FRINGE: context-aware FaiRness engineerING in complex software systEms

Fabio Palomba
University of Salerno
Italy
fpalomba@unisa.it

Andrea Di Sorbo
University of Sannio
Italy
disorbo@unisannio.it

Davide Di Ruscio
University of L'Aquila
Italy
davide.diruscio@univaq.it

Filomena Ferrucci
Università di Salerno
Italy
fferrucci@unisa.it

Gemma Catolino
University of Salerno
Italy
gcatolino@unisa.it

Giammaria Giordano
University of Salerno
Italy
giagiordano@unisa.it

Dario Di Dario
University of Salerno
Italy
ddidario@unisa.it

Gianmario Voria
University of Salerno
Italy
gvoria@unisa.it

Viviana Pentangelo
University of Salerno
Italy
vpentangelo@unisa.it

Maria Tortorella
University of Sannio
Italy
tortorella@unisannio.it

Arnaldo Sgueglia
University of Sannio
Italy
sgueglia@unisannio.it

Claudio Di Sipio
University of L'Aquila
Italy
claudio.disipio@univaq.it

Giordano D'Aloisio
University of L'Aquila
Italy
giordano.daloisio@graduate.univaq.it

Antinisca Di Marco
University of L'Aquila
Italy
antinisca.dimarco@univaq.it

Abstract

Machine learning (ML) is essential in modern technology, driving complex data-driven decisions. By 2025, daily data generation will exceed 463 exabytes, increasing ML's influence and ethical risks of data exploitation and discrimination. The European Union's Artificial Intelligence Act highlights the need for ethical AI solutions.

Project Fringe (context-aware FaiRness engineerING in complex software systEms) addresses software fairness in ML-intensive systems that collect data through interconnected devices. Fringe aims to provide software engineers, data scientists, and ML experts with methodologies and software engineering solutions to improve fairness in ML systems. The goals of the project include developing a metamodel for ML fairness, a fairness-aware monitoring infrastructure, contextual solutions for identifying fairness issues, and automated recommendation systems to design fairness properties throughout the software development lifecycle.

*Main contributor; †Presenter.

ESEM '24, October 24–25, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-1047-6/24/10
https://doi.org/10.1145/3674805.3695394

CCS Concepts

• **Software and its engineering** → **Software design engineering**; **Extra-functional properties**.

Keywords

Software fairness engineering; Ethical Artificial Intelligence; Software Engineering for Artificial Intelligence.

ACM Reference Format:
Fabio Palomba, Andrea Di Sorbo, Davide Di Ruscio, Filomena Ferrucci, Gemma Catolino, Giammaria Giordano, Dario Di Dario, Gianmario Voria, Viviana Pentangelo, Maria Tortorella, Arnaldo Sgueglia, Claudio Di Sipio, Giordano D'Aloisio, and Antinisca Di Marco. 2024. FRINGE: context-aware FaiRness engineerING in complex software systEms. In *Proceedings of the 18th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '24), October 24–25, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3674805.3695394

1 Information about the project

Project full name. FRINGE: context-aware FaiRness engineerING in complex software systEms

Acronym. FRINGE

Duration. November 30, 2023 - November 30, 2025.

Participants. University of Salerno (Italy), University of Sannio (Italy), University of L'Aquila (Italy).

Funding Agency: European Union - Next Generation EU through the Italian Ministry of University and Research, under the PRIN 2022 PNRR program (Contract: P2022553SL).

URL. https://fringe-prin-project.github.io

2 Summary of the Project Objectives

Fringe studies solutions to engineer software fairness in Machine Learning (ML)-intensive systems. ML is increasingly being used to create data-driven decision-making systems for which an algorithmic solution is not feasible due to the complexity of the problem. By 2025 over 463 exabytes of data per day will be produced [8]. Such data are increasingly retrieved by heterogeneous interconnected smart objects and real-time monitoring devices. The volume of data, relations among the pieces of information collected, and continuous data acquisition and learning activities conducted to improve the accuracy of ML-intensive systems make it infeasible for developers to properly verify that such systems act according to the ethical principles raised by the European Artificial Intelligence Act and make predictions that do not perpetuate discrimination against sensitive groups [19]. Recent reports showed the implications of fairness concerns for companies and society: over 40% of the world population has been affected by unfair decisions of ML-intensive systems in 2020 [13], and 71% of customers would stop dealing with companies that unethically treat sensitive data [8]. The relevance of software fairness in ML-intensive systems has also been made popular by infamous incidents happened to the recruitment instrument employed by Amazon [12] and the criminal recidivism predictions made by the commercial risk assessment software COMPAS [1].

To explain the motivations and challenges faced by Fringe, let us consider the following example, designed with one of the partner software companies that will support the research team during this project. This is an Italian provider of AI-based software solutions for medical IoT. According to the experience of the partner company, a key issue in medical AI is data labeling during clinical assessment. Samulowitz et al. [14] showed that, due to gender stereotypes, women are over-diagnosed for diseases like depression and under-diagnosed for other diseases like cancer, and Seyyed-Kalantari et al. [15] found that women are diagnosed later than men for most diseases. As such, if the data labels in health registries are affected by disparities, then the ML models built on top of these data are likely to keep perpetuating such inequalities [11]. Therefore, novel instruments able to assist developers in dealing with fairness-related properties would reduce the risk of biased predictions, positively affecting society.

While initial scientific contributions to fairness engineering have been made, recent experience demonstrates that developing fair ML-intensive systems is, to a large extent, an open challenge, as discussed in recent surveys [7, 16]. Indeed, methods and techniques that support the analysis of fairness requirements, design of fair ML algorithms, and monitoring of fairness properties over time are still under-explored, and significant improvements are needed. Fringe aims to develop innovative solutions to tackle most of the aforementioned problems and deliver solutions for the creation of *fair ML-intensive systems*. More particularly, Fringe focuses on four main challenges, described as follows:

C1. Context-dependent definition of software fairness: Defining software fairness is intrinsically challenging because it might depend on the specific use case implemented by the ML-intensive system, thus requiring context-aware methods to diagnose fairness-related properties.

C2. Multi-disciplinary context: Engineering fairness requires a broad variety of skills that are not always available inside teams, including knowledge about SE practices, ML algorithms, and ethical and societal design and principles.

C3. Data and feature heterogeneity: ML-intensive systems are often trained with data collected from heterogeneous sources and engineered through the definition of features having different natures (e.g., audio, video, text, etc.). The lack of methods and instruments to ensure fairness-aware data and feature extraction and cleaning makes it hard for practitioners to deal with fairness as part of the data and feature engineering process.

C4. Limited knowledge on fairness design and verification. Little is known about how to design ML-intensive systems for fairness, which factors can impact the level of fairness of those systems, and how fairness-related properties can be verified, especially in a continuous data acquisition and learning scenario.

These challenges must be addressed in a holistic manner, with an engineered approach that can address them by introducing proper methodologies, techniques, and tools.

In response to these challenges, Fringe aims to provide methodologies, techniques, and approaches to analyze fairness requirements and assist the design and monitoring of fairness-related properties of ML-intensive systems. Fringe will target interdisciplinary teams consisting of software engineers, data scientists, and ML experts, reducing the distances among such specialists and facilitating their collaboration. The project objectives are:

O1. Support to context-aware fairness definition analysis: Definition of automated tools that can extract relevant information to help practitioners analyze the context the ML-intensive system is being developed for and the fairness definition to use.

O2. Support to context-aware fairness metrics and analytics: Definition of solutions to monitor software fairness properties over time, e.g., by measuring the level of fairness at a given point of the evolution or reasoning on the compromise between fairness and other ML quality requirements.

O3. Support to context-aware fairness requirements engineering: Definition of (semi-)automated methods, techniques, and instruments to allow practitioners to deal with fairness-related requirements while engineering features and gathering data for ML training.

O4. Context-aware fairness design: Provision of methodologies and tools to support the design of ML-intensive systems with respect to fairness in a given context.

O5. Consolidation of developed solutions: Definition of a fairness metamodel and set of prototypes to support practitioners in the engineering processes to handle fairness throughout the lifecycle of ML-intensive systems.

The objectives are organized in the six Work Packages (WPs) depicted in Figure 1. One of the work packages (WP1) concerns

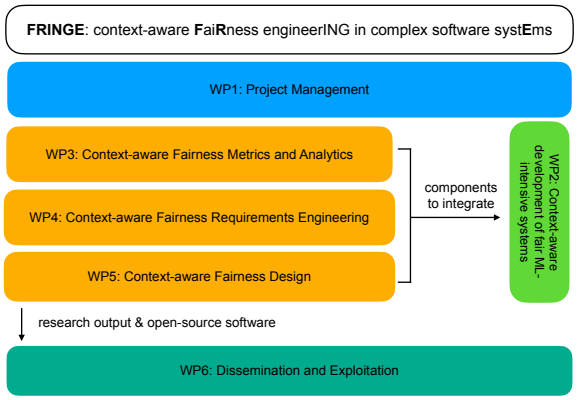

Figure 1: Work packages of the Fringe project.

project management, four (WP2-WP5) map onto O1-O5, and one (WP6) regards dissemination and exploitation.

3 Expected Tangible Outputs

The successful completion of the Fringe project will produce several tangible outputs, which are described in the following.

Cataloging Fairness Definitions and Monitoring Infrastructure. Addressing objectives O1 and O5, WP2 will create a *comprehensive catalog of fairness definitions*, outlining guidelines for their application in various contexts. A key deliverable will be a *fairness definition recommender system*, which will utilize the catalog to suggest the most appropriate fairness definitions for different scenarios. Additionally, this WP will develop a *monitoring infrastructure* designed to continuously assess the fairness of ML-intensive systems, ensuring they adhere to defined fairness specifications throughout the software lifecycle.

Taxonomy of Fairness Metrics and Analytical Framework. Fulfilling objective O2, WP3 aims to develop a detailed *taxonomy of fairness metrics*, which will be instrumental in the creation of a *fairness metrics recommender system*. This system will guide users in selecting the most relevant metrics for their specific needs. The WP will also produce a *comprehensive fairness analytical framework*, enabling systematic evaluation and improvement of fairness in machine learning applications.

Fairness Requirements Processing Framework. Aligned with objective O3, WP4 will establish a *framework for processing fairness requirements*. This includes supporting the elicitation, specification, and ongoing maintenance of context-aware fairness requirements. A significant output is a *fairness requirements recommender*, which will assist in accurately defining and managing fairness requirements tailored to specific contexts.

Design Patterns and Antipatterns for Fairness. To meet objective O4, WP5 will develop new *taxonomies of design patterns and antipatterns related to fairness* in software systems. It will also create semi-automated tools to help developers adopt beneficial patterns and avoid harmful ones. The primary output is a *fairness design recommender*, which will provide actionable insights and support for implementing fair design practices in software development.

4 Relevance of the Project to the ESEM Community

The Fringe project is designed to tackle the engineering of software fairness within ML-intensive systems, a significant concern given the increasing use of ML in complex data-driven decision-making processes. As vast amounts of data are generated and utilized, ensuring ethical and fair use of this data becomes paramount. The project's focus on developing methodologies, techniques, and tools to ensure fairness directly addresses the need for empirical evaluation and measurement of fairness in software systems, aligning with the core interests of the ESEM community.

More specifically, the Fringe project will first contribute to the ESEM community by providing empirically validated tools, methods, and frameworks that address fairness in ML-intensive systems. Secondly, the project's interdisciplinary approach will generate novel datasets related to the engineering of fairness-related properties, allowing researchers to build on top of the current body of knowledge and further the research on the building of fair ML-intensive systems. Additionally, the project's focus on real-world applications, such as the example involving medical AI, ensures that its contributions are practical and relevant, demonstrating the impact of empirical software engineering methods in practice.

5 Current Status and Intermediate Results

The project is currently ongoing and approaching its midpoint. The partners have already achieved various intermediate results for each objective. In the following, we summarize the current progress.

O1. Support to context-aware fairness definition analysis: We started the development of automated tools aiding practitioners in tailoring fairness definitions to specific contexts of ML-intensive systems [4]. A key component of this effort involves a model-driven approach utilizing the EMF ecosystem.[1] Compared to existing approaches [3, 18], our solution allows the definition of *(i)* bias given the application domain and *(ii)* custom fairness metrics and their composition. Fairness analysis tailors specific bias definitions to particular datasets, with defined scopes and associated fairness metrics. These metrics can be established in existing literature or custom-defined by users. The output is a model that encapsulates definitions of both bias and corresponding fairness analysis. From this model, an implementation is automatically generated using a code generator based on Acceleo technology.[2] Specifically, it produces Python code that checks the fairness of a given dataset using the specified fairness analysis information. Furthermore, we developed an initial domain-specific language (DSL) by relying on the Xtext technology[3] to facilitate the specification of the whole process.

O2. Support to context-aware fairness metrics and analytics: To start fulfilling O2, we surveyed the literature [2, 9, 10], collecting information about the defined fairness metrics in a novel taxonomy focusing on the specific fairness definition and data formats each metric applies to. By also exploring the collective knowledge shared in discussions on Q&A services (e.g., StackOverflow), this taxonomy aims at providing guidelines to

[1] https://projects.eclipse.org/projects/modeling.emf.emf
[2] https://marketplace.eclipse.org/content/acceleo
[3] https://eclipse.dev/Xtext/

developers on when and how to use each metric featured. More specifically, for each metric, the taxonomy summarizes the collected details according to 22 attributes belonging to five distinct dimensions, namely *(i) description and classification*, *(ii) representation*, *(iii) interpretation*, *(iv) applicability and usage*, and *(v) interoperability and integration*. The information about fairness metrics is progressively collected in an easily extensible knowledge base aimed to promote knowledge sharing across software communities. This represents a basic infrastructure on top of which we plan to develop a recommender system to help software engineers select the best metrics to adopt for monitoring specific fairness requirements, providing the right measurement strategies, the software libraries implementing the identified metrics, and code examples showing how to compute them.

O3. Support to context-aware fairness requirements engineering: We first developed ReFair [6], a context-aware requirements engineering framework designed to identify and classify sensitive features from User Stories early in the software development lifecycle. By leveraging natural language processing and word embedding techniques, ReFair facilitates the early consideration of fairness-related concerns by recommending context-specific sensitive features for ML tasks. This approach enhances the feasibility of integrating fairness considerations from the outset, ensuring that ethical principles are embedded throughout the development process. To help developers maintain fairness properties during the whole software development lifecycle, we are also exploring fairness-related issues and pull requests that occur in heterogeneous software projects leveraging AI-based solutions. Indeed, fairness constraint violations may emerge during the operation phase of a complex software system, signaled by users and developers using issue reports and pull requests. With the aim of defining a taxonomy of fairness issues, we collected issue reports and pull requests from large ML-based systems and started enumerating the observed behaviors, the likely causes, and the implemented solutions for the different documents. The fairness issues collected will also be used to experiment with AI-based solutions automatically identifying and classifying the various types of fairness issues. In addition, we focused on the integration of Large Language Models (LLMs) in software engineering tasks [17], addressing the ethical responsibilities associated with their deployment. We proposed a conceptual model that outlines ethical, social, and cultural considerations essential for guiding the development and validation of LLM-based approaches. Finally, we explored the fair requirements engineering processes required to develop emotion recognition systems [5], addressing ethical concerns such as consent, privacy protection, and algorithmic bias. By synthesizing existing literature and proposing guidelines, we laid the groundwork for responsible deployment of these systems in educational settings.

O4. Context-aware fairness design: In this respect, we worked toward the definition of a novel catalog of software engineering practices to handle fairness within ML-intensive system development and evolution. The catalog was established through two complementary research methods. First, we conducted a systematic mapping study to extract and categorize a set of 28 practices proposed in the literature. Second, we defined a survey study involving three samples of ≈50 practitioners each with experience in machine learning engineering, in which we validate the practices by assessing their frequency, impact, and application effort in practice. The results achieved so far will be employed as a basis for the development of recommendation systems which may suggest, according to the context, the best array of software engineering practices to incorporate fairness-related properties throughout the development process.

O5. Consolidation of developed solutions: We conceived an initial version of the metamodel for defining bias and fairness specification and assessment [4]. It underpins a comprehensive framework designed to systematically address fairness in machine learning systems. The metamodel has been devised by categorizing various types of biases, such as statistical biases (e.g., selection and measurement biases) and representational biases. The metamodel also permits the definition of fairness metrics tailored to different contexts and data types, emphasizing the necessity of contextual analysis to ensure relevance to specific use cases. Additionally, the metamodel is the key component to devise guidelines for implementing bias mitigation strategies, which can range from pre-processing data to modifying algorithms and post-processing outputs. Importantly, the metamodel integrates fairness considerations throughout the entire life-cycle of an ML system (from data collection to model training and deployment) ensuring that ethical principles are embedded at every stage of the ML system at hand. In particular, we plan to combine ReFair and the MDE framework to assist users throughout the whole fairness assessment process, i.e., from the elicitation of sensitive variables to the actual development.

6 Conclusion

In this paper, we introduced Fringe, a two-year project aiming at engineering fairness-related properties throughout the lifecycle of ML-intensive software systems. We reported about the project main objectives, the expected tangible outcomes, and the intermediate results achieved so far. The project started in November 2023 and, to date, already obtained significant results that we plan to further extend by the end of the project in November 2025.

Acknowledgments

This work has been partially supported by the European Union - NextGenerationEU through the Italian Ministry of University and Research, Projects PRIN 2022 PNRR, grant n. P2022553SL.

References

[1] AI Incident Base. 2016. AI Incident: Compas Recidivism Risk Assessment. https://incidentdatabase.ai/cite/40 Accessed: 2024-07-10.

[2] Simon Caton and Christian Haas. 2024. Fairness in machine learning: A survey. *Comput. Surveys* 56, 7 (2024), 1–38.

[3] Giordano d'Aloisio, Antinisca Di Marco, and Giovanni Stilo. 2023. Democratizing Quality-Based Machine Learning Development through Extended Feature Models. In *Fundamental Approaches to Software Engineering*, Leen Lambers and Sebastián Uchitel (Eds.). Springer Nature Switzerland, Cham, 88–110.

[4] Giordano d'Aloisio, Claudio Di Sipio, Antinisca Di Marco, and Davide Di Ruscio. 2024. How fair are we? From conceptualization to automated assessment of fairness definitions. arXiv:2404.09919 [cs.SE] https://arxiv.org/abs/2404.09919

[5] Dario Di Dario, Viviana Pentangelo, Maria Immacolata Colella, Fabio Palomba, and Carmine Gravino. 2024. Collecting and Implementing Ethical Guidelines for Emotion Recognition in an Educational Metaverse. In *Adjunct Proceedings*

of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. 549–554.

[6] Carmine Ferrara, Francesco Casillo, Carmine Gravino, Andrea De Lucia, and Fabio Palomba. 2024. ReFAIR: Toward a Context-Aware Recommender for Fairness Requirements Engineering. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. 1–12.

[7] Alina Köchling and Marius Claus Wehner. 2020. Discriminated by an Algorithm: A Systematic Review of Discrimination and Fairness by Algorithmic Decision-Making in the Context of HR Recruitment and HR Development. *Business Research* 13, 3 (2020), 795–848. https://doi.org/10.1007/s40685-020-00134-w

[8] McKinsey & Company. 2022. Data Ethics: What It Means and What It Takes. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/data-ethics-what-it-means-and-what-it-takes Accessed: 2024-07-10.

[9] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[10] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–44.

[11] A. Rajkomar, M. Hardt, M.D. Howell, G. Corrado, and M.H. Chin. 2018. Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine* 169, 12 (2018), 866–872. https://doi.org/10.7326/M18-1990

[12] Reuters. 2018. Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G Accessed: 2024-07-10.

[13] Harvard Business Review. 2020. AI Fairness Isn't Just an Ethical Issue. https://hbr.org/2020/10/ai-fairness-isnt-just-an-ethical-issue Accessed: 2024-07-10.

[14] A. Samulowitz, I. Gremyr, E. Eriksson, and G. Hensing. 2018. 'Brave Men' and 'Emotional Women': A Theory-Guided Literature Review on Gender Bias in Health Care and Gendered Norms towards Patients with Chronic Pain. *Pain Research and Management* 2018 (2018), 6358624. https://doi.org/10.1155/2018/6358624

[15] L. Seyyed-Kalantari, G. Liu, M. McDermott, I.Y. Chen, and M. Ghassemi. 2021. CheXclusion: Fairness Gaps in Deep Chest X-ray Classifiers. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*. 232–243.

[16] Christopher Starke et al. 2022. Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. *Big Data & Society* 9, 2 (2022), 20539517221115189. https://doi.org/10.1177/20539517221115189

[17] Gianmario Voria, Gemma Catolino, and Fabio Palomba. 2024. Is Attention All You Need? Toward a Conceptual Model for Social Awareness in Large Language Models. In *Proceedings of the 2024 IEEE/ACM First International Conference on AI Foundation Models and Software Engineering*. 69–73.

[18] Alfa Yohannis and Dimitris Kolovos. 2022. Towards Model-Based Bias Mitigation in Machine Learning. In *Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems* (Montreal, Quebec, Canada) *(Models '22)*. Association for Computing Machinery, New York, NY, USA, 143–153. https://doi.org/10.1145/3550355.3552401

[19] Jie M. Zhang and Mark Harman. 2021. "Ignorance and Prejudice" in Software Fairness. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 134–145. https://doi.org/10.1109/ICSE43902.2021.00023