# Universiteit Leiden

## ICT in Business and the Public Sector

A Framework for Responsible AI Use in HR

Name: Charlotte Eijkelkamp
Student-no: s3844080

Date: 20/09/2024

1st supervisor: Drs. N. van Weeren
2nd supervisor: Prof.dr.ir. J.M.W. Visser

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

**Abstract**

### Background

The use of AI is on the rise, with increasing adoption in the HR industry. As a result, its impact on society is also expanding. New regulations, such as the European AI Act, provide guidance for this, but there is ongoing debate about whether these measures are sufficient to protect fundamental human rights.

### Aim

The goal of this thesis is to develop guidelines for transparency and human oversight in AI systems used for HR processes and to identify effective ways to convey these guidelines. Ultimately, the aim is to encourage organizations to adopt responsible AI practices that go beyond mere compliance.

### Method

A responsible AI framework was developed using the Design Science research methodology. The design was based on principles focusing on a minimal interface and a focus on content delivery. To evaluate the framework participants were asked to interact with the framework and provide feedback in real time, followed by a questionnaire to measure usability and quality and to identify areas for improvement.

### Results

The created framework conveys guidelines for transparency and human oversight, derived from related literature and group interviews with candidates. The framework includes both generic guidelines and application-specific guidelines tailored to identified HR processes. The framework that we designed was demonstrated and evaluated with seven potential end-users. The evaluation resulted in an above-average usability score for the framework (76,4, $\sigma = 7.9$). Additionally, participants were generally positive about the guideline content and found it readable and implementable. Practical improvements for the framework were also identified during this process, such as ways to embed the framework in an organization.

### Conclusion

The design and evaluation of our framework have demonstrated that adopting responsible AI practices for the themes of Transparency and Human Oversight in HR can be encouraged and made more accessible, without solely emphasizing compliance.

## Acknowledgements

I would like to express my deepest gratitude to my first supervisor, Niels van Weeren, for being a patient and supportive listener and advisor when I encountered challenges. Your guidance has undoubtedly helped me grow throughout this period, and you consistently challenged me to think in terms of opportunities and possibilities rather than limitations.

I would also like to thank my second supervisor, Joost Visser. Your supervision allowed me to bring together all the loose ends in my thoughts in a comprehensive manner. I believe that your guidance has truly enabled me to elevate my thesis to a higher level.

A special thanks to Randstad for hosting my thesis internship. I am especially grateful to my supervisors there, Anisha Nadkarni and Martin Woodward. Thank you for your time, support, and for sharing your expertise with me.

And last but not least my friends, family, and boyfriend. I could not have done it without your support.

# Contents

# 1 Introduction

The widespread use of AI is undeniable. AI is increasingly being used in situations with greater and greater impact. However, this comes with significant risks. To mitigate these risks there has been an ongoing discussion about best practices, such as those outlined in the Ethics Guidelines for Trustworthy AI written by the European High-Level Expert Group on AI [1]. Now, this discussion is becoming more formalized with various proposals for laws and regulations.

## 1.1 Background

One of the main pieces of legislation on AI is the European Union AI Act [2]. The AI Act aims to protect the safety, health, and fundamental rights of EU citizens and has been in force since August 1, 2024. The requirements will begin to apply gradually over time. Two key features of the AI Act highlighted here are the risk-based approach and the CE certifications.

#### Risk-based approach

For the risk-based approach, the AI Act differentiates multiple levels of risk. These are, from highest to lowest risk level: unacceptable risk, high-risk, limited risk, and minimal risk. General purpose AI is noted as a separate risk level and falls right between limited and high risk. In systems identified as posing unacceptable risks, usage is prohibited, while systems with minimal risk require only the signing of a code of conduct. It is the systems falling within the middle two risk levels that entail numerous obligations. Furthermore, the AI Act distinguishes between the provider and the deployer of an AI system, where both are subject to different requirements. For example, with a chatbot, a limited-risk AI system, the provider has an obligation of transparency. Whereas in another limited risk system, such as a system that generates deepfake or synthetic content, this obligation falls on the deployer. So it depends on the level of risk and the role of the organization (provider/deployer), which obligations the system must meet.

For high-risk systems, a list of essential requirements has been established. These requirements include transparency, human oversight, data (governance), and accuracy, among others. An AI system falls into the high-risk category in two situations. First, when it is (a component of) a product already governed by either the New Legislative Framework (NLF), roughly all products that currently have a CE mark, or by other EU regulations. Second, when it is used in one of the eight high-risk areas outlined in the AI Act [2, 3]. A List of these areas can be found in table 1. Where the classification of risk-level is more straightforward, the requirements do leave some room for interpretation, complicating the process of becoming AI Act compliant. However, it is expected that some of this ambiguity will diminish once technical standards for implementing the AI Act are established. These technical standards are developed by private European Standardization Organizations (ESO's) such as CEN or CENELEC [4].

#### CE certification and Technical Standards

A product, in this case, an AI system, can obtain a CE certification by conducting a self-assessment based on these technical standards. This CE certification gives a presumed label of conformance with the EU legislation but is not a proof of conformance nor a quality indication

| | |
|---|---|
| 1. | Biometrics, insofar as their use is permitted under relevant Union or national law |
| 2. | Critical infrastructure |
| 3. | Education and vocational training |
| 4. | Employment, workers management and access to self-employment |
| 5. | Access to and enjoyment of essential private services and essential public services and benefits |
| 6. | Law enforcement, insofar as their use is permitted under relevant Union or national law |
| 7. | Migration, asylum and border control management, insofar as their use is permitted under relevant Union or national law |
| 8. | Administration of justice and democratic processes |

Table 1: List of high risk areas [2]

[5]. The concept behind CE certification is that a product can move freely throughout the European market, and ensure consumers benefit from the same level of protection [4]. Although the technical standards provide convenience in implementing the AI Act, their utilization has faced significant criticism. They centralize a lot of policy-making power within the ESOs [6]. Additionally, the CE markings, originally intended for ensuring safety and health, are now being used to safeguard fundamental rights in the AI Act [4].

These technical standards primarily seem to provide a technical baseline; compliance with a standard does not necessarily mean that a system is truly in line with the European values underlying the AI Act. Moreover, the essential requirements for high-risk systems are so ambiguous that their implementation will involve numerous considerations and trade-offs.

## 1.2 Problem Statement

The problem lies in the difficulty of capturing the responsible use of artificial intelligence in laws or technical standards, particularly concerning the human aspects of its utilization. While the legislation sets a baseline for responsible AI use, it often reduces the approach to mere compliance. This can result in organizations viewing responsible AI use as a means to reach compliance, rather than a goal in itself. A more self-driven, intrinsically motivated, and value-oriented approach may be better achieved through voluntary guidelines that promote awareness and empowerment. This thesis explores how such an approach can be applied in the context of HR.

## 1.3 Scope

The scope of this research entails the themes of Transparency and Human Oversight, as seen in Chapter III Section 2: Requirements for High-Risk AI Systems of the AI Act. The decision to focus on these two requirements is made because of the importance of the human factors. Additionally, to contextualize and define the types of systems, AI systems in employment, worker management, and access to self-employment will be considered, because of their impacts on fundamental rights. This corresponds to area (4) described as high-risk in Annex III in the AI Act [2]. This entails AI systems, used for processes such as hiring or promoting employees. For readability and convenience, this is described as AI systems used in HR processes. A further

motivation for this scope is read in the following two paragraphs.

The choice to focus this research on the requirements of Transparency and Human Oversight is based on the fact that the human aspects play a big role here. It is likely these areas, in which technical standards may fall short in achieving meaningful implementation. For example, Article 14(4)(a) specifies that the user assigned with human oversight should be enabled by the system to recognize system dysfunctions [2]. A proficient user likely has different needs in this regard compared to a non-proficient user. Ideally, the aim is to accommodate everyone in a way that does not unnecessarily drain resources.

The specific systems under examination are AI systems used in HR processes. These are processes that many will encounter in their professional lives. Moreover, in this context, fundamental rights are often at stake, such as the right to fair treatment in a job application, rather than primarily concerns about health and safety. Fundamental rights are precisely the aspect that is less effectively captured in technical standards [4].

## 1.4 Research Questions and Framework Objectives

This thesis will aim to answer the following research question:

*In what way, can organizations be guided on ethics and legislation in applying AI systems in HR processes, to responsibly embrace the opportunities of AI?*

To answer this question, three subquestions are defined:

- **RQ1:** What guidelines will help in reaching meaningful Transparency, in AI systems used in HR processes?
- **RQ2:** What guidelines will help in reaching meaningful Human Oversight, in AI systems used in HR processes?
- **RQ3:** What is the best-suited method for conveying the defined guidelines?

The goal of the research is to create a framework for responsible AI use, in which relevant legislation, as well as ethics, are taken into account. In reaching this goal, the design science research model (DSRM) is used [7]. This model is chosen because of the guidance it provides in applying Design Science research for information systems. The DSRM is made up of six steps: problem identification and motivation, defining objectives of a solution, design & development, demonstration, evaluation, and lastly communication. To try to solve the defined problem, the framework should adhere to the following objectives:

- **Facilitate progress towards responsible AI:** The framework should provide clear guidance for organizations to advance toward responsible AI usage. This includes offering practical tools and methodologies that can be implemented to ensure ethical considerations are integrated throughout the AI development lifecycle.
- **Establish a common starting point for organizations:** The framework should serve as a reference for organizations, ensuring that all stakeholders share a common understanding of what responsible AI entails and how each stakeholder contributes to it. Responsible AI is a multi-stakeholder endeavor, and having a common starting point enables stakeholders to hold each other accountable.

- **Inspire and activate responsible AI practices:** The framework should offer a balance between concrete, actionable steps, and higher-level, abstract guidance. While the framework should offer clear, practical advice that organizations can immediately implement, it should also include more inspirational guidance that encourages a broader understanding of responsible AI. These higher-level guidelines may not be directly actionable, but they are essential for conveying ethical AI principles and practices. Limiting the framework to only concrete, actionable steps risks reducing responsible AI to a mere checkbox exercise. Therefore it is an objective to strike a balance between these two.
- **Guidelines in line with AI Act:** While the goal is to go beyond compliance, one should first reach a point before one can go beyond this. Since the AI Act is new legislation, many organizations might also benefit from guidance in becoming AI Act compliant. However, especially since the scope only entails transparency and human oversight, the finished framework can not provide enough guidance in becoming AI Act compliant.

## 1.5 Thesis Outline

In this section, first, the outline of the thesis is described. This is followed by an indication of where the activities from the design science research model are applied [7].

In section 2 the related work about both transparency and human oversight is assessed. This is followed by the methods in section 3, where the creation of the framework and guidelines are described. The actual guidelines and framework, including a demonstration, are found in section 4. This is followed by an evaluation in section 5, the discussion in section 6, and finally a conclusion in section 7.

The way the design science research activities are applied is the following. The first activity of identifying a problem is described in section 1.2, the objectives of the solution are defined in section 1.4. The design and development are split up, in first the development in section 3 and then the design in section 4. The demonstration is found in section 4.3. Finally, the evaluation is found in section 5.

# 2 Related Work

In this section related research regarding the concepts of transparency and human oversight are examined. These concepts relate to two of the requirements for high-risk AI Systems, described in Chapter III, Section 2 of the AI Act [2]. The specific requirements for Transparency and Human Oversight are mentioned in Articles 13 and 14 respectively. Additionally, a brief overview of the concept of explainability is provided. Although not explicitly designated as a requirement in the AI Act, explainability is closely associated with both transparency and human oversight and is therefore included in the related work.

## 2.1 Transparency

Transparency and responsible AI usage appear to be nearly synonymous in contemporary discourse. In 2019, [8] found that transparency is the most prevalent concept in guidelines or principles for ethical AI. Similarly, at the European level, transparency is highly valued. The EU Commission's High-Level Expert Group on AI (AI HLEG) identified transparency as one of the seven key requirements for achieving trustworthy AI [1]. It is therefore not surprising that transparency has been incorporated into the AI Act. For instance, it is one of the requirements for high-risk AI systems (Article 13) and an obligation for providers and users of certain AI systems and GPAI models (Article 50) [2].

While very prevalent in literature, transparency is not something for which there exists a singular definition. For instance, [8] identified terms such as explainability, disclosure, or communication as aspects of transparency. Hence, something does not necessarily need to explicitly bear the name 'transparency' to contribute to transparent AI usage. Therefore, this section examines transparency as an overarching theme, including how other parts of the AI Act might contribute to transparency. This is achieved by first exploring various aspects and perspectives on transparency, followed by an examination of both the positive and negative aspects, and finally an analysis of how transparency is addressed in the AI Act.

### 2.1.1 Definitions and aspects of Transparency

Transparency emerges across various scientific disciplines, each with its own interpretation, making it challenging to clearly define this concept in the context of AI [9]. [10] defines transparency as a process in which actor A informs actor C about a decision, incorporating the process and a justification. From this definition, several aspects emerge.

Firstly, there is an informational aspect, wherein transparency serves as a means to address information asymmetry [11,12]. Secondly, it involves a relational aspect, requiring the presence of two or more parties [13,14]. Thus, a party must consider the needs and characteristics of the receiving party. Therefore, a single party cannot be transparent, as transparency is always in relation to another party. A third aspect, not explicitly addressed in the definition by [10] but described by [13], is the systemic aspect. This encompasses the institutional context and the manner in which transparency is embedded, including any legal and organizational measures influencing its implementation.

These three mentioned aspects, informational, relational, and systemic, can also be classified in another manner. [15] distinguishes between two types of approaches to transparency: verifiability and performability. Verifiability largely aligns with the informational aspect and focuses

Table 2: Different aspects of transparency compared to eachother

| Aspects by [10, 13] | Aspects by [15] |
|---|---|
| Informational | Verifiability |
| Relational | Performability |
| Systemic | |

solely on providing information. Performability encompasses more of the other two aspects, relational and systemic, and considers transparency more holistically. It is rather about the way transparency is 'performed' [13]. The way these aspects relate to each other can be seen in table 2.

### 2.1.2 Multiple perspectives on Transparency

The concept of transparency is encountered across multiple disciplines, each giving it a distinct meaning. Within the context of AI regulations, two key perspectives emerge: A legal perspective and a computer science perspective. A description of transparency from a legal perspective is "a quality of complex socio-technical interactions between the AI and its users, developers, owners, and wider society," [16][p. 1] while the computer science perspective is defined as "an algorithmic property that offers practical solutions but through a limited, technology-focused scope." [16][p. 1]. The disparity between these perspectives, with the legal viewpoint being more process-oriented and the computer science viewpoint being more technical, is termed the "transparency gap" by [16].

A similar distinction is described by [17], who notes that the AI Act (representing a legal perspective) concerns itself with transparency in an AI system, whereas the (computer science) literature describes transparency in an AI model. The difference between an AI model and an AI system lies in the fact that an AI model is a component of an AI system. Thus, an AI system encompasses an AI model and other (non-AI-based) software components [17], that would entail for instance an AI model and a UI. An example of this distinction is the AI system ChatGPT, which is based on the AI model GPT.

#### Transparency from a legal perspective

The legal perspective aims at achieving AI system transparency [17] and views transparency as a means rather than an end [16]. It is about having clear and accessible information about the underlying processes within systems [14]. The aim is therefore to increase trust, legitimacy [10] or accountability [13]. Another aim could be to increase system usability, in a type of transparency called 'Transparency in Use' by [14]. This entails information about how to use a system, or how to interpret the system outcomes.

#### Transparency from a Computer Science perspective

Transparency from a computer science perspective, or an Explainable AI (XAI) perspective, is more focused on the inherent opacity of certain AI systems, also known as black-box systems. These are systems that often exhibit high accuracy but low interpretability [18]. From this perspective, transparency is more an end goal, rather than a means to reach different goals [16]. More about explainability and XAI can be read in section 2.3.

### 2.1.3 Positive outcomes and aspects of transparency

There are several positive outcomes of transparency, they include:

- **Trust:** Transparency is expected to positively affect trust among both affected persons and users, yet the correlation between transparency and trust is under debate [12, 13]
- **Increased autonomy:** Having a transparent system can support autonomy and control for the user of the system [13].
- **Inspectability (or verifiability/traceability):** Transparency allows a third party to examine a system and ensure it meets defined standards [13].
- **Accountability:** Gaining insights into the workings of an AI system helps in understanding how a decision came about, facilitating the identification of an accountable party [13, 16].

Additionally, [11] highlights the importance of transparency in AI because it involves processes that are not generally understood by the general public, yet have significant impacts on many people.

### 2.1.4 Downsides and Challenges of Transparency

There are several downsides to transparency as well as challenges in reaching transparency. These are:

- **Information overload:** Excessive transparency can lead to occluding effects, such as information overload [10, 12]. For instance, having access to complicated source code may result in negative effects, such as a reduced sense of empowerment [10]. Another negative effect of information overload is user-responsibilization. Here transparency can be used to shift responsibility onto ill-equipped users. For example in informed consent, where most of the information presented ends up getting ignored by the reader as it is not meaningful information [13].
- **Complexity of stakeholders:** The complexity of stakeholders and their diverse expectations must be considered. It is essential to address the demand side of transparency; if the audience cannot leverage certain information due to its complexity or accessibility issues, there is no increase in autonomy or control [12, 13].
- **Ambiguity in accountability:** The relationship between transparency and accountability is not always clear. A fully transparent process does not necessarily mean that an accountable agent can be identified [13].
- **Privacy concerns:** Too much transparency can inhibit privacy. For instance, if the underlying training data contains personal data, full transparency (sharing the data) conflicts with privacy [11, 13].

### 2.1.5 Transparency in the AI Act

In this section, we examine how transparency is explicitly and implicitly addressed in the AI Act. It is a non-exhaustive list, primarily highlighting several examples to illustrate the breadth of this theme.

In the AI Act, two articles explicitly address transparency: Article 13 (Transparency and Provision of Information to Deployers) and Article 50 (Transparency obligations for providers and deployers of certain AI systems) [2]. Article 13 mentions an appropriate type and degree of transparency for deployers to "interpret the system's output and use it appropriately", as well as instructions for use including "concise, complete, correct and clear information that is relevant, accessible and comprehensible to users" [2]. In Article 50, transparency takes shape through disclosure aimed at informing natural persons that they are interacting with an AI system or viewing artificially generated or manipulated content (e.g., deepfakes).

In these explicit mentions of transparency, it seems to be the end-users that are mostly 'protected'. By for instance allowing them to properly interpret the output of the system as well as making them aware of their interactions with AI. However, other stakeholders may benefit from transparency. Two of these stakeholders are affected persons and external auditors. They are accounted for in the AI Act in the following ways.

Affected persons are addressed, for instance, in Article 86 (Right to Explanation of Individual Decision-Making). A person who has been subject to automated decision-making can subsequently inquire about how this decision was made. Another article in which affected persons are addressed is 26(11), which notes that natural persons who are subject to the use of high-risk AI should be notified about that. Article 26(7) highlights a similar approach, specifically for deployers who are employers. They should inform their employees before putting the AI system into service. Examples of where external auditors are addressed are Articles 11 (Technical Documentation) and 12 (Record-keeping). Here, transparency mainly takes the form of traceability, and the goal seems primarily compliance-oriented.

## 2.2 Human Oversight

Human oversight refers to the practice of having a natural person oversee the functioning of an AI system. It is a requirement for high-risk AI systems in the AI Act (Article 14) [2], where oversight primarily involves monitoring the operation or recognizing malfunctions or anomalies in an AI system. The key aim is to prevent or minimize risks to health, safety, and human rights.

This section delves into the literature on human oversight, first examining the various reasons for implementing it, as well as its negative aspects and flaws. Additionally, different approaches to human oversight are explored. Finally, it is discussed how human oversight is addressed in the AI Act.

### 2.2.1 Positive aspects and outcomes of Human Oversight

AI systems are certainly not perfect; there are already several examples of discriminatory and biased outcomes. Humans can be deployed to monitor such malfunctions or other undesirable outcomes [19]. Moreover, it is mentioned that a human face is important for feeling heard and seen [20]. Furthermore, not every process is suitable for automation; sometimes the human touch is crucial. Ideally, one would have the best of both worlds: the accuracy of an algorithm and the discretion of a human operator [21].

### 2.2.2 Downsides and Challenges of Human Oversight

Human oversight appears to be a good way to maintain a human touch and monitor AI

systems. However, there are several drawbacks to human oversight. The most significant disadvantage is that human oversight policies lack empirical evidence [21]. Thus, human oversight does not necessarily lead to a fairer process or higher accuracy. Additionally, people struggle to interpret and effectively use algorithms [22] and there is limited understanding of how humans and algorithms interact precisely [23]. What is known about the way human and computers (or AI) interact is automation bias, where individuals tend to blindly follow algorithmic recommendations [21], presentation bias, where the way information is presented can affect the way the user behaves [24], or the use of algorithms as moral buffers, where people may feel less responsible for a decision if an automation tool is involved [23].

Next to that, human oversight does not provide a direct solution to the accountability problem. Although the human overseer may be seen as the accountable person in this case, they are not the only individual responsible for the functioning of the algorithm. The frontline workers should not become scapegoats due to a shift in responsibility [21].

### 2.2.3 Approaches to Human Oversight

When thinking about human oversight, one can think about the way human oversight should be facilitated, and about the way human oversight should be performed.

To facilitate human oversight involves actively considering what human oversight should look like and what it represents in a specific situation [23], or simply reflecting on the use of an algorithm [21]. In performing human oversight, one could examine how the algorithm and a person interact. A study by Green and Chen [22] utilized various conditions. Examples include the 'update' condition, where individuals first made a choice without the algorithm's input, then the algorithm's prediction was presented, and they had to make the choice again. Another condition is 'explanation', where the algorithm's prediction is shared along with an explanation to the user before making a decision. There were minor differences between conditions, but notably, in all conditions, participants were unable to evaluate their accuracy, and participants still acted in a biased way. Other ways to think about the way human oversight is performed, could for instance be training of the relevant users [24].

### 2.2.4 Human Oversight in the AI Act

In contrast to transparency, human oversight is a less multifaceted concept. Therefore, human oversight appears less spread throughout the AI Act. When it comes to human oversight, it specifically pertains to Article 14, which outlines the essential requirement of human oversight [2].

What emerges from Article 14 is that the purpose of human oversight is to prevent or minimize risks to health, safety, and fundamental rights. Furthermore, the following points are highlighted:

- Consideration of facilitating human oversight from the design phase onwards.
- Human oversight measures should be proportionate to the risk, level of autonomy, and context of the AI system.
- Measures can be either built-in by the provider, implemented by the user, or both, but should always be identified by the provider.

- The design of the system must include: information about capacities and limitations, mechanisms to prevent automation bias, possible usage of interpretation tools to aid in correctly interpreting output, an option to not use the system, and finally, a way to intervene in the operation of the AI system.

## 2.3 Where Transparency and Human Oversight meet - Explainability

While not explicitly mentioned in the AI Act, explainability is undoubtedly related to both transparency and human oversight. Transparency and explainability are sometimes used interchangeably. Furthermore, explainability could be considered a precondition for human oversight. Moreover, explainability is a recurring theme in discussions about AI usage, particularly in the field of explainable AI (XAI) [18, 25]. In this section, we delve into the difference between interpretability and explainability and briefly discuss the role of explainability in the AI Act.

### 2.3.1 Interpretability vs. Explainability

When dealing with a black-box AI model, there are generally two different approaches, one is making it interpretable, and the other is making it explainable [16–18]. Interpretable AI is understandable through direct examination and lets the user directly see into the model's decision-making process [16–18]. Explainable AI is more about explaining the black box, it comes without constraints at the original model, but rather helps in explaining the outcome [16–18].

### 2.3.2 Explainability in the AI Act

Explainability is not explicitly outlined as a requirement for high-risk AI systems in the AI Act. However, there seems to be an indirect role for explainability. According to [17], the AI Act holistically addresses explainability by incorporating all the different requirements. One of the requirements that play a key role is the risk management system, as it helps in justifying the trade-offs that are for this specific AI system [17]. Another Article that might hint at explainability is Article 86 (Right to Explanation of Individual Decision-Making). Yet this Article does not indicate how extensive this explanation should be, or how explainable the AI system used in this decision should be.

# 3 Methods

In this section, the process leading up to the design is described. First, the creation of the framework is described. After that a description of the creation of the guidelines can be found.

## 3.1 Creation of the framework

This sections describes the steps done in preparation of creating the framework. This starts by identifying the relevant stakeholders, then discusses the way the applications for the application-specific guidelines came about. Finally the creation of the framework application is explained.

### 3.1.1 Stakeholder Identification

Three different groups of stakeholders are identified; the users of the framework, transparency audiences, and human oversight actors. A diagram of all stakeholders and their relation to both the AI System and the Framework is found in fig. 1.

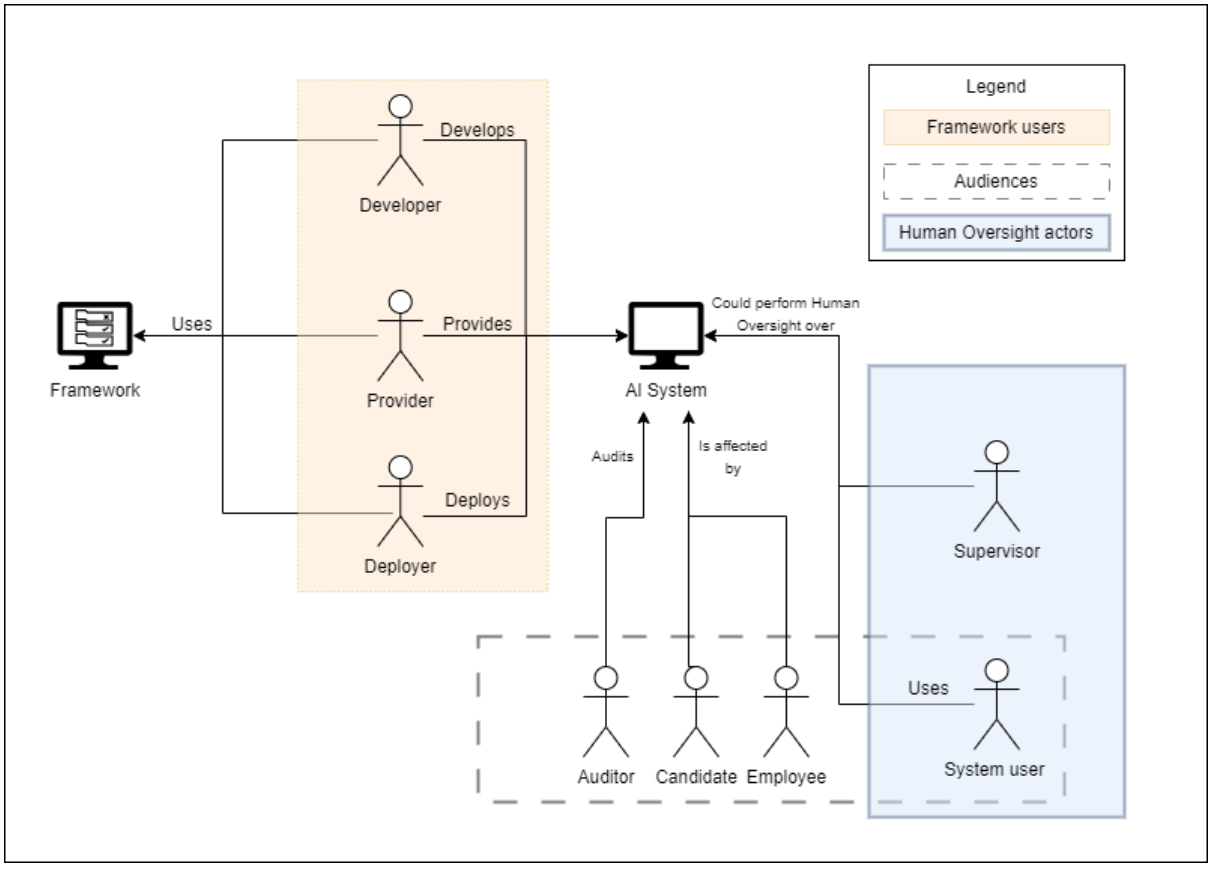



Figure 1: Stakeholder diagram

#### 3.1.1.1 Framework users

The group 'framework users' describes the parties that have accountability over the AI system and therefore should be guided by the framework in handling that accountability responsibly.

Three different framework users are defined: a developer, a provider, and a deployer. These users roughly correspond to the AI Act, which also defines deployers and providers. Additionally, the provider side is split up between a provider and a developer, this split is based on whether someone is involved on a technical level or a business/organizational level. A more in-depth explanation including some examples for every user is provided in the following paragraphs. Note that the system user is not defined as a framework user, even though they might have some accountability over (the performance of) the AI system. This is done intentionally to prevent user responsibilization and focus more on the development and implementation of responsible AI systems.

**Developer**

The framework user 'developer' is defined as everyone involved in the technical side of the development of the AI system. The developer can be seen as a subgroup of the provider. Examples of roles in this group are data scientists, model engineers, system developers, software engineers, or AI designers.

**Provider**

A provider is defined in the AI Act as "a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge" [2]. Examples of roles in this group are product owners, project managers, ethics and compliance officers, or customer success managers.

**Deployer**

A deployer is defined in the AI Act as "any natural or legal person, including a public authority, agency or other body, using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity. Depending on the type of AI system, the use of the system may affect persons other than the deployer." [2]. Examples of roles in this group are product owners on the user side, procurement, or ethics and compliance officers.

#### 3.1.1.2 Audiences

Audiences are the people and parties on the receiving end of transparency. As described in section 2.1, transparency is a relational concept, in which the needs and characteristics of the receiving party should be considered. These parties are (in)directly interacting with the AI system, and therefore are relevant parties to consider in 'being transparent'. In the context of AI systems used for HR processes, the following parties are defined as audiences: a system user, an employee, a candidate, and an auditor. For every audience, first, the role is defined, followed by some of their characteristics that might shape transparency. Lastly, some examples are given of their interactions with AI systems.

**System user**

A system user is a person directly interacting with an AI system, while also having some sort of decision power. Characteristics of a system user are that they likely have a very high domain knowledge, yet a lower level of technical knowledge. Examples of people in these roles would be recruiters or HR personnel. A recruiter might for instance have a lot of experience with interviewing candidates and forming an opinion, but less experience with interpreting an AI-based video-analysis of an interview.

**Employee**

An employee is someone within the deployers organization, that interacts with an AI actively or passively. However, there is no decision power at all. Characteristics of an employee are that they likely have little domain knowledge, nor a lot of technical knowledge. This may result in a purely technical explanation being insufficient, and potentially requiring the inclusion of additional domain-specific information. Mind that domain knowledge refers to the HR domain, as that is the domain in which the AI system is operating. An example of a direct interaction could be an employee using an AI tool that makes personalized career paths. An example of an interaction would be an employee being the subject of an AI-based advancement prediction. In both cases, this affects the employee and their career opportunities.

**Candidate**

A candidate is a job-seeker, that gets involved in the recruiting and hiring portion of HR. Candidates are generally people with no formal relation to the hiring organization (leaving internal candidates aside). Characteristics of candidates are that they have little insight into the use of AI and might have little technical knowledge. Similar to an employee, this might result in purely technical explanations being insufficient. An example of a direct interaction with an AI system a candidate might have is chatting with a qualifications chatbot. An indirect interaction might be a video interview of this candidate being analyzed by an AI system.

**Auditor**

An auditor is the party evaluating the AI system, this might be to assess for conformity with legislation, or for getting certifications for meeting (international) standards. In the context of the AI Act, the party performing the conformity assessments is a national 'notified body' [2]. The main characteristic of an auditor would be a higher level of technical knowledge. An example of an interaction with the AI system would be an audit, albeit not a direct interaction. The information required for an audit is clearly defined and is likely far more extensive than what would be expected from a candidate or system user.

#### 3.1.1.3 Human Oversight Actors

Human oversight actors are the individuals responsible for overseeing the functioning of the AI system. Human oversight is only happening after the AI system is put to market or into service. According to the AI Act, human oversight involves having a natural person supervise the AI system's operations [2]. However, there is no single prescribed method or designated role for carrying out human oversight. For this framework, two distinct types of human oversight actors

are defined, a system user and a supervisor. First, the actor is defined, and then the scale on which they would perform oversight is described.

**System user**

The system user signifies the actor actively using the system, and making decisions based on or with the AI system. The system user is an actor on the deployer side. The scale on which this actor is involved is in making individual decisions, such as hiring a specific candidate. They perform oversight on a more local level.

**Supervisor**

The supervisor on the other hand is the person overseeing the AI system on a broader scale. The supervisor can either be someone on the deployer side or the provider side. The scale on which they operate surpassed individual decisions, as they can see multiple decisions from multiple system users. They perform oversight on a more global level.

### 3.1.2 Creation of the process-, and application list

For the creation of the guidelines, a (non-exhaustive) list of applications within the HR context is created. The applications are grouped on the processes within HR. The list of HR processes is split between hiring & recruitment and current employees. This split relates to annex III, 4(a): Recruitment or selection of natural persons and 4(b): Decisions affecting terms of work-related relationships [2].

In the first half, recruitment and selection, the hiring funnel described by [26] forms the baseline. In addition to this, onboarding is added as a final process. This leads to five processes, ordered from most to least candidates involved. Next to the processes, the applications used in these processes are found in [26–28]. Secondly is the focus on current employees. Based on [27] and [28] five processes, each with applications are found.

All the processes, grouped per category can be found in table 3. All applications grouped per process are found in table 4. The full list, including descriptions and explanations for each process and application, can be found in appendix A. The applications are depicted at a high level rather than mentioning specific applications due to the many different applications/vendors. Therefore, the focus is on the application's purpose, with different implementations or levels of automation possible.

| **Processes in recruitment & selection** | **Processes for current employees** |
|---|---|
| Sourcing | Training and Skills Development |
| Screening | Performance Management |
| Interviewing | Advancement and Career Paths |
| Selection | Retention |
| Onboarding | Salary Evaluation and Employee Benefits |

Table 3: HR processes grouped per category

| Process | Applications |
|---|---|
| Sourcing | Job Description Generator |
| | Job Description Enhancer |
| | Targeted Advertisements |
| | Matching |
| | Headhunting |
| Screening | CV Parsing |
| | Qualifications Chatbot |
| | Pre-employment assessment |
| Interviewing | Video based Interviews |
| Selection | Background checks |
| | Offer generation |
| Onboarding | AI Aided learning |
| Training and Skills Development | AI Aided learning |
| | Skills identification |
| Performance Management | Performance Analysis |
| Advancement/Career Paths | Personalized Career Paths |
| | Advancement prediction |
| Retention | Retention Management |
| Employee Benefits/Salary Evaluation | Identifying important benefits |

Table 4: Processes and apps

### 3.1.3 Creation of the framework application

To convey the guidelines in the best way possible a web-based application is created. Next to properly conveying the guidelines, the goal is a usable application, therefore a user-centric design approach is equipped. A prototype is developed based on a few high-level requirements. Then this prototype undergoes a series of evaluations and improvements to lead up to the final version of the framework. First, the creation of the initial prototype is described, followed by an explanation of how the evaluations influenced the subsequent versions.

For the initial prototype, two key design principles, as outlined by [29], serve as the foundation. Other design principles are addressed less explicitly and will therefore not be discussed here. The first principle emphasizes a minimal interface, prioritizing content over visual elements. The second principle involves interface visualization, where complex data is transformed into easily understandable information. These two design principles are highlighted because of their focus on the content. Additionally, a hierarchical structure has been chosen, aligning with the organization of the guidelines and the underlying data, as illustrated in the simplified diagram shown in fig. 2.

After creating the initial prototype, three rounds of evaluations and improvements were conducted. This iterative process results in the addition of features such as icons, tags for supplementary information, and more intuitive filtering options. Additionally, a Frequently Asked Questions (FAQ) section and a landing page are added to enhance accessibility and better guide users in navigating the framework. Finally, certain information is removed from the framework to maintain clarity and conciseness.

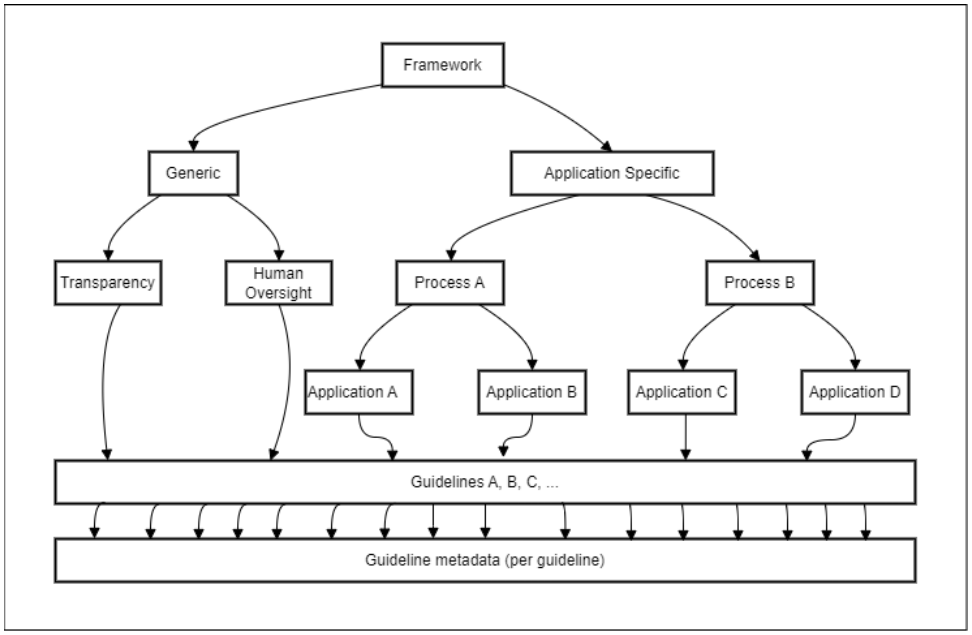



Figure 2: Simplified hierarchy of framework content

## 3.2 Creation of the guidelines

A set of guidelines is created for both transparency and human oversight. Next to that, a split is made between generic guidelines as well as application-specific guidelines. Generic guidelines apply to any AI system, and having application-specific guidelines might help users further concretize responsible AI use. The guidelines are based on both academic and non-academic literature, as well as group interviews with affected parties (candidates/talents). The interview questions are found in section B, and a summary of the interview results is found in section C. The guidelines are created in an iterative fashion, where first drafts of the guidelines are evaluated by end-users and improved accordingly.

### 3.2.1 Transparency

The guidelines for transparency are created based on the AI Act itself [2], the interview results, and the literature. For transparency, the AI Act mainly focuses on the instructions for use. While the contents are explicitly defined in the Act itself, additional best practices and examples can be found in [30]. The transparency-by-design framework offers higher-level insights into maintaining transparency as a fundamental principle [13]. The foundation model transparency index, though strictly aimed at foundational models and therefore not directly applicable to all AI systems, includes many relevant and useful items for achieving transparency. These items tend to be factual, addressing fewer relational aspects [11]. Lastly, the contractual terms for transparency from the city of Amsterdam [31], help in a higher-level grouping of transparency elements. Next to that, these contractual terms have a bigger focus on the relational aspects of transparency in comparison to [11]. So [31] focused more on *how* to share, whereas [11] has a bigger focus on *what* to share.

### 3.2.2 Human Oversight

Similar to transparency, the guidelines for human oversight are created based on the AI Act itself [2], the interview results, and the literature. The AI Act is mostly aimed at how to design and develop the AI system to enable a natural person to effectively oversee the system. For instance, enabling the user to understand the relevant capacities and limitations of the system and allowing them to address anomalies, dysfunctions, and unexpected performance. Next to that, it mentions to remain aware of automation bias [2]. However, as seen in section 2.2 automation bias is not the only flaw in human oversight that the user (among others) needs to be aware of. Therefore the guidelines for human oversight will not only focus on how to facilitate human oversight but also on awareness creation and how to perform human oversight. The baseline for this is the flaws found in section 2.2. Then sources more directed at these specific flaws, either further explaining the concept or flaw, or providing solutions, are used for the guidelines. This is done for presentation bias [24, 32], automation bias [33] and finally human AI interaction [34, 35].

# 4 Design

This section describes the design created for this research. First, the guidelines and their attributes are discussed, this is followed by an explanation of the decisions made in the creation of the framework. Lastly, a demonstration of the framework is given.

## 4.1 Guidelines

For both themes transparency and human oversight a set of generic guidelines and a set of application-specific guidelines is created. This section describes the structure and attributes of individual guidelines. The full set of guidelines can be found in appendix E. A diagram displaying all the different attributes per type of guideline is found in fig. 3

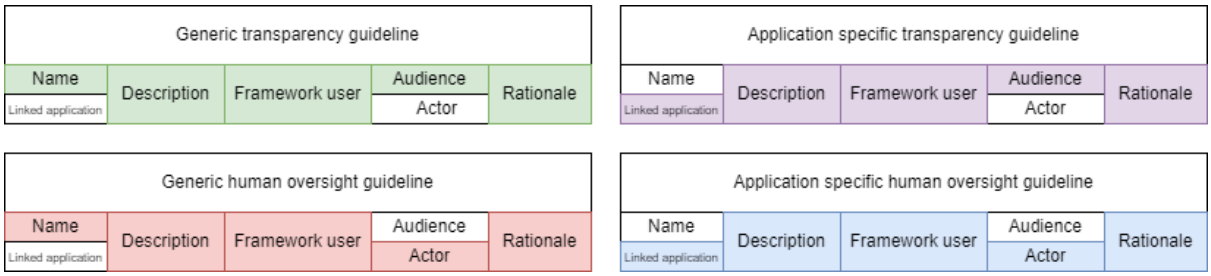



Figure 3: Attributes of all different types of guidelines

#### Name or Linked Application

Each generic guideline comes with a name, whereas application-specific guidelines come with the name of the linked application. The names of the guidelines indicate the theme of the guideline and are descriptive rather than activating. Because the generic guidelines are displayed in large numbers on a page, a name helps with scanning through them. For application-specific guidelines, this necessity is less important or not needed, as they are often not shown all at once.

#### Description

Every guideline has a description, describing what the guideline entails. For human oversight guidelines, the descriptions also indicate what type of guideline it is, the different types are 'Informative' 'To perform oversight' and 'To facilitate oversight'. These types can help the reader indicate how to approach implementing the guideline. For instance, a guideline of type informative is less actionable than the other two types and is mostly in place for creating awareness.

#### Framework user

The attribute 'framework user' identifies the individual or group for whom the guideline is relevant. The different types of framework users are 'Deployer', 'Provider', and 'Developer', which were identified and described in detail in section 3.1.1.1. This user is typically the one most accountable for implementing the guideline. However, the intention is not to place responsibility

solely on specific roles, as accountability should be shared across the board. Therefore, in some cases, multiple framework users are designated for certain guidelines.

#### Audience

All transparency guidelines have an audience. This is the actor to whom you are being transparent, so you have to consider their needs and characteristics in communication. The different types of audiences are 'System users', 'Employees', 'Candidates', and 'Auditors'. These were identified and described in detail in section 3.1.1.2.

#### Actor

For most of the human oversight guidelines, an actor is assigned. The different types of actors are 'System user' and 'Supervisor', which were identified and described in detail in section 3.1.1.3. The actor refers to the person performing human oversight, and they come into play whenever the AI system is in use.

#### Rationale

Lastly, each guideline is accompanied by a rationale. For the generic guidelines, the rationale typically includes the source(s) utilized, along with the relevant information derived from them, or a reference to a specific article of the AI Act. For the application-specific guidelines, the rationale either points to an AI Act article or refers back to a generic guideline. These guidelines are more specific and often serve as an instantiation of the broader generic guidelines.

## 4.2 Framework

This section discusses the decisions made in creating the design. The framework is created with node.js and a MySQL database. The view engine is EJS and the styling is done with CSS. These tools are chosen because of earlier experiences with them. Apart from the technical decisions made, certain styling decisions are made, that will be explained in the following sections.

### 4.2.1 Styling decisions

The most important goal of the framework is conveying the guidelines. Therefore, only the styling decisions for the guideline portion of the framework are motivated here. All other pages (home, FAQ, Documentation, and AI Act articles) are just auxiliary pages, containing information that might help the user browse the framework. First, the generic guideline page is addressed, followed by the application-specific portion.

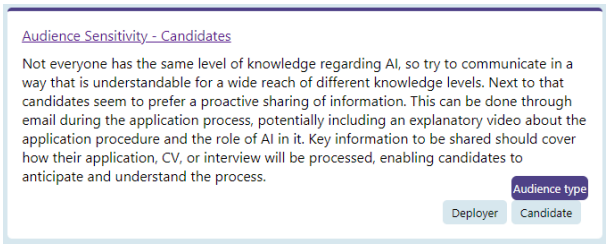



Figure 4: Hover over tags when displaying guideline

#### Generic guidelines

For the generic guidelines, the transparency and human oversight guidelines are displayed side by side. A screenshot of this page is found in fig. 5. Both columns with guidelines offer the option to filter for the relevant user. A guideline by itself is marked with a clickable name, a description, and some tags. When hovering over the tag, the type of tag is displayed. E.g. when hovering over a tag that says 'Candidate', a tooltip is shown that says 'Audience type', see fig. 4. When clicking the name of the guideline, brings you to a more detailed page of the guideline, also showing the rationale.

- **Two columns of guidelines:** Having two columns of guidelines creates a clear distinction between transparency and human oversight guidelines. Additionally, since both types of guidelines are equally important, displaying them in two columns has the advantage that both types of guidelines are immediately visible. This eliminates the need to scroll to view the human oversight guidelines.
- **Filtering for relevant user:** By providing filtering options in this manner, users can easily focus on guidelines relevant to their specific role. However, all guidelines are displayed by default, allowing users to access guidelines for other roles as well. This feature is particularly important because responsibilities are sometimes shared, and having visibility into the expectations of other roles can facilitate accountability and collaboration.
- **Guideline tags and hovering:** Adding tags to the guidelines helps with displaying extra information about the guideline, without using too much text. Additionally, presenting information in this standardized format can help users in mentally grouping or comparing guidelines more effectively.
- **Clicking the guideline for more details:** Clicking on a guideline to access a more detailed overview provides users with additional information that may be of interest, but is not considered the most crucial content.

#### Application-specific guidelines

Browsing for application-specific guidelines starts with finding the relevant application. See fig. 6. For searching, the user can filter on the relevant HR process or the higher-level processes (Recruitment & Current Employees). All the applications are also color-coded based on these two higher-level processes. For every application in the 'Recruitment' process (dark pink), a small blue icon is visible. These icons visualize the volume of candidates associated with these types of applications, and with that follow the flow of the hiring funnel (so more candidates in earlier phases in comparison to the later phases).

After selecting an application, the user is shown a description of the application, as well as all the application-specific guidelines relevant to this application. All the guidelines come with the same tags as the generic guidelines, but now an additional tag is found, indicating whether it is a transparency or a human oversight guideline. This results in a less explicit split between human oversight and transparency, in comparison to the generic guidelines. Furthermore, the guidelines are not clickable for extra information.

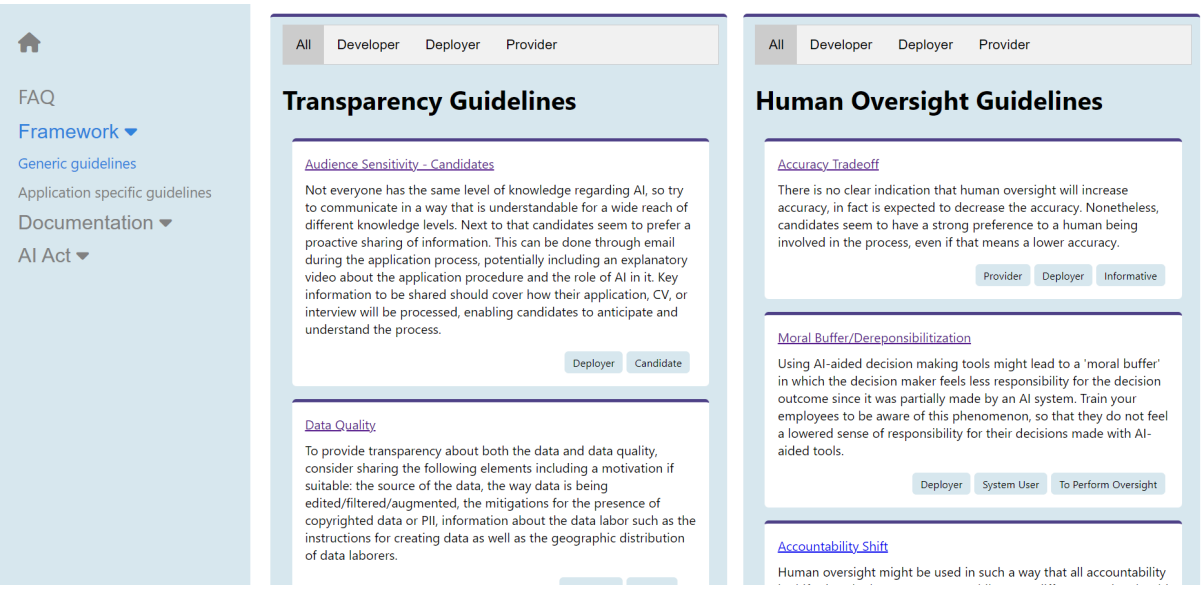



Figure 5: Generic guideline overview

- **Multiple filtering options:** By providing a filtering option, users can more easily find the information that is relevant to them. With two filter options available, users have the flexibility to choose the level of detail they need. In short, they can either divide the data into two categories (higher-level processes) or opt to display just about one-tenth of it (normal processes).
- **Color-coding of applications:** The color coding of the applications gives a further visual distinction between two different types of AI applications used for HR processes. It might help the user in digesting the information, as it allows them to mentally split the list up into smaller chunks.
- **Icons for candidate volumes:** The icons for candidate volumes are there to visualize the flow of the hiring funnel. It can indicate what applications would be used in earlier stages, and what applications might follow after that. It can help the user understand how the applications relate to each other, without needing to read up on the different HR processes.
- **Less explicit transparency/human oversight split:** There are just a few guidelines per application, so in most cases, all of them can be displayed without the need for additional scrolling. Additionally, for most of the application-specific guidelines, the distinction between transparency and human oversight is less pronounced. For example, making the process behind an outcome explainable to a user can both enhance transparency and serve as a necessary step to facilitate human oversight.
- **Non-clickable guidelines:** Whereas the generic guidelines allow you to click on their name to view further information, this is not the case for the application-specific guidelines. This has two reasons. The first one is a design decision, as there are no names for the application-specific guidelines there is less of a dedicated space for such a clickable link.

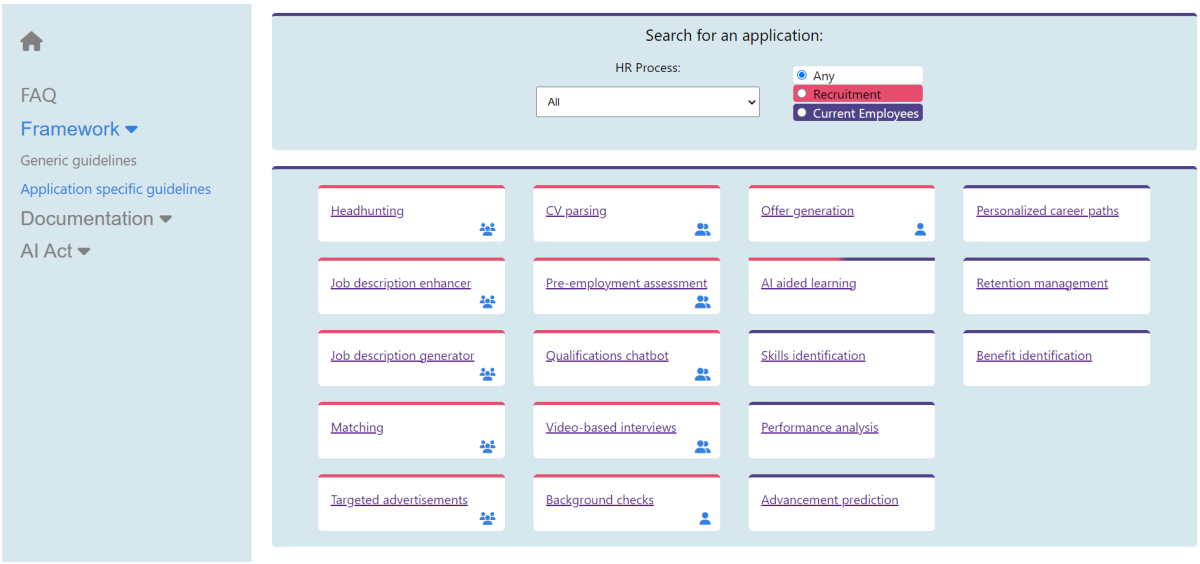



Figure 6: Application selection

The second reason is, that based on the intermediate evaluations the need for an extra layer of information is very little for application-specific guidelines. The user is already in a deeper layer compared to the generic guidelines.

## 4.3 Demonstration

To demonstrate the framework, a specific use case is highlighted: an organization that wants to enhance its hiring process using AI-based assessments. Before partnering with a provider, they first want to explore how to use such an application responsibly.

The user opens the framework and lands on the homepage fig. 8.

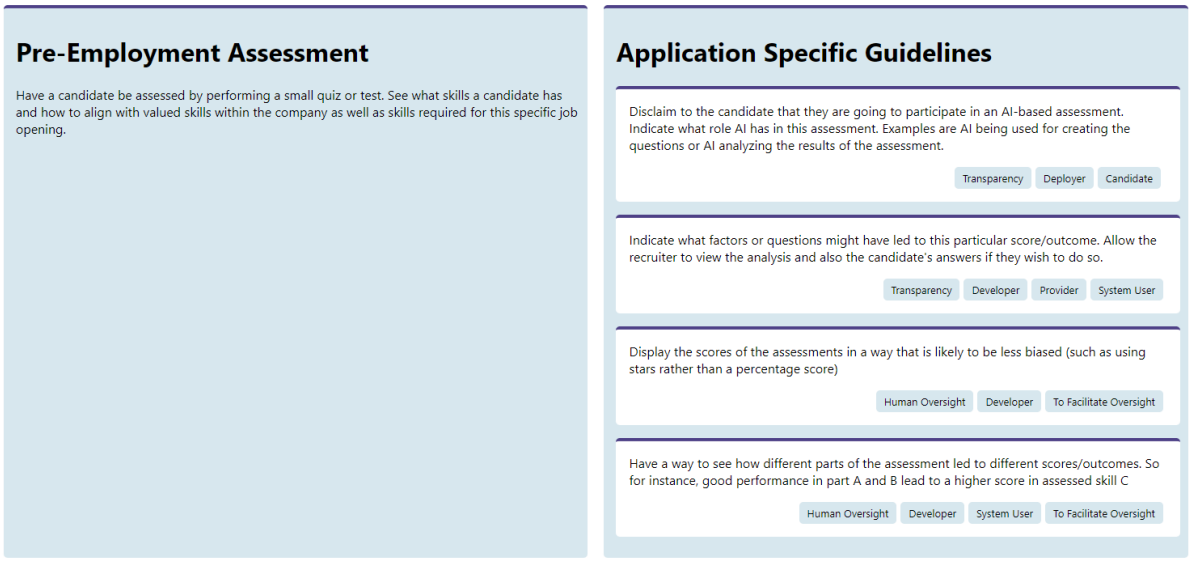



Figure 7: Application specific overview

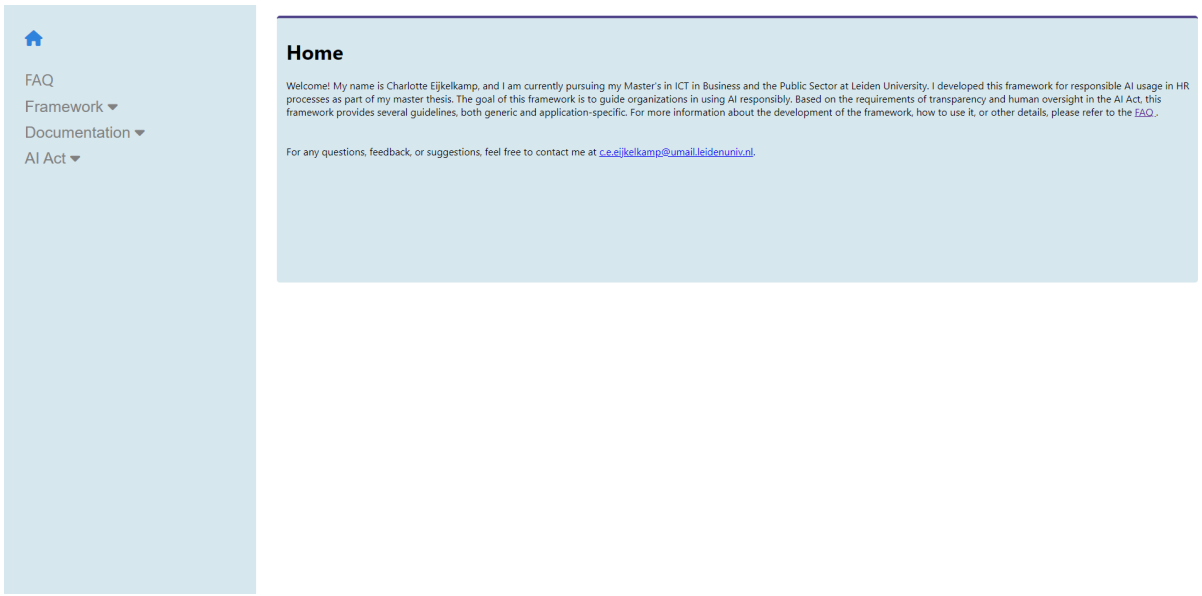



Figure 8: Home page

To see what the relevant guidance is for this application, the user goes to the application-specific guidelines fig. 9.

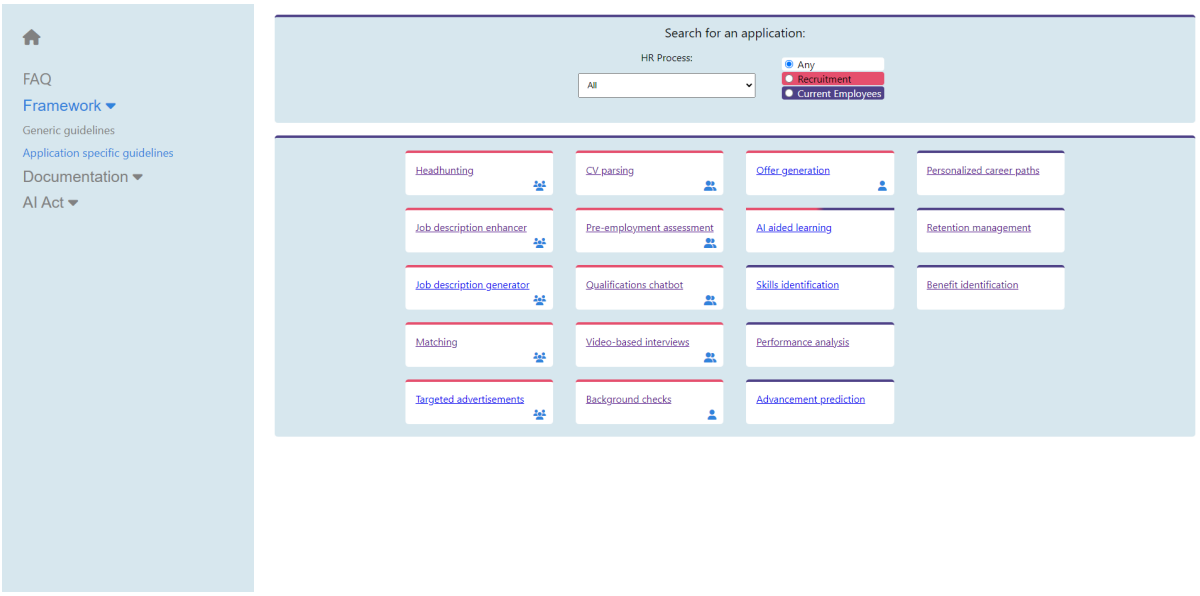



Figure 9: All applications in application overview

The user further filters the applications, to only display the recruiting applications fig. 10. Now that there are fewer applications to go through, the user can easily select the application relevant to them: 'Pre-employment assessment'.

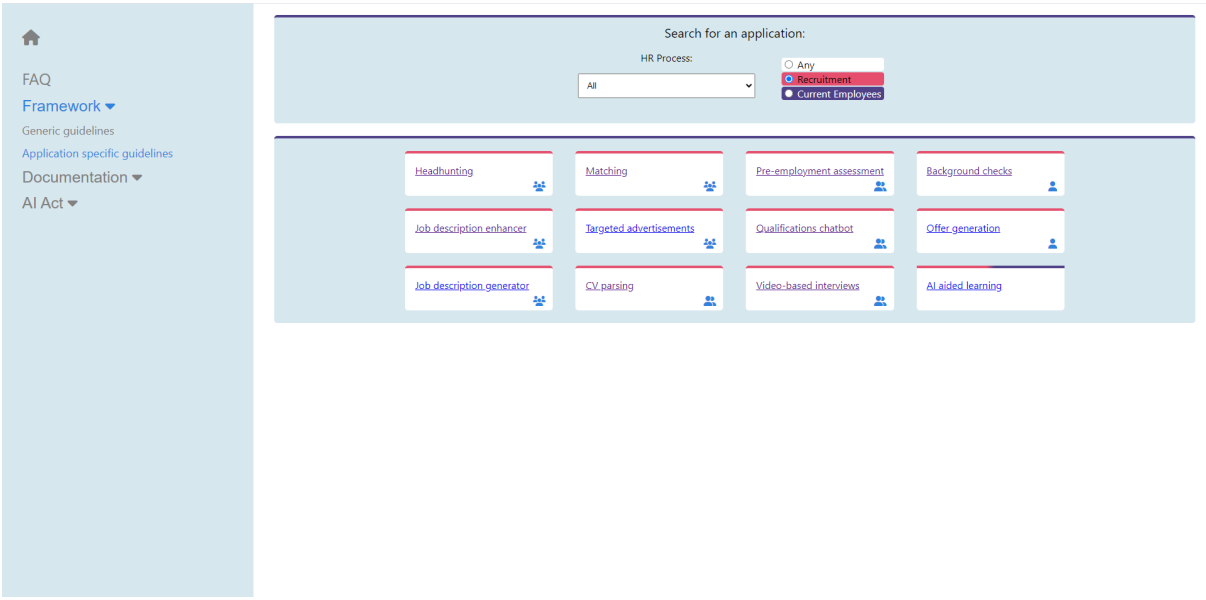



Figure 10: Filtered application overview

The detailed overview of the application is now visible fig. 11. The user can read the description of the application and assess if that is indeed the right application. The user can see the guidelines relevant to them, as they are marked with the tag 'deployer'. Next to that, the user can also read the other guidelines, as there is not an overwhelming amount of guidelines displayed. From reading this the user might learn what they can expect, or ask from a provider.


Figure 11: Detailed application view

# 5 Evaluation

## 5.1 Evaluation setup

The final framework is evaluated to test its usability and quality. The evaluation is performed individually with participants in different roles, that are all potential end-users of the framework. The evaluation is split into two phases. The first phase consists of interacting with the framework and answering some questions in person, the second phase consists of a digital self-administered questionnaire. The phases are executed right after each other. A more detailed set-up of the evaluation, including all the questions and questionnaire contents can be found in appendix D.

#### Phase 1: System interaction and open questions

In the first phase, the participant is asked to interact with the framework and comment as they go. After this interaction, a few open questions are asked. This phase is performed in person, as it allows the participant to give non-standardized answers, and reduce the unwillingness to take the time to answer (written) open questions [36].

#### Phase 2: Questionnaire

The second phase of the evaluation consists of a self-administered digital questionnaire. The questionnaire tests for usability by including the questions of the system usability scale (SUS), as it is a proven and simple way to measure perceived usability [37]. Next to that, a few questions are asked with regard to the quality of the content. This phase is done by self-administration, allowing for a more swift execution.

## 5.2 Evaluation results

The evaluation was conducted with seven participants, each holding a different role. An overview of the participants is found in table 5. The evaluation results are then divided into two phases: Phase 1 and Phase 2. The results from Phase 1 are presented as statements that were mentioned multiple times, ranging from general observations to specific suggestions. These statements are further categorized into 'information access and quality' and 'usability.' Lastly, the questionnaire results are presented.

| Participant | Framework role | Formal role |
|---|---|---|
| P1 | deployer | Legal counsel |
| P2 | provider | Product owner |
| P3 | developer | System architect |
| P4 | developer | Developer |
| P5 | deployer | Ethics officer |
| P6 | provider | Product owner |
| P7 | developer | Developer |

Table 5: Evaluation participants

### 5.2.1 Results from phase 1

#### 5.2.1.1 Information access and quality

The statements from the evaluations regarding information access and quality per participant can be found in table 6. A more extensive discussion per statement is found below.

| | p1 | p2 | p3 | p4 | p5 | p6 | p7 | Total |
|---|---|---|---|---|---|---|---|---|
| Positive about guideline content overall | x | x | x | x | x | x | x | 7 |
| Positive about the amount of information present | x | x | x | x | x | x | x | 7 |
| Hard time understanding the terminology | x | x | x | x | | x | x | 6 |
| Positive about filtering for framework user | | | x | x | x | x | x | 5 |
| Positive about having a rationale | x | x | x | x | | | | 4 |
| Suggested improvements about the actionability of the guidelines | | x | | | | x | | 2 |

Table 6: Statements about information access and quality per participant

**Positive about guideline content overall**

Seven out of seven participants indicated that overall they were positive about the guideline content. The participants indicated for example that they were easily readable or that they appeared to be logical and relevant.

**Positive about the amount of information present**

Seven out of seven participants indicated that no information was missing, nor any irrelevant information was present. However, some candidates did indicate that not every piece of information was present at the place they had hoped to find it. For example, the explanation of the processes currently found under 'Documentation' can better be embedded in the application-specific part of the guideline. As that is the place where the processes are used.

**Hard time understanding the terminology**

Six out of seven participants indicated having a hard time understanding the terminology. Most confusion came from the different types of stakeholders and tags. For example, the term 'audience' is mentioned as being a bit ambiguous in this context. It might be the affected parties to whom you should be transparent, or the people currently reading the tags. Even though the stakeholders are explained in the documentation, no participant first consulted the documentation before delving into the framework. As mentioned before, a better job could have been done in explaining terminology directly when it is used, rather than on a separate page.

**Positive about filtering for framework user**

Five out of seven participants were positive about the option to filter for framework user so that only information relevant to that user was visible. Next to that one participant suggested adding an option to filter based on the use case, so for instance only show the guidelines relevant for preparing for an audit. Another participant suggested also adding an option to filter for the other tags and not just the framework user.

**Positive about having a rationale**

Four out of seven participants were positive about having a rationale. The participants indicated that it helps in building confidence in the guidelines or in persuading other people to implement the guidelines.

**Suggested improvements about the actionability of the guidelines**

Two out of seven participants suggested the guidelines should be more actionable. Both of these participants are product owners. They both would like to see a clear action someone has to perform to implement the guideline. One participant said the guidelines might be too wordy and a bulleted list was preferred.

#### 5.2.1.2 Usability

The statements from the evaluations regarding information access and quality per participant can be found in table 7. A more extensive discussion per statement is found below.

| | p1 | p2 | p3 | p4 | p5 | p6 | p7 | Total |
|---|---|---|---|---|---|---|---|---|
| Suggested improvements for flow through app | x | x | x | x | x | | | 5 |
| Positive about having an application-specific section | x | x | | | | x | x | 4 |
| Suggested improvements about how to embed in the organization | | x | x | | | x | | 3 |
| Suggested a restructuring of the menu | x | | | | | x | | 2 |
| Suggested to include more visual aids | | | | x | | | x | 2 |
| Suggested to include a search bar | | | | | x | x | | 2 |

Table 7: Statements about usability per participant

**Suggested improvements for flow through app**

Five out of seven participants suggested improvements for the flow through the app. Different ideas were suggested such as a wizard or walkthrough, or a more extensive explanation on the home page. One participant mentioned missing a 'return' button.

**Positive about having an application-specific section**

Four out of seven participants had a positive response to the application-specific section. One participant indicated that it is a great help in translating legislation into tangible steps. Two other participants indicated that they find it pleasant to have a way to only view the relevant guidelines.

**Suggested improvements about how to embed in the organization**

Three out of seven participants made suggestions on how to further integrate the system into the organization. Some recommendations were more about streamlining usage, where others are about making sure the framework will be implemented effectively. For example, suggestions

for streamlining include adding checkboxes, a way to track progress, prioritizing the guidelines, grouping and exporting guidelines, and providing a space to store evidence. Recommendations for effective implementation are, for example, adding a statement on the homepage to emphasize the importance of the guidelines and encourage users in using them, designating a contact person for users to reach out to with questions, or providing a method for submitting suggestions, such as proposing new applications.

**Suggested a restructuring of the menu**

Two out of seven participants suggested to restructure the menu. The current menu structure can be seen in fig. 12. Participant number one proposed to put the documentation higher up, so that users would first go over the documentation before proceeding to the guidelines. Participant seven proposed to put the framework higher, and also to put the application specific guidelines above the generic guidelines. This was suggested because they expected that users would probably use those more.

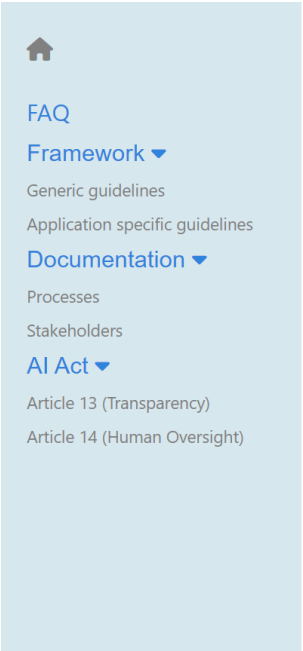



Figure 12: Current menu structure

**Suggested to include more visual aids**

Two out of seven participants suggested to use more visual aids. One suggestion was to further visualize the hiring funnel in the application overview. Another was to add more diagrams and figures such as the stakeholder diagram (see fig. 1) which is included in the stakeholder part of the documentation.

**Suggested to include a search bar**

Two out of seven participants suggested to add a search bar. Either to search for relevant guidelines based on key words, or to browse through the relevant AI Act articles.

### 5.2.2 Results from phase 2 (Questionnaire results)

The questionnaire aimed to measure perceived (content) quality and perceived usability. All questions were answered on a 5-pt likert scale. To determine the final score, the answer 'Strongly Disagree' gets rewarded 1 point, 'Strongly Agree' gets 5 points. For content quality three questions were answered, with an average score of 4.5 and a standard deviation of 0.5. The question "I was able to understand the guidelines" scored the highest (4.7 with a standard deviation of 0.5), followed by "Overall the guidelines presented were easy to read" (4.6 with a standard deviation of 0.5) and the lowest score was awarded to "I could see myself implementing the guidelines" (4.3 with a standard deviation of 0.5). For usability the questions of the SUS [37] were used. For positively worded questions 1 point got deducted, for negatively worded questions the score contribution is 5 minus the answer. All scores get added and multiplied by 2.5, bringing the range to 0-100. This resulted in an average SUS score of 76.4, with a standard deviation of 7.9.

# 6 Discussion

In this section, the results will be discussed, by assessing how well this research was able to answer the research questions, and how well the framework was able to meet the set objectives. Followed by the limitations of the research and suggestions for future research.

## 6.1 Research questions

Three research questions were defined. Due to their similarity and the joint approach in answering them, RQ1 and RQ2 will be discussed in the same section. Lastly, RQ3 is assessed.

- **RQ1:** What guidelines will help in reaching meaningful Transparency, in AI systems used in HR processes?
- **RQ2:** What guidelines will help in reaching meaningful Human Oversight, in AI systems used in HR processes?
- **RQ3:** What is the best-suited method for conveying the defined guidelines?

### 6.1.1 RQ1 and RQ2

Two sets of generic and application-specific guidelines were created, for both transparency and human oversight. This section discusses that the aspects of these guidelines, as well as the contents of these guidelines, resulted in answering both RQ1 and RQ2. This is done by first explaining which elements were added to the guidelines and why they are suitable. This will first be explained in the context of the overall situation, followed by a detailed discussion of the specific approaches for transparency and human oversight. After that, the content of the guidelines themselves will be highlighted, including what works well and what could be improved.

#### Overall situation

First, an overview of different AI systems used in HR processes is created to answer these research questions. There were 18 different applications identified. By providing guidance for specific applications, the guidelines will be more actionable and can therefore have a bigger contribution towards reaching both meaningful transparency and human oversight.

In addition, three types of framework users were identified. These are the stakeholders responsible for the execution and safeguarding of the guidelines. The framework users include the deployer, provider, and developer. Deployer and provider are terms defined in the AI Act, while developer is defined as a subcategory of provider and refers to the party responsible for the technical implementation. Having this developer subcategory allowed us to further target the more technical guidelines.

#### Transparency

To explore what contributes to meaningful transparency, the related work on transparency was examined. This review identified various aspects and perspectives on transparency, as well as its potential benefits and drawbacks. Based on the relational aspect of transparency, a specific

stakeholder group was identified: the Audience. This group consists of four different stakeholders: candidates, employees, auditors, and system users. A categorization of the stakeholders as well as an explanation of why they are suitable is given below.

- Candidates: These are individuals subject to decisions made by high-risk AI systems. According to Article 26(11) of the AI Act, they should be notified when they are affected by such decisions.
- Employees: Identified according to Article 26(7) of the AI Act, this group includes employees who must be informed by their employers (the deployers) when they are subject to a decision made by a high-risk AI system.
- Auditors: These are formal entities tasked with checking or assessing AI systems, such as the notified bodies defined in Article 31 of the AI Act.
- System Users: This group aligns with the concept of 'Transparency in Use' as defined by [14]. Transparency in use refers to improving system usability by, for example, sharing information on how to use the system effectively.

**Human Oversight**

To determine what contributes to meaningful human oversight, the related work on human oversight was reviewed. The key takeaways primarily highlight several downsides and challenges associated with human oversight, which the framework user should be aware of. Additionally, the stakeholder group identified for human oversight includes two distinct actors: the system user and the supervisor.

The decision to include these two actors is influenced by the concept of an accountability shift described by [21]. This idea emphasizes that human line workers should not be made scapegoats for the outcomes of decisions made by AI systems. To address this, another level of oversight is introduced—the supervisor—who provides accountability and oversight at a more global level, serving as a safeguard for the system users.

**Guideline contents**

So, with the framework users, transparency audiences, human oversight actors, and the list of different applications in mind the guidelines were designed. In total, there are 10 generic and 25 application-specific guidelines for transparency, and 9 generic and 21 application-specific human oversight guidelines created.

For both sets of guidelines, evaluations were conducted, and the results showed that participants were generally positive about them. However, one aspect that could not be tested in the evaluation was the completeness of the guidelines, as participants could only assess what they were presented with. The evaluations did highlight the actionability of the guidelines, though opinions on this were divided.

Many of the human oversight guidelines, compared to the transparency guidelines, are less actionable and tend to be more informative. This could be because human oversight is a less prevalent topic in the literature. There does not seem to be enough research on this theme yet, leading to general uncertainty about the best ways to implement human oversight.

Most participants were positive about the application-specific guidelines, particularly because they make the guidelines more applicable. However, two participants, both product owners, felt that the guidelines were not actionable enough, which might be influenced by their specific perspective as product owners. This raises the concern that valuable information might be lost if the guidelines are overly simplified, so it is worth considering how much these users should be accommodated.

### 6.1.2 RQ3

To determine the best method for conveying the defined guidelines, a framework was developed and evaluated. This process not only established a starting point but also provided insights for improving the framework and with that the approach to conveying the guidelines.

As part of the evaluation, the participants answered the questions of the System Usability Scale (SUS). Taking the average of all the participants' scores, resulted in a score of 76.4 which according to [38] indicates an above-average score ($> 68$). This might indicate that the created framework offers a decent way of conveying the guidelines, but that there is still room for improvement.

Alongside the SUS questions, the in-person statements and responses to the open-ended questions from the first evaluation phase provided additional insights into the suitability of the framework. Notably, many participants were positive about how the framework enabled them to navigate through the guidelines. Elements of the framework that elicited this positive reaction include the ability to filter guidelines based on the type of framework user and the presence of an application-specific section. What emerged in both cases was that it was very helpful to display only the information relevant to the user and their specific situation.

On the other hand, some areas for improvement were identified. Two significant improvements will be highlighted here. The first issue is the placement of the right information in the right place. Many participants struggled with understanding the terminology used in the framework, such as the definitions of different stakeholders. Although the framework provided explanations for each term, these were often located in different sections than where the user was currently browsing. A better solution would have been to present this information as close as possible to where the terms are used. This is where the current version of the framework fell short, and it is something that should be addressed in future iterations.

The second issue concerns accommodating different types of users. It was noted that both product owners approached the framework with a highly practical mindset, preferring to use it as a checklist. They expressed a desire for less text and a clearer prioritization of the guidelines, explaining that they often have limited time and cannot afford to read through lengthy sections without concrete actions to take away. However, this feedback was not voiced by other participants. The framework, therefore, failed to meet the specific needs of product owners. This does raise the question of whether it is even possible to do so. Some ethical practices are not easily reduced to a simple checklist, which may limit how effectively the framework can cater to such practical demands.

## 6.2 Framework Objectives

Four different objectives of the framework were defined. This section addresses to what extent these objectives were achieved.

- **Guidelines in line with AI Act:** The framework did reach this objective the following way. Clear mentions of transparency and human oversight in the AI Act are in articles 13 and 14 respectively. Both articles are translated into guidelines, for instance by providing guidelines about how to create the Instructions for Use, which is required by Article 13. Furthermore, other articles of the AI Act, such as Article 86 have found their way back into the guidelines.
- **Facilitate progress towards responsible AI:** The framework was partially successful in achieving this objective. On the one hand, the *guidelines* were rated as practical, particularly due to the application-specific section. On the other hand, the *framework* seemed to have a certain level of optionality, which might make it more difficult to use effectively. Suggestions for improvement include assigning a responsible contact person or including references to relevant documentation or forms within the organization.
- **Establish a common starting point for organizations:** The framework successfully provided a common starting point, particularly by offering easy filtering options for different types of users. This feature allows users to switch between guidelines aimed at various roles. However, since the evaluation was conducted in an experimental setting, it remains uncertain whether this functionality will be utilized in practice.
- **Inspire and activate responsible AI practices:** The framework was able to offer both inspirational and actionable guidelines. For example, the guideline 'Automation Bias' (generic human oversight) provides very practical tips, such as reducing the number of on-screen details or offering supportive information to users instead of directives. In contrast, the guideline 'Accountability Shift' (also generic human oversight) is more informative, raising awareness of the phenomenon but not offering direct solutions. Although there were participants who expressed a desire for more actionable guidelines, they also acknowledged the value of the additional information provided by the framework in these more 'inspirational' guidelines.

## 6.3 Limitations

This section addresses the limitations associated with the research process. It first examines the process of guideline and framework creation, followed by the limitations of the evaluation process. Finally, it considers the theoretical limitations of the created framework.

### 6.3.1 Process limitations

The development of the guidelines was based on the results of group interviews and relevant literature. The group interviews involved a total of seven participants, in three different groups. This sample size may be too small and might not provide a sufficiently generalizable perspective. Additionally, a potential limitation is that participants might have been influenced too strongly by their peers in this group setting. However, most findings were consistent with the existing literature.

For the guidelines derived from the relevant literature, many findings were drawn from the field of algorithmic decision-making, as this area is better researched compared to AI decision-making. However, there is limited literature on human-AI interaction, so it cannot be conclu-

sively stated that findings from algorithmic decision-making are fully applicable to AI decision-making.

The evaluation process also had some limitations. Participants were asked to explore the framework in a way that felt natural to them and were not given specific tasks to perform. This approach may have resulted in not uncovering all practical limitations or, conversely, highlighting issues that might not arise in a real-world setting. Additionally, the experimental context of the evaluation presented its own limitations. Themes such as time constraints and the ease of finding specific guidelines were not factors in the evaluation, yet these are critical issues that would likely play a role in practical applications. Another possible limitation is that all participants worked in a big organization and the results might not be generalizable for smaller organizations.

### 6.3.2 Theoretical limitations of the framework

This section examines the theoretical limitations of the framework, as opposed to the practical limitations already highlighted in the evaluation.

The framework was designed to encourage users to look beyond compliance, specifically in the areas of transparency and human oversight. Objectives for this framework were for example "establishing a common starting point for organizations." Two significant limitations emerge in this context: the lack of an upper boundary and the challenge of accommodating various types of users.

Firstly, the absence of an upper boundary. The guidelines aim to achieve meaningful transparency and human oversight, but no clear upper limit is defined. For example, in the case of transparency, it can be debated whether there should be transparency about aspects that were previously not fully disclosed. In an ideal scenario, one might share extensive information with a candidate about the process they will undergo. However, in situations without AI, candidates were not always aware of every detail of the process either. This raises questions about how far transparency should go.

Secondly, the challenge of accommodating different types of users. Different users have varying goals when using the framework. Ideally, the framework would cater to all these needs. However, the evaluation revealed that some users, particularly product owners, preferred more actionable guidelines. This preference contrasts with the nature of some guidelines that encompass ethical standards, which tend to be broader and less actionable. This raises the question of whether the framework can effectively meet the needs of all these different types of users.

## 6.4 Future work

First, some ways to further refine the framework will be discussed in this section. Followed by some suggestions to further research.

### 6.4.1 Framework refinements

This section outlines four major areas for refining the framework: adding additional AI Act themes, incorporating risk levels, indicating specific risks, and implementing practical improvements based on evaluation feedback.

Firstly, currently, only the themes of human oversight and transparency are addressed. However, the AI Act contains multiple other requirements that could be effectively incorporated into

this framework. For example, the Risk Management System requirement described in Article 9 could be included. Expanding the framework to cover more themes would create a more comprehensive tool, helping users to achieve compliance and beyond.

Secondly, the framework is built around several applications that span various risk levels. Some applications are high-risk, requiring more stringent compliance with AI Act requirements compared to limited-risk systems. Currently, the framework does not explicitly indicate the risk level of each system. Adding this information could provide two main benefits. First, it would help users differentiate between mandatory requirements and recommendations, enabling them to prioritize guidelines more effectively. Second, explicitly mentioning risk levels would offer greater clarity, helping users to reason more easily about new situations (e.g., an application not included in the list). This would be possible because users would have seen examples of what qualifies as high-risk or limited-risk, and would have a clearer understanding of the specific requirements associated with these risk levels.

Thirdly, indicating specific risks associated with each application could be a valuable enhancement. Currently, many potential risks are not highlighted, as the framework focuses on transparency and human oversight. However, there are important questions that remain underemphasized, such as whether to use social media data from a candidate or what potential biases might exist within the system [26]. Including these considerations could help users identify critical issues that require attention, or assist users in deciding whether they should proceed with providing or deploying a particular application at all.

Finally, practical improvements suggested during evaluations should be considered. These include adding a wizard or a clearer explanation of how to use the framework, providing explanations as close as possible to where they are needed, incorporating more visual aids, and adding a search function.

### 6.4.2 Suggestions for future research

Three suggestions for future research are provided, focusing on exploring different sectors, various organization sizes, and framework utilization in practice.

Firstly, regarding different sectors, the framework aims to encourage organizations to look beyond mere compliance, considering not only their business interests but also the interests of the affected persons. These fundamental principles could be applied to investigate how a similar framework might function in other sectors. Research could explore which elements—beyond the specific applications currently used—would need to be adapted to ensure the framework's effectiveness in different sectors.

Another option is to evaluate the framework in smaller organizations, particularly those that may not have the resources and people to streamline responsible AI practices. These organizations might benefit a lot from such a framework, making it a worthwhile area of investigation.

Finally, further research could focus on how the framework is utilized in practice. For example, studies could examine how frequently and at what stages of the process the framework is used, as well as which types of users make the most use of it. This would help determine whether the framework is having the desired impact in real-world settings.

# 7 Conclusion

This research aimed to develop a framework for guiding organizations in the responsible use of AI within HR processes, with a particular focus on the themes of transparency and human oversight. While legislation like the AI Act is making responsible AI use more concrete, it may not fully encompass all the ethical dimensions involved.

To provide this guidance, a design science research methodology was employed, resulting in the creation of two sets of guidelines and a framework to present them. The guidelines were tailored specifically for AI applications in HR processes, as well as more general guidelines applicable across all applications. These guidelines were developed based on a review of existing literature and insights gathered from group interviews with prospective candidates. The framework was designed to house these guidelines, along with auxiliary information, allowing users to easily browse and filter through the content.

This research identified a list of AI applications potentially used within HR processes, relevant stakeholders for responsible AI use, and key aspects needed to enhance the guidelines, such as defining an audience for transparency guidelines or specifying an oversight actor for human oversight guidelines. In total, 10 generic and 25 application-specific transparency guidelines, along with 9 generic and 21 application-specific human oversight guidelines, were developed.

The framework was then evaluated with seven participants, focusing on both its usability and the quality of the guidelines. The framework demonstrated slightly above-average usability, as measured by the System Usability Scale. Practical improvements were identified, such as the need to place auxiliary information as close to the relevant content as possible. Overall, participants found the guidelines understandable and useful. The inclusion of application-specific guidelines was particularly well-received. However, reactions to the actionability of the guidelines varied depending on the participant's role.

In the future, the framework could be refined by incorporating additional themes beyond human oversight and transparency, clearly indicating risk levels, or highlighting specific risks associated with particular applications. Further research could explore which elements of the framework might be applicable in other sectors, or by testing the current framework more thoroughly across different organization sizes or in more practical, real-world settings.

# 8 References

[1] AI HLEG, High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy ai," 2019. https://ec.europa.eu/digital-singlemarket/en/news/ethics-guidelines-trustworthy-ai.

[2] The European Parliament and the Council of the European Union, "Artificial intelligence act," 2024.

[3] L. Edwards, "The EU AI Act: a summary of its significance and scope," *Artificial Intelligence (the EU AI Act)*, vol. 1, 2021.

[4] M. Gornet, "The European approach to regulating AI through technical standards." Oct. 2023.

[5] ANEC, "ANEC Position Paper on CE Marking," Nov 2012.

[6] M. Veale and F. Zuiderveen Borgesius, "Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach," *Computer Law Review International*, vol. 22, no. 4, pp. 97–112, 2021.

[7] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems*, vol. 24, no. 3, pp. 45–77, 2007.

[8] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature machine intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[9] S. Larsson and F. Heintz, "Transparency in artificial intelligence," *Internet Policy Review*, vol. 9, no. 2, 2020.

[10] K. de Fine Licht and J. de Fine Licht, "Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy," *AI & society*, vol. 35, pp. 917–926, 2020.

[11] R. Bommasani, K. Klyman, S. Longpre, S. Kapoor, N. Maslej, B. Xiong, D. Zhang, and P. Liang, "The foundation model transparency index," *arXiv preprint arXiv:2310.12941*, 2023.

[12] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux, "Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns," *Big Data & Society*, vol. 6, no. 1, p. 2053951719860542, 2019.

[13] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux, "Towards transparency by design for artificial intelligence," *Science and Engineering Ethics*, vol. 26, no. 6, pp. 3333–3361, 2020.

[14] G. Andrada, R. W. Clowes, and P. R. Smart, "Varieties of transparency: Exploring agency within ai systems," *AI & society*, vol. 38, no. 4, pp. 1321–1331, 2023.

[15] O. B. Albu and M. Flyverbom, "Organizational transparency: Conceptualizations, conditions, and consequences," *Business & society*, vol. 58, no. 2, pp. 268–297, 2019.

[16] B. Gyevnar, N. Ferguson, and B. Schafer, "Bridging the transparency gap: What can explainable ai learn from the ai act?," *arXiv preprint arXiv:2302.10766*, 2023.

[17] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, *et al.*, "The role of explainable ai in the context of the ai act," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1139–1150, 2023.

[18] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, p. 101805, 2023.

[19] K. Kyriakou and J. Otterbacher, "In humans, we trust: Multidisciplinary perspectives on the requirements for human oversight in algorithmic processes," *Discover Artificial Intelligence*, vol. 3, no. 1, p. 44, 2023.

[20] R. Koulu, "Proceduralizing control and discretion: Human oversight in artificial intelligence policy," *Maastricht Journal of European and Comparative Law*, vol. 27, no. 6, pp. 720–735, 2020.

[21] B. Green, "The flaws of policies requiring human oversight of government algorithms," *Computer Law & Security Review*, vol. 45, p. 105681, 2022.

[22] B. Green and Y. Chen, "The principles and limits of algorithm-in-the-loop decision making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–24, 2019.

[23] M. Busuioc, "Accountable artificial intelligence: Holding algorithms to account," *Public Administration Review*, vol. 81, no. 5, pp. 825–836, 2021.

[24] R. Baeza-Yates, "Bias on the web," *Communications of the ACM*, vol. 61, no. 6, pp. 54–61, 2018.

[25] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[26] M. Bogen and A. Rieke, "Help wanted: An examination of hiring algorithms, equity, and bias," 2018.

[27] P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in human resources management: Challenges and a path forward," *California Management Review*, vol. 61, no. 4, pp. 15–42, 2019.

[28] Q. Jia, Y. Guo, R. Li, Y. Li, and Y. Chen, "A conceptual artificial intelligence application framework in human resource management," 2018.

[29] A. Blair-Early and M. Zender, "User interface design principles for interaction design," *Design Issues*, vol. 24, no. 3, pp. 85–107, 2008.

[30] F. Heymans, T. Gils, and W. Ooms, "From policy to practice: Prototyping the eu ai act's transparency requirements.," 2024.

[31] City of Amsterdam, "Contractual terms for algorithms," 2021.

[32] G. Adomavicius, J. Bockstedt, S. Curley, and J. Zhang, "Reducing recommender systems biases: An investigation of rating display designs," *MIS Quarterly*, vol. 43, no. 4, pp. 19–18, 2019.

[33] K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: a systematic review of frequency, effect mediators, and mitigators," *Journal of the American Medical Informatics Association*, vol. 19, no. 1, pp. 121–127, 2012.

[34] Q. Yang, A. Steinfeld, C. Rosé, and J. Zimmerman, "Re-examining whether, why, and how human-ai interaction is uniquely difficult to design," in *Proceedings of the 2020 chi conference on human factors in computing systems*, pp. 1–13, 2020.

[35] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, *et al.*, "Guidelines for human-ai interaction," in *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–13, 2019.

[36] B. C. Choi and A. W. Pak, "A catalog of biases in questionnaires," *Preventing chronic disease*, vol. 2, no. 1, 2005.

[37] J. R. Lewis, "The system usability scale: past, present, and future," *International Journal of Human–Computer Interaction*, vol. 34, no. 7, pp. 577–590, 2018.

[38] J. Sauro, "Measuring usability with the system usability scale (SUS)," 2011.

# Appendices

# Appendix A Process and application list

## A.1 Sourcing

Sourcing involves attracting potential candidates to apply for open positions through advertisements, job postings, and individual outreach efforts [26].

### A.1.1 Job Description Generator

Generate job descriptions from provided requirements. Helps in automating a frequent task, and can aid in ensuring a clear and readable text is presented to the candidate.

### A.1.2 Job Description Enhancer

Employ AI to optimize job descriptions for attractiveness, readability, and inclusivity.

### A.1.3 Targeted Advertisements

Have your advertisements placed in such a way that they reach the people who are most likely to apply to the job posting.

### A.1.4 Matching

Finding a match between a candidate and a job opening. This is done by both matching a candidate to multiple suitable jobs and having multiple qualified candidates matched to a job opening.

### A.1.5 Headhunting (Flight risk)

Have AI aid you in seeking passive candidates, by making predictions of who is likely to leave their existing job, and/or who is a good fit for the position.

## A.2 Screening

Screening refers to employers assessing candidates, both before and after they apply, by analyzing their experience, skills, and characteristics [26].

### A.2.1 CV Parsing

Parse a CV and turn it into qualifications or skills, based on natural language processing.

### A.2.2 Qualifications chatbot

Have candidates interact with the chatbot to pre-screen the candidates based on their qualifications.

### A.2.3 Pre-employment assessment

Have a candidate be assessed by performing a small quiz or test. See what skills a candidate has and how it aligns with valued skills within the company as well as skills required for this specific job opening.

## A.3 Interviewing

Interviewing continues the assessment of candidates in a direct and personalized manner, by planning physical or digital conversations [26].

### A.3.1 Video based interviews

Have the interview recording parsed and analyzed for, for instance, facial expressions or vocal indications, etc.

## A.4 Selection

The selection process involves employers making final hiring and compensation decisions [26].

### A.4.1 Background checks

Automated background checks that utilize public data to assess whether a candidate may pose a risk for misconduct.

### A.4.2 Offer generation

Generate a salary offer based on market insights.

## A.5 Onboarding

Onboarding is the process of integrating a new employee into an organization with the aim of quickly maximizing their productivity [27].

### A.5.1 AI aided learning

Have AI improve training and learning by delivering personalized learning experiences. Potential functionalities could be analyzing employee progress, tailoring content to individual needs, and providing real-time feedback.

## A.6 Training and Skills Development

About offering training programmes to employees, that suit their individual needs. Partially overlaps with onboarding, in terms of learning/training. So the application mentioned in A is also used in this process. Next to that, the following application is also applicable in the training and skills development domain:

### A.6.1 Skills identification

Use AI to assess and analyze the skill sets within an organization, identifying areas for improvement.

## A.7 Performance Management

The process of identifying both good and poor performance and determining factors that could enhance job performance [27].

### A.7.1 Performance Analysis

Evaluate employee or company performance to identify top performers, areas for improvement, and factors influencing performance.

## A.8 Advancement/Career Paths

This process involves creating individualized career paths that align with the ambitions and capabilities of employees, as well as overseeing and forecasting successful career advancements.

### A.8.1 Personalized Career Paths

Create career development plans for employees based on their skills, interests, and goals. It could recommend customized learning or advancement pathways.

### A.8.2 Advancement prediction

Employ AI to forecast which employees are likely to excel in new roles within the organization. By analyzing historical performance data, skills assessments, and other relevant factors, this tool identifies individuals with the potential to thrive in different positions.

## A.9 Retention

The process of predicting who is likely to leave, and managing the level of retention [27].

### A.9.1 Retention management

Utilize AI to predict which employees are at risk of leaving the organization, often by analyzing factors such as LinkedIn activity, job satisfaction surveys, and historical retention data.

## A.10 Employee Benefits

The process involves identifying the most valued benefits among employees to inform decisions on offerings and recommendations. Additionally, it assesses the effects of these benefits, such as their impact on recruitment and retention [27].

### A.10.1 Identifying important benefits

Use AI to identify the benefits that hold the highest significance for employees. By analyzing data such as employee feedback, surveys, and market trends, this tool pinpoints the most valued benefits within the workforce.

# Appendix B Interview Protocol

In total three different group interviews were held, all with two or three participants per group. The interviews are conducted in a group setting, enabling participants to engage in open discussions and delve beyond surface-level beliefs. Since this may be a relatively new or unfamiliar field, participants might not have given it much thought yet. The group setting encourages deeper reflection on the topic. The interview is semi-structured, with few predetermined questions, allowing for a more flexible and exploratory conversation.

**Introduction**

- Who are you and what is your role?
- What is your base knowledge about AI?
- What is your base knowledge about AI in HR processes?

**Transparency**

- Are there any things you regard as intransparent with regards to AI?
- Are there any things you regard as intransparent with regards to AI used in HR processes?
- If AI was used in your job application process, what would you like to know about this? (think about how it was developed, how it works, what are the risks etc.)
- What information would you like to have access to?
- In what form would you like to access this information?
- How much effort would you put into finding this information?

**Human Oversight**

- To what extent do you want human involvement in your AI-based job application process?
- ***explain to the interviewees the downsides of human oversight (such as automation bias, a decreased accuracy etc.)***
- Does this change your view on the role of a human in the process?
- What is more important to you, accuracy or a human touch?

# Appendix C Interview results summary

Seven different participants, spread out over three groups were interviewed. All participants were at the start of their careers and therefore prospective candidates. The participants had varying levels of AI knowledge, but no participant considered themself an AI expert. Overall, the participants who reported having higher levels of AI knowledge seemed to have stricter requirements or expectations. Likely because they were more aware of the risks associated with using AI.

In summary, the participants did not have the trust in AI to be able to fully capture the human aspect of recruitment. Participants did see opportunities for using it in less 'human' processes, such as checking qualifications. Participants are not fully aware of whether or where AI is used but would like to be informed proactively about this. Participants want to know how their job application is affected by AI and how they should take that into consideration when applying.

## Takeaways

**Doubts and concerns of the participants:**

- AI not being able to capture your personality.
- AI feels too much like a checklist and lacks nuance.
- Candidates with a non-typical profile have fewer chances.
- In preparing for a job application, the candidates would rather focus on the content and not on 'how to beat/game' the AI.

**Opportunities/Positive sides:**

- Using AI for high-volume applications or earlier in the process when more applicants are involved.
- Using AI for standard inclusion or exclusion criteria such as a language requirement or a specific driver's license.
- Using AI for onboarding, specifically the factual and informational parts.
- Using AI might allow for less biased procedures and for instance anonymous applications.

## Transparency takeaways

**Things considered untransparent about AI:**

- Not knowing whether AI is used at all.
- Not knowing the way AI is influencing the process.
- Not knowing the way AI is making decisions or a lack of understanding of the inner workings of a system.

**Things the participants would like to know:**

- Whether AI is used in your job application, and in which part of your application that is.
- The way AI affects your job application, and how to take that into account.
- Tips and examples about how to best prepare for your application or interview
- Examples of what would be a good application or a bad application, e.g. an analysis of a video interview that would flag looking away too much as a bad interview and maintaining eye contact as a good interview.
- E.g. structuring your CV in a certain way or including certain keywords.
- The risks associated with the AI system used.
- Whether there is a way to opt out of being subject to an AI system.

**The way this information should or could be shared:**

- Proactive communication is preferred:
- During the start of your job application, via mail or an explanatory video.
- In case of rejection, notify the way AI played a role in this and why you got rejected.
- Mention it in the vacancy
- Some information on the organization's website about the way AI is included in the hiring process.

### Human oversight takeaways

- Human touch is more important than accuracy.
- Human touch remains very important, especially to see whether you as a person match with the organization.
- Preference for having a face/person/name of someone responsible for handling your application.
- Use AI as an advisor, e.g. give the recruiter multiple options or a ranking, instead of only one candidate.
- Fully automated rejections only in the early phases.

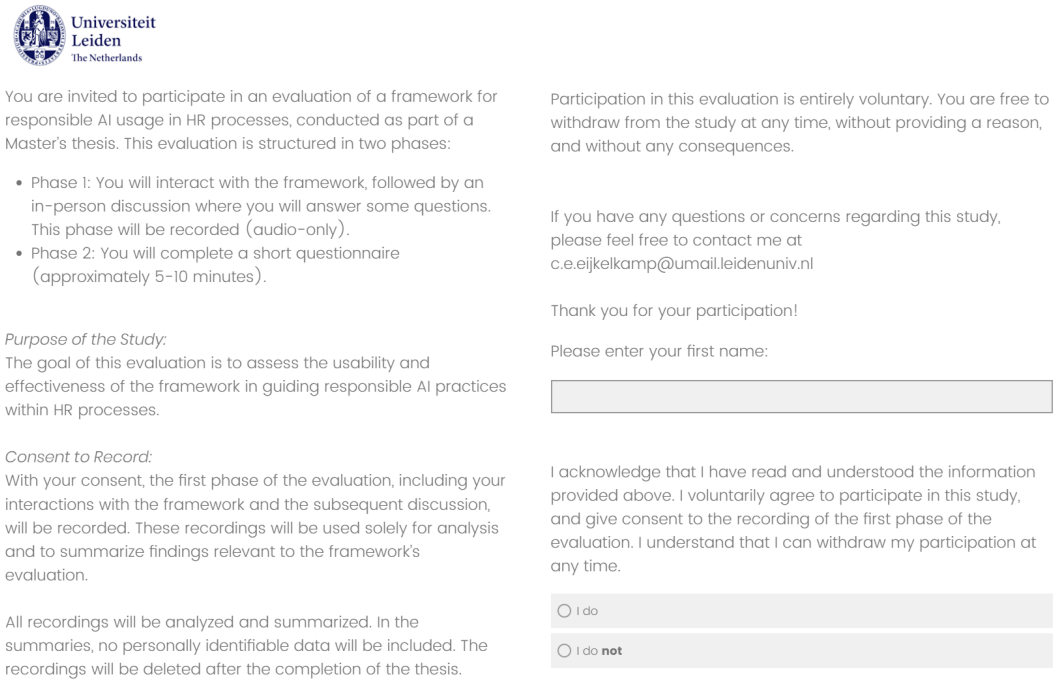
# Appendix D Evaluation set-up

## D.1 Phase 1: System interaction and open questions

The participant should first read and sign the consent form in fig. 13, before proceding to the interaction with the framework. After this interaction the following open questions were asked.

**Open questions:**

- Is there any information that appears to be missing?
- Is there a feature that is missing?
- What stands out negatively?
- What stands out positively?


Universiteit Leiden
The Netherlands

You are invited to participate in an evaluation of a framework for responsible AI usage in HR processes, conducted as part of a Master's thesis. This evaluation is structured in two phases:

- Phase 1: You will interact with the framework, followed by an in-person discussion where you will answer some questions. This phase will be recorded (audio-only).
- Phase 2: You will complete a short questionnaire (approximately 5-10 minutes).

*Purpose of the Study:*
The goal of this evaluation is to assess the usability and effectiveness of the framework in guiding responsible AI practices within HR processes.

*Consent to Record:*
With your consent, the first phase of the evaluation, including your interactions with the framework and the subsequent discussion, will be recorded. These recordings will be used solely for analysis and to summarize findings relevant to the framework's evaluation.

All recordings will be analyzed and summarized. In the summaries, no personally identifiable data will be included. The recordings will be deleted after the completion of the thesis.

Participation in this evaluation is entirely voluntary. You are free to withdraw from the study at any time, without providing a reason, and without any consequences.

If you have any questions or concerns regarding this study, please feel free to contact me at c.e.eijkelkamp@umail.leidenuniv.nl

Thank you for your participation!

Please enter your first name:

I acknowledge that I have read and understood the information provided above. I voluntarily agree to participate in this study, and give consent to the recording of the first phase of the evaluation. I understand that I can withdraw my participation at any time.

○ I do

○ I do **not**

Figure 13: Consent form

## D.2 Phase 2: Questionnaire

Indicate to what extent you agree with the statements below.

| | Strongly disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Strongly agree |
|---|---|---|---|---|---|
| I think that I would like to use this system frequently. | ○ | ○ | ○ | ○ | ○ |
| I found the system unnecessarily complex. | ○ | ○ | ○ | ○ | ○ |
| I thought the system was easy to use. | ○ | ○ | ○ | ○ | ○ |
| I think that I would need the support of a technical person to be able to use this system. | ○ | ○ | ○ | ○ | ○ |
| I found the various functions in this system were well integrated. | ○ | ○ | ○ | ○ | ○ |
| I thought there was too much inconsistency in this system. | ○ | ○ | ○ | ○ | ○ |
| I would imagine that most people would learn to use this system very quickly. | ○ | ○ | ○ | ○ | ○ |
| I found the system very cumbersome to use. | ○ | ○ | ○ | ○ | ○ |
| I felt very confident using the system. | ○ | ○ | ○ | ○ | ○ |
| I needed to learn a lot of things before I could get going with this system. | ○ | ○ | ○ | ○ | ○ |

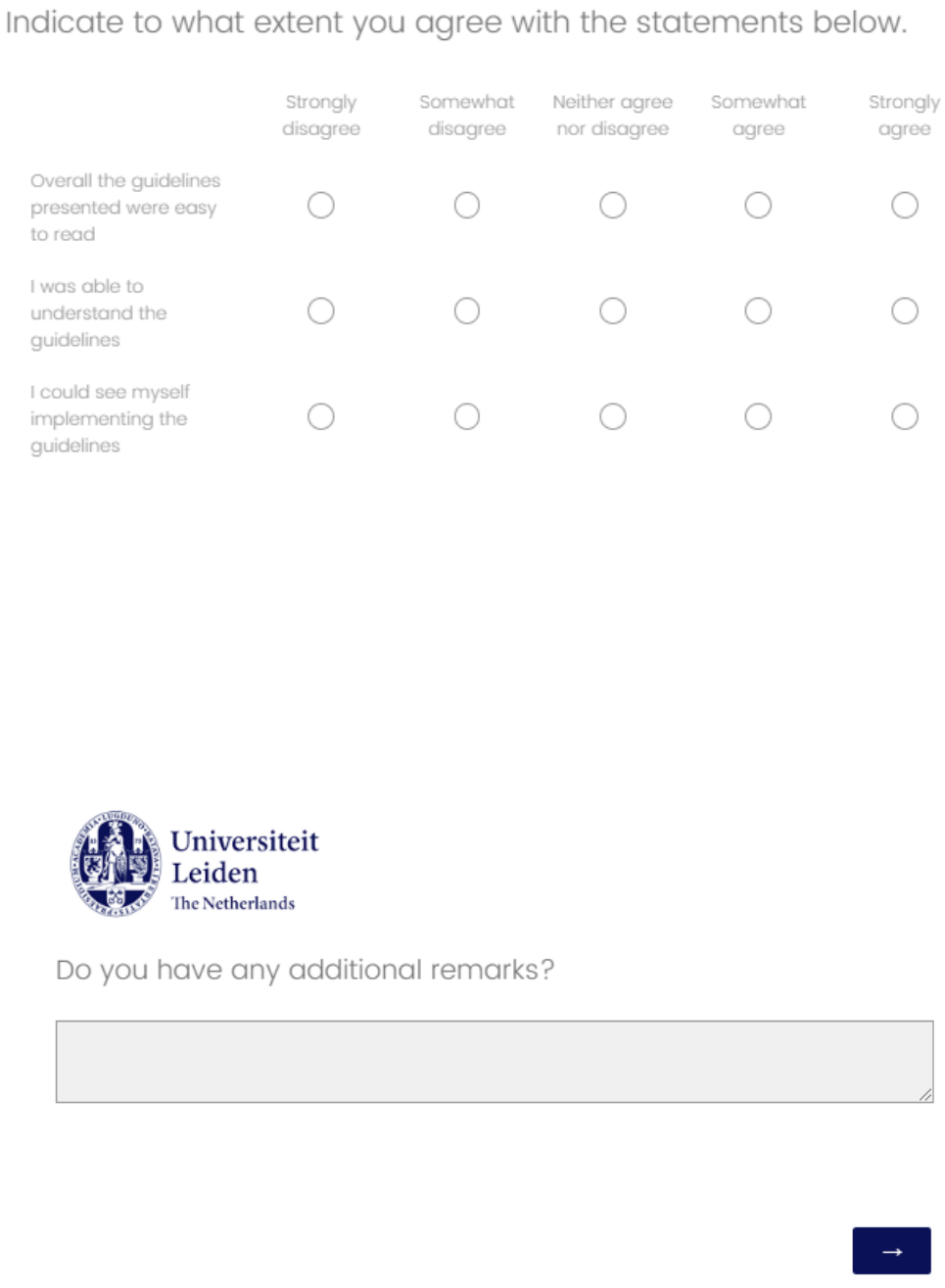



Figure 14: Questionnaire contents

# Appendix E Guidelines

## E.1 Transparency guidelines

### E.1.1 Generic guidelines

Table 8: Generic transparency guidelines

| Name | Description | Framework user | Audience | Rationale |
|---|---|---|---|---|
| Audience sensitivity - candidates | Not everyone has the same level of knowledge regarding AI, so try to communicate in a way that is understandable for a wide reach of different knowledge levels. Next to that candidates seem to prefer a proactive sharing of information. This can be done through email during the application process, potentially including an explanatory video about the application procedure and the role of AI in it. Key information to be shared should cover how their application, CV, or interview will be processed, enabling candidates to anticipate and understand the process. | Deployer | Candidates | A combination of the performability aspect of transparency [15] and results from interviews with (prospective) candidates |
| Data quality | To provide transparency about both the data and data quality, consider sharing the following elements including a motivation if suitable: the source of the data, the way data is being edited/filtered/augmented, the mitigations for the presence of copyrighted data or PII, information about the data labor such as the instructions for creating data as well as the geographic distribution of data laborers. | Developer | Auditors | A combination of AI Act articles 13.3(b).VI (Transparency) and 10 (Data and data governance) [2] and the foundational model transparency index (FMTI) categories 'data', 'data labor', 'data mitigations' and 'data access' [11] |
| System quality | Indicating the quality of the system entails showing that the system performs accurately and correctly and is suitable for the intended use. Sharing the following information, including a motivation where suitable, could help in reaching this goal: what development methods are used during the creation of the system (e.g. Agile or feature-driven development), what frameworks are used, and what additional dependencies are required? Also, include information about the performance of the system. | Developer | Auditors | A combination of AI Act article 11 (technical transparency, annex IV) [2], the standard clauses for procurement of algorithmic systems article 4 (quality of the algorithmic system) [31], and the foundational model transparency index (FMTI) category 'methods' [11] |

Table 8: continued

| Name | Description | Framework user | Audience | Rationale |
|---|---|---|---|---|
| Technical transparency | Take into account that the source code might need to be shared with third-party auditors, but only on demand, not by default. Next to source code other information that can be disclosed are the system's architecture and components. Take into account that the source code might need to be shared with third-party auditors, but only on demand, not by default. Next to source code other information that can be disclosed are the system's architecture and components. | Developer | Auditors | A combination of AI Act article 11 (technical transparency, annex IV) [2] and the standard clauses for procurement of algorithmic systems article 5.2 (technical transparency) [31] |
| Compute | In the case of creating and training a model, share the required compute as well as the energy usage and broader environmental impacts resulting from the model creation. | Provider | Auditors | A combination of AI Act article 13(3)e [2] and the foundational model transparency index category 'compute' [11] |
| Procedural transparency | In being transparent about the procedures involved in creating the AI system, the following information can be shared: What choices and assumptions are made in creating the system (e.g. what is the assumed data distribution, or assumptions about how the system will be used), as well as the parties involved in the development of the system and their roles, and the way risks are identified and mitigated. | Provider | Auditors | AI Act article 9 (risk management system) [2] and the standard clauses article 4 (procedural transparency) [31] |
| Instructions for Use (IFU) | Article 13 of The AI Act requires providers of AI systems to include instructions for use (IFU) with their systems. The IFU must be in an appropriate digital format or other suitable form and should contain concise, complete, correct, and clear information. The instructions must be relevant, accessible, and easy to understand for deployers [2]. | Provider and Developer | System user | Taken from AI Act article 13(2) (transparency) [2]. |

Table 8: continued

| Name | Description | Framework user | Audience | Rationale |
|---|---|---|---|---|
| IFU best practices | Ensure the Instructions for Use (IFU) document is specifically targeted to the primary users. Begin by describing who the primary user is, this can be done by for instance detailing their level of knowledge, role within the organization, or their relation to the AI system. The document should follow a logical structure and be written in clear, concrete language. Consider adapting the language of the IFU to the mother tongue of the target user for better comprehension. Enhance understanding by including visual elements such as images, tables, or graphs. Finally, include a FAQ or troubleshooting section to address common issues and questions. | Provider | System user | Taken from AI Act article 13(2) (transparency) [2] and enhanced with best practices from policy prototyping paper [30] |
| IFU content | In the Instructions for Use, the following information should at least be included: Who the providing party is and their contact details. The intended purpose of the system and the way human oversight is set up (what is the human role?). The way the system is keeping logs. Specifications on the input data and the data used for e.g. training and testing. The accuracy, robustness, and cybersecurity of the system, as well as the metrics used to determine them. The risks to fundamental rights that might be prevalent when using the system in a 'regular' way. Next to that, the IFUs should contain information about the (planned) changes to the system. An example of such an IFU can be found in [30], and the full list of requirements can be found in the AI Act [2]. | Provider | System user | AI Act article 13(2) (transparency) [2] and policy prototyping paper [30] |
| Right to Explanation of Individual Decision-Making | In article 86 of the AI Act, the right to explanation of individual decision-making is described. It applies whenever a decision made by the deployer, based on the outcomes of a high-risk AI system, has significant legal effects or adverse impact on the fundamental rights of the person subject to this decision. In this situation, the affected person has a right to obtain a clear and meaningful explanation of the role of the AI system in the decision-making procedure and the main elements of the decision taken. | Deployer | Candidates and Employees | Taken from AI Act Article 86 [2]. |

### E.1.2 Application specific guidelines

Table 9: Application specific transparency guidelines

| Application | Description | Framework user | Audience | Rationale |
|---|---|---|---|---|
| Job Description Generator | Whenever the underlying model of the system is trained on a known dataset, it is important to inform the user about potential malfunctions that may arise due to the dataset containing misinformation or being skewed. For instance, if there was very little training data in a specific language, there is a risk that the data may be gender-biased, potentially resulting in job descriptions that better align with male candidates. Another scenario could involve outdated data referencing obsolete working laws, increasing the risk of misinformation in the generated job description. | Developer and Provider | System user | Based on generic transparency guideline 'Data quality', also required to inform the user about the performance and accuracy in the Instructions for Use [2] |
| Job Description Enhancer | Consider explaining why certain changes are recommended. Try to present the user with examples leading to this specific recommendation or score (E.g. specific words or sentences that result in a lower readability score). | Developer | System user | Transparency by making the system more explainable and easier to use |
| Targeted Advertisements | Explain to a candidate why they are seeing this advertisement. Next to that consider sharing what advertisements this candidate might miss out on. | Developer | Candidates | Informing affected parties about the way their exposure to job openings is affected by AI. |
| Matching | Indicate what is done to minimize biased outcomes in matching candidates to jobs and vice versa. | Provider | System user | Based on generic transparency guideline 'Procedural transparency' |
| | Indicate that the results shown are based on a matching module, and explain what elements this candidate or job got matched on. This can for instance be done by highlighting what similarities are found between this candidate and a candidate that is deemed suitable, or an ideal candidate profile as described by recruiters. | Developer and Provider | System user | Transparency by making the system more explainable and easier to use |
| | Strike a balance between clarifying what factors contribute to a match or a favorable score, without sharing too much information that could be exploited by candidates to game the system. | Developer and Provider | Candidates | AI Act article 26.11 describing the obligation of deployers to inform people whenever they are subject to the use of the high-risk AI system [2]. |

Table 9: continued

| Application | Description | Framework user | Audience | Rationale |
|---|---|---|---|---|
| | Provide transparency regarding the creation of an ideal profile, including the data sources utilized, like exclusively in-company data. If such is the case, communicate about the measures implemented to mitigate biases. | Provider and Deployer | Candidates | Based on generic transparency guideline 'Procedural transparency' |
| Headhunting | Clarifying what factors contribute to a match or a favorable score, without sharing too much information that could be exploited by candidates to game the system. | Developer and Provider | Candidates | AI Act article 26.11 describing the obligation of deployers to inform people whenever they are subject to the use of the high-risk AI system [2]. |
| | Provide transparency regarding the creation of an ideal profile, including the data sources utilized, like exclusively in-company data. If so, communicate about the measures implemented to mitigate biases. | Provider and Deployer | System user | Based on generic transparency guideline 'Procedural transparency' |
| CV Parsing | Allow a candidate insight into what their CV got parsed into. Next to that give the candidate the chance to change the information when incorrect. | Developer and Provider | Candidates | AI Act article 26.11 describing the obligation of deployers to inform people whenever they are subject to the use of the high-risk AI system and article 86 describing the right to an explanation of individual decision making [2]. |
| Qualifications Chatbot | The AI Act mandates the inclusion of a disclaimer when using chatbots. The objective is to ensure users are aware of interacting with a chatbot, especially if it's not already evident from the context. | Provider | Candidates | AI Act article 50.1 [2] |
| | Inform the candidate what the output of the process is. In case of rejection explain why the candidate got rejected, and give the candidate the chance to revisit this decision. | Provider | Candidates | AI Act Article 86 describing the right to individual decisionmaking [2]. |
| Pre-employment assessment | Disclaim to the candidate that they are going to participate in an AI-based assessment. Indicate what role AI has in this assessment. Examples are AI being used for creating the questions or AI analyzing the results of the assessment. | Deployer | Candidates | AI Act article 26.11 describing the obligation of deployers to inform people whenever they are subject to the use of the high-risk AI system [2]. |
| | Indicate what factors or questions might have led to this particular score/outcome. Allow the recruiter to view the analysis and also the candidate's answers if they wish to do so. | Developer and Provider | System user | Transparency by making the system more explainable and easier to use |

Table 9: continued

| Application | Description | Framework user | Audience | Rationale |
|---|---|---|---|---|
| Video based Interviews | Inform the candidate that the video will be analyzed by AI and provide details on the methods employed for this analysis. For instance by vocal analysis or an analysis of facial features. Next to that disclose what exactly the role of the AI is, by for instance sharing what parts of this process are automated and what parts are not. Share where in the process humans are involved and the way they interact with the AI system, e.g. what they will do with the outputs of the analysis. | Provider and Deployer | Candidates | AI Act article 26.11 describing the obligation of deployers to inform people whenever they are subject to the use of the high-risk AI system [2]. |
| Background checks | Inform the candidate that an automated background check will take place, and indicate which data will be used for this check. Next to that disclose what specific data points or features might have led to the outcome of this candidate's background check. | Provider and Deployer | Candidates | AI Act article 26.11 describing the obligation of deployers to inform people whenever they are subject to the use of the high-risk AI system [2]. |
| Offer generation | Inform the user about which features or parameters were used to come to specific offers. | Developer | System user | Transparency by making the system more explainable and easier to use |
| AI Aided learning | Indicate what specific factors lead to the recommended learning path, next to that include a way to see what different learning options could have been (like trainings that did not get recommended) | Developer | Employees | AI Act article 26.7 describing the obligation of deployers who are employers to inform their employees whenever they are subject to the use of the high-risk AI system [2]. |
| Skills identification | Be open about how the ideal combination of skill sets within an organization came about, and indicate what is done to ensure the organization is not missing out on new skills. | Provider and Deployer | Employees | Based on generic transparency guideline 'Procedural transparency' |
| Performance Analysis | Communicate what factors contribute to indicating high performance, and share what data is used to analyze performance. For instance financial data relevant to a specific team. | Developer and Provider | Employees | Based on generic transparency guideline 'Procedural transparency' as well as transparency by making the system more explainable |

Table 9: continued

| Application | Description | Framework user | Audience | Rationale |
|---|---|---|---|---|
| Personalized Career Paths | Indicate why specific steps on the path are recommended for this employee. What personal data of this employee is used, and how are personal preferences reflected in this career path? Consider indicating the different options, currently not displayed to the employee. | Developer | Employees | AI Act article 26.7 describing the obligation of deployers who are employers to inform their employees whenever they are subject to the use of the high-risk AI system [2]. As well as providing transparency by making the system more explainable. |
| Advancement prediction | Share which factors contributed to the prediction of advancement for this specific employee. | Developer | System user | Transparency by making the system more explainable and easier to use |
| Retention Management | Indicate what specific factors or data points have led to the prediction of an employee at risk of leaving the organization. Share what type of data was used for this prediction. For example, LinkedIn activity. | Developer | System user | Transparency by making the system more explainable and easier to use |
| | When performing actions based on the prediction of leaving employees, explain the way the AI system has influenced these actions. | Deployer | Employees | AI Act article 26.7 describing the obligation of deployers who are employers to inform their employees whenever they are subject to the use of the high-risk AI system [2]. |
| Identifying important benefits | When indicating what might have been important benefits, also disclose how accurate this prediction is and the percentage of employees this will likely be important to. For instance, the benefit of receiving a company vehicle might be a very important benefit to the ones who received one, but might not be applicable to every employee. | Developer | System user | Based on generic transparency guidelines 'Data quality' and 'System quality' |

## E.2 Human oversight guidelines

### E.2.1 Generic guidelines

Table 10: Generic human oversight guidelines

| Name | Description | Framework user | Actor | Rationale |
|---|---|---|---|---|
| Accuracy trade-off | (Informative) There is no clear indication that human oversight will increase accuracy, in fact, it is expected to decrease accuracy. Nonetheless, candidates seem to have a strong preference to a human being involved in the process, even if that means a lower accuracy. | Provider and Deployer | - | Statement about no increase in accuracy from [21], candidate preferences from the interviews. |
| Moral buffer or dereponsibilitization | (To perform oversight) Using AI-aided decision-making tools might lead to a 'moral buffer' in which the decision maker feels less responsibility for the decision outcome since it was partially made by an AI system. Train your employees to be aware of this phenomenon, so that they do not feel a lowered sense of responsibility for their decisions made with AI-aided tools. | Deployer | System user | The concept of a moral buffer as explained by [23] |
| Accountability shift | (To perform oversight) Human oversight might be used in such a way that all accountability is shifted to the human operator, while many different parties should hold accountability for the functioning and the outcomes of the AI system. A human operator should not be used as a scapegoat, nor should human oversight be used as a way to fix or legitimize using a flawed algorithm. Be wary of providers trying to escape accountability by shifting all accountability to the human operator. | Provider and Deployer | - | Accountability shift and human used as scapegoats as described by [21] |
| Human Oversight in the AI Act | (Informative and to perform oversight) The AI Act requires the provider to facilitate human oversight for the deployer. This entails building in measures or identifying measures that should be implemented by the deployer, as well as providing the deployer with the right tools and methods to understand the capacities, interpret the outcome, manage the outcome (by e.g. overruling or reversing) or halting the functioning of the AI system. Next to that the natural person performing human oversight should be made aware of automation bias. | Provider and Deployer | - | AI Act Article 14 (Human Oversight) [2] |
| Human AI interaction | (Informative) The interaction between humans and AI systems is a topic that is currently underrepresented in literature. There still is a limited understanding of the way these interactions take shape. Therefore it might sometimes be hard to predict the suitability or effectivity of certain human oversight measures implemented. | Provider | - | Taken from [21] as one of the flaws of human oversight |

Table 10: continued

| Name | Description | Framework user | Actor | Rationale |
|---|---|---|---|---|
| Design for usability | (To facilitate oversight) To properly facilitate human oversight, an AI system should be designed with usability in mind. In case of human AI interactions different things to take into account could be offering support to efficient invocation, dismissal or corrections. Offering multiple options when the outcome is unsure. Encouraging feedback and making sure it is clear the way this feedback is getting incorporated. | Developer | System user | Combination of guidelines related to system-design from the human-ai guidelines from [35] |
| Training and awareness creation | (To perform oversight) In addition to technical measures, human oversight can be more effectively implemented by natural persons who possess the appropriate skills, awareness, and knowledge. Training programs can help raise awareness by introducing individuals to the various flaws and pitfalls of human oversight. Furthermore, these programs should also provide training in understanding and using the specific AI system in question. | Provider and Deployer | System user and Supervisor | AI Act Article 26.2 about assigning HO to competent people [2] and more info about training from [24] |
| Automation bias | (Informative and to facilitate oversight) Automation bias is the tendency to over-rely on the outcomes of the AI system. This might result in accepting false positives or dismissing false negatives, while it should be the other way around. Without taking automation bias into account, human oversight is just a checkbox. To minimize automation bias, one could reduce the number of on-screen details, offer supportive information to users instead of directives, and actively ask the user to share the reasoning behind their decisions. | Developer | System user and Supervisor | Information about automation bias from [33] |

Table 10: continued

| Name | Description | Framework user | Actor | Rationale |
|---|---|---|---|---|
| Presentation bias | (Informative and to facilitate oversight) Presentation bias is the type of bias caused by the way the outcomes of a system are presented. Examples of this could be ranking bias, in which higher ranked items are more likely to be selected, or the effect that content placed near images is also attracting more attention. To counter the effects of presentation bias, one could randomize the order of the outcomes when all are equally relevant, so that there is not one specific item being shown on top of the ranking at all times. Furthermore, when using recommender systems, consider making use of a visual recommendation rather than a numerical one (e.g. displaying a number of stars instead of showing '80%'), as that might have positive effects on reducing bias. | Developer | System user and Supervisor | Information about different types of presentation bias [24] and the tips to minimize from [32] |

### E.2.2 Application specific guidelines

Table 11: Application specific human oversight guidelines

| Application | Description | Framework user | Actor | Rationale |
|---|---|---|---|---|
| Job Description Generator | (Facilitative) Ensure that the user of the application is aware of the flaws or shortcomings of the system so that they are stimulated to actively check and reread the generated text for errors. | Provider and Deployer | System user | Based on generic human oversight guideline 'Training and awareness creation' |
| Job Description Enhancer | (Facilitative) Aside from an explanation about why a change is recommended, provide the user with a few alternatives for changes as well as an option to dismiss the change. | Developer | System user | Based on generic human oversight guideline 'Design for usability' |
| Targeted Advertisements | (Performative) From time to time have someone check in on the advertisement and the way the targeting is behaving. In this way, one can oversee from a higher level whether everything is still going as expected. | Deployer | System user | Example of how human oversight can be approached for this application |

Table 11: continued

| Application | Description | Framework user | Actor | Rationale |
|---|---|---|---|---|
| Matching | (Facilitative) Make sure multiple matches are shown to both, so a recruiter gets to see multiple candidates. This allows for an active input of the recruiter, rather than just a 'checking the box' when accepting one specific recommended candidate. | Developer | System user | Based on generic human oversight guideline 'Design for usability' |
| | (Facilitative) Make sure there is a way to check the higher-level performance of all the matches. This allows checking whether the way the matching system is adapting is positive or whether there is any undesired behavior such as a very biased matching outcome. | Developer | Supervisor | Stimulate the framework user to view human oversight in different ways and not only local or only global. |
| Headhunting | (Facilitative) Make sure that the outcomes, such as the suitability of a candidate or the likelihood of leaving come with a certainty score, as well as an explanation as to why these candidates seem suitable or likely to leave. | Developer | System user | Based on generic human oversight guideline 'Design for usability' |
| CV Parsing | (Performative/Facilitative) On a higher level see whether a lot of candidates were agreeing with what their CV got parsed into. Also include a randomized trial in which a manual check is needed for a random CV to see whether everything is still working as expected. Such a randomized approach strikes a balance between still benefiting from the efficiency of the parser vs still checking the functionality | Provider and Deployer | Supervisor | Stimulate the framework user to view human oversight in different ways and not only local or only global. |
| Qualifications Chatbot | (Performative/Facilitative) Include randomized checks every once and a while to keep sight of the functioning of the chatbot. Next to that have a way of checking on a high level whether candidates generally seem to agree or disagree with the way the chatbot is handling their conversation and turning it into qualifications. | Provider and Deployer | Supervisor | Stimulate the framework user to view human oversight in different ways and not only local or only global. |
| Pre-employment assessment | (Facilitative) Display the scores of the assessments in a way that is likely to be less biased (such as using stars rather than a percentage score) | Developer | - | Based on generic human oversight guideline 'Presentation bias' |
| | (Facilitative) Have a way to see how different parts of the assessment led to different scores/outcomes. So for instance, good performance in parts A and B leads to a higher score in assessed skills C | Developer | System user | Based on generic human oversight guidelines 'Design for usability' and 'Automation bias' |

Table 11: continued

| Application | Description | Framework user | Actor | Rationale |
|---|---|---|---|---|
| Video based Interviews | (Facilitative) Make sure that the way the analysis of the video interview is presented to the recruiter includes an explanation. so that the human overseer is able to make informed decisions. | Provider and Developer | System user | Based on generic human oversight guidelines 'Design for usability' and 'Automation bias' |
| | (Facilitative) Inform the user about the different areas in which the video analysis might fall short. Examples could be a specific accent or ethnicity underrepresented in the training data. | Provider and Developer | System user | Based on generic human oversight guideline 'Design for usability' |
| Background checks | (Performative/Facilitative) Allow for a way to oversee background check performance on a higher level. So that is not on an individual level, but rather on a bigger group of people. Are there any unwanted patterns or results in the outcome of the background checks? That is, there might be a structural disadvantage to people of a certain background, or any weird outliers that are not supposed to be there | Provider and Developer | Supervisor | Stimulate the framework user to view human oversight in different ways and not only local or only global. |
| Offer generation | (Facilitative) Allow the user to change or dismiss the offer. Ways of changing the offer could for instance also be changing the importance of specific factors and regenerating, or just changing specific outputs. | Developer | System user | Based on generic human oversight guideline 'Design for usability' |
| AI Aided learning | (Facilitative) An HR employee responsible for the learning paths should be enabled to oversee the performance of the AI system. This could, for instance, entail a dashboard that displays the overall employee satisfaction of the recommended learning paths, or a bias assessment in which is assessed whether a certain level/type of job is recommended to not only the same type of people) | Provider and Developer | Supervisor | Stimulate the framework user to view human oversight in different ways and not only local or only global. |
| Skills identification | (Facilitative) Allow the user to give feedback to the system, for example when certain skills are shown that are very likely not relevant. Not only will this help in involving the user and having them (critically) assess the system, but this will help improve the functioning of the system as well. | Developer | System user | Based on generic human oversight guideline 'Design for usability' |
| Performance Analysis | (Facilitative) The outcome of the performance analysis should come with both explanations and certainty scores, so that decisions made based on these outcomes are well-informed. | Developer | System user | Based on generic human oversight guidelines 'Design for usability' and 'Automation bias' |

Table 11: continued

| **Application** | **Description** | **Framework user** | **Actor** | **Rationale** |
|---|---|---|---|---|
| Personalized Career Paths | (Facilitative) The employee to whom a personalized career path is offered, should be able to give feedback to the AI system. For example, whether certain steps on the path are in line with the wishes and ambitions of the employee. There should be an overarching place where the dedicated HR employee can oversee both this feedback and the overall functioning of the system. | Provider and Developer | Supervisor | Based on generic human oversight guideline 'Design for usability' |
| Advancement prediction | (Facilitative) Predicted advancement should come with both an explanation and a certainty score. In this way certain outcomes with for instance a lower certainty score can be further inspected when deemed necessary. | Developer | System user | Based on generic human oversight guidelines 'Design for usability' and 'Automation bias' |
| Retention Management | (Performative/Facilitative) Next to predictions of who is likely to leave, offer not just one, but multiple possible follow-up actions. In that way, the system will not steer the user into one specific direction but allows the user to make a decision themselves. | Provider and Developer | System user | Based on generic human oversight guidelines 'Design for usability' and 'Automation bias' |
| Identifying important benefits | (Performative/Facilitative) Make sure there is a way to offer feedback on the given outcomes so that there is still room for manual inputs and improvements. Such as a specific benefit that is not included in the system's outcome, but is very likely to be an important benefit to many. The system user should be able to indicate the importance of said benefit. | Provider and Developer | System user | Based on generic human oversight guideline 'Design for usability' |