AI Fairness-by-Design Multi-Stakeholder Methodology

-

A Comprehensive Framework for Fair AI Design and Development

AEQUITAS
unbias AI

ANNEX III: Fundamental Rights Impact Assessment for AI Fairness (FRIA-F)

Fair-by-Design sub-methodology

Abbreviation	Meaning
AFF	Affectees
AIU	AI Users
DDM	Development Decisionmakers
DE	Domain Experts
EGTAI	Ethics Guidelines for Trustworthy AI
FbD	Fair-By-Design
FDCGM	Fair Data Collection, Governance, and Management
FMM	Fair Model Methodology
FOIM	Fair Output Interpretation Methodology
FRIA-F	Fundamental Rights Impact Assessment for Fairness
GDM	Governance Decisionmakers
SIM	Stakeholder Identification Methodology
TAIRA	Trustworthy AI Readiness Assessment
MAP	Multistakeholder Approach to AI Fairness-by-Design
ML	Machine learning
NLP	Natural Language Processing

AEQUITAS
unbias AI

Contents

Introduction

The proposed Fundamental Rights Impact Assessment (FRIA) for AI fairness (FRIA-F) has been constructed as one of the 'Fair-by-Design Sub-Methodologies.'

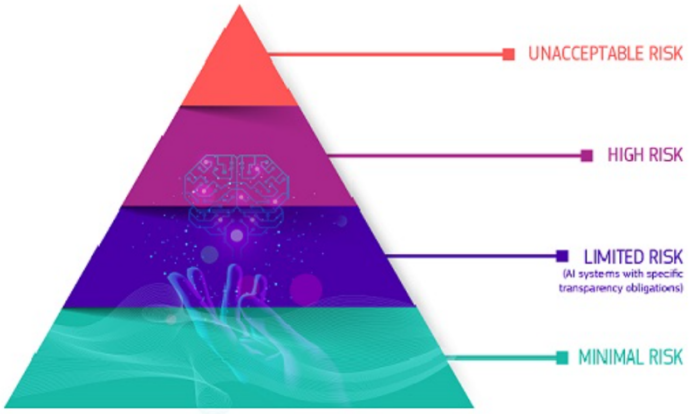
By conducting FRIAs, all parties involved in the AI lifecycle can proactively identify and eliminate or mitigate potential unfairness in their AI systems, safeguarding fundamental rights. Conducting FRIA-Fs helps prevent potential biases, as well as unintended consequences and engenders the protection of rights through the understanding of the consequences that technical, organisational or even commercial choices might have for the protection of fundamental rights.

First, this document explores the fundamentals: what risk impact assessments are, the concept of risk of AI, and some relevant legislations that mandate FRIAs. Next, we'll delve into a brief comparison of the different forms of FRIA and the essential elements that make up a robust FRIA. Finally, we'll culminate with a practical model for conducting a FRIA specifically focused on achieving fairness in AI systems.

FRIA – the fundamentals

This section contains a glossary of the most relevant terms and their meaning for the FRIA-F. The provided descriptions are derived primarily from the law, particularly the European Union's AI Act, and legal academic literature that considers the terms' interpretations.

Risk (Rights Impact) Assessment	In context of AI, Risk/Rights Impact Assessments are methodical, structured, and systematic processes that aim to identify, analyse, assess, and manage the effects that result from the development, implementation and use of an AI system. Risk/Rights Impact Assessments can either be mandated by the law or executed voluntarily. The assessment may also include other (sometimes overlapping) assessments such as for data protection, children's rights, environmental impact and corporate sustainability. While Risk/Rights Impact Assessments are often seen as upfront evaluations (ex-ante), their effectiveness hinges on ongoing monitoring. Monitoring deployment should continuously inform the identified risks, mitigation strategies, and ultimately lead to adjustments as needed.
Harm	In the context of the AI Act (Article 1), harm is seen as the effect of AI that jeopardizes a high level of protection of health, safety, fundamental rights enshrined in the Charter, including democracy, the rule of law and environmental protection.
Risk	According to the Article 3(2) of the AI Act, risk refers to the "combination of the probability of an occurrence of harm and the severity of that harm". From this perspective, risk can be conceptualised as the potential rather than the certainty of harm.[2] As Mireille Hildebrandt argues, the violation of rights may

	not necessarily demand that harm is realized, but rather the potential for such harm can be enough.[3] While health, safety, and fundamental rights are widely recognized as high-risk areas for AI development and deployment, risk can manifest in other categories. For instance, the use of AI systems is a significant factor that aggravates climate change and ecosystem disruptions. Nevertheless, the scope of this document centres around the analysis of risks to fundamental rights, whereby any risks posed thereto may be considered as high.
Risk categorisation	Also referred to as the risk pyramid, the AI Act categorizes AI practices and AI systems with an intended purpose, based on the acceptability and unacceptability of their risk to health, safety and fundamental rights. The correct categorisation of a developed or used AI system into one of the risk categories (unacceptable, high-risk, limited or minimal risk) is essential for the identification of legal prohibitions or requirements associated with the AI system. Illustrated by the European Commission, the risk categories under the AI Act are as follows:[4] 
Unacceptable Risk	Unacceptable risk poses a significant threat of violating fundamental rights. The severity of potential violations is deemed too high to allow the AI system to enter the EU internal market. This is affirmed by AI Act, which notes that the list of prohibited AI practices includes “all those AI systems whose use is considered unacceptable as contravening Union values, for instance by violating fundamental rights.” A list of prohibited AI systems is provided under Article 5 of the AI Act. Prevention of unacceptable harm is also mentioned in the Ethics Guidelines of Trustworthy AI, under the principle of robustness. The Guidelines stipulate that “AI systems be developed (…) in a manner such that they reliably behave as intended while minimizing unintentional and unexpected harm and preventing unacceptable harm.”[5]

High-Risk	According to the AI Act and the European Union's risk classification for AI systems, high-risk systems are those AI applications that pose a significant risk of harm to health, safety or fundamental rights and would subsequently have to abide by a series of requirements and obligations before they can be put on the EU internal market.[6] The criteria for high-risk AI system categorization under the AI Act goes beyond those based on fundamental rights considerations and are outlined in Article 6. Article 6 focuses on (1) the intended purpose and product category of the system, and (2) its use cases and potential to harm individuals' fundamental rights, (3) the specific functionalities or uses of the system.
Limited Risk	
Residual Risk	The leftover risk that persists despite the application and implementation of mitigation strategies and efforts. Residual risk can be considered tolerable when eliminating all risk entirely is not possible, when the residual risk is minor, and the benefits of AI deployment largely outweighs it.
Minimal Risk	AI systems that raise 'minimal risk' to health, safety or fundamental rights.
Systemic risk at Union level	According to Art 3.65 'means a risk that is specific to the high-impact capabilities of **general purpose** AI models, having a significant impact on the internal market due to its reach, and with actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain.
Impact	Impact can be conceptualised as an observable or measurable consequence that is caused by, in the context of this document, the activities led within the AI system's lifecycle. Impacts can be either beneficial or detrimental to an individual. Importantly, impact perception plays a crucial role in how we understand the intensity of an impact, even beyond the attributes (positive, negative) of the impact itself. While the scope, duration, and target of an impact can, to some extent, be measurable, its perceived strength is often subjective and shaped by those affected and the socio-political context in which the impact materialises.
Significant Risk of Harm	Seen in relation to the notion of 'harm severity' that stems from the definition of risk in the AI Act (see above), a significant risk harm is assumed for all high-risk AI systems that fall under the high-risk categorisation of the AI Act (see different deliverable). According to Articles 6.2 and 6.3 of the Act and the accompanying relevant Recitals (Recital 53), high risk AI systems are deemed not to hold a significant risk of harm if they intend to do either one of the following:

	(1) perform only a narrow procedural task (e.g., data structuring, document classification) (2) improve the results of a completed human activity (e.g., language improvement in documents, language styling, brand alignment) (3) detect (deviations from) decision-making patterns but without replacing or influencing the previous human assessment without the meaningful review of the results of this pattern identification exercise by a person (e.g., after human assessment, analysing grading patterns) (4) perform a preparatory task to an assessment relevant for the purpose of the use cases listed in Annex III (e.g., indexing files, search functions, text processing, data source allocation, translation) According to Recital 53 the systems that would fall under those categories do not (or, therefore, should not) either materially influence the decision-making pertaining to an individual's health, safety and fundamental rights, nor harm those three categories in a substantial manner. In other words, shall a material influence of decision-making or a harm to health, safety or fundamental rights occur, then the system will be considered high risk, and as posing significant harm. The Recital further clarifies that material influence means that that the outcome or the substance of either human or automated decision-making is affected. Because the exceptions to significant risk / harm are to be considered narrowly and only upon thorough documentation of the AI system provider, the departing assumption is that all (high-risk AI) systems maintain a significant risk of harm until proven to perform a strictly interpreted narrow function as described above. For the purposes of the present FRIA-F, it is important to merely retain that all harms presented to fundamental rights are deemed as significant, and that all systems that can be classified as high-risk AI are (regardless of the outcome of the FRIA and whether the risk manifests) believe to hold significant risk of harm.
Serious incident	Article 3.49 states that a 'serious incident' is any incident or malfunctioning of a general purpose AI system that directly or indirectly leads to any of the following: (a) the death of a person or serious damage to a person's health; (b) a serious and irreversible disruption of the management and operation of critical infrastructure. (c) breach of obligations under Union law intended to protect fundamental rights; (d) serious damage to property or the environment

High impact capabilities	Art 3.64 ‘high-impact capabilities’ in general purpose AI models means capabilities that match or exceed the capabilities recorded in the most advanced general purpose AI models.
Affected population	Refers to the number of individuals (1-to-n) who are exposed to the (potential) negative consequences or adverse effects of the development, use and deployment of AI systems, throughout or only during one phase of the AI lifecycle. The experience may be either direct or indirect, tied to a specific or multiple events or phenomena. The affected population may sometimes be easily observable (seen), or may only come into view by means of the cascading effect of a risk (unseen). In order to identify the stakeholders and the affected population, please also refer to the Stakeholder Identification Methodology (D6.3.)
Risk management system	Referred to in Article 9 of the AI Act, high risk AI providers must establish, maintain, implement a risk management system that spans the entire lifecycle of AI systems. Specifically, Article 9.2.(a) prescribes the assessment of known and the reasonably foreseeable risks that the high-risk AI system can pose to health, safety or fundamental rights (a FRIA) based on the assumption that the high-risk system is used in accordance with its intended purpose is one step of the risk management process. The requirements of the risk management system under Article 9, including in particular that of the FRIA, concern AI providers, and should also be considered separately from the FRIA that is to be performed by some AI system deployers according to the specifications of Article 27. Though the FRIA itself may follow similar considerations or logic, the stakeholders behind each of the analysis are different. In this document, we propose a FRIA model that can be replicated in both cases, subject to the critical assessment of each party.
Risk elimination and mitigation requirements	Risk mitigation strategies are proactive measures and precautions that are identified and subsequently enacted to minimise or mitigate risk. The EU's AI Act exemplifies this concept by outlining specific requirements for high-risk AI systems (Articles 9-15). These requirements encompass various aspects, including data governance practices, technical documentation standards, and the crucial role of human oversight. Furthermore, the Act emphasizes the importance of conformity assessments, adherence to certification obligations, and ongoing evaluation of safeguard effectiveness

FRIA for AI Fairness

In 2019, the AI High Level Expert Group on AI presented Ethics Guidelines for Trustworthy AI[1]. These guidelines define trustworthy AI as being lawful, ethical and socio-technically robust. For the ethical element of trustworthy AI, the guidelines explicitly take fundamental rights as a basis for AI ethics. This stemmed from the recognition that AI never operated in a lawless world and that existing laws, regulations and treaties apply to AI as much as they apply to any other technology. In the landscape of digital EU law, the upcoming AI Act is one of the most recent regulations that is set to govern the placement on the market and put into use of AI systems in the EU. The principal aim of this regulation is to safeguard health, safety and fundamental rights from the adverse effects and risks derived from and within the AI lifecycle.

The regulation proposes a categorisation of AI systems based on the perceived risks they pose to health, safety and fundamental rights. The higher the risk, the stricter the rules for such systems to enter the EU market.[7]

Once an AI-system is categorised as high risk, providers and public sector deployers of such a system must perform a fundamental rights impact assessment (FRIA) (article 9(2)(a) and art. 27 of the AI Act respectively).

Though no further explanation is provided on how to perform a FRIA, providers need to document detailed information regarding the foreseeable unintended outcomes and sources of risks to health and safety, fundamental rights and discrimination in view of the intended purpose of the AI system (art. 11 jo. ANNEX IV para. 3), while public sector deployers must describe and analyse the following elements (art. 27(1)):

- The processes wherein the AI system will be used according to its intended purpose
- Temporality of AI system deployment (time and frequency)
- Affected population
- The risks of harm that are likely to impact the identified affected population
- Human oversight measures aligned with the instructions relayed by the provider
- Response plan in case of risk materialisation and attributed internal governance structure including complaint mechanisms

We note that, while the AI Act does not explicitly prescribe documenting these elements for providers, they're still crucial considerations.

Here, the described interconnection between the (risk-prevention) provisions of the AI Act and the rights granted in the ECFR are important, as one cannot be achieved without respect of the other. In entangling the two regulations, the EU strives to sustain a preventative approach to risk management, therefore applying the precautionary principle.

Under the preventative approach to risk management, we understand the proactive process that seeks to minimize risk before their materialisation and includes the following key aspects that are also seen in the FRIA process: risk identification, risk analysis, and risk management

[1] 'Ethics Guidelines for Trustworthy AI | Shaping Europe's Digital Future', 8 April 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

(elimination, reduction, acceptance, monitoring). This ex-ante approach seeks to reduce costs in the long run by preventing incidents, improve efficiency by addressing a streamlined solution management system, and promotes a culture whereby safety is considered throughout the organisation and AI lifecycle. This preventative approach is underpinned by the precautionary principle.

Hence, the risk elimination and mitigation strategies identified as part of the preventative approach to shielding fundamental rights in the context of AI systems and in accordance with the requirements of the AI Act, should be interpreted through the lens of the precautionary principle.

The final crucial consideration regarding fundamental rights, and FRIAs for AI is scalability. Fundamental rights apply equally to all individuals in the EU, regardless of the severity of a potential violation. Even when a single person is impacted by a fundamental rights violation, however minor this violation in itself may seem to be, it is significant when it comes to risk/impact assessments.

Imagine an AI system that erroneously restricts welfare distribution to a given individual. This restriction can have severe (non-)observable long-term consequences on this person's livelihood, health, and ability to exercise other fundamental rights. While scalability of violations, impact or risk might be one factor in assessing likelihood and severity, it should not be the sole focus. Even a single person affected is a concern, and the associated risk should be eliminated or sufficiently mitigated.

First and foremost, we note that all fundamental rights enshrined in the ECFR may be equally, and concomitantly affected by the design, development and deployment of AI systems.[9] The primary focus of the FRIA-F however, lies on fairness related legal norms that stem from the AI Act and the Charter of Fundamental Rights of the European Union (ECFR). Further inspiration is drawn from the Council of Europe's European Convention on Human Rights for the Protection of Human Rights and Fundamental Freedoms (ECHR) and the recently adopted Council of Europe's Convention on the Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law (AI Convention).

Fairness related fundamental rights

Fundamental rights which are most relevant to fairness include:

- Prohibition of Discrimination, art. 14
- Equality before the law, art. 20
- Non-discrimination, art. 21
- Cultural, religious and linguistic diversity, art. 22
- Equality between women and men, art. 23
- The rights of the child, art. 24
- The rights of the elderly, art. 25
- Integration of persons with disabilities, art. 26

Concomitantly, other fundamental rights could be affected because of infringement of the above equality, non-discrimination and diversity rights, such as:

- Right to life, liberty and security art. 2 and 6
- Fair and just working conditions, art. 31
- Presumption of innocence, art. 48
- Access rights to:
  - Education, art. 14
  - Social Security, art. 34
  - Healthcare, art. 35
  - Services of general economic interest, art. 36

It is essential to note that the proposed FRIA-F structure can be replicated for all (other) fundamental rights to achieve a complete overview of an AI's system's impact and the respective risk management techniques.

In contrast to existing FRIA's, that often only ask whether a system poses a risk to a certain fundamental right, FRIA-F utilizes reverse engineering tactics to encourage critical thinking towards fundamental rights protection. It hence aims to be a useful tool to aid people without in-depth knowledge of fundamental rights to understand whether a technical or organisational aspect of their system could negatively affect which fundamental right and if so, how this could be addressed. When used alongside other 'fair-by-design' methodologies, this approach promotes the development of fair AI systems and sustain fair activities throughout the AI lifecycle.

Fundamental Rights Risk Identification - reverse engineered

The following methodology is meant to establish an if-this-then-that logic, employing reverse engineering tactics, by identifying first any risk-triggers in the AI systems' design, development and use, and connecting those triggers to potential fairness related fundamental rights violations. Because the AI Act aims to protect against these violations, the methodology also helps identifying the relevant prohibited practices (art. 5) or risk elimination and mitigation requirements (art. 9 – 15 AI Act) related to the trigger. At this stage, we document only focusses on prohibited and high risk AI systems, and not on any other requirements and obligations (such as for limited risk AI systems and general purpose AI models.

The document focuses on the AEQUITAS uses cases, however the methodology could be generally applied to other use cases as well. The FRIA-F is best conducted at an early stage during the AI lifecycle. In the Multi-Stakeholder AI Fair-by-Design Methodology (meta-methodology), the FRIA-F is placed at the "Risk Assessment" stage.

Recruiting and Candidate Selection				
Trigger	FR affected	Clarification	AI Act Connection	Elimination/mitigation measure
The system does not provide for (the possibility of) human oversight, or interpretation, understanding and explanation	Human dignity	The system could lead to de-humanization of the recruiting process both for the recruiter and the applicant	Art. 14	Use an explainable AI model

of the system's functioning or output by the user				
The AI system's training data volume is excessive and goes beyond what is necessary for the intended purpose, containing redundant and unnecessary information that exceeds the data minimization principle	Human dignity	Excessive data collection increases the risk of capturing and amplifying historical biases and discriminatory patterns in hiring practices, while also unnecessarily processing sensitive personal information about candidates' cultural, religious, and demographic characteristics. The overcollection of data creates a more complex and opaque system where discriminatory effects become harder to detect and correct, particularly affecting vulnerable groups like elderly workers, persons with disabilities, and cultural minorities.	Art 10.2 & 10.3	Apply the data minimalisation principle and reduce the volume of data to what is strictly necessary for the systems intended purpose. (Refer to Annex IV for further instruction)
The AI systems' training data include information that is arbitrary, or otherwise irrelevant for the position, such as by using irrelevant or arbitrary features or variables	Human dignity	The arbitrary features in training data may inadvertently correlate with protected characteristics (like race, gender, or age), leading to discriminatory recruitment outcomes while obscuring the source of bias. This threatens human dignity by potentially perpetuating systemic inequalities. Take for example a recruitment system that considers whether an applicant went to a 'prestigious' university. Including features like	Art. 10.3 (relevance)	Critically assess each separate data feature in the training data for relevance; examine what assumptions are embedded within each feature and whether these assumptions are fair (Refer to Annex IV for relevant tools to do so)

		prestigious university attendance or specific educational backgrounds, risks discriminating against qualified candidates from less privileged backgrounds who may not have had access to these opportunities		
The AI systems' training data includes features that are protected against discrimination: sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation	Non-discrimination (art. 22) Cultural, religious and linguistic diversity (art. 22) Equality between women and men (art. 23) Rights of the elderly (art. 25) Integration of persons with disabilities (art. 26)	In a recruitment algorithm, using protected characteristics (like gender, race, age, disability) in training data would perpetuate discriminatory hiring practices and unfairly limit job opportunities for certain groups, violating fundamental rights to non-discrimination in employment. The AI could learn from historical biases in hiring data, creating automated barriers that especially disadvantage women, minorities, older workers, and persons with disabilities, all while making discrimination harder to detect since it's embedded in complex algorithmic decisions.	Art 10.5	Anonymise protected features, only include them in training data if bias detection and correction cannot be fulfilled by processing other data, delete protected data once bias has been corrected
The AI system's training data does not fully reflect the population (geographical, demographical, social) the system's is intended to be used for	Non-discrimination (art. 21) Cultural, religious and linguistic diversity (art. 22) Equality between women and men (art. 23)	Insufficient data volume and variety across different groups creates systematic bias in hiring where underrepresented candidates face unfair treatment due to the recruitment AI's limited understanding of their qualifications and potential. Take for example, female applicants applying for jobs in male dominated fields[10] . When training	Art 10.2, 10.3 & 10.4	Ensure that sampling is consistent across various demographics, remove any variable and proxy-variables that could lead to discrimination (e.g. gender, race, etc) (Refer to Annex IV for further instructions)

	Rights of the elderly (art. 25) Integration of persons with disabilities (art. 26)	data predominantly contains male candidates' profiles and success patterns, the AI system may unfairly disadvantage qualified women applicants by treating typical male career paths and qualifications as the default standard. This creates a self-perpetuating cycle where women continue to be underrepresented in these fields because the AI recruitment system, trained on historically skewed data, fails to recognize and fairly evaluate their potential.		
The AI system uses secondary data sources that were originally collected for different purposes and contexts	Non-discrimination (art. 21) Cultural, religious and linguistic diversity (art. 22) Equality between women and men (art. 23) Rights of the elderly (art. 25) Integration of persons with disabilities (art. 26)	Using recruitment data from different contexts or purposes (like transferring hiring practices from one country to another) fails to account for important cultural, social, and regulatory differences in employment practices and qualifications. Thus, when applied in a different context, the system might produce biased outputs	Art 10.2 Art 10.4 Annex IV (a)	Retrain models with data from the intended context rather than directly transferring; create culturally-adaptive parameters that can adjust to different contexts (See Annexes IV and V for further mitigation tools and strategies)
The AI system's data processing methods (cleaning, normalization, standardization,	Non-discrimination (art. 21) Cultural, religious and	Standard data processing methods can inadvertently discriminate by treating valid but uncommon characteristics of minority candidates as outliers to	Art 10.2	Regular bias audits at each processing stage (cleaning, normalization, standardization);

labelling, etc) introduce or amplify biases by failing to account for legitimate variations across different groups.	linguistic diversity (art. 22) Equality between women and men (art. 23) Rights of the elderly (art. 25) Integration of persons with disabilities (art. 26)	be removed or normalized away, particularly affecting cultural expressions, non-traditional career paths, or gaps in employment history. Automated cleaning and standardization processes might disproportionately impact candidates from underrepresented groups by failing to recognize legitimate variations in how different groups express qualifications, experience, or skills.		Test processed data for representation levels of different groups; Validate that legitimate variations are preserved after processing (See Annex IV for more mitigation tools and strategies)
The system uses a prescriptive model that provides action recommendations without integrating human-oversight measures in the decision making process		A prescriptive AI system might automatically reject candidates based on predetermined criteria without considering unique circumstances or allowing human recruiters to exercise professional judgment, leading to qualified candidates being overlooked due to rigid algorithmic rules.	Art. 14 Annex IV (2e)	Design the system to present recommendations as discussion starting points rather than final decisions, implement mandatory human review processes, and maintain clear documentation of the reasoning behind each recommendation.
The system uses a predictive model and the training data, or one or more of its features are poor predictors of the system's intended outcome.	Human dignity Non-discrimination (art. 21)		Art.15	Do not use a predictive model Critically assess whether the features in the training data (both individually as well as in combination) are good or poor predictors of the intended outcome
The system uses Natural Language Processing	Non-discrimination (art. 21)	A resume screening system fails to properly interpret and evaluate qualifications written in	Art. 15	Implement multilingual model training and cultural

techniques that fail to account for linguistic variations and cultural expressions, leading to systematic misinterpretation of user inputs.	Cultural, religious and linguistic diversity (art. 22)	non-standard English or containing culturally specific expressions, systematically disadvantaging candidates from diverse backgrounds.		adaptation layers in the NLP pipeline.
The system employs Computer Vision algorithms trained predominantly on limited demographic datasets, resulting in reduced accuracy for underrepresented groups.	Non-discrimination (art. 21) Equality between women and men (art. 23) Rights of the elderly (art. 25) Integration of persons with disabilities (art. 26)	A video interview analysis system performs poorly when analysing candidates with different skin tones, facial features, or cultural dress choices, leading to biased assessments.	Art.15	Ensure training data includes diverse representation and implement regular bias testing across different demographic groups. Consider whether a computer vision algorithm is a necessary part of the recruitment process (See Annex V)
The system's continuous learning capabilities create feedback loops that amplify biases over time without detection or correction mechanisms	Human dignity Non-discrimination (art. 21)	An AI system that continuously learns from hiring decisions starts to amplify subtle biases, creating a self-reinforcing cycle where initial small biases grow stronger over time	Art. 14 Art. 15	Implement bias detection mechanisms, regular model evaluation, and controls on feedback loops to prevent bias amplification. (See Annex V)
The system produces outcomes that systematically differ across demographic groups, showing statistically significant disparities	Non-discrimination (art. 21) Equality between women and men (art. 23)	An AI recruitment tool selects male candidates at a rate of 60% while selecting female candidates at only 30% for technical positions, demonstrating statistical disparity that exceeds legal thresholds for disparate impact.	Art.15 (accuracy) Annex IV (2b)	Implement statistical parity constraints during model training and regularly calculate disparate impact ratios (See Annex V)

	Rights of the elderly (art. 25) Integration of persons with disabilities (art. 26)			
The system suffers from concept drift where model performance degrades over time as societal patterns change	Non-discrimination (art. 21) Equality between women and men (art. 23) Rights of the elderly (art. 25) Integration of persons with disabilities (art. 26)	An AI recruitment system trained on historical data fails to adapt to changing job requirements and skill sets, disadvantaging candidates with newer qualifications or non-traditional career paths.	Art.15 (accuracy) Annex IV (2f), Annex IV (3) & Annex IV (6)	Implement continuous model monitoring and regular retraining protocols to adapt to changing patterns.
The system is deployed in a context different from its originally intended use case	Human dignity Non-discrimination (art. 21) Equality between women and men (art. 23) Rights of the elderly (art. 25) Integration of persons with disabilities (art. 26)	An AI system designed to screen entry-level technical positions is inappropriately used for senior management roles, where the evaluation criteria and success factors are substantially different, leading to incorrect assessments.	Annex IV (2f), Annex IV (3) & Annex IV (6)	Implement strict context validation protocols and require new validation studies before deploying systems in different contexts.
The system consists of multiple AI components	Human dignity	A complex AI recruitment system uses multiple models (resume screening, video interview	Art. 14 Annex IV (2c) & Annex IV (6)	Create detailed system architecture maps showing

feeding data into each other sequentially but lacks documentation about how errors or biases in early processing stages might compound through the pipeline.	Non-discrimination (art. 21) Equality between women and men (art. 23) Rights of the elderly (art. 25) Integration of persons with disabilities (art. 26)	analysis, and skills assessment) that feed into each other sequentially, but there's no documentation about how errors or biases in early stages might compound through the pipeline, potentially amplifying discrimination against certain candidates.[11]		data flows and model interactions, document how outputs from each component influence subsequent stages, and implement bias testing at each transition point (See Annexes IV and V)
The system is used at a critical decision point, despite being designed for low-stakes preliminary screening	Human dignity	AI system designed for initial resume sorting is inappropriately used to make final hiring decisions, bypassing important human evaluation stages and denying candidates fair consideration	Art.14 Annex IV (2e)	Map decision points and their impact levels, restrict system usage based on impact level, and require additional safeguards for high-stakes decisions (See Annex V)
The system makes recommendations based on demographic patterns from one geographical or cultural context in a different context	Non-discrimination (art. 21) Cultural, religious and linguistic diversity (art. 22)	An AI recruitment system trained on data from one country is used in another where educational systems, job titles, and career progression patterns differ significantly, leading to systematic disadvantages for local candidates	Annex IV (2f), Annex IV (3) & Annex IV (6)	Require localization studies and cultural adaptation before deploying systems in new geographical or cultural contexts
The system's output cannot be interpreted properly by the user	Non-discrimination (art. 21) Human dignity	Potential biased or unfair outcomes remain invisible to the user	Art. 14	Do not use opaque systems/black box Ensure interpretability
The system's output, functionality or UI might trigger	Non-discrimination (art. 21)			

automation bias with the user				
The system fails to provide mechanisms for users to challenge or question its outputs	Human dignity Non-discrimination (art. 21)	When an AI system flags a candidate as "unsuitable" based on pattern matching, there is no clear process for recruiters or candidates to understand and potentially contest this assessment.	Art. 14	Establish clear procedures for questioning and appealing system outputs, with documented review processes.

Conclusion

The FRIA-F methodology presented here offers a comprehensive framework for assessing and addressing fundamental rights impacts in AI recruitment systems through a reverse engineering approach. Rather than using traditional checklist-based methods, this approach starts by identifying potential risk triggers in the AI system's design, development, and deployment, and the organisational structures around it, then mapping these to specific fundamental rights impacts and corresponding mitigation measures.

Several key principles emerge from this methodology:

1. The importance of continuous assessment throughout the AI lifecycle, rather than treating FRIA as a one-time compliance exercise
2. The recognition that no AI system is entirely risk-free, and that even impacts affecting a single individual warrant serious consideration and mitigation
3. The value of examining systems holistically, considering how different components interact and how risks may compound throughout the pipeline
4. The need for context-specific evaluation, recognizing that AI systems may perform differently across different cultural, geographical, and organizational settings
5. The critical role of human oversight and interpretability in ensuring fair outcomes and maintaining accountability

For effective implementation, organizations should integrate this FRIA-F methodology into their broader risk management frameworks while remaining flexible enough to adapt to evolving societal needs and regulatory requirements. Success requires ongoing collaboration between technical teams, legal experts, and domain specialists to identify, assess, and mitigate fundamental rights impacts throughout the AI system's life cycle.

The methodology also emphasizes that fairness in AI systems cannot be achieved through technical solutions alone - it requires careful consideration of social, cultural, and ethical dimensions, along with robust governance structures and clear procedures for addressing concerns when they arise.

Moving forward, this framework should be viewed as a living document that can evolve based on practical implementation experiences and emerging challenges in AI fairness. Organizations should use it not just as a compliance tool, but as a foundation for building more equitable and inclusive AI systems that respect and protect fundamental rights.