Appendix A: Use case HR2 detailed findings

A.1 Fair Data Governance and Management Methodology (FDCGM)

A.1.1 Use Case and Process Summary:

This report summarizes the group sessions conducted in collaboration with a hiring company to implement and evaluate the Fair Data and Management Methodology in the AI Fairness-by-Design AEQUITAS project. A key component of this initiative focuses on data governance and management, driven by the requirements set forth in the EU AI Act, which outlines specific obligations for AI providers. These obligations include standards related to the origin, collection process, relevance, and processing of data. The methodology translates regulatory obligations into practical questions, forming the basis for multi-disciplinary, hands-on sessions that assess training, testing, and validation data for fairness. This evaluation incorporates technical, legal, ethical, and social considerations while also proposing potential mitigation measures for bias. It additionally aims for a comprehensive assessment that aligns data management practices with legal standards like the GDPR.

These sessions were organized with a hiring company which deploys engineering talent and tech expertise to clients. This organization uses an AI hiring tool to match client needs with appropriate candidates and the tool's dataset contains candidate information collected to support the hiring process. We conducted three working group sessions to evaluate the fairness of this dataset in accordance with the Fair Data Governance and Management Methodology:

- Session 1 (08/11/24): General scoping and assessment of the entire dataset.
- Session 2 (15/11/24): Focused analysis of specific data features, specifically sex (data feature G) and "candidate state" (data feature C).
- Session 3 (22/11/24): Continued the data feature analysis with "candidate state" (data feature C) and "dynamism" (MM).

Throughout each session, there was a diverse range of participants, some from the hiring company itself, but the majority came from different disciplines and fields. All participants, guided by a moderator, provided insights based on their individual expertise. The process culminated in the documentation of results, findings, and a discussion of options to avoid or mitigate bias regarding the data features, accomplished by filling out the associated Data Dictionary.

In addition to encouraging stakeholders to think about compliance and familiarize themselves with the requirements of the EU AI Act, these sessions provided ALLAI with valuable feedback on the Fair Data Governance and Management Methodology, revealing areas for improvement in both the methodology and the legal interpretation of the AI Act. The sessions also served as a practical exercise in bridging language gaps

Funded by
the European Union

and resolving interpretation challenges related to legal texts and requirements. This was important for ensuring that the diverse range of participants, both those with technical backgrounds and those without, could clearly understand and engage with the requirements.

A.1.2 Methodology

The methodology was structured into two parts, each designed to assess the dataset and its individual data features at the developmental stage of the AI system's life cycle.

Part One:

The first part consisted of a high-level analysis via the questionnaire of the entire dataset to establish a comprehensive understanding of the data being used. This step is important during the AI system's development phase, as it lays the foundation for fair data collection, governance, and design before moving on to algorithm selection and design. Analyzing the dataset at this early stage helps to identify any biases or unfairness in the data. The documentation of findings was completed in the Data Dictionary: Data Set Overview (due to time constraints, this exercise was undertaken outside of the workshop sessions by the ALLAI team).

Part Two:

The second part delved into the specific data features within the dataset, consisting of three-sub-parts: part a) participants were asked to identify the relevant information regarding the individual datapoint via the questionnaire, part b) participants were asked to identify and assess potential bias/unfairness regarding the particular datapoint, and part c) the documentation of findings in the Data Dictionary and the discussion of options to avoid/mitigate bias regarding the datapoint (due to time constraints, part C was undertaken by the ALLAI team).

To undertake the exercise, each participant received the following documents: Three worksheets, each covering Part 1 and Part 2, the Dataset, and a README file to provide context and guidance for the analysis.

A.1.3 Session 1

Attendees

- Technical Expert from UNIBO (University of Bologna)
- Social Sciences Expert from UNIBO (University of Bologna)
- Technical Expert from ITI (Instituto Tecnológico de Informática)
- Worker Representative from Eurocadres
- Socio-technical Expert from UMU (Umeå University)
- Legal and Ethical Expert from ALLAI
- Moderator from ALLAI
- Technical Decision-makers from the hiring company

A.1.3.1 Summary of Session 1

The goal of this session was to complete Part 1 the Fair Data Governance and Management Exercise. The session began by answering questions to elicit a general overview of the dataset’s characteristics, analyzing how data was collected, processed, and the implications this had on fairness.

A.1.3.2 Results

Dataset Composition

The dataset consists of candidate data collected over five years (2018-2023), primarily comprising CVs and other recruitment data from the hiring company’s historical operations. Since the dataset was built using pre-existing data, questions arose as to whether the dataset was sufficiently representative of the broader population and as to whether it adequately reflected the workforce, especially in STEM fields. In addition, the dataset included datapoints that seemed irrelevant given the system’s intended purpose. The technical team from the hiring company clarified that the primary purpose of the AI system was to assist clients in identifying suitable candidates for specific job roles, and the recruitment process relied on this dataset to match candidates with client needs. As such, data features like Sex (G) and Age Range (D) in a hiring context raised concerns about arbitrariness. This is particularly problematic under Article 10.2(d) of the EU AI Act, which mandates that any assumptions regarding the dataset and its data points must be directly relevant to what the dataset or model aims to represent and measure. The discussion ultimately highlighted that the dataset was not specifically designed or curated for the AI system’s intended purpose. Instead, it was a collection of data accumulated over several years, and it was later decided to repurpose it for training the AI system. In this way, it seemed the AI system was adapted to fit the existing dataset, rather than the dataset being tailored to meet the specific requirements of the AI system.

Data Characteristics

The dataset contained 21,000 rows, which the technical participants argued was sufficient to train, validate, and produce reliable results. However, participants such as the Social Sciences Expert from UNIBO highlighted that the "appropriate" data volume is highly domain specific. The Technical Expert from UNIBO concurred, noting that different fields, such as healthcare versus HR, might have varying data requirements, and domain-specific benchmarks could help determine the necessary dataset size.

Data Collection and Sampling

Throughout the sessions, we identified significant fairness concerns related to the data collection methods and associated sampling strategies. The dataset was compiled from various sources, including the hiring company’s internal CV database, LinkedIn profiles, university recruitment processes, and internal interview notes. This diversity in data sources raised issues about the uniformity of data collection, as the different methods could introduce biases. For example, HR staff searching for candidates on LinkedIn might have access to more detailed candidate profiles than what appeared on CVs,

Funded by
the European Union

leading to inconsistent evaluations between candidates. Furthermore, certain candidates had additional data features, such as communication (KK), maturity (LL), dynamism (MM), and mobility (NN), which were not uniformly applied across the dataset. These extra features posed several problems. Firstly, they were assessed through standardized interview questions that relied heavily on the interviewer's subjective judgment, particularly for traits like "maturity", "dynamism" and "communication." Additionally, these skills were rated on a 1-4 scale, while other skills in the dataset used a 1-3 scale. This inconsistency in rating systems is problematic because it can lead to unequal weightings of skills, potentially skewing the AI's evaluation of candidates.

Finally, the sessions revealed methodological issues in candidate data collection, characterized by inconsistent data density and sampling approaches. Mandatory platforms, such as CV upload systems, required specific categorical responses including details about gender, region, and age. In contrast, platforms like LinkedIn allowed candidates more discretion in providing information, where candidates freely entered information about their professional skills rather than selected from a predefined list.

Participants highlighted that the optional nature of certain fields created inconsistencies in the data density for each candidate. The Socio-technical Expert from UMU pointed out that optional fields could result in missing values, as candidates could easily skip non-mandatory questions. The Technical Decision-Makers from the hiring company clarified that while some fields (like age and skills) were mandatory, others were optional, leading to variability in candidate profiles.

Another problematic factor was the variation in what was considered "mandatory" based on a candidate's nationality. For Italian candidates, data on town, province, and region were collected, while for international candidates, only the country of origin was recorded. This raised concerns about fairness and potential discrimination, as geographic proximity could influence candidate selection, leading to biases based on location. It was also noted that excluding candidates based on location is problematic because some might be willing to relocate.

The non-uniform data collection methods, along with the resulting data gaps and inconsistencies in candidate data density, raised fairness concerns. Candidates with more comprehensive profiles could inadvertently receive preferential treatment, as their richer data allows for more accurate matching to client requirements. This inconsistency in data collection risks introducing biases into the AI system, potentially disadvantaging candidates with less complete profiles and undermining the fairness of the recruitment process.

Statistical Properties and Representativeness:

Throughout the session, there was debate surrounding the topic of demographic representation and the statistical properties of the dataset. This discussion was driven by the ambiguity in the EU AI Act (Articles 10.2.f, 3, and 10.4), which requires AI system providers to consider the "appropriate statistical properties" of the dataset, taking into account the individuals or groups the high-risk AI system is intended to serve. However,

Funded by
the European Union

the legal term “appropriate statistical properties” remained unclear, leading to a broader conversation on what constitutes the right level of demographic representation needed to mitigate bias.

A central tension emerged regarding the inclusion of protected categories (such as race, gender, etc.) in the dataset. Some participants argued that these categories should be included, as doing so would enable organizations and clients to actively target underrepresented groups, thereby helping to meet diversity goals and address societal inequalities. Conversely, other participants felt that including protected categories in hiring contexts could lead to biased associations within the AI system, such as linking certain genders to specific professions, thus reinforcing stereotypes and potentially generating biased hiring outcomes.

This debate expanded to address the issue of systemic underrepresentation in the dataset, which reflected societal biases. For instance, there were significantly fewer female candidates than male candidates. The challenge was how to address this bias while acknowledging that this imbalance mirrored the real-world underrepresentation of women in STEM fields. The dilemma, therefore, was how to balance statistical "accuracy" (ensuring the dataset mirrored the accurate gender (im)balance in the field0 with the need for representational fairness in AI-driven hiring processes. Several strategies emerged during the discussion. The Social Sciences Expert from UNIBO proposed using census data and demographic benchmarks to provide a statistical foundation for evaluating and potentially correcting the systemic disparities in the dataset. The Legal and Ethical Expert from ALLAI suggested excluding all protected characteristics to avoid reinforcing biases, while the Technical Decision-Maker from the hiring company recommended omitting gender-specific data features, a proposal supported by the hiring company’s HR team. However, there was push-back on these suggestions of ‘fairness by unawareness’ due to the deletion potentially reinforcing bias by making developers lose access to measuring variables of bias.

Ultimately, the participants agreed that addressing representational issues in the dataset depends heavily on the AI system’s intended purpose. However, for a hiring tool designed for the general working population, the system should aim to promote equality and fairness. Simply mirroring current labor market demographics could entrench historical inequities, reinforcing existing biases. Therefore, the value of a dataset for a hiring context should not lie in reflecting current demographic realities but in addressing and mitigating systemic biases in order to ensure AI systems are designed in a way that does not perpetuate or exacerbate structural inequalities in professional environments.

Error Handling:

This session additionally provided insight into the technical approaches to data management, specifically focusing on error handling and data quality assessment within the dataset. A key observation was that technical participants demonstrated a prevalent perspective characterized by data maximization (making the most of the available data) rather than data optimization (choosing the right data for the AI’s intended purpose). Indeed, the Technical Expert from UNIBO revealed that the typical workflow involved

Funded by
the European Union

working with available data, including potential errors, with comprehensive error detection and outlier analysis typically occurring only at the final stages of data analysis. We found this approach revealed a broader tendency to equate dataset size with dataset quality, with participants often advocating for larger volumes of data (albeit with potentially more errors) under the assumption that more data improves AI system performance. However, a closer examination of the dataset highlighted multiple qualitative issues that challenged this quantity-focused perspective. For instance, numerous duplicate entries were found, undermining the perceived value of dataset volume. These duplications not only inflated the dataset size but also risked introducing systematic biases into the model. In turn, while the dataset contained a large volume of data, it was found that some data features such as sex were not evenly distributed relative to the broader population.

A.1.4 Session 2

Summary of Session 2:

The goal of this session was to engage in Part 2 of the Data Governance and Management Methodology. The moderator guided the group through each question in Part 2 of the Fair Data Governance and Management Exercise. In Part A, the first data feature examined was sex (Data Feature G). Once the discussion around sex was completed, the moderator moved on to Part B to examine the second data feature, "candidate state" (Data Feature C). Each participant contributed their insights, allowing for a comprehensive evaluation of the chosen features. However, due to time constraints, Part B could not be completed during the session and was carried over to session 3.

Attendees:

- Management Decision-Maker from the hiring company
- Technical Expert from Philips
- Technical Expert from UNIBO (University of Bologna)
- Economist from ULL (Universidad de la Laguna)
- Legal and Ethical Expert from ALLAI
- Moderator from ALLAI
- Socio-technical Expert from UMU (Umeå University)
- Technical Expert from UMU (Umeå University)

A.1.4.1 Results

Meta-Methodological

Delving into the questionnaire for the data features revealed that many questions were not appropriate at the individual feature level and should instead be considered at the dataset level. For example, Article 10.4 of the EU AI Act stipulates that “data sets shall consider characteristics or elements particular to the specific geographical, contextual, behavioral, or functional settings within which the high-risk AI system is intended to be used.” Initially, we attempted to evaluate these requirements at the feature level, but this approach proved problematic. Participants found the terminology ambiguous, particularly

Funded by
the European Union

since the legal text does not provide clear guidance on how to translate these concepts into technical terms. Furthermore, evaluating these considerations at the feature level proved to be overly granular. The suggestion was made to shift these evaluations to the dataset level, where they could be more effectively addressed in the context of the broader system.

Part A: Sex as a Data Feature

Relevance:

When discussing the relevance of "sex" as a data feature, participants expressed conflicting viewpoints and highlighted key trade-offs, echoing the debates from previous sessions. Some participants argued that retaining sex in the dataset was essential for developers to measure and mitigate bias, emphasizing its utility as a diagnostic variable. Conversely, others maintained that sex was irrelevant in the hiring context and advocated for its omission. They supported approaches like "fairness through unawareness," where the algorithm is designed to ignore protected characteristics entirely. However, the Technical Expert from UNIBO noted that removing sex as a feature could decrease both accuracy and fairness, highlighting the importance of understanding the potential trade-offs of data-processing decisions.

Ultimately, participants found that any assumptions or information which were supposed to be represented or formulated by including gender would prove to be unfair, rendering it inappropriate as a data feature for the hiring tool. No unanimous agreement was reached on whether gender always disadvantages fairness, yet participants acknowledged that its role is highly context specific. The core takeaway was the importance of aligning the dataset with the AI system's intended purpose, ensuring that no extraneous or arbitrary features are included that could inadvertently introduce bias.

Scale:

A problematic aspect of the dataset emerged when examining the scale of the sex feature. The dataset's gender classification was limited to a binary male/female categorization, which we found to be a reductive approach. By restricting gender to these two categories, the dataset inherently introduces bias, excluding candidates who do not conform to traditional gender classifications. When discussing whether the measurement scale of gender could introduce bias or unfairness, technical participants emphasized that the purpose of the AI system plays a key role in defining and measuring fairness in relation to a feature like gender. Some participants argued that if the system is intended solely to analyze male/female differences in a hiring context, this binary representation might be acceptable for that specific purpose. However, if the system is expected to capture a broader range of gender identities, limiting the dataset to just male and female categories would introduce bias.

Volume:

We found the volume of candidates specific to the data feature of sex comprised 22% females and 78% males. Participants agreed this unequal distribution could lead to biases that impact the system's fairness in the hiring context.

Type of Data Feature:

The participants acknowledged that "sex," as a data feature, is classified as a special category of personal data under the GDPR. Moreover, they recognized that sex falls under the prohibited grounds for discrimination as outlined in Article 21 of the Charter of Fundamental Rights of the European Union (ECFR). In this context, the inclusion of sex as a data feature poses significant concerns, as it could undermine equal hiring opportunities for men and women. Unlike proxy variables, which might indirectly enable discrimination, gender is inherently a directly discriminatory variable.

Statistical Properties and Representativeness:

This session revisited many similar themes from the previous discussion around statistical properties and representativeness, albeit with different participants continuing to express differing views on how representational fairness should be defined. Some argued that the dataset should reflect existing population distributions as closely as possible, seeing the current 22/78% sex split as a statistically accurate representation of the contemporary labor market. Others disagreed, asserting that this approach would merely perpetuate historical discriminatory patterns entrenched in existing social structures. Some participants stressed the need for maintaining accurate population reference points to ensure statistical validity, while others advocated for proactively creating more balanced distributions, such as aiming for 50/50 or 40/60 gender ratios. This led to further discussions about the potential use of synthetic data to balance the dataset, with the consensus being that its efficacy would depend on the specific context and fairness criteria applied. The Technical Expert from Philips raised the salient point that the dataset is primarily Italian and that there could be difficulties when transferring the learned model to other parts of Europe where male/female labor distribution could be different (issue of distribution shift in the context of transfer learning). As such, this session's participants emphasized that for the purpose of equality of opportunity, the distribution should be shifted to 50/50 even if it is not an accurate reflection of 'reality'.

Part B: Candidate State as a Data Feature

Relevance:

All participants agreed that "candidate state" was a relevant feature for the hiring tool, as it measured and represented whether a candidate was hired or not which seemed necessary and proportionate to the intended purpose. Moreover, it was established that "candidate state" was not only relevant but also the target variable driving the AI system's algorithm. As the target variable, it aimed to classify this outcome out of all the potential candidates and the client's needs (not predict). From a legal perspective, the choice of target variable is of utmost importance because it defines the 'intended purpose' of an AI system, which is essential for determining its classification and corresponding obligations under the AIA. However, the designation of "candidate state" as the target

Funded by
the European Union

variable raised some concerns. The dataset employed a binary classification system, categorizing candidates as either "hired" (marked as 1) or "not hired" (marked as 0). This approach introduced a limitation whereby candidates who were still in the hiring process, under active consideration, or awaiting final decisions were categorized into the "not hired" category. This oversimplified classification introduced inaccuracies into the dataset, especially since the value attributed to the candidate has not been updated since the data was sampled.

A.1.5 Session 3

A.1.5.1 Summary of Session 3

The goal of this session was to complete Part 2 the Fair Data Governance and Management Exercise. The session started as a continuation of Part B, examining the second data feature of "candidate state" (Data Feature C). Upon completion, the moderator moved on to examine "dynamism" (Data Feature MM).

Attendees:

- Technical Expert from UPV (Universitat Politècnica de València)
- Gender Equality Advocate from Period Think Tank APS
- Technical Expert from UNIBO (University of Bologna)
- Technical Expert from UNIBO (University of Bologna)
- Technical Expert from UNIBO (University of Bologna)
- Socio-technical Expert from UMU (Umeå University)
- Socio-technical Expert from UMU (Umeå University)
- Legal and Ethical Expert from ALLAI

A.1.5.2 Results

Part B (continued: Candidate State as a Data Feature)

Relevance:

While participants agreed that "candidate state" is a proportionate and necessary data feature for the hiring tool, some participants raised concerns about its potential misuse as a proxy for discrimination. Specifically, "candidate state" could inadvertently influence judgment by revealing historical patterns in a candidate’s hiring processes that might bias decision-making. For example, if a candidate had not progressed in this hiring process or in previous hiring processes, this data could disadvantage the candidate by indicating reduced employment potential. In addition, participants raised concerns as to whether “candidate state” could act as a proxy that may compromise dataset anonymity. Participants such as the Technical Expert from UNIBO and the Socio-technical Expert from UMU claimed that performing statistics and correlating “candidate state” as a data feature with other demographic data features (such as nationality, gender, etc.) could potentially be-anonymize the dataset. In this session, participants such as the Gender Equality Advocate from Period Think Tank APS suggested that “candidate state” should not be used as a target variable due to its interaction with other data features. Indeed,

Funded by
the European Union

they claimed using it could lead the system to learn and reinforce biased traits linked to specific candidate profiles, potentially skewing the hiring process.

Volume:

Participants such as a Technical Expert from UNIBO claimed that when it comes to data, more is better, especially given the high number of features in a dataset. With 40 data features, many of which are categorical or serve as containers for other features, the dataset has a high dimensionality. Despite this, the dataset itself is only medium-sized, making a larger data volume necessary to ensure accurate results.

Statistical Properties and Representativeness:

Participants acknowledged that the distribution of hires in this context was uneven, which could pose fairness concerns. For instance, examining the "hired" category revealed that the majority of successful candidates were Italian, while only a small number of Brazilian, Egyptian, or Moroccan candidates were hired. If certain nationalities, such as Moroccans, are almost absent from the "hired" category in the dataset, it could implicitly suggest that candidates from these backgrounds are not suitable for hiring. In this way, we see how uneven categorical distribution, even on the basis of nationality, could disadvantage certain groups, reinforcing biased outcomes. In a broader sense, participants agreed that “candidate state” was not ‘measuring’ anything but rather is a categorical variable that represents the candidate's hiring status within the system.

Part B: Dynamism as a Data Feature

Relevance:

Dynamism proved to be a challenging data feature to analyze due to the differing interpretations between the technical members of the hiring company’s team in Session 1 and the participants in Session 3. In Session 1, participants described dynamism (along with traits like maturity, mobility, and communication) as being assessed through standardized interview questions. These questions relied on the interviewer’s subjective judgment, with skills rated on a 1-4 scale, implying the candidate had already progressed to the interview stage. In contrast, participants in Session 3, such as the Socio-technical Expert from UMU, viewed dynamism as a trait or skill typically assessed via a questionnaire to evaluate how candidates perform outside of work experience. Under their interpretation, it was unclear when this data was collected during the hiring process. However, it was emphasized that if the hiring tool is primarily focused on skill matching and competence for clients, that these traits (such as maturity and dynamism) become highly relevant. Indeed, they elucidated how these traits are often tied to personality assessments and are used as ‘matching variables,’ with the competence profile for a job specifying the required level of dynamism, for example, to match the right candidate to the role. A concern emerged regarding the subjectivity of soft skill requirements. While requiring technical competencies, for example, a certain programming language, represent relatively objective, measurable skills, soft skill assessments introduce significant potential for unconscious bias because there could be bias on what soft skills are required for the job.

Funded by
the European Union

Participants raised a concern similar to the issues identified with using sex as a data feature. The problem with using a trait like “dynamism”, which is subjective as a data feature, is that the AI system may learn and perpetuate the historical biases associated with it. For example, if the system learns that candidates with a dynamism score of 4 are more likely to be hired, this could introduce bias, reinforcing discriminatory patterns. Additionally, participants pointed out the issue with clients specifying requirements such as a score of 4 for dynamism or mobility. This approach is inherently discriminatory, as it excludes candidates with lower scores for seemingly arbitrary reasons, limiting opportunities based on subjective criteria.

When asked whether “dynamism” could serve as a proxy for bias, participants such as the Socio-technical Expert from UMU argued that if 'dynamism' scores were obtained in a biased manner, then it could indeed become a proxy for bias. This point was reinforced by participants like the Gender Equality Advocate from Period Think Tank APS, who highlighted the inherent problems with data features like “dynamism”, noting that such traits can be subject to stereotyping. For instance, there could be biased associations, such as the stereotype that men are more 'dynamic' than women. This raised the socio-technical issue about how societal bias can seep in through the subjective assessment of candidates based on these traits, rather than through the AI system’s processing of the data itself. Ultimately, however, the participants unanimously agreed that, under the hypothetical condition where the data on a candidate's “dynamism” is accurately collected and processed, “dynamism” as a data feature would not pose a fairness issue for the AI system

Statistical Properties and Representativeness:

Participants also identified “dynamism” as problematic in the dataset due to the presence of missing values, with only a few candidates having a recorded “dynamism” score. This reinforced the session 1 concerns about discrepancies in the data sampling methodology, highlighting that the dataset was collected from a diverse group of candidates at different stages of the hiring process and at varying points in time.

A.1.6 Findings

In taking on positions as both moderators and participants in the Fair Data Governance and Management exercise sessions, ALLAI was able to identify key overarching themes throughout these sessions. In addition, ALLAI’s focus on the legal, societal, and ethical perspectives also allowed us to identify some points which were not always elicited throughout the sessions.

A.1.6.1 Key Themes Throughout:

Throughout each session, it was challenging to pinpoint the exact intended purpose of the AI hiring tool. Despite the hiring company’s initial statements about the system being designed to help clients identify suitable candidates for specific roles, the operational mechanisms of the system (at a ‘model’ and algorithmic level) remained unclear as there was a lack of unified understanding among technical participants. Each session identified potential operational frameworks for the AI hiring tool, each carrying distinct methodological and fairness implications. For example, it was uncertain whether the AI system would function as a learning recommender system that learns from data to suggest the best candidates. Key fairness concerns arose as to whether this type of model could introduce biased feedback loops of learning and selection. An alternative conceptualization was that the AI system would be a simple matchmaking model aligning candidates with client needs, or a hybrid matchmaking and ranking model that assigns a ranked score to each recommended candidate. The absence of clarity concerning a defined technical model and purpose compromised the ability to undertake a comprehensive fairness assessment at both the data feature and dataset level because considerations are dependent on the context of the AI system's function and operation.

Secondly, confusion regarding the Statistical Properties and Representativeness required for fairness was persistent throughout each session. We acknowledged the significance of this issue due to its inclusion within the AI Act and we agreed that further clarification should be sought from the AI Office. We seek to undertake this step as doing so would help resolve ambiguities in the legal text and enhance how these concepts are addressed within the Fair Data Management and Collection Methodology.

Thirdly, throughout the sessions we identified an on-going meta-narrative with political undertones underlying the technical discussions. Debates around trade-offs (such as whether to prioritize 'accurate' demographic representativeness or include certain protected data features) highlighted the fact that decisions about collection and management of data are not merely technical but political, often carrying societal implications. These discussions underscored that a dataset is not a neutral collection of information but instead represents historical, societal, and cultural biases. In this way, it became evident that both the dataset and the AI hiring tool could perpetuate and even amplify these societal discriminatory patterns, highlighting the need to scrutinize and address the underlying biases inherent in the data.

While participants demonstrated an awareness of how datasets could reflect societal biases, ALLAI identified a blind spot regarding the potential for bias in data features beyond the traditionally recognized protected categories (such as gender, race, and ethnicity). For example, none of the participants questioned the use of personality assessments (whether through questionnaires or interviews) as a tool for evaluating traits like "dynamism" or "communication." In fact, in Session 3, it was even concluded that if these scores were hypothetically collected under "perfect" conditions, they would present no fairness concerns. However, there was a failure to recognize that the concept of "dynamism" itself is problematic, as it attempts to transform subjective human characteristics into seemingly objective assessment criteria. This overlooks how discrimination can be embedded in the process of quantifying inherently qualitative traits.

Historically, personality and trait assessments have been used in ways that perpetuate systemic racism, sexism, and classism. These assessments often have culturally biased roots, as they are frequently based on norms established by majority populations thereby marginalizing those from different cultural or social backgrounds. As such, this oversight underscores the need to examine all data features especially when such criteria are drawn from historically biased constructs of personality and competence.

Finally, after examining the potential biases inherent in various data features and their impact on fairness, it appeared that no data feature could be entirely free from bias, raising the possibility that the dataset might not be suitable for use in an AI-driven hiring tool. This conclusion was met with apprehension and push-back by technical participants. Some cautioned against the impulse to eliminate every potentially problematic feature, warning that doing so could result in a dataset devoid of any usable data. Other participants claimed bias was going to be inevitable because if every data feature is viewed as interpretable through the lens of human judgment, then every feature could potentially be biased. Others argued that consigning certain data features as biased should not be interpreted as proof that everything is biased.

Instead, it reflects the uncertainty surrounding the specific use case. While these points highlighted the complexities of data governance and management, especially from a technical perspective, the Legal and Ethical Expert from ALLAI offered a contrasting view. They asserted that if thorough evaluation of each data feature using fair data governance principles revealed persistent fairness issues, then the resulting empty dataset must be accepted as the necessary outcome. In such cases, it might be necessary to conclude that AI systems are simply unsuitable for certain applications, particularly when the risks to fairness are too great to justify their use (especially in the legal context mandated by the AIA). In this way, we wish to emphasize that the current way of working the dataset to fit the AI system is untenable, and rather, that both the AI system and dataset should be adapted to fit legal, societal, and contextual needs.

A.1.7 Areas for Improvement:

Several recommendations emerged from these sessions. First, there is a need for greater clarity in translating complex legal concepts from the EU AI Act into practical and implementable principles. Indeed, many participants lacked the legal expertise to interpret these mandates, underscoring the importance of making legal language more interpretable to both technical and non-technical stakeholders.

The exercises also revealed a knowledge gap between technical and non-technical participants, making it difficult to sustain productive interdisciplinary dialogue. To bridge this gap, stakeholders should partake in the development of shared resources such as data dictionaries to ensure mutual understanding of key concepts.

Finally, we recommend that stakeholders must be prepared to critically evaluate every data feature for potential bias and prioritize fairness over technical convenience, as required by the EU AI Act. The prevailing industry practice of trying to make datasets fit AI systems must be reconsidered. Instead, the focus should shift toward adapting both the AI system and the dataset to address societal, legal, and contextual requirements, ensuring that fairness remains central to AI development.

Funded by
the European Union