 14 December 2020.

Peel J (2004) Precaution - A Matter of Principle, Approach or Process? Melb. J. Int. Law 5 (2):483–501. http://www.austlii.edu.au/au/journals/MelbJlIntLaw/2004/19.html. Accessed 4 February 2017.

Peel K, Tretter E (2019) Waterfront Toronto: Privacy or Piracy? https://osf.io/xgz2s. Accessed 28 December 2020.

Pieters W (2011) Security and Privacy in the Clouds: A Bird's Eye View. In: Gutwirth S, Poullet Y, de Hert P, Leenes R (eds) Computers, Privacy and Data Protection: An Element of Choice. Springer, Dordrecht, pp 445-457.

Powell AB (2021) Undoing optimization : civic action in smart cities. Yale University Press, New Haven.

Raab C (2004) The future of privacy protection. Cyber Trust & Crime Prevention Project. https://www.piawatch.eu/node/86. Accessed 28 April 2017.

Raab C, Wright D (2012) Surveillance: Extending the Limits of Privacy Impact Assessment. In: Wright D, De Hert P (eds) Privacy Impact Assessment. Springer, Dordrecht, pp 363-383.

Raab CD (2020) Information Privacy, Impact Assessment, and the Place of Ethics. 37 Computer Law & Security Review DOI: https://doi.org/10.1016/j.clsr.2020.105404.

Raso F, Hilligoss H, Krishnamurthy V, Bavitz C, Kim L (2018) Artificial Intelligence & Human Rights Opportunities & Risks. https://cyber.harvard.edu/sites/default/files/2018-09/2018-09_AIHumanRightsSmall.pdf?subscribe=Download+the+Report. Accessed 28 September 2018.

Reisman D, Schultz J, Crawford K, Whittaker M (2018) Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability. https://ainowinstitute.org/aiareport2018.pdf. Accessed 29 June 2018.

Salcito K, Wielga M (2015) Kayelekera HRIA Monitoring Summary. http://nomogaia.org/wp-content/uploads/2015/10/KAYELEKERA-HRIA-MONITORING-SUMMARY-10-5-2015-Final.pdf. Accessed 20 February 2021.

Scassa T (2020) Designing Data Governance for Data Sharing: Lessons from Sidewalk Toronto. Technology & Regulation, Special Issue: Governing Data as a Resource, Technology and Regulation 44–56.

Scheinin M, Molbæk-Steensig H (2021) Pandemics and human rights: three perspectives on human rights assessment of strategies against COVID-19. https://cadmus.eui.eu/handle/1814/69576. Accessed 25 February 2021.

Selbst AD (forthcoming) An Institutional View Of Algorithmic Impact Assessments. 35 Harvard Journal of Law & Technology. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3867634. Accessed 7 August 2021.

Selbst AD, boyd d, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and Abstraction in Sociotechnical Systems. In: Proceedings of the Conference on Fairness, Accountability, and

Transparency (ACM 2019). https://doi.org/10.1145/3287560.3287598. Accessed 4 January 2020.

Shasha S, Mahmoud M, Mannan M, Youssef A (2019) Playing With Danger: A Taxonomy and Evaluation of Threats to Smart Toys. IEEE Internet of Things Journal 6(2):2986-3002.

Side Walk Labs (2019) Toronto Tomorrow. A new approach for inclusive growth. MIDP.

Spiekermann S (2016) Ethical IT Innovation: A Value-Based System Design Approach. CRC Press, Boca Raton.

Steeves V (2020) A dialogic analysis of Hello Barbie's conversations with children. Big Data & Society, 7(1), doi: https://doi.org/10.1177/2053951720919151.

Stirling A, Gee D (2002) Science, precaution, and practice. Public Health Reports 117(6):521–533.

The Danish Institute for Human Rights (2014) The AAAQ Framework and the Right to Water: International indicators for availability, accessibility, acceptability and quality, Copenhagen. https://www.humanrights.dk/sites/humanrights.dk/files/media/migrated/aaaq_international_indicators_2014.pdf. Accessed 24 June 2019.

The Danish Institute for Human Rights (2020a) Guidance and Toolbox. https://www.humanrights.dk/sites/humanrights.dk/files/media/dokumenter/udgivelser/hria_toolbox_2020a/eng/dihr_hria_guidance_and_toolbox_2020a_eng.pdf. Accessed 20 February 2021.

The Danish Institute for Human Rights (2020b) Guidance on HRIA of Digital Activities. Phase 1: Planning and scoping. Copenhagen. https://www.humanrights.dk/sites/humanrights.dk/files/media/document/HRIA%20Toolbox_Phase%201_ENG_2020b.pdf. Accessed 20 February 2021.

The Danish Institute for Human Rights (2020c) Guidance on HRIA of Digital Activities. Phase 2: Data Collection and context analysis. https://www.humanrights.dk/sites/humanrights.dk/files/media/document/Phase%202_Data%20Collection%20and%20Context%20Analysis_ENG_accessible.pdf. Accessed 20 February 2021.

The Danish Institute for Human Rights (2020d) Guidance on HRIA of Digital Activities. Phase 4: Impact prevention, mitigation and remediation. https://www.humanrights.dk/sites/humanrights.dk/files/media/document/Phase%204_%20Impact%20prevention%20mitigation%20and%20remediation_ENG_accessible.pdf. Accessed 20 February 2021.

The Danish Institute for Human Rights (2020e) Guidance on HRIA of Digital Activities. Phase 5: Reporting and Evaluation. https://www.humanrights.dk/sites/humanrights.dk/files/media/document/HRIA%20Toolbox_Phase%205_ENG_2020e.pdf. Accessed 20 February 2021.

The Danish Institute for Human Rights (2020f) Human Rights Impact Assessment – Durex and Enfa value chains in Thailand. https://www.humanrights.dk/publications/human-rights-impact-assessment-durex-enfa-value-chains-thailand. Accessed 2 March 2021.

The Danish Institute for Human Rights (2020g) Scoping practitioner supplement. Human rights impact assessment guidance and toolbox. https://www.humanrights.dk/sites/humanrights.dk/files/media/document/HRIA%20Toolbox_Phase%201_Scoping%20Prac%20Sup_ENG_2020g_0.docx. Accessed 2 October 2021.

Tosun J (2013) How the EU Handles Uncertain Risks: Understanding the Role of the Precautionary Principle. JEPP 20(10):1517-1528.

UN Committee on Economic, Social and Cultural Rights (CESCR) (1981) General Comment No. 1: Reporting by States Parties.

UN Human Rights Committee (HRC) (1996), CCPR General Comment No. 25: The right to participate in public affairs, voting rights and the right of equal access to *public* service (Art. 25), CCPR/C/21/Rev.1/Add.7.

UNICEF (2018) Children's Online Privacy and Freedom of Expression. https://www.unicef.org/csr/files/UNICEF_Childrens_Online_Privacy_and_Freedom_of_Expression(1).pdf. Accessed 18 December 2020.

UNICEF, The Global Compact, Save the Children (2012) Children's Rights and Business Principles. https://d306pr3pise04h.cloudfront.net/docs/issues_doc%2Fhuman_rights%2FCRBP%2FChildrens_Rights_and_Business_Principles.pdf. Accessed 30 November 2020.

United Nations (2011) Guiding Principles on Business and Human Rights. https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf. Accessed 8 December 2020.

Veale M (2020) A Critical Take on the Policy Recommendations of the EU High-Level Expert Group on Artificial Intelligence. European Journal of Risk Regulation, 1-10, doi:https://doi.org/10.1017/err.2019.65.

Verbeek P-P (2011) Moralizing Technology. Understanding and Designing the Morality of Things. The University of Chicago Press, Chicago.

Wachter S, Mittelstadt B, Russell C (2021) Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. West Virginia Law Review 123(3): 735-790.

Winner L (1980) Do Artifacts Have Politics? Daedalus 109(1):121–136.

Winner L (1983) Technē and Politeia: The Technical Constitution of Society. In: Durbin PT, Rapp F (eds) Philosophy and Technology. Springer, Dordrecht, pp 97-111.

World Bank, Nordic Trust Fund (2013) Human Rights Impact Assessments: A Review of the Literature, Differences with other forms of Assessments and Relevance for Development. World Bank and Nordic Trust Fund, Washington.

Wright D (2010) A framework for the ethical impact assessment of information technology. Ethics Inf. Technol. 13:199–226.

Wylie B (2020) In Toronto, Google's Attempt to Privatize Government Fails – For Now. Boston Review, 13 May.

Zuiderveen Borgesius FJ (2020) Strengthening legal protection against discrimination by algorithms and artificial intelligence. Int. J. Hum. Rights 24(10):1572-1593.