

Human Rights by Design for AI Technologies: Towards Outlining and Operationalizing an Emerging Field of Research and Practice

Leonardo Iwaya¹, Samuel Wairimu¹, Tara Mehrabi², Annika Rudman², Simone Fischer-Hübner^{1,3}

¹Department of Mathematics and Computer Science, Karlstad University, 651 88 Karlstad, Sweden

²Department of Social and Psychological Studies, Gender Studies, Karlstad University, 651 88 Karlstad, Sweden

³Chalmers University of Technology, 412 96 Gothenburg, Sweden
and the University of Gothenburg

leonardo.iwaya@kau.se

Abstract

The notion of human rights has gained traction in the context of complex sociotechnical systems, especially those involving Artificial Intelligence (AI), as it provides an established normative baseline for enforcement, rather than the existing myriad of vague ethical frameworks, codes of conduct, and guidelines. This movement has led to the emergence of the concept of Human Rights by Design (HRbD), denoting that the design of policies, technologies, and services ought to respect human rights from the outset and by default. To better shape this emerging field, this paper draws on existing academic literature to propose a definition of HRbD, articulate principles, and identify effective engineering strategies. Furthermore, an approach for operationalizing HRbD is presented, grounded on industry-academia collaboration and Participatory Action Research (PAR). This work also connects the field of HRbD to important initiatives, such as the European AI Act and ethics guidelines for trustworthy AI, and details the practical operationalization of HRbD across multiple PAR case studies as part of the researchers' ongoing work. In conclusion, this work advocates embedding HRbD into the design of new and existing AI systems, highlighting the important role of software engineers in the responsible development of technology.

Introduction

The deployment and use of Artificial Intelligence (AI) systems, while offering significant societal benefits, raise important concerns due to their potential impact on human rights. These concerns include embedded bias and discrimination, lack of transparency, and disruption amplifiers, whereby small errors may escalate into severe consequences, such as misdiagnosis in healthcare (Bansal et al. 2023; Leavy, Siapera, and O'Sullivan 2021; Macrae 2025). Such risks are particularly critical in high-risk AI systems, including those used in education, justice, and health and social care, where violations may affect rights enshrined in the Universal Declaration of Human Rights (UDHR) (United Nations 1948) and the European Convention on Human Rights (ECHR) (Council of Europe 1950).

A growing number of scholars and practitioners have therefore advocated for a "Human Rights by Design"

(HRbD) approach that embeds human rights into the design process of technological systems (Penney et al. 2018; Yeung, Howes, and Pogrebna 2020). This emerging field of research and practice draws heavily on the concept of privacy by design (Cavoukian 2009), which engineers are already familiar with, and expands its scope to encompass all dimensions of human rights (privacy included).

However, as an emerging field, the current literature still lacks a unified definition of HRbD that would help delineate the concept's scope and its main pillars. Furthermore, although current research also points to an increasing number of theories, frameworks, and methods related to HRbD, as well as to concrete engineering strategies for operationalizing it, the state of evidence remains incipient. Currently, most HRbD engineering strategies lack rigorous empirical studies that evaluate and/or validate their effectiveness in real-world practice, as shown in the systematic review by Wairimu et al. (2026).

With that in mind, this paper aims to define the concept of HRbD and summarize its principles and main engineering strategies, based on existing literature. As a result, this helps shape the field of HRbD by identifying influential work and scholars, comparing conceptualizations, and further articulating their positions on the topic. Additionally, considering the current gap of empirical research, this paper also proposes an approach to operationalizing HRbD methods (i.e., threat modeling methods, Human Rights Impact Assessments (HRIAs), and Trustworthy AI assessments), leveraging industry-academia collaboration and Participatory Action Research (PAR) case studies (Cornish et al. 2023; Runeson and Höst 2009) that are part of the researchers' ongoing work. Hence, this study also describes this approach, discussing how gamified threat modeling tools and impact assessment frameworks are translated into actionable steps. We also critically reflect on the application and limitations of our findings and approach in general.

This paper is structured as follows: Section 2 provides background on HRbD and presents influential work on the topic. Section 3 describes existing conceptualizations of HRbD and proposes a unified definition. Section 4 summarizes the core principles across the literature surveyed. Section 5 presents existing human rights engineering strategies, helping engineers with more concrete methods for HRbD. Section 6 details the researchers' approach to operationaliz-

ing HRbD in real-world industry case studies. Section 7 describes the limitations of the current study. Lastly, Section 8 presents the conclusion and future work.

Human Rights by Design: an Emerging Field

By acknowledging that information technology (IT) systems can both promote freedom and oppression, the impact of IT on human rights has been a topic of intense debate for decades (Metzl 1996). For instance, Brown, Clark, and Trossen (2010) have advocated that values such as human rights, with the UDHR as a primary example, should be “baked into” the design of the Internet architecture. Scholars also emphasize that technology has never been neutral (Cath and Floridi 2017), and engineers need to acknowledge that systems not *designed* to combat prejudice will often amplify it (Suzor et al. 2019).

With that in mind, there is an emphasis on addressing problems from the outset, at the earliest stages of system design, when changes are less costly and less complex. In recent years, this emphasis has been reflected in the term “by design,” which has also gained currency as a legal concept (Shwartz Altshuler et al. 2025). For instance, as enshrined in the EU General Data Protection Regulation, Article 25 stipulates the need for “data protection by design and by default” (European Parliament 2016).

The term data protection by design, in turn, originated in the notion of privacy by design, as proposed by Cavoukian (2009), explicitly based on the Fair Information Practice Principles (FIPPs) in 1973, and incorporating principles such as user control, transparency, and data minimization into proactive design processes. The seven foundational principles of privacy by design, defined by Cavoukian (2009), can be summarized as follows:

1. **Proactive not reactive; preventive not remedial:** Anticipates and prevents privacy-invasive events.
2. **Privacy as the default setting:** Personal data are automatically protected without requiring user action.
3. **Privacy embedded into design:** Privacy is a core functionality, not an “add-on” after the fact.
4. **Full functionality – positive-sum, not zero-sum:** Rejects false dichotomies (e.g., privacy vs. security) for “win-win” solutions.
5. **End-to-end security – full lifecycle protection:** Ensure privacy protection from data collection to destruction, the entire lifecycle.
6. **Visibility and transparency – keep it open:** Keeps component parts and operations open for independent verification.
7. **Respect for user privacy – keep it user-centric:** Keeps interests uppermost through user-centric defaults and empowering options.

Accordingly, many authors in the area of HRbD have acknowledged that the concept borrows heavily from privacy by design (Penney et al. 2018; Shwartz Altshuler et al. 2025; Kjaerum 2018), yet covers a much broader scope of rights, with privacy being just one of them. However, some

academics prefer to articulate HRbD from the notion of “legal protection by design” (Yeung, Howes, and Pogrebna 2020), originally proposed by Hildebrandt (2011), which also has original roots from privacy by design nonetheless. For a more detailed account of this evolution among the “by design” terms, including HRbD, we refer readers to Shwartz Altshuler et al. (2025).

Throughout our research, we have identified specific influential works that have substantially helped to articulate HRbD as an emerging concept and field of research and practice. Among the first, it is the work of Penney et al. (2018), articulating HRbD as a “*principle whereby they [i.e., businesses] commit to designing tools, technologies, and services to respect human rights by default, rather than permit abuse or exploitation as part of their business model.*” In this work, the authors explicitly modeled HRbD after the privacy by design paradigm, following the same rationale for HRbD. We also noticed that some authors addressed the concept of HRbD in their work without explicitly describing it, but referring to it as “*an ongoing commitment to ensure that human rights concerns are identified, considered, and addressed in the design and deployment of technological tools*” (Suzor et al. 2019).

A more comprehensive proposal comes from the work of Yeung, Howes, and Pogrebna (2020), who proposed a “human rights-centered design, deliberation and oversight” framework for AI governance. In their work, this framework is described as a “*governance regime for AI and other relevant automated systems (understood as complex socio-technical systems which include the data upon which they rely for training and operation) seeks to ensure that these systems will be human-rights compliant and reflect the core values that underpin the rule of law. It entails systematic consideration of human rights concerns at every stage of system design, development, and implementation (making interventions where this is identified as necessary).*” The proposed framework integrates a range of methods for ethical design strategies, tools, and techniques for software systems design, verification, testing, and auditing, as well as social and organizational measures for effective and legitimate governance. It is also structured based on four core principles, which can be briefly summarised as follows:

1. **Design and Deliberation:** Integrating human rights standards into technical architecture while providing opportunities for stakeholders to express views on potential impacts.
2. **Assessment, Testing, and Evaluation:** Regularly checking system behavior and datasets against human rights standards throughout the development process.
3. **Independent Oversight, Investigation, and Sanction:** Mandating external oversight by a properly resourced, independent regulator with legal powers to enforce compliance.
4. **Traceability, Evidence, and Proof:** Designing systems to provide an audit trail that serves as concrete evidence of human rights due diligence.

Together, the work of Penney et al. (2018) and the framework proposed by Yeung, Howes, and Pogrebna (2020)

can be seen as the central studies that helped conceptualize HRbD. Even though the work of Yeung, Howes, and Pogrebna (2020) does not explicitly mention the term “human rights by design” in its description, the adopted term “human rights-centered design” can be considered equivalent, as other scholars have noted (Shwartz Altshuoler et al. 2025). On the other hand, the work of Penney et al. (2018) explicitly uses and articulates the term HRbD, and has also been cited as a main reference by other scholars (Donahoe and Metzger 2019). Nonetheless, in this study, we aim to develop further an improved definition of HRbD that better outlines the meaning and scope of this emerging field.

Towards Defining HRbD

In addition to the works of Penney et al. (2018) and Yeung, Howes, and Pogrebna (2020), it is important to include other scholars who have discussed and articulated HRbD and to take their views into account when developing our definition. Table 1 provides a summary of the main conceptualizations of HRbD identified by our research team. We want to point out that we do not claim this is an exhaustive list of publications, but rather a set of key sources our team has identified over the past few years working on the topic.

As shown in Table 1, we highlight parts of the quotes that emphasize the main keywords and excerpts that we wanted to capture in a workable definition. Our team of researchers comprises computer scientists with expertise in trustworthy AI and privacy, as well as legal scholars and social scientists with expertise in human rights-based approaches, post-colonial perspectives, gender, and feminist technoscience. As part of an industry-academia collaboration, we are also co-producing our research with three business partners (both SMEs and large corporations) and one non-profit association, particularly in the digital health sector, alongside medical professionals, lawyers, software engineers, and bioinformaticians.

Based on the works identified (see Table 1) and the authors’ discussions and reflections on the topic, we derived the following definition for HRbD:

Human rights by design is the proactive and systematic embedding of international human rights standards into the core architecture and the entire lifecycle of sociotechnical systems – including their design, development, implementation, and real-world dynamic operational conditions and interactions with other systems. It requires that human rights concerns are identified, considered, integrated into the system design, and protected by default through preventative and mitigating technical and organizational measures, such as engineering strategies, rigorous impact assessments, and independent oversight, rather than through reactive, post-deployment fixes. Similarly, human rights by design can also be applied to the design of all policy- and law-making.

As a brief commentary, the above definition reflects a technology-agnostic approach, not limited to AI systems (Penney et al. 2018; Kjaerum 2018; Suzor et al. 2019; Donahoe and Metzger 2019; Yeung, Howes, and Pogrebna

2020), that stresses the need to embed human rights proactively into system design and its lifecycle. We also wanted to convey the notion of sociotechnical systems (Yeung, Howes, and Pogrebna 2020), emphasizing the interaction between people and technology and the need to consider both technical and organizational elements. The importance of HRIAs is also recognized as a key strategy for achieving HRbD (Donahoe and Metzger 2019; Yeung, Howes, and Pogrebna 2020; Council of Europe 2020). Additionally, the suggestion from Kjaerum (2018) regarding the HRbD for policy- and law-making was also integrated into the definition.

Core Principles of HRbD

In addition to the above definition, we aimed to summarize the core principles of HRbD presented in the identified literature. As the seven foundational principles of privacy by design, defined by Cavoukian (2009), have served as the conceptual basis, they are intrinsically reflected in the HRbD principles. Similarly, the work of Yeung, Howes, and Pogrebna (2020) defines four core principles for their human rights-centered approach to design, deliberation, and oversight that encompass fundamental governance activities.

With that in mind, we sought to compile principles from all sources into a unified list. This list is not meant to be interpreted as an exhaustive or fixed set of HRbD principles, but rather a living proposition that researchers and practitioners can further develop. Figure 1 illustrates how the main influential sources, i.e., Cavoukian (2009) and Yeung, Howes, and Pogrebna (2020), can also be mapped onto the unified list of principles. Here, it is also important to note that these mappings are just examples, as principles are, by nature, “soft and rubbery” and can overlap or be interpreted in nuanced ways.

Human Rights by Design and by Default & Systematic Integration

Many scholars advocate that human rights “by design” and “by default” should be systematically upheld throughout the entire system lifecycle - from data acquisition, to design, modeling, and deployment (Penney et al. 2018; Yeung, Howes, and Pogrebna 2020; Kjaerum 2018; Suzor et al. 2019; Donahoe and Metzger 2019; Council of Europe 2020; Shwartz Altshuoler et al. 2025). The authors often describe as “baked into” (Brown, Clark, and Trossen 2010; Cath and Floridi 2017), “embedding” (Cavoukian 2009; Shwartz Altshuoler et al. 2025) or the “hard-wiring” (Yeung, Howes, and Pogrebna 2020) of safeguards into the architecture of sociotechnical systems from the outset rather than treating them as add-ons or afterthoughts (Nemitz 2018; Cavoukian 2009; Penney et al. 2018; Donahoe and Metzger 2019).

However, it is also worth distinguishing between the terms “by design” and “by default,” a point that the identified literature did not sufficiently articulate. For example, Article 25 of the EU GDPR (European Parliament 2016) defines the notions of “by design” and “by default” as complementary obligations. They both address integrating data protection into systems from the outset, yet focus on different stages

Ref.	Main quotes	Notes
Cath and Floridi (2017)	<i>“the IETF [i.e., Internet Engineering Task Force] should enable the actualisation of human rights through the protocols and standards it designs by implementing a responsibility-by-design approach to engineering” (p. 449) [...] “a strategy of responding to value-sensitive and human rights-by-design questions in technical terms” (p. 453)</i>	(a) Cites Brown, Clark, and Trossen (2010) on the idea that universal values, e.g., human rights, to be “baked into” the Internet architecture.
Penney et al. (2018)	<i>“a ‘human-rights-by-design’ principle whereby they [i.e., businesses] commit to designing tools, technologies, and services to respect human rights by default, rather than permit abuse or exploitation as part of their business model.” (p. 106)</i>	(a) Provides an explicit description for HRbD.
Nemitz (2018)	<i>“The answer proposed here is that we need a new culture of technology and business development for the age of AI which we call ‘rule of law, democracy and human rights by design’” (p. 2) [...] “in AI and a three-level technological impact assessment for new technologies” (p. 2) [...] “And AI will only pass this test if by design, the principles of democracy, rule of law and compliance with fundamental rights are incorporated in AI, thus from the outset of program development.” (p. 13)</i>	(a) Describes the notion of “rule of law, democracy and human rights by design.” (b) HRbD is related to the conduction of HRIAs.
Kjaerum (2018)	<i>“The concept of ‘human rights by design’ is slowly emerging as the way forward. The concept has been borrowed from the IT sector, where programmers were challenged to introduce privacy aspects at the very beginning of the construction of a new IT tool or program.¹⁰ EU lawmakers can apply the ‘human rights by design’ concept in a similar fashion to all policy-making and law-making.” (p. 316)</i>	(a) Describes HRbD by drawing from privacy by design.
Suzor et al. (2019)	<i>“an ongoing commitment to ensure that human rights concerns are identified, considered, and addressed in the design and deployment of technological tools” (p. 87)</i>	(a) HRbD is only explicitly used in the title.
Donahoe and Metzger (2019)	<i>“Another new idea that is now drawing the attention of product developers is the concept of ‘human rights by design.’¹⁸ This rests on a willingness to ask the question “What will this technology mean for the human rights of the people who use it?” This work of reflection must take place as technologies are being developed – by means of human-rights-impact assessment processes, for instance – not after they have already been adopted. In order to fulfill their obligation to protect citizens, public authorities also need to step up their vigilance regarding what private companies are doing with AI.” (p. 124)</i>	(a) Cites Penney et al. (2018). (b) HRbD is related to the conduction of HRIAs.
Yeung, Howes, and Pogrebna (2020)	<i>“Our proposed governance regime for AI and other relevant automated systems (understood as complex socio-technical systems which includes the data upon which they rely for training and operation) seeks to ensure that these systems will be human-rights compliant and reflect the core values that underpin the rule of law. It entails systematic consideration of human rights concerns at every stage of system design, development and implementation (making interventions where this is identified as necessary).” (p. 8-9)</i>	(a) Describes their framework for “human rights-centered design.” (b) HRbD is related to the conduction of HRIAs.
Council of Europe (2020)	<i>“member states should conduct HRIAs before any AI system is designed, developed or deployed. For that process to be effective, it is essential to consider that human rights harms do not only stem from the use of AI systems but can be inflicted at any point during their lifecycle, including the data acquisition and modelling phases. In addition, it is important to pay attention to the dynamic conditions of the real-world environments in which the systems will be operating and interacting with other systems that are simultaneously at play.” (p. 9)</i>	(a) HRbD is only explicitly used in the title. (b) HRbD is related to the conduction of HRIAs.
Shwartz Altshuler et al. (2025)	<i>“The ‘By design’ legal framework was developed to embed fundamental rights such as privacy, autonomy, dignity, and fairness into technology architecture. Initially conceptualised as Privacy by design (PbD), this principle gradually expanded into broader paradigms such as Human Rights by design (HRbD), Ethics by design (EbD), and Responsible AI – each emphasising proactive governance through safeguards built into technologies rather than relying on post-deployment enforcement.” (p. 4) [...] “As AI technologies spread, policymakers recognised that privacy alone was insufficient. HRbD added principles of participation, accountability, and non-discrimination (Yeung et al., 2020);” (p. 5)</i>	(a) Cites Yeung, Howes, and Pogrebna (2020). (b) Describes HRbD by drawing from privacy by design.

Table 1: Summary of the main conceptualizations connected to HRbD.

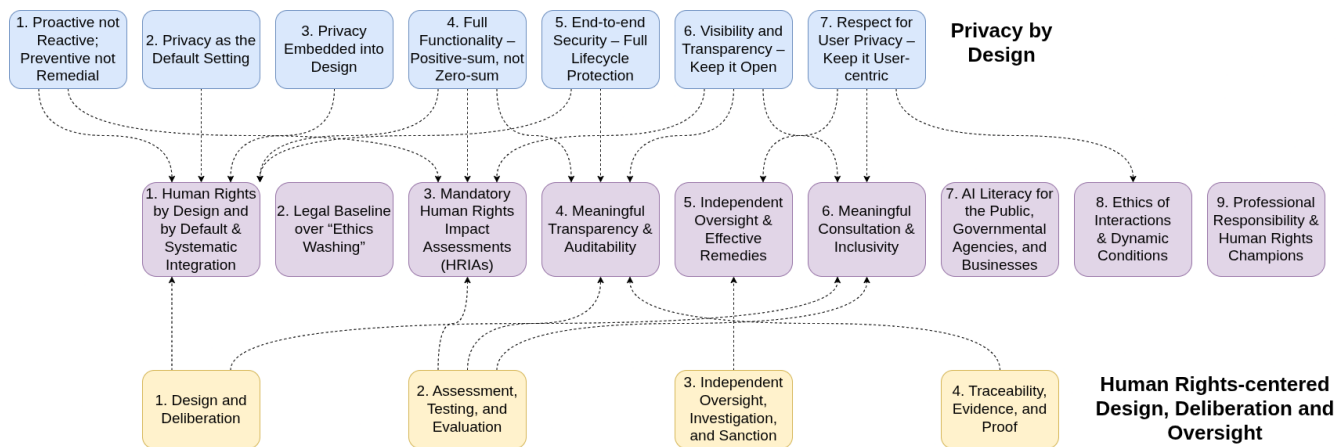


Figure 1: Examples of the mapping of principles from Cavoukian (2009) and Yeung, Howes, and Pogrebnina (2020) on a unified list of HRbD principles.

of the system lifecycle. “By design” (Article 25 I) refers to proactively embedding considerations into the architecture and development of systems from the initial planning. Ensuring the principles are built into the system’s core functionality rather than added as an afterthought, and identifying and mitigating privacy risks throughout the entire data processing lifecycle. On the other hand, “by default” (Article 25 II) focuses on the configuration and operational settings of those systems to ensure maximum protection without requiring user intervention. Only the necessary data for specific purposes is used, ensuring that the strictest settings are automatically enabled, preventing unnecessary data sharing or processing.

Legal Baseline over “Ethics Washing”

Many scholars expressed concerns about vague ethical frameworks, advocating for established international human rights law (e.g., the Universal Declaration of Human Rights (United Nations 1948) and the European Charter of Human Rights (Council of Europe 1950)) as normative baselines that are legally binding and enforceable (Cath and Floridi 2017; Nemitz 2018; Yeung, Howes, and Pogrebnina 2020; Donahoe and Metzger 2019; Council of Europe 2020). As stressed by Yeung, Howes, and Pogrebnina (2020), this principle seeks the end “Ethics Washing,” where companies use voluntary codes of conduct and guidelines to avert external regulation, by focusing on legally mandated standards and the rule of law. This is particularly necessary because tech giants lack strong incentives to self-regulate, and there is overwhelming evidence that they cannot be relied upon to honor their voluntary commitments (Yeung, Howes, and Pogrebnina 2020). It can also be added that the legal baseline for HRbD is seen as a strong motivation point, as its concrete historical roots and jurisprudence are also desirable when dealing with trade-offs and tensions among protection goals and when balancing rights must be justified (Bertaina et al. 2025; Janssen, Seng Ah Lee, and Singh 2022; Cosentini et al. 2025).

Nonetheless, Nemitz (2018) also highlights that the role

of ethical rules for technology still has a place, especially as precursors to the law – orienting toward the possible content of legal rules. He also notes that intra-company ethics and codes of conduct are good practices for engineers and leaders of major businesses (Nemitz 2018), even though they are not sufficient on their own.

Mandatory Human Rights Impact Assessments

HRIAs have been recognized as the primary tool for operationalizing HRbD, offering an *ex-ante* (or *ex-post*) systematic methodology for evaluating potential harms and impacts across the full spectrum of rights (Nemitz 2018; Donahoe and Metzger 2019; Yeung, Howes, and Pogrebnina 2020; Council of Europe 2020). HRIAs not only offer a proactive process but also support other principles, such as transparency, accountability, compliance, and stakeholder trust. HRIAs have also become mandatory in Europe for certain high-risk systems under Article 27 of the EU AI Act (European Parliament 2024), referred to as Fundamental Rights Impact Assessments (FRIAs). Given their importance in the surveyed scholarship, we expand on HRIAs and FRIAs in Section 5, which covers engineering strategies.

Meaningful Transparency & Auditability

Meaningful transparency, visibility, and openness mean disclosing the logic, decision-making processes, and purposes behind algorithms, including being clear about their accuracy and error rates and the potential trade-offs made during the design (e.g., model accuracy vs. data minimization) (Cavoukian 2009; Cath and Floridi 2017; Nemitz 2018; Donahoe and Metzger 2019; Council of Europe 2020). This is especially relevant for today’s AI systems, which are increasingly opaque, if not complete black boxes. In tandem with transparency, systems must be auditable, with traceable and accountable processes that leave an “audit trail” (Yeung, Howes, and Pogrebnina 2020) of evidence throughout the design process, operations, and automated decisions so that they can be effectively reviewed and contested by humans.

Furthermore, the AI Act (European Parliament 2024) also requires that providers clearly communicate to users and be transparent whenever a system is based on AI. For instance, letting users know whether they are talking to a human or an AI chatbot, or when a picture is generated by AI, etc. So this is an even simpler requirement for transparency than making the decision process transparent or explainable.

Independent Oversight & Effective Remedies

HRbD welcomes the establishment of public supervisory authorities with legal power to independently verify, monitor compliance, investigate abuses, and handle complaints (Cavoukian 2009; Yeung, Howes, and Pogrebna 2020; Council of Europe 2020). When harms occur, affected people (in particular vulnerable groups, including children) must have the right to redress and effective remedies, which may include compensation, reinstatement, the deletion of datasets, or the destruction of systems resulting from unlawful data processing (Council of Europe 2020; Suzor et al. 2019).

Meaningful Consultation & Inclusivity

HRbD deals with sociotechnical systems and is inherently an interdisciplinary area, for which systems should be designed through multistakeholder collaboration that involves integrative consultation with groups most likely to be affected (Cavoukian 2009; Council of Europe 2020). This process should ensure that affected groups are adequately consulted about the proposal and given opportunities to express their views (Yeung, Howes, and Pogrebna 2020). Furthermore, stakeholder consultation needs to be inclusive, especially when systems may impact vulnerable, minority, and marginalized populations (Council of Europe 2020).

AI Literacy for the Public, Governmental Agencies, and Businesses

Scholars also highlight the need to raise awareness and promote AI literacy among the public and government agencies, so they can effectively understand and participate in consultations when risks arise at the individual or collective levels (Donahoe and Metzger 2019; Council of Europe 2020). AI literacy is also needed among young technologists and engineers, and schools and universities should prioritize it (Donahoe and Metzger 2019). For the governmental agencies, it is also particularly important to consider HRbD as part of public procurement processes (Council of Europe 2020). Worth mentioning, Article 4 of the EU AI Act also defines that states and companies that create and use AI systems must make sure their employees and anyone else who operates or uses these systems on their behalf are well-educated about AI (European Parliament 2024).

It is worth adding that AI literacy requirements pose a significant research challenge for industry and academia, especially given the vast amount of information that must be carefully tailored for different industry sectors and application contexts. Such awareness-raising and training materials should not be limited to the software engineering teams but should be leveraged as a part of the organization's human

rights culture. Human rights should ultimately be part of organizations' espoused values and culture, with visible artifacts (e.g., AI policies and governance) and practices (e.g., technical and organizational measures) strongly promoted by senior leaders and top management (Relinger 2014). Creating a human rights culture requires extensive engagement of employees involved in the AI engineering process, ensuring they are well informed, with a change in mindset first (i.e., knowledge, attitudes, and perceptions) and behavior later (i.e., actions and practices). Thus, AI training and awareness programs can be seen as crucial for enabling this change in organizational culture.

Ethics of Interactions & Dynamic Conditions

Drawing specifically from the work of Schwartz Altshuler et al. (2025) on phygital spaces (i.e., physical and digital), the connection between HRbD and their proposed concept of "ethics of interactions" is highly relevant for the design of sociotechnical systems. In particular, this concept recognizes that immersive or spatial technologies (e.g., smart glasses, virtual and augmented reality) are disrupting social norms and collective dynamics due to the multiplicity of interaction modes (i.e., person-to-person, person-to-space, person-to-reality, person-to-machine, and person-to-platform). In turn, new scenarios emerge that can violate human rights (e.g., peer-to-peer surveillance, emotional inference, and corporate and government control over augmented realities). The ethics of interactions is also corroborated by other scholars who have highlighted the need to consider the "dynamic conditions" of real-world environments in which systems operate and interact, as well as the other systems (and people) that are simultaneously at play (Council of Europe 2020).

Here, it can also be added that the ethics of interactions should address the issue of humans overly trusting AI or using it as a kind of human companion. For instance, this refers to the phenomenon of automation bias in decision-making (Cummings 2017), in which humans tend to disregard or fail to seek contradictory information when a computer-generated solution is accepted as correct, and this can be exacerbated in time-critical domains. Another phenomenon is known as the ELIZA effect, defined by Hofstadter (1995) as the tendency of people to read far more into the output (a string of symbols and words) generated by computers than is warranted. Hence, again, transparency and a constant reminder that someone is communicating with a machine (not a human) are important.

Professional Responsibility & Human Rights Champions

The work of Cath and Floridi (2017) blends human rights as key values under their proposed approach of "responsibility-by-design." Penney et al. (2018) also refer to technologists as the "guardians of business ethics," reflecting on their individual decisions and ability to do good or cause harm when designing and operating systems. Given that, scholars underscore the need to foster and cultivate a "human rights consciousness" (Yeung, Howes, and Pogrebna 2020) into

engineering education and professional codes of ethics, so that engineers internalize human rights compliance as part of their core beliefs and values, and hopefully act as human rights champions within their organizations. Furthermore, connected to the idea of FRIAs, Nemitz (2018) also argues that they can help to instill a new “culture of responsibility” of technology for democracy, the rule of law, and fundamental rights. In turn, this connects back to the reflections on the principles of AI literacy and building a human rights culture within organizations.

Engineering Strategies for HRbD

It is also important to translate the notion of HRbD and its core principles into engineering strategies that technologists and software development organizations can adopt more concretely. Analogous to privacy by design evolving into privacy engineering (Spiekermann and Cranor 2008; Gürses, Troncoso, and Diaz 2011; Gürses and Del Alamo 2016; Iwaya, Babar, and Rashid 2023), it would be natural for HRbD to branch into a specific subject of “human rights engineering.”

According to the surveyed scholarship, engineers have long been familiar with the concepts of privacy by design and security by design and have incorporated privacy and security measures into system architecture for many decades (Brown, Clark, and Trossen 2010; Cath and Floridi 2017; Yeung, Howes, and Pogrebna 2020; Shwartz Altshuler et al. 2025). HRbD is therefore simply expanding these ideas into a broader set of requirements grounded in human rights.

Two other important engineering considerations suggested by scholars are the ideas of “desirable inefficiency” (Ohm and Frankle 2018) and “friction-in-design” (Frischmann and Benesch 2023). Their rationale is to introduce deliberate “speed bumps” (e.g., waiting periods, multi-step confirmations, or consent prompts) into AI systems to slow down high-risk decisions, encouraging further reflection, accountability, and informed user participation (Shwartz Altshuler et al. 2025).

Another aspect highlighted was the need to reject the ideology of value-neutral technological development, acknowledging that technology can be misused against people (e.g., Internet censorship, invasions of privacy, psychological manipulation) and that engineers should proactively design systems to combat such misuse (Cath and Floridi 2017; Suzor et al. 2019). In this regard, Yeung, Howes, and Pogrebna (2020) also highlights the need for more “value-sensitive design,” which was proposed by Friedman (1996) as a way to integrate human and moral values into system design, connecting architects and engineers with those affected by the systems and other stakeholders.

Furthermore, recent work by Wairimu et al. (2026) has specifically addressed the topic of “HRbD engineering strategies” through a systematic review. The work proposes a new taxonomy of human rights engineering strategies (e.g., frameworks, methods, models, and tools) based on 36 studies to support software engineers in HRbD. Their taxonomy is organized along three broad categories of engineering strategies: (1) *assessment and evaluation frameworks*,

comprising HRIAs and FRIAs, risk analysis methods, and trustworthy AI assessments (e.g., Mantelero (2018); Raman, Chabot, and Popken (2025)); (2) *conceptual frameworks and models for software design* that help integrate HRbD in the requirements engineering and system modeling stages (e.g., (Aizenberg and Van Den Hoven 2020; Hayama et al. 2025)); and (3) *ethical and trustworthy AI guidelines* for developers that specifically address human rights dimensions (e.g., Leavy, Siapera, and O’Sullivan (2021)).

Particularly in the context of impact assessments, recent work by Mantelero (2024) presents a detailed review of the development of FRIAs. Examples were the FRIA methodology for digital activities by the Danish Institute for Human Rights (DIHR 2025) and the Dutch government’s FRAIA (Government of the Netherlands 2024). Furthermore, other scientific/research-based FRIA/HRIA methodologies have been proposed (Mantelero 2022; Janssen, Seng Ah Lee, and Singh 2022; Casaburo and Marsh 2024; Bertaina et al. 2025), many of which leverage prior work on Data Protection Impact Assessments (DPIAs). However, most existing methodologies have not been evaluated in real-world settings (Wairimu et al. 2026). The scientific empirical evidence remains thin that such FRIA/HRIA methodologies will “deliver the goods” and truly support practitioners in identifying risks and integrating countermeasures into AI systems. There is a lack of evaluation of existing approaches, with HRIAs sometimes regarded as “policy-level documents” (Jørgensen 2019), unsuitable to effectively communicate impacts and measures to software practitioners and other stakeholders.

This significant communication gap between the legal and software engineering worlds may make it challenging to operationalize an HRbD strategy beyond mere legal compliance and to truly support companies that embed human rights throughout the design process. Suppose existing HRIAs methodologies are considered too vague and abstract (Janssen, Seng Ah Lee, and Singh 2022). In that case, software development teams will find them impractical, which can result in a legal team conducting assessments in isolation, defeating the purpose of influencing the system’s design.

All in all, the existing work reveals a critical mass of scientific studies in the emerging areas of HRbD. Still, it also highlights significant gaps in the evaluation and validation of engineering strategies in practice. Therefore, this study also aimed to present an approach for operationalizing and evaluating existing HRbD methods in collaboration with industry, as described below.

An Approach to Operationalize HRbD

Aiming to bridge the operationalization gap between human rights and engineering, we propose an approach that addresses the early stages of the design process and is rooted in risk and impact assessment methods. Figure 2 illustrates our three-step approach to operationalizing HRbD, comprising: (i) a gamified awareness-raising activity, (ii) the conduction of an interdisciplinary HRIA, and (iii) a broader trustworthy AI assessment.

The approach was co-produced with three industry partners, one non-profit association, and academia, and is grounded in action research (Petersen et al. 2014; Hegazy et al. 2025), drawing on three ongoing case studies of AI-based systems in the digital health and social care sectors – the complete reports on these case studies are part of our future work.

In particular, we follow the Participatory Action Research (PAR) approach (Cornish et al. 2023), which involves collaboration between researchers and industrial participants working closely to address real-world challenges. Following PAR, the approach is further developed through iterative cycles (Cornish et al. 2023) consisting of reflecting, planning, acting, and observing, which occur during workshops, meetings, and stakeholder consultations throughout the process. We also follow the case study guidelines proposed by Runeson and Höst (2009), ensuring adherence to best practices in empirical software engineering. In the following subsections, we discuss the three-step approach in detail, including the methodologies selected in each step.

It is worth noting, however, that the proposed approach is not comprehensive to the full scope of HRbD, which we further explain in Section 7 on the study's limitations. That is, it is expected that HRbD approaches will take further steps beyond the impact and risk assessments, also addressing the technical design. For instance, the work by Gürses, Troncoso, and Diaz (2011) on engineering privacy by design also includes steps such as *multilateral requirements analysis and implementation and testing of the design*, which should be part of an overarching design phase.

Step 1: Gamified Awareness-Raising

This first step of the approach is crucial as it raises awareness of AI-related threats and ethical considerations associated with the design and deployment of AI systems. Leveraging gamified approaches, especially card-based methods, has been shown not only to raise awareness but also to encourage discussion and promote critical reflection (Elsayed-Ali et al. 2023a).

Within our industry-academia collaboration, we introduced the Practical Library of Threats 4 Artificial Intelligence (PLOT4AI) (PLOT4AI 2025) as a concrete, gamified awareness-raising method for this step. PLOT4AI is a threat modeling methodology designed to identify and mitigate AI-related threats under eight main categories: (i) Ethics & Human Rights, (ii) Data & Data Governance, (iii) Transparency & Accessibility, (iv) Privacy & Data Protection, (v) Cybersecurity, (vi) Safety & Environmental Impact, (vii) Bias, Fairness & Discrimination, (viii) Accountability & Human Oversight. The methodology requires gathering relevant stakeholders to support the process, with an emphasis on multidisciplinary teams (e.g., engineers, legal, marketing, finance, and end-users). Nevertheless, it is worth noting that other methods could also be used instead of PLOT4AI, such as the responsible & inclusive cards (Elsayed-Ali et al. 2023b), the impact assessment cards (Bogucka et al. 2025), and Judgment Call (Ballard, Chappell, and Kennedy 2019), which are all card-based methods aimed at explaining AI risks and assessing ethical concerns, respectively.

During recurring research project meetings with industry partners, and based on their AI solutions, relevant stakeholders were identified to participate in the threat modeling activity. As mentioned, PLOT4AI consists of 138 cards, grouped into eight categories, each presenting a potential threat on one side and corresponding mitigations on the reverse. Given our focus on operationalizing HRbD, the cards under “Ethics & Human Rights” were especially emphasized during the workshop. However, other categories, such as “Bias, Fairness & Discrimination” and “Accountability & Human Oversight,” were also typically discussed.

The threats in this category are posed as reflective questions that encourage all participants (industry and academia) to consider what could go wrong from a sociotechnical perspective when ethics and human rights impacts are overlooked. By framing the threats as questions, participants are encouraged to engage in discussion, document their thoughts and reflections, and consider potential human rights violations associated with the deployment and use of their AI products.

Following this interactive exercise, the outputs of the first step are expected to include increased awareness of human rights-related AI threats, a preliminary assessment, and a summary of threat modeling, all of which serve as important inputs for the second step in our approach. We also collect field notes and documentation at every step to support the multiple PAR case studies underway.

Step 2: Interdisciplinary HRIA

The second step involves a formal and interdisciplinary process, as shown in Figure 2. While the first step involves a workshop in which stakeholders engage in open and reflective discussions, the second step entails conducting a structured HRIA with an interdisciplinary reference group comprising medical professionals, privacy and human rights lawyers, computer scientists, software engineers, and industry professionals. This means that the university team (external to the company) and reference group members take the lead in conducting the assessment independently, while engaging with the industry partner through periodic consultation meetings. These meetings help us validate our HRIA assessment phases, enabling further questioning and clarification.

To operationalize this interdisciplinary HRIA in a structured manner, aligned with the EU AI Act, we adopted the Danish Institute for Human Rights (DIHR) FRIA framework (DIHR 2025). Although this methodology was selected for the ongoing case studies, other HRIA/FRIA methodologies could also be used, such as the Human Rights, Ethical and Social Impact Assessment (HRESIA) (Mantelero 2018), the Dutch Fundamental Rights and Algorithmic Impact Assessment (FRAIA) (Government of the Netherlands 2024), or the Human Rights, Democracy, and Rule of Law Impact Assessment (HUDERIA) (Committee on AI 2024).

The DIHR FRIA consists of 5 phases: (i) planning and scoping, (ii) assessing and mitigating negative impacts on fundamental rights, (iii) deployment decision and public reporting, (iv) monitoring and review, and (v) consulting affected groups and stakeholders. The latter is cross-cutting,

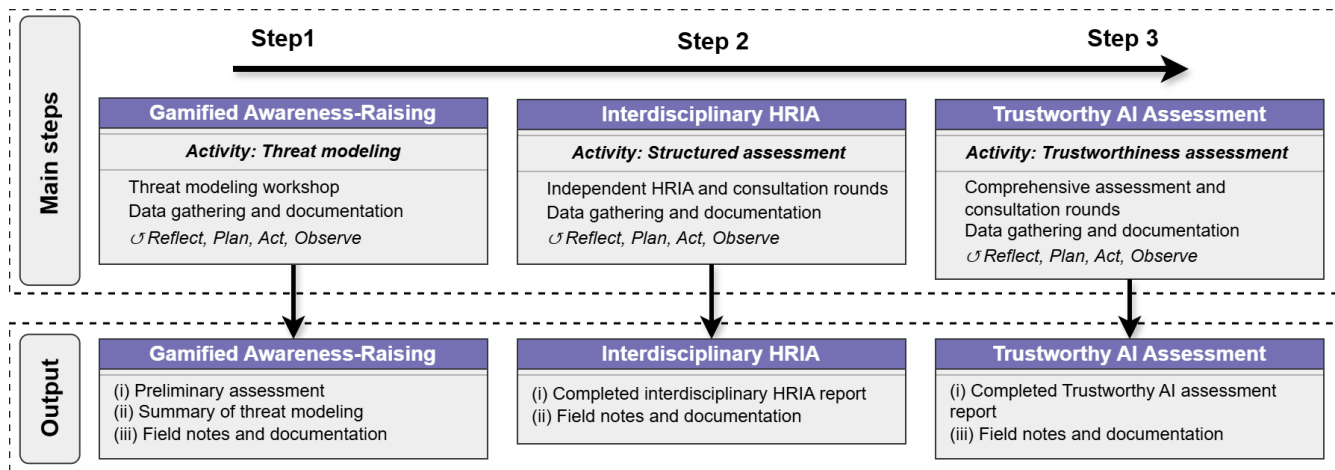


Figure 2: Proposed approach for operationalizing HRbD through PAR case studies and industry-academia collaboration.

meaning that stakeholder consultation occurs throughout all phases of the assessment process. While the assessment phases are conducted sequentially, meetings are held with industrial partners at the end of each phase to reflect on findings, gather additional information, and document relevant findings. This may include information that is not publicly available. The output of this step is a completed interdisciplinary FRIA for the companies.

Step 3: Trustworthy AI Assessment

The third step of the proposed approach assesses whether the AI systems satisfy established principles of “Trustworthy AI,” i.e., the High-Level Expert Group’s (HLEG) ethics guidelines and principles for Trustworthy AI¹. According to the ethical guidelines, trustworthy AI should be lawful, ethical, and robust (European Commission 2019). Therefore, this step assesses the trustworthiness of AI systems to determine whether they satisfy the seven key requirements for trustworthy AI (i.e., human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental well-being, and accountability).

This step also involves a comprehensive assessment, supported by consultations with the multidisciplinary team of experts (university team, reference group, and industry partners), data collection, and documentation activities to gather evidence related to the AI system and to support the evaluation against the trustworthy AI requirements. Similar to the previous steps, the process incorporates reflection and interactive discussion throughout the assessment, facilitated by a series of consultation meetings.

For this activity, we adopted the Z-Inspection process (Zicari et al. 2021) to guide our team in co-producing a comprehensive assessment for trustworthy AI. As in the previous steps, we acknowledge that alternative methods can also be used, such as the Assessment List for Trustworthy AI (AL-TAI) (European Commission 2020). Z-Inspection follows a

¹Ethics guidelines for trustworthy AI (<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>)

process in which ethical issues and tensions are identified and mapped to the requirements set out in the EU guidelines for trustworthy AI. Understanding how to handle trade-offs among multiple protection goals is also a research challenge we seek to address in our case studies. For instance, to protect privacy, developers may adopt data minimization strategies, but by collecting and training models with less data, they may also lose robustness and accuracy.

Research Findings Translation & Open Data

As part of our research approach, we also plan to translate research findings into HRbD guidelines for broad dissemination within the industry, benefiting researchers and practitioners. The objective is for our research findings to be openly available (i.e., in line with open science best practices, with data made FAIR – Findable, Accessible, Interoperable, and Reusable). Therefore, a dedicated online repository² has been created to archive guidelines, industry briefs, and other research artifacts (e.g., data, documentation, gamified approaches, and other learning tools), enhancing research impact and dissemination.

Limitations

Regarding the limitations of our research, it is important to note that HRbD is a field in its embryonic stage. As it evolves, more scientific studies and evidence will emerge, potentially leading to additional considerations that may affect the proposed definition and the set of core principles. Therefore, we are not proposing “the definition,” but rather “a” definition together with a set of principles for HRbD derived from the identified literature. With that in mind, we welcome changes to our proposals regarding HRbD, and we hope this study will help shed light on the topic and more clearly outline its scope and dimensions as an emerging field of research and practice.

²Open Repository for the Operationalizing Human Rights in the Design of AI-supported Digital Health (OPEHRA) project: <https://codeberg.org/OPEHRA>

Another important limitation concerns the scope of our proposed approach to operationalizing HRbD. In this approach, we primarily focus on threat modeling, impact analysis, and risk analysis methods, concentrating efforts in the early stage of the design process. However, the full scope of HRbD includes many more processes and activities. As revealed in the core principles of HRbD, several other governance activities could be addressed (e.g., audibility, independent oversight, and AI literacy). Even though our approach aims to provide a starting point for HRbD operationalization, we acknowledge that much more research and a range of approaches are needed to address HRbD in its full scope.

Conclusion and Future Work

As many scholars advocate, relying on vague ethical guidelines is insufficient; we should move beyond the “Ethics Washing” trap. This means shifting to HRbD, grounded in legally binding standards (e.g., the UDHR and the ECHR), to create enforceable guardrails. It is not reasonable to expect businesses to self-regulate through voluntary ethical guidelines and codes of conduct, as incentives to prioritize human rights may conflict with financial interests. In this regard, the EU AI Act takes an important step towards establishing independent oversight bodies with the authority to investigate harms, impose remedies (e.g., compensation, data deletion), and destroy unlawful systems – especially when they impact vulnerable groups (e.g., children, ethnic minorities, refugees).

For technologists and engineers, HRbD is a familiar concept that draws on established areas, such as “privacy by design” and “security by design.” Engineers are responsible for designing systems that integrate human rights protections as a core architectural principle, not an afterthought. As a start, they can leverage existing frameworks, such as threat modeling methods and HRIAs or FRIAs, to systematically evaluate risks across the system lifecycle. However, they also need to prioritize meaningful transparency and auditability, which requires developers to provide explainable or interpretable machine-generated decisions, disclose accuracy metrics, and clearly document trade-offs that affect human rights. Systems must also be designed with traceable “audit trails” to enable scrutiny by independent bodies and affected stakeholders. Without these safeguards, the risk of opacity and unexplainable “black boxes” threatens to exacerbate biases, discrimination, and unintended harms.

As part of our future work, we are evaluating our proposed approach to operationalizing HRbD through a series of PAR case studies underway. In addition to empirically assessing our approach, we are taking this opportunity to evaluate the specific methods selected (i.e., PLOT4AI, DIHR FRIA, and Z-Inspection), helping to fill a significant evidence gap in human rights engineering. Another research pathway we are pursuing involves creating tailored AI literacy materials aligned with the EU AI Act (European Parliament 2024) to support industry and non-profit partners in the health and social care sectors.

Acknowledgments

This work is funded by the Swedish Knowledge Foundation (KK-stiftelsen) via the Hög program, financing the Operationalizing Human Rights in the Design of AI-supported Digital Health (OPEHRA) research project. Leonardo H. Iwaya, Tara Mehrabi, and Simone Fischer-Hübner are also partly funded by Region Värmland (Grant: RUN/230445) and the European Regional Development Fund (ERDF) (Grant: 20365177) in connection with the DHINO 2 project, and by Vinnova (Grant: 2018-03025) via the DigitalWell Arena project. Simone Fischer-Hübner is also partly funded by the Wallenberg AI, Autonomous Systems and Software Program (WASP) by the Knut and Alice Wallenberg Foundation, and the EU Horizon Europe project TRUMAN on TRUstworthy and HuMAN-centric AI (Grant: 101214000).

We want to thank the industry and public-sector partners, AB Mavatar, Visiba Group AB, Pfizer AB, inTechrity i Karlstad AB, and the Swedish Association of Local Authorities and Regions (SALAR), for their feedback and discussions as part of the OPEHRA project reference group. The reference group participants are:

- **Fredrik Barrenäs** (AB Mavatar), Founder and Ph.D. in bioinformatics research, specializing in immunology, disease pathogenesis, and pharmaceuticals.
- **Parham Pourbozorgi** (AB Mavatar), Ph.D. in medical biochemistry with several years of experience in R&D informatics.
- **Hannah Gibson** (Visiba Group AB), Chief Commercial Officer, working with digital and AI patient consultation systems.
- **Linda Moris** (Pfizer AB), MD, Health Manager, working with healthcare access and innovation.
- **Sara Gandrén** (inTechrity), Founder and Lawyer, specializing in IT law and data protection regulation.
- **Birgitta Persdotter** (Karlstad University), Docent working at the R&D Welfare Värmland group, leading projects such as the National follow-up of the social services transition (NUSO) in collaboration with SALAR.

References

- Aizenberg, E.; and Van Den Hoven, J. 2020. Designing for human rights in AI. *Big Data & Society*, 7(2): 2053951720949566.
- Ballard, S.; Chappell, K. M.; and Kennedy, K. 2019. Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, 421–433.
- Bansal, C.; Pandey, K. K.; Goel, R.; Sharma, A.; and Jan-girala, S. 2023. Artificial intelligence (AI) bias impacts: Classification framework for effective mitigation. *Issues in Information Systems*, 24(4): 367–389.
- Bertaina, S.; Biganzoli, I.; Desiante, R.; Fontanella, D.; In-verardi, N.; Penco, I. G.; and Cosentini, A. C. 2025. Fundamental rights and artificial intelligence impact assessment: A new quantitative methodology in the upcoming era of AI Act. *Computer Law & Security Review*, 56: 106101.

- Bogucka, E. P.; Constantinides, M.; Šćepanović, S.; and Quercia, D. 2025. Impact Assessment Card: Communicating Risks and Benefits of AI Uses. *Proceedings of the ACM on Human-Computer Interaction*, 9(7): 1–42.
- Brown, I.; Clark, D. D.; and Trossen, D. 2010. Should specific values be embedded in the Internet architecture? In *Proceedings of the Re-Architecting the Internet workshop*, 1–6.
- Casaburo, D.; and Marsh, I. 2024. Ensuring fundamental rights compliance and trustworthiness of law enforcement AI systems: the ALIGNER Fundamental Rights Impact Assessment. *AI and Ethics*, 4(4): 1569–1582.
- Cath, C.; and Floridi, L. 2017. The design of the internet’s architecture by the Internet Engineering Task Force (IETF) and human rights. *Science and engineering ethics*, 23(2): 449–468.
- Cavoukian, A. 2009. Privacy by design: The 7 foundational principles. Information and privacy commissioner of Ontario, Canada.
- Committee on AI. 2024. Methodology for the risk and impact assessment of artificial intelligence systems from the point of view of human rights, democracy and rule of law (HUDERIA). Council of Europe.
- Cornish, F.; Breton, N.; Moreno-Tabarez, U.; Delgado, J.; Rua, M.; de Graft Aikins, A.; and Hodgetts, D. 2023. Participatory action research. *Nature Reviews Methods Primers*, 3(1): 34.
- Cosentini, A.; Pollicino, O.; De Gregorio, G.; Ermellino, A.; Fontanella, D.; Inverardi, N.; Paolucci, F.; Penco, I.; Regoli, D.; and Trapani, S. T. 2025. When Fundamental Rights Meet Artificial Intelligence: an Extended Framework for Impact Assessment in the Age of Generative AI. In *2025 International Joint Conference on Neural Networks (IJCNN)*, 1–9. IEEE.
- Council of Europe. 1950. The Convention for the Protection of Human Rights and Fundamental Freedoms. Council of Europe.
- Council of Europe. 2020. Human rights by design: future-proofing human rights protection in the era of AI. European Commission.
- Cummings, M. L. 2017. Automation bias in intelligent time critical decision support systems. In *Decision making in aviation*, 289–294. Routledge.
- DIHR. 2025. A guide to Fundamental Rights Impact Assessments under the EU Artificial Intelligence (AI) Act. The Danish Institute for Human Rights.
- Donahoe, E.; and Metzger, M. M. 2019. Artificial intelligence and human rights. *Journal of Democracy*, 30(2): 115–126.
- Elsayed-Ali, S.; Berger, S. E.; Santana, V. F. D.; and Becerra Sandoval, J. C. 2023a. Responsible & Inclusive Cards: An online card tool to promote critical reflection in technology industry work practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–14.
- Elsayed-Ali, S.; Berger, S. E.; Santana, V. F. D.; and Sandoval, J. C. B. 2023b. Responsible & Inclusive Cards: An Online Card Tool to Promote Critical Reflection in Technology Industry Work Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.
- European Commission. 2019. Ethics guidelines for trustworthy AI. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.
- European Commission. 2020. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. European Commission.
- European Parliament. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).
- European Parliament. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance).
- Friedman, B. 1996. Value-sensitive design. *interactions*, 3(6): 16–23.
- Frischmann, B.; and Benesch, S. 2023. Friction-in-design regulation as 21st century time, place, and manner restriction. *Yale JL & Tech.*, 25: 376.
- Government of the Netherlands. 2024. Fundamental Rights and Algorithms Impact Assessment (FRAIA). Government of the Netherlands.
- Gürses, S.; and Del Alamo, J. M. 2016. Privacy engineering: Shaping an emerging field of research and practice. *IEEE Security & Privacy*, 14(2): 40–46.
- Gürses, S.; Troncoso, C.; and Diaz, C. 2011. Engineering privacy by design. *Computers, Privacy & Data Protection*, 14(3): 25.
- Hayama, N.; Yamagata, Y.; Nishihara, H.; and Matsuno, Y. 2025. A gsn-based requirement analysis of the eu AI regulation. In *International Conference on Computer Safety, Reliability, and Security*, 183–196. Springer.
- Hegazy, S.; Elsner, C.; Bosch, J.; and Holmström-Olsson, H. 2025. Overcoming experimentation challenges in software ecosystems of large product and service organizations: A participatory action research study. *Journal of Systems and Software*, 112550.
- Hildebrandt, M. 2011. Legal Protection by Design: Objections and Refutations. *Legisprudence*, 5(2): 223–248.
- Hofstadter, D. R. 1995. *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. Basic books. <https://archive.org/details/douglas-r.-hofstadter-collected->

- works-pdf/Fluid%20Concepts%20and%20Creative%20Analogies%20by%20Douglas%20R.%20Hofstadter/ (accessed 2026-05-19).
- Iwaya, L. H.; Babar, M. A.; and Rashid, A. 2023. Privacy engineering in the wild: Understanding the practitioners' mindset, organizational aspects, and current practices. *IEEE Transactions on Software Engineering*, 49(9): 4324–4348.
- Janssen, H.; Seng Ah Lee, M.; and Singh, J. 2022. Practical fundamental rights impact assessments. *International Journal of Law and Information Technology*, 30(2): 200–232.
- Jørgensen, R. F. 2019. *Human Rights in the Age of Platforms*. MIT Press. <https://mitpress.mit.edu/9780262039055/human-rights-in-the-age-of-platforms/> (accessed 2026-05-19).
- Kjaerum, M. 2018. Human rights: Early days or coming to an end? *Netherlands Quarterly of Human Rights*, 36(4): 311–318.
- Leavy, S.; Siapera, E.; and O'Sullivan, B. 2021. Ethical data curation for AI: An approach based on feminist epistemology and critical theories of race. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 695–703.
- Macrae, C. 2025. Managing risk and resilience in autonomous and intelligent systems: Exploring safety in the development, deployment, and use of artificial intelligence in healthcare. *Risk Analysis*, 45(4): 910–927.
- Mantelero, A. 2018. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*, 34(4): 754–772.
- Mantelero, A. 2022. Beyond data. In *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI*, 1–43. Springer.
- Mantelero, A. 2024. The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, legal obligations and key elements for a model template. *Computer Law & Security Review*, 54: 106020.
- Metzl, J. F. 1996. Information technology and human rights. *Human Rights Quarterly*, 18(4): 705–746.
- Nemitz, P. 2018. Constitutional democracy and technology in the age of artificial intelligence. *Philosophical transactions of the Royal Society A: Mathematical, physical and engineering sciences*, 376(2133): 20180089.
- Ohm, P.; and Frankle, J. 2018. Desirable inefficiency. *Fla. L. Rev.*, 70: 777.
- Penney, J.; McKune, S.; Gill, L.; and Deibert, R. J. 2018. Advancing human-rights-by-design in the dual-use technology industry. *Journal of International Affairs*, 71(2): 103–110.
- Petersen, K.; Gencel, C.; Asghari, N.; Baca, D.; and Betz, S. 2014. Action research as a model for industry-academia collaboration in the software engineering context. In *Proceedings of the 2014 international workshop on Long-term industrial collaboration on software engineering*, 55–62.
- PLOT4AI. 2025. Practical Library Of Threats 4 Artificial Intelligence: A threat modeling library to help you build Trustworthy AI. PLOT4AI.
- Raman, V.; Chabot, C.; and Popken, B. 2025. Assessing Human Rights Risks in AI: A Framework for Model Evaluation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2107–2120.
- Relinger, R. 2014. *Embedding the Corporate Responsibility to Respect Human Rights within Company Culture*. John F. Kennedy School of Government.
- Runeson, P.; and Höst, M. 2009. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2): 131–164.
- Shwartz Altshuler, T.; Aridor Hershkovitz, R.; Mikulinsky, R.; and Müller, B. 2025. Governing phygital spaces: Human rights by design meets speculative design. *Internet policy review: Journal on internet regulation*, 14(4).
- Spiekermann, S.; and Cranor, L. F. 2008. Engineering privacy. *IEEE Transactions on software engineering*, 35(1): 67–82.
- Suzor, N.; Dragiewicz, M.; Harris, B.; Gillett, R.; Burgess, J.; and Van Geelen, T. 2019. Human rights by design: The responsibilities of social media platforms to address gender-based violence online. *Policy & internet*, 11(1): 84–103.
- United Nations. 1948. Universal Declaration of Human Rights. Office of the United Nations High Commissioner for Human Rights.
- Wairimu, S.; Iwaya, L. H.; Mehrabi, T.; Rudman, A.; and Fischer-Hübner, S. 2026. Human-Rights-by-Design Engineering Strategies for AI Systems: a Systematic Mapping Study. *ResearchGate*. <https://doi.org/10.13140/RG.2.2.28231.46244> (accessed 2026-05-21).
- Yeung, K.; Howes, A.; and Pogrebná, G. 2020. AI Governance by Human Rights–Centered Design, Deliberation, and Oversight: An End to Ethics Washing. In Dubber, M. D.; Pasquale, F.; and Das, S., eds., *The Oxford Handbook of Ethics of AI*. Oxford University Press. ISBN 9780190067397.
- Zicari, R. V.; Brodersen, J.; Brusseau, J.; Düdler, B.; Eichhorn, T.; Ivanov, T.; Kararigas, G.; Kringen, P.; McCullough, M.; Möslin, F.; et al. 2021. Z-Inspection®: a process to assess trustworthy AI. *IEEE Transactions on Technology and Society*, 2(2): 83–97.