SPATIAL: Practical AI Trustworthiness with Human Oversight

Abdul-Rasheed Ottun[1], Rasinthe Marasinghe[1], Toluwani Elemosho[1], Mohan Liyanage[1], Ashfaq Hussain Ahmed[1], Michell Boerger[3], Chamara Sandeepa[4], Thulitha Senevirathna[4], Vinh Hoa La[5], Manh-Dung Nguyen[5], Claudio Soriente[7], Samuel Marchal[6], Shen Wang[4], David Solans Noguero[8], Nikolay Tcholtchev[9, 3], Aaron Yi Ding[2] and Huber Flores[1]

[1]University of Tartu, Estonia; [2]Delft University of Technology, Netherlands; [3]Fraunhofer Institute for Open Communication Systems, Germany; [4]University College Dublin, Ireland; [5]Montimage, France;

[6]VTT Technical Research Centre of Finland Ltd, Finland; [7]NEC Labs, Germany; [8]Telefónica Research, Spain;

[9]RheinMain University of Applied Sciences, Germany

firstname.lastname@{ut.ee, tudelft.nl, fraunhofer.de, ucd.ie, montimage.com, vtt.fi, neclab.eu, telefonica.com}

***Abstract*—We demonstrate SPATIAL, a proof-of-concept system that augments modern applications with capabilities to analyze trustworthy properties of AI models. The practical analysis of trustworthy properties is key to guaranteeing the safety of users and overall society when interacting with AI-driven applications. SPATIAL implements AI dashboards to introduce human-in-the-loop capabilities for the construction of AI models. SPATIAL allows different stakeholders to obtain quantifiable insights that characterize the decision making process of AI. This information can then be used by the stakeholders to comprehend possible issues that influence the performance of AI models, such that the issues can be resolved by human operators. Through rigorous benchmarks and experiments in a real-world industrial application, we demonstrate that SPATIAL can easily augment modern applications with metrics to gauge and monitor trustworthiness. However, this, in turn, increases the complexity of developing and maintaining the systems implementing AI. Our work paves the way towards augmenting modern applications with trustworthy AI mechanisms and human oversight approaches.**

***Index Terms*—Practical Trustworthiness, Artificial Intelligence, Fairness, Human oversight, Industrial Use Cases**

I. Background and Motivation

All regulatory and economic frameworks worldwide recognize artificial intelligence as a pivotal technology to support the functionality of emerging modern applications [1]–[4]. However, challenges such as lack of transparency, resilience, and accountability have led to the imposition of strict regulations on its usage [5], [6]. The primary goal is to ensure best practices and minimize risks in developing AI-based software. Consequently, trustworthy AI has evolved from traditional trustworthy computing to specifically address the safety of AI software and responsible societal deployment [7]. Traditional methods, however, are not directly applicable to AI-based software. As AI continues to integrate into every aspect of human life, new methods are required to gauge, adjust, and monitor the trustworthiness of AI inference capabilities.

Modern applications have evolved from basic client-server architectures to more complex architectures that incorporate machine learning (ML) [8] and distributed machine learning, e.g., Federated Learning (FL) [9]. These system architectures implement AI pipelines to build models that learn and improve over time from data contributed by end-users. This allows for AI-based recommendations and guidance to enhance user experiences. With emerging regulatory guidelines emphasizing transparency and requiring greater human control and oversight, there is a regained focus on methods like Explainable AI (XAI). These methods aim to make the workings of AI more understandable and to integrate human feedback directly into AI systems. Adopting such approaches is crucial not only for developing new opportunities and markets but also for safeguarding the fundamental rights and liberties of individuals who depend on AI.

In this demonstration, we present SPATIAL [9], a proof-of-concept system architecture that augments AI components with mechanisms to gauge and monitor the inference capabilities of AI and its performance in practice. SPATIAL does this by characterizing AI using different trustworthy properties. Conceptually, SPATIAL uses AI sensors and dashboards to abstract the complexity [10]. An AI sensor is instrumented within an application to monitor a specific trustworthy property, e.g., fairness, and this results is then shown in the AI dashboard. Simply put, an AI dashboard shows to users quantifiable metrics extracted by AI sensors [10]. Based on this, SPATIAL is designed following a micro-service pattern architecture in the back-end, and a dashboard showing the computed results in the front-end. Through a rigorous evaluation, in which SPATIAL augments a real-world industrial application, we demonstrate that AI models can be characterized and their quantifiable characteristics can be shown to users without introducing much overhead in existing architectures. We also demonstrate how human oversight can improve the understanding of users over AI. However, greater engagement methods are required to foster the active participation of users. Our work paves the way towards implementing practical AI trustworthiness in modern

applications.

II. System design and implementation

We begin by explaining the augmented software architecture in which SPATIAL builds upon. After this, we describe the system implementation and deployment of SPATIAL.

A. Software design

Figure 1 shows the latest design of a modern architecture implementing distributed machine learning (FL). It is possible to observe that expected functionality in the architecture is linked to a design concern, requiring a specific human expertise. From the figure, it is also possible to observe the components that SPATIAL uses to augment the architecture [10]. These components implement routines to analyze the AI algorithm and data used in the construction of machine/deep learning models (AI models). SPATIAL augments system architectures with two types of components, one located at the server side, and other located at the client end.

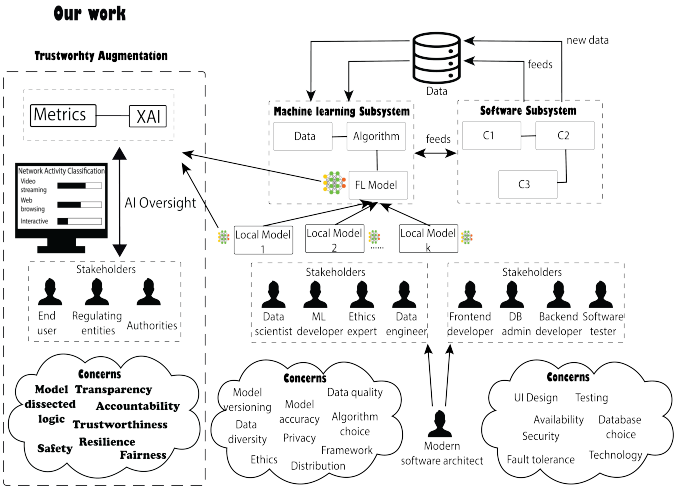

Fig. 1: SPATIAL software architecture.

AI sensors in SPATIAL: These depict virtual sensors that monitor and characterize a trustworthy property over time. Each trustworthy property is linked to an AI sensor. These sensors are instrumented within the target application whose trustworthiness is measured. Instrumented as APIs within applications, these sensors enable the quantification of AI compliance against available requirements, offering insights into the model's compliance with desired specifications. The key advantage of API instrumentation is that if the functionality of an AI sensor requires heavy computation, then the functionality is outsourced/offloaded to the server.

AI dashboards in SPATIAL: This is a user interface that enables human oversight in the analysis of AI models, for both data and AI models after training. The AI dashboards present all the quantifiable measurements extracted by the AI sensors to users visually. This allows human experts to collaborate in overseeing model development and ensuring tuning of the AI system to address trade-offs in trustworthiness properties while complying with regulatory requirements.

B. System deployment

Figure 2 shows the deployment of our augmented software architecture. Next, we provide a detailed description of each component implementation.

Back-end implementation: SPATIAL follows a micro-service pattern to estimate AI trustworthiness based on combined metrics and services. The key idea is that each micro-service specializes in characterizing a specific, trustworthy property, e.g., micro-service for fairness, micro-service for privacy. Micro-service patterns enable easy replacement of metrics for quantifying trustworthiness. This is beneficial as, currently, there is a mismatch between legal and technical trustworthiness. Thus, metrics that align better with legal requirements can be easily updated in SPATIAL. Node.js serves as a foundational runtime environment in our architecture, preceding the API Gateway. It is employed for building scalable server-side applications, leveraging its asynchronous, event-driven programming model to handle concurrent requests efficiently. We rely on open-source Kong technology for our API gateway, which supports easy extensions through OpenAPI and configurations for continuous integration. The API Gateway orchestrates communication, ensuring each micro-service receives the necessary input, processes it, and delivers the correct response. We used NGINX to define Upstreams and API addresses in the configuration file to target particular URL paths to route to the corresponding micro-services. Metrics and services quantifying trustworthiness as micro-services are containerized (using Docker) and follow a request/response scheme. To aggregate metric/service in SPATIAL, a virtual machine is first created, followed by pushing Docker images encapsulating all the dependencies and configurations into the virtual machine. Deployment through Docker containers simplifies the procedure and provides a standardized, isolated environment, ensuring seamless deployment experiences across different instances. Our SPATIAL deployment is located in the High-Performance Computing (HPC) Center [11]. Affiliated with the University of Tartu, and part of the LUMI supercomputer.

Current micro-services include, XAI services (LIME, Occlusssion sensitivity and SHAP), fairness metric over data using IBM AIF360 that quantifies demographic disparity, network traffic service applying impact and complexity metrics on AI models, differential privacy service obfuscating data, medical data analysis service implementing visualization methods for explanations, security diagnosis service implementation detection methods of model stealing and data poisoning attacks, LLM service implementing Llama LLM for adapting explanations to specific stakeholder terminology and the ML component implement traditional training functionality for different ML algorithms [9], [12].

Front-end implementation: SPATIAL frontend is implemented using React, providing users with an intuitive interface to seamlessly integrate with SPATIAL features. Node.js serves as the required runtime environment for React's development tools, including Babel and Webpack. The Bootstrap 5

framework is utilized for responsive design, while Tailwind CSS is employed for customized styling, resulting in visually appealing UI components. Additionally, the SPATIAL client integrates Okta for identity management, ensuring secure and robust authentication and authorization capabilities for access control. For dataset management and responsive chart visualization, we utilize D3.js, Chart.js, and Papaparse for parsing CSV data.

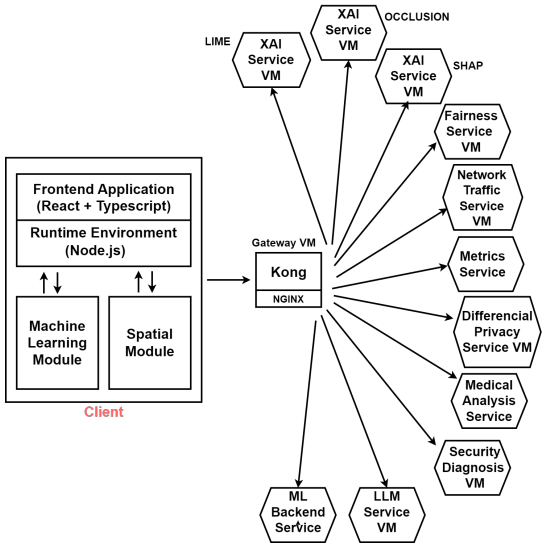

Fig. 2: SPATIAL system deployment.

III. Human oversight: SPATIAL Analysis and Usage

New regulatory requirements to make AI trustworthy and responsible are transforming the role that humans play when interacting with AI, and consequently, humans are required to be involved in the construction process of building, using and deploying AI models in applications. We next describe how SPATIAL supports human oversight, its applicability in a real-word industrial application.

A. SPATIAL usage

Figure 3 shows the overall flow of usage of SPATIAL. First, a user (aka stakeholder) login into the SPATIAL. Here, the user can then select the type of stakeholders, such that the LLM component can adjust the generated explanations from SPATIAL metrics and services based on the expertise of the user. In step 1, after the user is logged in, the user can build AI model using the ML component, see 3.1. To do this, the user has to upload a dataset or provide a link to retrieve the dataset. Alternatively, it is possible for the user to upload its own serialized version of the AI model. Once an AI model is available, the AI model can be analyzed using SPATIAL back-end metrics and services. Thus, the AI model is passed to the AI dashboard. Next is step 2; at this point, both the AI model and the data will be passed to the respective micro-services to characterize a specific, trustworthy property (Explainability and fairness properties are considered in this demo, see 3.2(a) & 3.2(b)). Each result provided by a micro-service will be visually presented in the AI dashboard, either as a diagram or a text explanation. The AI dashboard is used by the user to understand the quantifiable trustworthy characteristics of the AI model. After this, the AI dashboard can be used to facilitate changes on either the AI model or data. In this process, a new version of the model or data is created, such that changes can be applied, and SPATIAL can be reapplied in the new versions. SPATIAL also provides a comparison tool feature, such that different trustworthy properties from different AI model or data version can be compared side by side.

B. SPATIAL demonstration in industrial use case

Application: Medical e-calling application: It is a mobile application, part of an e-calling system, that uses accelerometer data to detect the falling of an elderly person. As the falling event is detected, the application triggers an emergency call to request medical assistance.

Dataset: The UniMiB dataset is used to train an AI model that classifies different types of activities based on accelerometer data. This dataset serves as a benchmark for human activity and fall detection, containing 11,771 acceleration samples from 30 subjects comprising both male and female genders. It encompasses nine classes representing activities of daily living (ADL) and eight classes representing falls.

SPATIAL applicability: Since a key trustworthy characteristic is to determine whether the AI model can be used on any individual, we rely on SPATIAL to perform fairness analysis. The service performs individual fairness analysis and group fairness analysis. It then generates results relating to model consistency in assigning labels to similar instances, class imbalance, disparate impact, equal opportunity, and equalized odds. As described previously, to perform this analysis, after the user has logged into SPATIAL, the user just has to upload the dataset and then pass it to the AI dashboard and SPATIAL micro-services are called. Notice that our use case places emphasis on fairness micro-service for this demonstration, but other micro-services work in the same manner.

AI dashboard results: Figure 3 also shows the results estimated by the fairness component. From the figure, it is possible to observe that all the results are presented as visually generated graphs and text explanations. These explanations can be adjusted based on the type of stakeholder using the latest advancements in LLM. As a result, it is easy to interpret the following bias result in figure 3 for the provided datasets. In one of the fairness analysis results (bottom right), both Age and Gender features exhibit relatively high consistency metrics of 0.740 and 0.755, respectively, suggesting consistent treatment across different instances. However, the class imbalance metric reveals an imbalance in the Age feature (0.505), while Gender shows a slight imbalance (-0.12), indicating potential disparities in the Age representation of different groups as the Age feature favors the majority class while the Gender

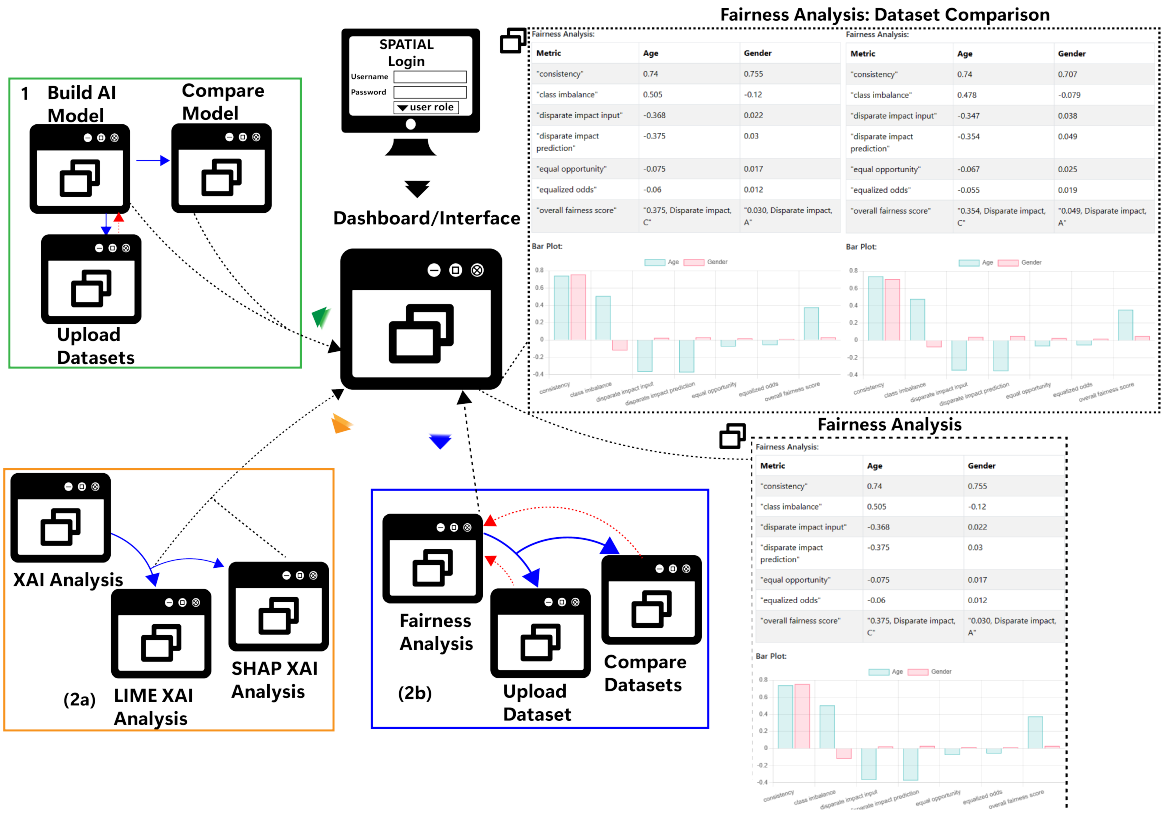

Fig. 3: Overall flow of SPATIAL usage and its applicability (fairness only) over a use case application.

favors the minority class. Furthermore, the disparate impact input metric for Age (-0.368) and Gender (0.022) indicates unequal treatment of the groups by the model's predictions. Suggesting that one group has an advantage over another in the model prediction. Disparate prediction input metrics indicate somewhat similar outcomes, with Age (-0.375) and Gender (0.03). However, the equal opportunity metric reveals negative values for Age (-0.075) and positive values for Gender (0.017), indicating disparities in opportunities for positive outcomes across different groups as the majority group is more favored. Finally, fairness score metrics show positive values for Age (0.375) and Gender (0.030), indicating fair practices with efforts to address bias in disparate impact and equal opportunity.

IV. Summary and Conclusions

We demonstrated a working implementation of SPATIAL, a proof-of-concept system that augments modern applications with capabilities to analyze and quantify trustworthy aspects of AI models. SPATIAL uses a micro-service and API gateway pattern to combine different methods for analyzing AI algorithms, its data, and the resulting AI model. SPATIAL also implements an AI dashboard to show the results of the analysis to users, introducing human-in-the-loop feedback that can be used to monitor and tune AI model behavior. Through rigorous benchmarks and analyses that consider a real-world industrial application, we demonstrated the performance and scalability of SPATIAL to perform practical AI trustworthiness.

Acknowledgment

This research is part of SPATIAL project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101021808.

References

[1] T. Babina *et al.*, "Artificial intelligence, firm growth, and product innovation," *Journal of Financial Economics*, vol. 151, p. 103745, 2024.
[2] Cio.gov - executive order (eo) 13960. [Online]. Available: https://www.cio.gov/policies-and-priorities/Executive-Order-13960-AI-Use-Case-Inventories-Reference
[3] E. Commission, *European approach to artificial intelligence*, Accessed March 1, 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence
[4] C. A. of China, *Interim Measures for the Management of Generative Artificial Intelligence Services*, Accessed March 1, 2024. [Online]. Available: http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm
[5] A. Goldfarb, "Pause artificial intelligence research? understanding ai policy challenges," *Canadian Journal of Economics/Revue canadienne d'économique*, 2024.
[6] H. Li, L. Yu, and W. He, "The impact of gdpr on global technology development," pp. 1–6, 2019.
[7] J. M. Wing, "Trustworthy ai," *Communications of the ACM*, vol. 64, no. 10, pp. 64–71, 2021.
[8] H. Muccini *et al.*, "Software architecture for ml-based systems: what exists and what lies ahead," in *2021 IEEE/ACM AI WAIN*, pp. 121–128.
[9] A.-R. Ottun *et al.*, "The spatial architecture: Design and development experiences from gauging and monitoring the ai inference capabilities of modern applications," in *Proceedings of IEEE ICDCS 2024*, 2024.
[10] H. Flores, "Ai sensors and dashboards," *IEEE Computer Magazine*, 2024.
[11] University of Tartu, "Ut rocket," 2018. [Online]. Available: share.neic.no

[12] M. Boerger *et al.*, "Deliverable (d3.4) - performance evaluation in controlled environments and guidelines to build the pilot studies in real testbeds," EU SPATIAL project, Tech. Rep., 2024. [Online]. Available: https://spatial-h2020.eu/