Towards Automated Context-Aware Bias Assessment

Fabian Linde[0009−0004−0856−8036]

Universidad Politécnica de Madrid, Madrid, Spain
f.linde@upm.es

Abstract. Algorithmic bias poses a major challenge to trustworthy and lawful AI, as required by frameworks such as the EU AI Act. While existing automated bias assessment tools provide statistical measures of bias in datasets, they neglect the socio-technical context in which AI systems operate. This abstraction from real-world use undermines responsible automation and complicates compliance with legal and ethical requirements. This research proposes an approach to integrate contextual information into automated bias assessments through a socio-technical, ontology-based framework. First, it conceptualizes and operationalizes the notion of "context" for bias assessment. Second, it models the information exchanges between AI providers and data holders—formalized through ontology engineering and aligned with standards like ISO 42001 and the Data Spaces Support Centre (DSSC) architecture. Third, it develops a proof-of-concept automation capable of selecting and executing context-appropriate bias metrics on datasets within data spaces while preserving data sovereignty. This research aims to bridge the gap between technical and conceptual approaches to bias by enabling context-aware, legally aligned, and interoperable bias assessment tools for responsible AI governance.

Keywords: Bias Assessment · AI Ethics · Ontology Engineering · EU AI Act.

1 Introduction

AI systems increasingly influence decisions impacting individuals and societies. However, algorithmic bias risks discrimination and unfairness within data and models. Soft law approaches to AI, like AI Ethics principles, highlight bias and fairness as vital for trustworthy AI [8], while regulations such as the EU AI Act[1] mandate bias assessments. However, current automated methods (e.g. [21,4,3]) often treat bias as a purely technical issue, focusing on statistical measures without considering the socio-technical context. This oversight prevents responsible automation and compliance with regulations such as the EU AI Act, which requires datasets to consider specific settings and examine impacts on health, safety, rights, and discrimination (Art. 10).

[1] Regulation (EU) 2024/1689

Conceptual accounts highlight the role of context in bias assessment [2,18,19], yet have not been translated into operational tools for automated context-aware bias assessments - research is needed to adapt these assessments to include context. Developing a socio-technical system[2] requires bridging technical and conceptual limitations through an interdisciplinary approach. This paper outlines my proposed research approach: operationalizing a definition of context, formalizing necessary information flows using ontological engineering, and developing a proof-of-concept for context-aware bias assessment automation. This system will aim to facilitate interaction between the AI provider, acquiring data for AI system development, and the data holder, who owns the relevant data. Context, provided by the AI provider, will inform the bias assessment on the data holder's dataset.

Common European Data Spaces enter the picture as a promising architectural context for automation tools, serving as a cornerstone of the European Commission's data strategy, aimed at facilitating future data exchanges. Therefore, tools for data processing should align with data spaces architecture. The Data Spaces Support Center's (DSSC) building blocks provide an environment for tools like automated bias assessment, which are considered value-added services within the DSSC's technical blueprint.[3]

This paper outlines a proposal for developing a context-aware automated bias assessment tool for datasets in data spaces, detailing research questions (section 2), current state of the art (section 3), preliminary ideas (4), methodology (5), and foreseen contributions (section 6).

Illustratory example

Company *ExampleHire* wants to develop a new AI-powered hiring tool, and is therefore searching for a dataset to train their model on. They find a promising asset in a data space catalogue offered by data holder *DataBank*, but, as providers of a high-risk use case under the AI Act, *ExampleHire* are required to assess the bias of their training dataset. They do not want to purchase the data only for potentially finding it is too biased, however, and *DataBank* do not want to share their data for free. Therefore, an automated tool located at *ExampleHire*'s data space connector could exchange information about the foreseen AI tool context with *ExampleHire*, perform an automated bias assessment on *DataBank*'s data, and subsequently share the result with *ExampleHire*.

[2] In this paper, I employ the notion of *socio-technical* quite loosely, without yet incorporating elements of socio-technical theory (as initiated by Trist and Bamforth [25]

[3] https://dssc.eu/space/BVE2/1071255173/Value-Creation+Services

2 Research Question

Primary research question RQ*: *How can the automated bias assessment of datasets incorporate the socio-technical context of the AI application built upon assessed dataset?*

This central research question will be addressed through four derivative secondary research questions:

RQ1: *What is the relevant "*context*" of an AI system for bias assessment?*

RQ2: *What* information *needs to flow between which* actors *to enable context-sensitive bias assessment, and how can it be formalized?*

RQ3: *How can a context-aware bias assessment process be* automated*, utilizing the formally represented information interchanges?*

RQ4: *How can context-aware bias assessments be* evaluated*?*

3 State of the art - problems and solutions

So far, I have identified 4 research problem (RP) clusters that are pertinent to my foreseen area of research.

RP1: Sustained research effort is being directed towards the area of *compliance automation* across sectors, including compliance automation in the context of AI and data applications. As each sector faces distinct regulatory requirements, compliance automation research focuses on different aspects of compliance automation in different sectors. For instance, compliance automation in healthcare aims at ensuring the accuracy and especially privacy of healthcare data [5], while in the construction industry, connected services allow for the verification of entire buildings against regulatory requirements [1]. In the context of AI, automated compliance encompasses approaches to compliance with both legal and ethical requirements. In the legal context, a large part of the research work examines compliance with the EU AI Act, due to it being one of the first hard laws regulating AI in the world, as well as due to its extensive set of requirements for high-risk AI systems, increasing compliance costs.

Several recent approaches have been proposed to address the problem of compliance automation, specifically for AI and the EU AI Act. Automation of requirements on documentation and record-keeping (EU AI Act Articles 11 & 12) is introduced through AI Cards [9] or Compliance Cards [14], by capturing metadata about AI systems as well as their risk documentation in a machine-readable manner using Semantic Web and ontology methods in order to automate analysis. For requirements on conducting Fundamental Rights Impact Assessment (FRIA), it has been suggested to automate them utilizing established Business Process Compliance approaches [15], or by developing tools supporting various steps, such as determining FRIA necessity, information gathering or notification of authorities [16]. More generally, beyond compliance with the EU AI Act specifically, semi-automate software models supporting AI ethics compliance have been outlined [7].

RP2: The second problem relevant research problem identified concerns the *automation of bias assessment*. This relates to and constitutes a more specific instance of the previous, more general problem of automated compliance, as the assessment and mitigation of bias is demanded in both hard (e.g. EU AI Act) and soft law (such as various AI Ethics principles [8]) regulations. Automatically assessing the bias of an AI system and/or its training dataset can therefore greatly support compliance efforts while saving cost and time.

To address this problem, multiple libraries exist that incorporate numerous statistical tools to calculate various bias measures on datasets, such as AI Fairness 360 (AIF360) [3], Fairlearn [4] or Aequitas [21,12], among others. All of them include algorithms for various metrics of bias assessment on datasets, allowing for a host of different values associated with bias and fairness for a single dataset. Beyond the numerical assessment of bias, resources have been developed to formalize information about bias in AI - most notably, the Doc-BiasO ontology, aimed at incorporating terminologies and relationships of biases and relevant metrics in the Trustworthy AI field.

RP3: The third problem commonly identified in research on fairness and bias in AI is the selection of the *appropriate bias metric based on the context* [18,19,2]. As presented above, there exist numerous measures of statistical bias in datasets - however, only a subset of available metrics are suitable to a specific use case in a specific context. Selecting the appropriate metrics given the context and normative constraints remains a relevant problem that requires an interdisciplinary approach between computer scientists, lawyers, philosophers, and beyond.

To my knowledge, there are not yet any solutions in existence that automatically incorporate knowledge about an AI system's context into the assessment of bias of its training dataset. Frameworks have been proposed to formalize the selection of the appropriate bias metric depending on the AI use case context, specifically for classification [19] or more generally [2]. However, as of today, no system has been developed that automates this selection. This is where my proposed research connects to: the development of a system selecting the suitable bias metric based on context input and executing the measurement through existing libraries, while ensuring data sovereignty through the appropriate integration into data spaces.

RP4: Finally, the fourth pertinent research problem concerns *data sovereignty*. While lacking a general consensus on its precise meaning, it generally involves some notion of an entity maintaining control or ownership over its data [23]. Recent literature reviews reveal a large volume of research work in this area [11], and due to the ongoing nature of both the expansion of data-intensive technologies as well as privacy debates, it can be safely assumed that this topic maintains its relevancy.

Proposed approaches to data sovereignty can be roughly divided into three aspects: governance-organisational, legal-political, and technical approaches. Expanding on the state of the art of each of these aspects would expand beyond the scope of this section. However, I wanted to highlight the initiative of Common

European Data Spaces, designed to address all three aspects to data sovereignty [22]. The European Commission funds the Data Spaces Support Centre to contribute towards the creation of these data spaces, publishing blueprints on business, organisational and technical building blocks. Data sovereignty is to be ensured inter alia through the enforcement of access and usage policies. Applied to the overall research proposal, such policies will also factor in safeguarding that any automated bias assessment occurring on a dataset in the data space will not subvert the data holder's control over it.

4 Preliminary ideas and proposed approach

Regarding each of the research sub-questions presented in section 2, I will now present the corresponding research objective and envisioned approach.

RQ1: *What is the relevant "*context*" of an AI system for bias assessment?*

When aiming for an automation of *context*-aware bias assessment, it is naturally important what the context in question includes. Answering this question leads to our the first research objective (RO):

RO1: *Conceptualize, define and operationalize the notion of "*context*" in bias assessment.*

In general, the "context" of an AI system in the setting of bias assessment should be understood to include any attribute of the system that influences its bias, such as the architecture of the model, or its intended scope and purpose.

While datasets, by virtue of the nature of the data they contain, are already situated in some context, they might be utilized to develop AI systems with different application areas. For example, a dataset containing annotations to images (such as COCO[13]) might be used to train an object detection algorithm for safety cameras, or an image captioning tool for accessibility. As individual datasets might possess bias along multiple dimensions[26], despite using the same dataset, these systems can exhibit different biases (for instance, unequal detection of persons with different clothing styles in the first case, or unequal accuracy of captioning scenes within different cultural settings). These different application areas are hence part of the (bias-forming) *context* of the AI system.

This explanation and short example should serve only as a very early first working definition of "context" to illustrate the notion. To achieve the first research objective of carving out the adequate definition, I propose the following approach. First, different general approaches to the notion of "context" should be examined and compared in order to arrive at a operational definition of context that can be utilized in the following stages. Perspectives to be included are from computer science, sociology, law, and philosophy.

It will be both especially important and presumably difficult to differentiate between the context of the AI application as determined by the provider and as determined by the deployer (both in the sense of the EU AI Act Art. (3)). As the overall approach to the context-aware bias assessment will be developed in the context of data spaces based on the interaction between AI provider and

data holder, the context of a potentially separate AI developer further down the value chain complicates things. Therefore, it will be essential to clearly delineate between these two areas of context and explicitly present the limits of the research scope.

RQ2: *What* information *needs to flow between which* actors *to enable context-sensitive bias assessment, and how can it be formalized?*

With an operational definition of context from RO2, the work on a framework to support the automated context-aware bias assessment can be started in this stage. In order to make the bias assessment context-aware, we need to combine information from the AI provider about the context with information about the dataset that the data holder owns. As we assume that AI provider and data holder are two distinct entities,[4] some model of the information exchange between them needs to be established. Therefrom derives the second research question:

RO2: *Identify relevant* actors *and* information *and develop a formal model of interchanges.*

Data Spaces provide a promising paradigm for establishing the relevant connections between actors, as they are inherently designed to bring together parties that own and parties that are in need of data. Blueprints developed by the Data Spaces Support Centre (DSSC) will be taken into account to enable interoperability. Another measure towards interoperability is the harmonization with actors and roles provided in relevant ISO standards (such as ISO 42001 on AI management systems or ISO 24027 on bias in AI). As a principle overall purpose of this work is to support AI providers in their compliance with the AI Act (specifically Art. 10(2)), its provisions (such as the connection of bias assessment to harms on health, safety, fundamental rights or discrimination) will be incorporated when mapping the necessary information flows.

To this end, existing ontologies for the AI Act (such as for the FRIA citerintamaki2024developing or the risk representation [10]) and for bias in AI [20] will be analyzed and potentially incorporated to finally obtain a formal model that builds upon existing resources to support interoperability and standardization.

RQ3: *How can a context-aware bias assessment process be* automated*, utilizing the formally represented information interchanges?*

Building on the formal model of actors and their necessary exchange of information, we can turn towards developing the prototype of an automation algorithm, and hence the third research objective:

RO3: *Develop a Proof-of-Concept for an automation solution that enables context-aware bias assessment between data holder and AI provider as data space participants.*

Fortunately, as we have seen above, there exist already a variety of software tools to examine the statistical bias present in datasets; combined, these libraries support the computation of dozens of bias metrics. The central task of the au-

[4] The considerations should also largely apply for the case where AI provider and data holder are a single entity. However, the focus lies on two distinct entities as this is the more general case.

tomation tool to be developed therefore revolves around selecting and applying the most appropriate of these metrics on the dataset owned by the data holder, given the context supplied by the AI provider. At the current stage, the specific implementation details remain to be explored, in order to find an architecture that is most suited to this task. Overall, the idea is to situate the solution as part of the data holder's data space connector, in order enable their sovereignty over their data.

The application could involve some agentic capabilities, a term broadly used and contemporarily considered to involve systems combining large language models (LLMs) with external tools for reasoning and planning [6,28]. In the context of automated bias assessment, agentic systems could process input from AI providers, automate the selection of existing assessment tools, execute the tools on datasets, and return results. Resources from RO2 may enhance the agentic system or its LLM through retrieval-augmented generation using ontologies [24].

Note that, as input from and interaction with the AI provider is required, the tool itself might not truly be classified as *automatic*. Depending on the final architecture and common definitions of the concepts, it might finally be closer to a *semi-* or *partially automated* approach.

RQ4: *How can context-aware bias assessments be* evaluated*?*

Finally, some form of evaluation is needed to assess the adequacy of the outcomes of RQs 1-3. Therefore, the final research objective is to:

RO4: *Develop an approach for* evaluation.

For RO1, this will be particularly difficult, as this work will take place on a conceptual level. Perhaps the most tangible outcome is the operationalization of the notion of context, and the results of RO1 can only be measured against their utility for later ROs. More specific evaluation methods will be explored, but for now, this remains an open question.

Evaluation for RO2 is a more specific and established practice, with common evaluation resources and metrics available for ontological resources. Tools like the Ontology Pitfall Scanner [17] automatically identify common issues in ontology engineering. Additionally, competency questions and SPARQL queries are used to test an ontology's fulfillment of its requirements [29,27].

For RO3, evaluation remains open. Results are more quantitative than RO1, yet they cannot rely on a single metric due to the system's focus on selecting and applying various bias metrics. Developing a new benchmark may be necessary to accurately assess performance, and consulting domain experts, like those in bias assessment, AI ethics, and law, could provide valuable insights.

5 Research methodology

This research envisions an interdisciplinary methodology that integrates conceptual analysis, formal modeling, and technical implementation to explore how the socio-technical context of AI systems can be incorporated into automated bias assessment. The framework proceeds through an iterative sequence of theory-

building, modelling, and empirical validation, allowing conceptual insights to inform formal structures and, ultimately, computational artefacts.

The **first** stage develops a conceptual foundation for understanding "context" in AI systems. Here, a literature review will be employed to analyze, interpret and compare frameworks of "context" from computer science, sociology, law and philosophy to assess their suitability and relevancy for the overall research question.

The **second** stage formalizes the actors and information exchanges for automated context-aware bias assessment. It is foreseen that this step mainly aligns with the established ontology engineering approach, including the specification and scheduling of the ontology, the search for ontological (AIRO[10], DocBiasO[20]) and non-ontological (ISO standards, EU regulation) resources as well as ontology design patterns, the construction of a conceptual model, and finally the implementation, improvement, and documentation of the ontology.

The **third** stage applies the model within a proof-of-concept automation system that aims to operationalize context-aware bias assessment. This involves implementing a Proof-of-Concept prototype capable of receiving contextual inputs, selecting appropriate bias metrics, and performing dataset evaluations within a data space environment. Developing he prototype might involve a comparison and selection of existing AI models (perhaps LLMs), potentially combined with training or fine-tuning. Other methods such as RAG might be employed, although, at this time, the specific nature of implementation is still an open question.

In the **final** stage, the evaluation has three goals: assessing the first stage's conceptualization, the second stage's formalization process, and the third stage's prototype. Evaluating the first stage is complex due to its qualitative results, requiring a thorough review of existing methods. The second stage's formal model is evaluated using standard procedures like pitfall analysis [17] and SPARQL-based competency questioning. Evaluating the automation prototype involves reviewing literature to find suitable benchmarks, and possibly developing a new method through comparing best practices and consulting experts.

6 Expected contributions and impact

As of now, the actual contributions offered towards the problem solution are minimal, since my research is in its earliest phases. The foreseen contributions include two main solutions: first, a machine-readable model of information exchange necessary for context-aware automated bias assessment, and second, a proof-of-concept (at technology readiness level 2) development an automated tool incorporating context into bias assessment.

Combined, these results are foreseen to impact the AI industry in Europe, specifically by supporting and facilitating the compliance with the AI Act (especially Art. 10), potentially providing proof of due diligence with regulatory requirements. Conceivably, the outputs might further support interoperability by providing standards in AI bias assessments. Finally, policy recommendations

might arise to influence requirements of bias assessment in future draftings or adaptations of regulation on AI and data.

Acknowledgments. This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Sklodowska-Curie Action Doctoral Networks grant agreement No. 101169409 (HARNESS).

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Amor, R., Dimyadi, J.: The promise of automated compliance checking. Developments in the Built Environment **5**, 100039 (2021)
2. Barr, C.J., Erdelyi, O., Docherty, P.D., Grace, R.C.: A review of fairness and a practical guide to selecting context-appropriate fairness metrics in machine learning. arXiv preprint arXiv:2411.06624 (2024)
3. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias (Oct 2018)
4. Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K.: Fairlearn: A toolkit for assessing and improving fairness in ai (2020)
5. Boda, V.V.R., Allam, H.: Automating compliance in healthcare: Tools and techniques you need. International Journal of Emerging Trends in Computer Science and Information Technology **2**(3), 38–48 (2021)
6. Brohi, S., Mastoi, Q.u.a., Jhanjhi, N.Z., Pillai, T.R.: A research landscape of agentic ai and large language models: Applications, challenges and future directions. Algorithms **18**(8) (2025)
7. Cappelli, M.A., Di Marzo Serugendo, G.: A semi-automated software model to support ai ethics compliance assessment of an ai system guided by ethical principles of ai. AI and Ethics **5**(2), 1357–1380 (2025)
8. Floridi, L., Cowls, J.: A unified framework of five principles for ai in society. Machine learning and the city: Applications in architecture and urban design pp. 535–545 (2022)
9. Golpayegani, D., Hupont, I., Panigutti, C., Pandit, H.J., Schade, S., O'Sullivan, D., Lewis, D.: Ai cards: Towards an applied framework for machine-readable ai and risk documentation inspired by the eu ai act. In: Annual Privacy Forum. pp. 48–72. Springer (2024)
10. Golpayegani, D., Pandit, H.J., Lewis, D.: Airo: An ontology for representing ai risks based on the proposed eu ai act and iso risk management standards. In: Towards a knowledge-aware AI, pp. 51–65. IOS Press (2022)
11. Hummel, P., Braun, M., Tretter, M., Dabrock, P.: Data sovereignty: A review. Big Data & Society **8**(1), 2053951720982012 (2021)
12. Jesus, S., Saleiro, P., Jorge, B.M., Ribeiro, R.P., Gama, J., Bizarro, P., Ghani, R., et al.: Aequitas flow: Streamlining fair ml experimentation. arXiv preprint arXiv:2405.05809 (2024)

13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
14. Marino, B., Chaudhary, Y., Pi, Y., Yew, R.J., Aleksandrov, P., Rahman, C., Shen, W.F., Robinson, I., Lane, N.D.: Compliance cards: Automated eu ai act compliance analyses amidst a complex ai supply chain. arXiv preprint arXiv:2406.14758 (2024)
15. Novelli, C., Governatori, G., Rotolo, A.: Automating business process compliance for the eu ai act. In: Legal Knowledge and Information Systems, pp. 125–130. IOS Press (2023)
16. Pandit, H.J., Rintamaki, T.: Towards an automated ai act fria tool that can reuse gdpr's dpia. arXiv preprint arXiv:2501.14756 (2024)
17. Poveda-Villalón, M., Gómez-Pérez, A., Suárez-Figueroa, M.C.: OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation. International Journal on Semantic Web and Information Systems (IJSWIS) **10**(2), 7–34 (2014)
18. Ruf, B., Boutharouite, C., Detyniecki, M.: Getting fairness right: Towards a toolbox for practitioners. arXiv preprint arXiv:2003.06920 (2020)
19. Ruf, B., Detyniecki, M.: Towards the right kind of fairness in ai. arXiv preprint arXiv:2102.08453 (2021)
20. Russo, M., Vidal, M.E.: Towards an ontology-driven approach to document bias. Journal of Artificial Intelligence Research **83** (2025)
21. Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., Ghani, R.: Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577 (2018)
22. Scerri, S., Tuikka, T., de Vallejo, I.L., Curry, E.: Common european data spaces: Challenges and opportunities. In: Data Spaces: Design, Deployment and Future Directions. pp. 337–357. Springer (2022)
23. von Scherenberg, F., Hellmeier, M., Otto, B.: Data sovereignty in information systems. Electronic Markets **34**(1), 15 (2024)
24. Sharma, K., Kumar, P., Li, Y.: Og-rag: Ontology-grounded retrieval-augmented generation for large language models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 32950–32969 (2025)
25. Trist, E.L., Bamforth, K.W.: Some social and psychological consequences of the longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. Human relations **4**(1), 3–38 (1951)
26. Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A., Russakovsky, O.: Revise: A tool for measuring and mitigating bias in visual datasets. International Journal of Computer Vision **130**(7), 1790–1810 (2022)
27. Wiśniewski, D., Potoniec, J., Ławrynowicz, A., Keet, C.M.: Analysis of ontology competency questions and their formalizations in sparql-owl. Journal of Web Semantics **59**, 100534 (2019)
28. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al.: The rise and potential of large language model based agents: A survey. Science China Information Sciences **68**(2), 121101 (2025)
29. Zemmouchi-Ghomari, L., Ghomari, A.R.: Translating natural language competency questions into sparql queries: a case study. In: WEB 2013: The First International Conference on Building and Exploring Web Based Environments. pp. 81–86 (2013)

Fabian Linde

Education

Universidad Politécnica de Madrid — Madrid, Spain
Ph.D. Artificial Intelligence — *Sep. 2025 – today*

Technical University Munich — Munich, Germany
M.Sc. Politics & Technology — *Oct. 2023 – Sep. 2025*

Ludwig Maximilian University — Munich, Germany
M.A. Philosophy — *Oct. 2022 – Mar. 2025*

Ludwig Maximilian University — Munich, Germany
B.A. Philosophy — *Oct. 2019 – May 2023*

Sorbonne Université — Paris, France
Erasmus - Computer Science — *Sep. 2017 – Jan. 2018*

Technical University Munich — Munich, Germany
B.Sc. Engineering Science — *Oct. 2015 – May 2020*

Research Experience

Universidad Politécnica de Madrid — Sep. 2025 – today
PhD Researcher — *Madrid, Spain*
- Research on context-aware automated bias assessment
- Research on legal ontologies

Supervised Program for Alignment Research — Sep. 2023 – Dec. 2023
Research Participant — *Berkeley, USA (Remote)*
- Technical research into sycophantic behavior of LLMs
- Fine-tuning and evaluating LLMs

Technical University Munich — Feb. 2025 – Aug. 2025
Research Assistant — *Munich, Germany*
- Research on transparency in AI and AI alignment

Work Experience

SAP SE — Jul. 2024 – Aug. 2025
Working Student AI Ethics Governance — *Munich, Germany*
- Governance of risk/impact management of AI use cases
- Support of adaption, rollout and implementation of AI Ethics policy
- Support in AI Act compliance

European Commission (DG CNECT) — Mar. 2024 – Jul. 2024
Blue Book Trainee — *Luxembourg, Luxembourg*
- Supporting the management of the procurement and development process of the European Language Data Space
- Research into technical aspects of data protection issues of language models

appliedAI — Oct. 2023 – Feb. 2024
Jun. Trustworthy AI Specialist — *Munich, Germany*
- Research, analysis and publication of emerging AI regulation (mostly EU AI Act)
- Work on AI Regulatory Sandboxes and relevant AI standards

0009-0004-0856-8036 | f.linde@upm.es | linkedin.com/in/fabianlinde