Responsible and Trusted AI: An Interdisciplinary Perspective (2026)

Kevin Baum[1,2,3,4,5(✉)], Thorsten Helfer[5,6], Felix Jahn[1,5], Sophie Kerstan[7], Andreas Sesing-Wagenpfeil[8], Timo Speith[5,9], and Sarah Sterz[5,10]

[1] German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
kevin.baum@dfki.de
[2] Hamburg University of Technology (TUHH), Hamburg, Germany
[3] Saarland Informatics Campus, Saarbrücken, Germany
[4] Oxford Internet Institute, Oxford, UK
[5] Algoright e.V., Saarbrücken, Germany
[6] CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
[7] Department of Psychology, University of Freiburg, Freiburg im Breisgau, Germany
[8] Institute for Legal Informatics, Saarland University, Saarbrücken, Germany
[9] Department of Philosophy, University of Bayreuth, Bayreuth, Germany
[10] Saarland University, Saarbrücken, Germany

Abstract. Regulation and industry practice have produced a sizeable inventory of safeguards for AI systems: humans placed in decision loops, oversight obligations, documentation duties, explanations, disclaimers. Whether the items in this inventory do what they are supposed to do is a separate question, and one that is easy to leave unasked, because a safeguard's presence is far cheaper to verify than its effect. The three contributions to the Responsible and Trusted AI track at AISoLA 2026 examine that gap at three different levels of an AI system: the causal structure of the decision process into which a human is placed, the coordination layer of a distributed AI-enabled system, and the model and the hardware it runs on. Each contribution identifies a safeguard that can be satisfied nominally while failing in substance, and each responds by relocating the demand into structure that is fixed before deployment rather than delegating it to a person at the point of use. This introduction sets out the shared problem, situates the four papers within it, and marks the tensions between them that we expect the on-site discussion to take up.

Keywords: Human oversight · AI Act · Interpretability · AI governance · Interdisciplinarity

1 Introduction

Two years of intense regulatory and standardisation activity have left us with a substantial inventory of safeguards for AI systems. Humans are placed in decision loops. Oversight functions are mandated. Technical documentation is required

© The Author(s) 2027
B. Steffen (Ed.): AISoLA 2026, LNCS 17146, pp. 107–114, 2027.
https://doi.org/10.1007/978-3-032-41269-0_7

and its contents specified. Explanations are demanded, and where they cannot be given, disclaimers are displayed. The inventory has grown faster than the evidence that its items work.

Consider what each of these safeguards permits. A human who must approve every output may approve every output. A completed documentation template may describe an architecture that has since drifted away from the code. An explanation may satisfy the person receiving it while bearing no reliable relation to the computation that produced the output it explains. A notice reading "AI can make mistakes" may sit beneath a conversation engineered to be found convincing. In each case the safeguard is present, auditable, and inert.

The gap between a safeguard's presence and its effect is the unanticipated thread running through the three contributions to this year's Responsible and Trusted AI track. The papers, however, differ in other regards. They come from philosophy, law, software engineering, machine learning, and the philosophy of science; they address a regulatory taxonomy, a coordination language for cyber-physical systems, and the hardware substrate of neural networks. What they have in common is, on a high level of abstraction, a diagnosis and a response. The diagnosis is that a requirement stated in terms of what can be observed from outside—a human is present, a document exists, an explanation is available—can be met without the property that motivated it. The response, in all cases, is to push the requirement down into structure that is settled before the system is used, rather than leaving it to whoever happens to be at the interface when something goes wrong.

Following the tracks at AISoLA 2023, 2024 [7], and 2025 [8], this year's edition again brings together researchers from philosophy, law, psychology, and computer science. The track is co-organised by the Centre for European Research in Trusted Artificial Intelligence (CERTAIN).

2 Nominal and Effective Safeguards

The gap we are describing is not an accident of drafting, and it will not be closed by more careful drafting alone (as was made clear by one of last year's contributions, cf. [9]). It arises from a structural feature of how safeguards are specified.

Safeguards have to be checkable. Whether a human sign-off occurred, whether a document was filed, whether an interface displays a notice—these are things an auditor can establish from records. The properties that actually matter are not like this. Whether the person signing off was in a position to withhold their signature; whether the document describes the system as built; whether the notice changes what users believe or feel: these are dispositional or counterfactual properties, sensitive to training, timing, organisational context, and commercial incentive. Specifications like these therefore reach for operationalisation, for observable proxies—and the proxy relation holds only under background conditions that the specification does not itself fix. For human oversight in particular, the distance between the proxy and the property it stands in for has been mapped in some detail [12].

When the proxy comes apart from the property, the result can be worse than having no safeguard at all. A nominal safeguard licenses the appearance of due care [13], and it allocates the residual risk. Someone was formally in charge, so the failure is theirs: the overseer who did not catch it, the user who should have known better than to believe the machine. Responsibility settles on the last human in the chain, who is typically the person with the least capacity to have prevented the outcome and the least influence over the conditions that produced it—the position Elish calls a *moral crumple zone* [3]. Neither half of this is new. Automation does not so much remove the human contribution as transform it, and what remains is often the harder job [1]; and where humans have actually been studied in the loop of algorithmic decisions, their presence has not reliably improved the decisions [6].

Closing the gap requires two moves, and the papers in this track make one or both. The first is *analytic*: state precisely which property the safeguard is supposed to secure, in terms that distinguish it from its proxy. The second is *constructive*: build that property into structure fixed in advance—the causal architecture of a decision process, the coordination layer of a system, the architecture of a model—so that satisfying the requirement is not left to a person's vigilance at the moment of use. Neither move suffices alone, and they may not suffice even together. A conceptual analysis with no purchase on implementation remains idle; an architecture built to satisfy a misdescribed requirement satisfies the wrong thing, precisely and verifiably. This is why the work has to be interdisciplinary, and not merely multidisciplinary: the analytic and the constructive move constrain each other, and the constraint runs in both directions.

A corollary is visible in all three papers, though none of them makes it the main thesis. If effectiveness depends on conditions that individuals at the point of use do not control, then responsibility has to migrate upstream, toward providers, architectures, and design decisions taken long before a user or an overseer encounters the system.

3 The Contributions

3.1 The Causal Position of the Human

Baum and Laux [2] address the vocabulary in which human involvement is specified. "Human-in-the-loop", "human-on-the-loop", and "human oversight" are used across computer science, law, philosophy, and psychology to pick out configurations that differ fundamentally, which obstructs both interdisciplinary work and regulatory operationalisation. Their proposal is to draw the distinction causally rather than spatially. A human-in-the-loop is *constitutive*: without their contribution no manifestation of the decision process can generate an output. A human-on-the-loop is *corrective*: the process produces outputs by default, and the human can prevent, modify, or override them from outside the primary causal chain. In their terms, the first is a gate and the second a switch. Within the corrective category they distinguish synchronous, asynchronous, and anticipatory modes, situated in a nested model of provider and deployer runtime, and they

add a second, orthogonal dimension of cognitive integration: whether the human forms judgements independently of the system or partly through its representations.

The payoff for our theme comes when descriptive categories are separated from the normative requirements they serve. Statutory human oversight, on their account, is a normative mode of the corrective configuration: it demands not merely the causal position but genuine preparedness and capacity for effective intervention. The consequence is pointed. Article 22 GDPR [4], read as prohibiting decisions "based solely on automated processing", appears to demand a constitutive human, and in that sense a rubber-stamp would satisfy it. Article 14 AI Act's requirement of effective oversight demands something a rubber-stamp cannot provide, and can therefore offer stronger protection than the provision that looks more demanding [5]. Where the same person occupies both roles—which is sometimes unavoidable, since the only plausible overseer may be the person already in the loop—the authors treat the resulting conflict as a design problem calling for layered oversight structures and mechanisms that counteract automation bias [10], rather than as a curiosity to be noted and set aside.

3.2 Compliance as an Architectural Property

Wagner, Bhimaraju, Borg, Runeson, and Lee [14] take up requirements that are usually discharged by producing documents. The high-risk obligations in Chapter III of the AI Act are system-level rather than model-level, and they apply to AI-enabled cyber-physical systems that are typically concurrent and distributed—exactly the setting in which reliable, reproducible, and testable behaviour is hardest to achieve. The authors show how Lingua Franca, a polyglot coordination language with deterministic semantics and explicit timing, turns several of these obligations into architectural primitives, demonstrated on an AI-enabled traffic light control system with three federated intersections, a centralised planner behind a safety gate, a local fallback planner, a health aggregator, and a human oversight console.

Four results are relevant here. The architecture diagram required by Annex IV(2)(c) is generated from the code rather than drawn alongside it, so the documentation cannot silently fall out of date. Determinism gives an unambiguous definition of correct behaviour, which is what makes incorrect behaviour detectable and hence testable at all. The coordination mechanisms—an `after` delay on the planner's output, `max wait` on the health aggregator—bound the latency with which a fault can be declared, which is the concrete content of the anomaly-monitoring capability that Article 14(4)(a) requires of an oversight interface. And the use of a logical timeline guarantees in-order timestamps for causally related events, addressing a standing obstacle to trustworthy logging in distributed systems and supporting the record-keeping obligation in Article 12. The paper's worked-out tradeoffs are worth attention in their own right: the authors accept a one-second delay on non-time-critical AI plans in order to preserve immediate reaction to emergency and manual-override triggers, and make the assumption explicit rather than leaving the timing to chance.

3.3 Interpretability and the Substrate

Seet, Bozenhard, and Ostermann [11] locate the same gap inside the notion of interpretability. The familiar definition—the degree to which a human can understand the cause of a decision—is satisfied by whatever a human observer finds satisfying, and a model that produces plausible explanations for the wrong reasons is not thereby safer. Mechanistic interpretability sets a higher bar: a causal account of the model's internal components, testable by interchange interventions and therefore falsifiable. Drawing on the philosophy of science, the authors connect this to decomposition and localisation as strategies of mechanistic explanation, and argue that a model's interpretability tracks how far it supports localisation. Superposition then appears not as a nuisance but as the systematic obstacle: under training pressure, networks represent more features than they have neurons, and no concept can be cleanly assigned to a part.

Their distinctive move is to ask why models of this kind are the ones we build. Part of the answer, they argue, is the hardware. Dense connectivity and the gradient traffic it generates are practical only because electronic interconnects offer high bandwidth at low latency; a substrate with lower bandwidth would force more localised models, which would in turn have to compensate with greater expressivity per node. The paper then assesses candidate substrates—analogue electronics, electrochemical neuromorphic architectures, chemical reaction networks, and molecular mechanical logic—against per-node expressivity, energy efficiency, and how far each is from being buildable, and is candid that the most promising on the first two counts is the furthest from practical synthesis. Along the way the authors distinguish interpretability from controllability, noting that activation steering already delivers reliable behavioural intervention without mechanistic understanding: a weaker property, but an attainable one. Of the three papers, this is the one that pushes the constructive move furthest down the stack, and it is correspondingly the most speculative.

3.4 Points of Contact and Tension

Placed side by side, the three papers address three layers: the causal structure of a decision process, the coordination layer of a distributed system, and the model and its substrate. None of them treats trustworthiness as something a system can be asked, or asserted, to have.

They do not agree, and the disagreements are as productive as the convergence. Wagner et al. take the stochastic model as given and build guarantees around it, in the surrounding architecture; Seet et al. argue that the model itself is where the constraint has to be imposed, and are prepared to contemplate changing the hardware to get it. Baum and Laux hold that oversight can be made effective if the overseer's causal position and capacities are specified correctly and supported architecturally. Whether these positions are complementary or in genuine competition is, we think, one of the more interesting questions the track can put to its participants.

4 Conclusion

Adding safeguards is not the same as securing responsible and trusted AI. What matters is whether the safeguards added do the work they were introduced to do, and establishing that requires saying precisely what the property in question is, which is conceptual, normative, and legal work, and then building it into structures fixed before deployment, which is engineering work. The three contributions to this year's track do this at three different levels, and their juxtaposition raises questions none of them settles alone. Which requirements can be discharged architecturally, and which depend irreducibly on people, organisations, and markets? Where effectiveness rests on background conditions, who is responsible for maintaining those conditions? And what can oversight mean for systems deployed directly to hundreds of millions of people, where the deployer and the person at risk are the same individual?

We look forward to putting these questions to the authors and to the wider AISoLA community in Kos, and hope the exchange will produce the kind of collaboration that neither philosophy, law, psychology, nor computer science can substitute for on its own.

Acknowledgments. We thank the authors for their contributions and the AISoLA organisers for hosting the track, which is co-organised by the Centre for European Research in Trusted Artificial Intelligence (CERTAIN). Kevin Baum, Sophie Kerstan, and Sarah Sterz were supported by the German Research Foundation (DFG) under grant No. 389792660, as part of TRR 248, see https://perspicuous-computing.science. Kevin Baum and Felix Jahn were supported by the German Federal Ministry of Education and Research (BMBF) as part of the project MAC-MERLin (Grant Agreement No. 16IW24007). Kevin Baum was further supported by the European Regional Development Fund (ERDF) and the Saarland within the scope of (To)CERTAIN (Project ID EFRE-AuF-0000942), and Sarah Sterz by the DFG project CARAT (Project No. 547583482).

References

1. Bainbridge, L.: Ironies of automation. Automatica **19**(6), 775–779 (1983). https://doi.org/10.1016/0005-1098(83)90046-8
2. Baum, K., Laux, J.: Constitutive vs. corrective: a causal taxonomy of human runtime involvement in AI systems. In: Steffen, B. (ed.) AISoLA 2026, LNCS, vol. 17146, pp. xx–yy. Springer, Cham (2027)
3. Elish, M.C.: Moral crumple zones: cautionary tales in human-robot interaction. Engag. Sci. Technol. Soc. **5**, 40–60 (2019). https://doi.org/10.17351/ests2019.260
4. European Parliament and Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1–88 (2016)

5. European Parliament and Council of the European Union: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). OJ L, 2024/1689, 12.7.2024 (2024)
6. Green, B., Chen, Y.: The principles and limits of algorithm-in-the-loop decision making. Proc. ACM Hum.-Comput. Interact. **3**(CSCW), 1–24 (2019). https://doi.org/10.1145/3359152
7. Helfer, T., Baum, K., Sesing-Wagenpfeil, A., Schmidt, E., Langer, M.: Responsible and trusted AI: an interdisciplinary perspective. In: Steffen, B. (ed.) Bridging the Gap Between AI and Reality. AISoLA 2024. LNCS, vol. 15217, pp. 35–39. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-75434-0_3
8. Kerstan, S., et al.: Responsible and trusted AI: an interdisciplinary perspective (2025). In: Steffen, B. (ed.) Bridging the Gap Between AI and Reality. AISoLA 2025. LNCS, vol. 16220, pp. 141–145. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-07132-3_9
9. Langer, M., Lazar, V., Baum, K.: On the complexities of testing for compliance with human oversight requirements in AI regulation. In: Steffen, B. (ed.) Bridging the Gap Between AI and Reality. AISoLA 2025. LNCS, vol. 16220, pp. 160–169. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-07132-3_11
10. Laux, J., Ruschemeier, H.: Automation bias in the AI Act: on the legal implications of attempting to de-bias human oversight of AI. Eur. J. Risk Regul. **16**(4), 1519–1534 (2025). https://doi.org/10.1017/err.2025.10033
11. Seet, I., Bozenhard, J., Ostermann, S.: Enhancing AI interpretability and safety through localised architectures. In: Steffen, B. (ed.) AISoLA 2026, LNCS, vol. 17146, pp. xx–yy. Springer, Cham (2027)
12. Sterz, S., et al.: On the quest for effectiveness in human oversight: interdisciplinary perspectives. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 2495–2507. Association for Computing Machinery, New York (2024). https://doi.org/10.1145/3630106.3659051
13. Wagner, B.: Ethics as an escape from regulation: from "ethics-washing" to ethics-shopping? In: Bayamlıoğlu, E., Baraliuc, I., Janssens, L., Hildebrandt, M. (eds.) Being Profiled: Cogitas Ergo Sum. 10 Years of Profiling the European Citizen, pp. 84–89. Amsterdam University Press, Amsterdam (2018). https://doi.org/10.1515/9789048550180-016
14. Wagner, M., Bhimaraju, S., Borg, M., Runeson, P., Lee, E.A.: Compliance-by-design with Lingua Franca: toward meeting EU AI Act requirements. In: Steffen, B. (ed.) AISoLA 2026, LNCS, vol. 17146, pp. xx–yy. Springer, Cham (2027)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

