Architecting Human Oversight in AI-Enabled Software Systems: From Principles to Implementation

George Manias
Digital Systems
University of Piraeus
Piraeus, Greece
gmanias@unipi.gr

Vrettos Moulos
Digital Systems
University of Piraeus
Piraeus, Greece
vrettos@unipi.gr

Anna Gatzioura
Department of Engineering
Universitat Pompeu Fabra
Barcelona, Spain
0000-0003-2532-0169

Tanja Zdolsek Draksler
Faculty of Electrical Engineering and Computer Science
University of Maribor
Maribor, Slovenia
0000-0002-6782-6275

Alenka Gucek
Department for Artificial Intelligence
Jožef Stefan Institute
Ljubljana, Slovenia
alenka.gucek@ijs.si

Dimosthenis Kyriazis
Digital Systems
University of Piraeus
Piraeus, Greece
dimos@unipi.gr

***Abstract*—Human oversight has become a foundational requirement for trustworthy AI-enabled software systems, particularly in high-stakes domains such as healthcare, public sector decision-making, and critical infrastructures. While regulatory frameworks (e.g., EU AI Act) and ethical guidelines emphasize human-in-the-loop and human-on-the-loop approaches, there is currently a significant gap between high-level principles and concrete software architecture practices that enable effective human oversight at design time and runtime. In parallel, the rapid adoption of Large Language Models (LLMs), agentic AI, and Digital Twins is reshaping software system architectures. However, many implementations treat oversight as an external governance layer rather than as a first-class architectural concern integrated into system design, verification, monitoring, and evolution. This creates risks related to automation bias, lack of explainability, unsafe autonomy escalation, and insufficient accountability mechanisms. This tutorial bridges software architecture, human-centered AI, and governance-aware system engineering and provides participants with a systematic approach to designing, implementing, and evaluating human oversight mechanisms across the software lifecycle.**

Keywords—human oversight, Trustworthy AI, Human-in-the-loop, Human-centered AI, Digital Twins, Governance-aware Software, LLM-enabled Systems

I. Sumarry

Software systems have undergone a long evolution from monolithic architectures to distributed service-based systems and, more recently, to AI-enabled and agentic software ecosystems [1-2]. This evolution is enabling higher levels of automation and system autonomy, but it also introduces new risks related to decision transparency, accountability, automation bias, and safe operation in high-stakes environments [3]. In domains such as healthcare, public administration, and cybersecurity, ensuring meaningful human oversight is becoming a core requirement rather than an optional system feature [4]. As a result, human oversight is emerging as a critical quality attribute for modern AI-enabled software systems [5]. On top of this, regulatory frameworks and ethical guidelines emphasize human-in-the-loop and human-on-the-loop approaches, there is currently limited guidance on how to translate these principles into concrete software architecture practices [6]. In many real-world systems, human oversight is implemented at the interface or governance level, rather than being embedded into system architecture, runtime monitoring, and decision orchestration mechanisms [7]. This often results in oversight mechanisms that are difficult to scale, difficult to validate, and ineffective in supporting human decision-making under uncertainty [8]. There is therefore a need for systematic methods that enable software architects and AI engineers to design oversight-aware systems from the ground up. At the same time, what is urgently needed is the AI Act to be enriched with correlated guidelines able to translate high level, international, legal and ethical standards into architecturally enforceable constraints. Developing this will require systematizing and analyzing use cases, especially from the high-risk domain (like algorithmic hiring), to tap into the varied and context-specific human rights, legislative, and administrative practices.

This tutorial aims to guide practitioners and researchers in designing AI-enabled software systems where human oversight is treated as an architectural property. It introduces oversight-by-design principles and demonstrates how oversight can be integrated across the system lifecycle, including data pipelines, model orchestration layers, decision workflows, and runtime monitoring infrastructures. It is also grounded in real-world experience from European research and innovation projects (i.e., AI4Gov, TWON, CERTAIN, COMFORTage, HumAIne, FINDHR, VIGILANCE) contributing expertise in different domains spanning from trustworthy AI deployment in public sector to human-centric AI-enabled decision support in healthcare and cybersecurity. The tutorial team combines expertise in software architecture, AI-enabled digital twins, human-centered AI, and deployment of AI solutions in regulated and high-risk environments.

II. Goals and Intended Audience

A. Learning Goals

With this tutorial participants will be able to learn about Trustworthy and Responsible AI Foundations (including bias overview, types of bias in AI, and bias mitigation strategies). Moreover, the foundations of human oversight in AI-enabled software systems and architectural patterns for human-in-the-loop and human-on-the-loop systems will be discussed. Participants will also be trained on designing escalation and intervention workflows in distributed AI systems, as well as on entegrating explainability, traceability, and accountability into software architectures, and evaluating oversight effectiveness using socio-technical metrics. These will be

2768-4288/26/$31.00 ©2026 IEEE
DOI 10.1109/ICSA-C68850.2026.00034

achieved through applying oversight-by-design principles to real-world use cases following examples from other projects.

B. Intended Audience

The intended participants of the event are: (i) Software architects designing AI-enabled systems; (ii) AI engineers integrating models into production software; (iii) Researchers in software architecture and human-centered AI; (iv) Public sector and healthcare digital solution developers; (v) PhD students working on trustworthy AI or socio-technical systems. Of course, without limiting any participation.

TABLE I. TUTORIAL STRUCTURE

Session 1: Foundations and Architectural Design (90')		
Topic	***Duration***	***Material to be used***
Introduction and Motivation (AI-enabled software, risks of autonomy, need for human oversight as architectural property)	15 mins	Slides, problem framing diagrams, real-world system examples
Foundations of Human Oversight in AI Systems (HITL, HOTL, HIC models, socio-technical risks, regulatory and governance context)	25 mins	Conceptual framework diagrams, oversight lifecycle model, reference architecture overview
Architectural Patterns for Oversight-by-Design (design-time vs runtime oversight, escalation workflows, monitoring and intervention patterns)	35 mins	Architecture pattern catalog, UML/component diagrams, reference architecture templates
Interactive Discussion / Mini Exercise (identify oversight points in a sample architecture)	15 mins	Sample system architecture diagram, guided discussion worksheet
Session 2: Real-World Systems and Hands-on Design (90')		
Topic	***Duration***	***Material to be used***
Project Case Studies (deployment architectures, oversight challenges, lessons learned)	35 mins	Case study architecture diagrams, workflow examples, socio-technical evaluation examples
Hands-on: Designing an Oversight-Ready AI Software Architecture (group-based guided design exercise)	40 mins	Architecture canvas template, design worksheet, scenario description, starter architecture diagrams
Wrap-up, Key Takeaways, and Q&A	15 mins	Key takeaway slides, oversight readiness checklist, further reading resources

III. TUTORIAL IMPLEMENTATION

The tutorial will be delivered as two 90-minute sessions combining lectures, case-based discussion, and a guided hands-on design exercise. The first session will introduce foundations of human oversight in AI-enabled software systems and present architectural patterns for integrating oversight mechanisms at design time and runtime. The second session will focus on real-world system implementations and lessons learned from the aforementioned EU-funded R&I projects followed by a hands-on group exercise where participants will design an oversight-aware architecture for an AI-enabled decision-support scenario.

IV. PRESENTERS

Dr. George Manias is an Assistant Professor in the Jheronimus Academy of Data Science, Tilburg University, Netherlands. He received his Ph.D. in Artificial Intelligence (AI) and an M.Sc. in Big Data and Analytics from the Department of Digital Systems, University of Piraeus, Greece, with a strong background in Data Science, Natural Language Processing (NLP), and software engineer.

Dr. Vrettos Moulos is a senior research software engineer in Institute of Communication and Computer Systems in Greece. He holds a PhD in secure microservice architecture patterns from the School of Electrical and Computer Engineering of the National Technical University of Athens (NTUA).

Dr. Tanja Zdolšek Draksler is a researcher at the Jožef Stefan Institute and the University of Maribor, specializing in the intersection of AI/LM and the social sciences. Her work centers on driving digital innovation, ensuring that data-driven systems remain human-centric and support the public interest.

Dr. Alenka Guček is a researcher and data visualization engineer at the Jožef Stefan Institute. She focuses on exploratory visualization tools that make complex AI services and data accessible for society.

ACKNOWLEDGMENT

The research leading to the results used in this tutorial has received funding from the Horizon Europe projects AI4Gov (GA no 101094905), TWON (GA no 101095095), FINDHR (GA no 101070212), COMFORTage (GA no 101137301), CERTAIN (GA no 101189650), HumAIne (GA no 101120218), and the Digital Europe Action project VIGILANCE (GA no 101249737).

REFERENCES

[1] Vaidhyanathan, K., & Muccini, H. (2025, September). Software Architecture in the Age of Agentic AI. In European Conference on Software Architecture (pp. 41-49). Cham: Springer Nature Switzerland.

[2] Katal, A., Prasanna, P., Birla, R., & Kunal. (2025). Evolution from Monolithic to Microservices Architecture: A New Era in Software Architecture. In Advancements in Optimization and Nature-Inspired Computing for Solutions in Contemporary Engineering Challenges (pp. 235-279). Singapore: Springer Nature Singapore..

[3] Bandi, A., Kongari, B., Naguru, R., Pasnoor, S., & Vilipala, S. V. (2025). The rise of agentic ai: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. Future Internet, 17(9), 404.

[4] Maranhão, J. J., & Guerra, E. M. (2024, July). A prompt pattern sequence approach to apply generative AI in assisting software architecture decision-making. In Proceedings of the 29th European Conference on Pattern Languages of Programs, People, and Practices (pp. 1-12).

[5] Kyriakou, K., & Otterbacher, J. (2023). In humans, we trust: Multidisciplinary perspectives on the requirements for human oversight in algorithmic processes. Discover Artificial Intelligence, 3(1), 44.

[6] Enqvist, L. (2023). 'Human oversight'in the EU artificial intelligence act: what, when and by whom?. Law, Innovation and Technology, 15(2), 508-535.

[7] Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review, 45, 105681.

[8] Sadovski, E., Aviv, I., & Hadar, I. (2025, September). Navigating the Human-oversight Dilemma in AI-based Systems. In 2025 IEEE 33rd International Requirements Engineering Conference Workshops (REW) (pp. 454-461). IEEE.