

# Trustworthiness and Responsibility in AI – Causality, Learning, and Verification

Vaishak Belle<sup>\*1</sup>, Hana Chockler<sup>\*2</sup>, Shannon Vallor<sup>\*3</sup>,  
Kush R. Varshney<sup>\*4</sup>, Joost Vennekens<sup>\*5</sup>, and Sander Beckers<sup>†6</sup>

- 1 University of Edinburgh, GB. vaishak@ed.ac.uk
- 2 King’s College London, GB. hana.chockler@kcl.ac.uk
- 3 University of Edinburgh, GB. svallor@ed.ac.uk
- 4 IBM Research – Yorktown Heights, US. krvarshn@us.ibm.com
- 5 KU Leuven, BE. joost.vennekens@kuleuven.be
- 6 University of Amsterdam, NL. srekcebrednas@gmail.com

---

## Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 24121 “Trustworthiness and Responsibility in AI – Causality, Learning, and Verification”. How can we trust autonomous computer-based systems? Since such systems are increasingly being deployed in safety-critical environments while interoperating with humans, this question is rapidly becoming more important. This Dagstuhl Seminar addressed this question by bringing together an interdisciplinary group of researchers from Artificial Intelligence (AI), Machine Learning (ML), Robotics (ROB), hardware and software verification (VER), Software Engineering (SE), and Social Sciences (SS); who provided different and complementary perspectives on responsibility and correctness regarding the design of algorithms, interfaces, and development methodologies in AI.

The purpose of the seminar was to initiate a debate around both theoretical foundations and practical methodologies for a “Trustworthiness & Responsibility in AI” framework that integrates quantifiable responsibility and verifiable correctness into all stages of the software engineering process. Such a framework will allow governance and regulatory practices to be viewed not only as rules and regulations imposed from afar, but instead as an integrative process of dialogue and discovery to understand why an autonomous system might fail and how to help designers and regulators address these through proactive governance.

In particular, we considered how to reason about responsibility, blame, and causal factors affecting the trustworthiness of the system. More practically, we asked what tools we can provide to regulators, verification and validation professionals, and system designers to help them clarify the intent and content of regulations down to a machine interpretable form. While existing regulations are necessarily vague, and dependent on human interpretation, we asked:

How should they now be made precise and quantifiable? What is lost in the process of quantification? How do we address factors that are qualitative in nature, and integrate such concerns in an engineering regime?

In addressing these questions, the seminar benefitted from extensive discussions between AI, ML, ROB, VER, SE, and SS researchers who have experience with ethical, societal, and legal aspects of AI, complex AI systems, software engineering for AI systems, and causal analysis of counterexamples and software faults.

**Seminar** March 17–22, 2024 – <https://www.dagstuhl.de/24121>

**2012 ACM Subject Classification** General and reference → General conference proceedings;  
Social and professional topics → Codes of ethics

**Keywords and phrases** responsible AI, trustworthy AI, causal machine learning, autonomous systems

**Digital Object Identifier** 10.4230/DagRep.14.3.75

---

\* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Trustworthiness and Responsibility in AI – Causality, Learning, and Verification, *Dagstuhl Reports*, Vol. 14, Issue 3, pp. 75–91

Editors: Vaishak Belle, Hana Chockler, Shannon Vallor, Kush R. Varshney, and Joost Vennekens



DAGSTUHL Dagstuhl Reports

REPORTS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Executive Summary


*Vaishak Belle (University of Edinburgh, GB, vaishak@ed.ac.uk)*

*Hana Chockler (King's College London, GB, hana.chockler@kcl.ac.uk)*

*Shannon Vallor (University of Edinburgh, GB, svallor@ed.ac.uk)*

*Kush R. Varshney (IBM Research – Yorktown Heights, US, krvarshn@us.ibm.com)*

*Joost Vennekens (KU Leuven, BE, joost.vennekens@kuleuven.be)*

License  Creative Commons BY 4.0 International license

© Vaishak Belle, Hana Chockler, Shannon Vallor, Kush R. Varshney, and Joost Vennekens

### Motivation and research area

How can we trust autonomous computer-based systems? Widely accepted definitions of autonomy take the view of being “independent and having the power to make your own decisions.” While many AI systems fit that description, they are often assembled by integrating many heterogenous technologies – including machine learning, symbolic reasoning or optimization – and correspondingly the notion of trust is fragmented and bespoke for the individual communities. However, given that automated systems are increasingly being deployed in safety-critical environments whilst interoperating with humans, a system would not only need to be able to reason about its actions, but a human user would need to additionally externally validate the behavior of the system. This seminar tackled the issue of trustworthiness and responsibility in autonomous systems by considering: notions of cause, responsibility and liability, and tools to verify the behavior of the resulting system.

In the last few years, we have observed increasing contributions in terms of manifestos, position papers, and policy recommendations issued by governments and learned societies, touching on interdisciplinary research involving AI ethics. This has primarily focused on “Fairness, Accountability, and Transparency” (FAT) with a majority focus on fairness, as individual and group fairness seems relatively easier to define precisely. On the other hand, DARPA’s XAI agenda has led to a resurgence in diagnostic explanations, but also ignited the question of interpretability and transparency in machine learning models, especially deep learning architectures. Our high-level motivation is that governance and regulatory practices can be viewed not only as rules and regulations imposed from afar but instead as an integrative process of dialogue and discovery to understand why an autonomous system might fail and how to help designers and regulators address these through proactive governance. But before that agenda can be approached, we need to resolve an important low-level question: how can we understand trust and responsibility of the components that make up an AI system? Autonomous systems will make ‘mistakes’, and accidents will surely happen despite best efforts. How should we reason about responsibility, blame, and causal factors affecting trustworthiness of the system? And if that is considered, what tools can we provide to regulators, verification and validation professionals, and system designers to help them clarify the intent and content of regulations down to a machine interpretable form? Existing regulations are necessarily vague, depending on the nuance of human interpretation for actual implementation. How should they now be made more precise and quantifiable?

The purpose of the seminar was to initiate a debate around these theoretical foundations and practical methodologies with the overall aim of laying the foundations for a “Trustworthiness & Responsibility in AI” framework – a framework for systems development methodology that integrates quantifiable responsibility and verifiable correctness into all stages of the software engineering process. As the challenge, by nature, is multidisciplinary, addressing it must involve experts from different domains, working on creating a coherent,

jointly agreed framework. The seminar brought together researchers from Artificial Intelligence (AI), Machine Learning (ML), Robotics (ROB), hardware and software verification (VER), Software Engineering (SE) and the Humanities (HUM), especially Philosophy (PHI), who provided different and complementary perspectives on responsibility and correctness regarding the design of algorithms, interfaces, and development methodologies in AI. From the outset, we wished to especially focus on understanding correctness for AI systems that integrate or utilize data-driven models (i.e., ML models), and to anchor our discussions by appealing to causality (CAU). Causality is widely used in the natural sciences to understand the effect of interventions on observed correlations, allowing scientists to design physical and biological laws. In ML too, increasingly there is recognition that conventional models focus on statistical associations, which can be misleading in critical applications demanding human-understandable explanations. The concept of causality is central to defining a notion of responsibility, and thus was a key point in our discussions.

### Directions identified and discussed

The seminar involved extensive discussions between AI, ML, ROB, VER, SE, PHI and HUM researchers who have experience in the following research topics:

- Ethical aspects of AI & ML algorithms: explainability and interpretability in AI algorithms, bias & fairness, accountability, moral responsibility. For example, there were discussions on large language models, their black box nature, and capabilities. There was also quite a bit of work on how explanations and causality might be related. Relevant papers that the participants identified included [10, 1].
- The moral and legal concepts of responsibility that underpin trust in autonomous systems, and how these relate to or can be aided by explainability or causal models of responsibility.
- Technical aspects of AI & ML algorithms: explainability and interpretability in AI algorithms, bias & fairness, accountability, quantification of responsibility. There were discussions regarding how visual input and human-in-the-loop models could provide the next frontier of explainability. Relevant papers identified by the participants included [11].
- Complex AI systems: robotics, reinforcement learning, integrated task and motion planning, mixed-initiative systems. There were discussions that suggest that incorporating high-level specifications from humans could considerably enhance the literature. Examples include recent loss function-based approaches and program induction-related directions for reinforcement policies [5, 4].
- Software engineering for AI systems: development methodologies, specification synthesis, formal verification of ML models, including deep learning architectures, software testing, causality. Outside of a range of recent approaches and looking at verifying the robustness properties of newer networks, there was a discussion on enhancing these perspectives by modeling trust. In fact, what exactly trustworthy machine learning might look like and the components it might involve were also discussed. Examples of relevant work include [9, 12, 8].
- Causal analysis of counterexamples and software faults. Causality was a central topic in the discussion, anchoring some of the key perspectives on how trustworthy AI, as well as explanations, could be addressed along with more nuanced notions such as harm. Following Joseph Halpern's talk on how harm could be formalized and related discussions, a number of relevant papers were identified as promising starting points for causal analysis [2, 3].

- Social aspects of AI & society, AI & law, AI & ethics. Examples of related literature include ideas on the types of ethical robots, the ironies of automation, and the notion of how empathy should apply to explainability among other related topics [7, 6]

### Open questions

Discussions between researchers from these different areas of expertise allowed us to explore topics at the intersection between the main areas, and to ask (and obtain partial answers on) the following questions:

- What sorts of explanations, and more generally, correctness notions are users looking for (or may be helpful for them)? How should these be generated and presented?
- How should we reason about responsibility, blame and causal factors affecting trustworthiness in individual components? How should that be expanded to the overall AI system?
- How do we define and quantify trust? Is trust achieved differently depending on the type of the user? Can trust in AI be achieved only using technology, or do we need societal changes?
- How do users reason about and handle responsibility, blame and cause in their day-to-day activities, and how do we interface those concepts with that of the AI system?
- Do our notions of responsibility and explanations increase user's trust in the technology?
- Who are the users of the technology? We envision different types of users, from policy makers and regulators to developers of the technology, to laypeople – the end-users. Should we differentiate the type of analysis for different categories of users?
- What tools can we provide to regulators, verification and validation professionals and system designers to help them clarify the intent and content of regulations down to a machine interpretable form?
- What tools are available to verify ML components, and do they cover the scope of “correct behavior” as understood by users and regulators?
- What SE practices are relevant for interfacing, integrating and challenging the above notions?
- How can properties of AI systems that are of interest be expressed in languages that lend themselves to formal verification or quantitative analysis?
- What kinds of user interfaces are needed to scaffold users to scrutinise the way AI systems operate?
- What frameworks are needed to reason about blame and responsibility in AI systems?
- How do we integrate research in causal structure learning with low-level ML modules used in robotics?
- How do we unify tools from causal reasoning and verification for assessing the correctness of complex AI systems?
- What challenges arise in automated reasoning and verification when considering the above mixed-initiative systems?
- Given a falsification of a specification, what kind of automated diagnosis, proof-theoretic and causal tools are needed to identify problematic components?
- How broadly will counterfactual reasoning (i.e., “what-if” reasoning) be useful to tackle such challenges?

**References**

- 1 Lisanne Bainbridge. Ironies of automation. *Automatica*, 19, 1983.
- 2 Sander Beckers, Hana Chockler, and Joseph Halpern. A causal analysis of harm. *Advances in Neural Information Processing Systems*, 35:2365–2376, 2022.
- 3 Ilan Beer, Shoham Ben-David, Hana Chockler, Avigail Orni, and Richard Trefer. Explaining counterexamples using causality. *Formal Methods in System Design*, 40:20–40, 2012.
- 4 Vaishak Belle and Andreas Bueff. Deep inductive logic programming meets reinforcement learning. In *The 39th International Conference on Logic Programming*. Open Publishing Association, 2023.
- 5 Craig Innes and Subramanian Ramamoorthy. Elaborating on learned demonstrations with temporal logic specifications. *arXiv preprint arXiv:2002.00784*, 2020.
- 6 William Kidder, Jason D’Cruz, and Kush R Varshney. Empathy and the right to be an exception: What llms can and cannot do. *arXiv preprint arXiv:2401.14523*, 2024.
- 7 Bran Knowles, Jason D’Cruz, John T. Richards, and Kush R. Varshney. Humble ai. *Commun. ACM*, 66(9):73–79, aug 2023.
- 8 Ekaterina Komendantskaya and Guy Katz. Towards a certified proof checker for deep neural network verification. *Logic-Based Program Synthesis and Transformation*, page 198.
- 9 Madsen and Gregor. Measuring human-computer trust. In *11th Australasian Conference on Information Systems*, 2000.
- 10 James H. Moor. Four types of ethical robot. *Philosophy Now*, 2009.
- 11 Spinner, Schlegel, Schäfer, and El-Assady. explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE Trans. on Visualization and Computer Graphics*, 2020.
- 12 Kush R. Varshney. *Trustworthy Machine Learning*.

## 2 Table of Contents

### Executive Summary

<i>Vaishak Belle, Hana Chockler, Shannon Vallor, Kush R. Varshney, and Joost Vennekens</i> . . . . .	76
--	----

### Overview of Talks

Responsible AI Control <i>Nadisha-Marie Aliman</i> . . . . .	81
Moral Responsibility for AI Systems <i>Sander Beckers</i> . . . . .	81
Are We Correct To Ascribe Conversational Agency to LLM-Based Chatbots? <i>Jan M. Broersen</i> . . . . .	82
The Simpson and Bias Amplification Paradoxes <i>Yanai Elazar</i> . . . . .	82
Trustworthy Autonomy <i>Michael Fisher</i> . . . . .	82
AI Safety and The EU AI Act <i>Leon Kester</i> . . . . .	83
Specification-based Falsification and Repair of DNN Controllers <i>Stefan Leue</i> . . . . .	84
Kaspar Causally Explains <i>Mohammad Reza Mousavi</i> . . . . .	84
BRIO: A Bias and Risk Assessment Tool <i>Giuseppe Primiero</i> . . . . .	85
Trustworthiness for Medical Diagnostics: What and How? <i>Ajitha Rajan</i> . . . . .	86
Some Challenges on the Path to Certifying AI-Enabled Autonomy <i>Subramanian Ramamoorthy</i> . . . . .	86
A Causal Analysis of Harm <i>Joseph Halpern</i> . . . . .	87
AI Governance and Agential Power: How Can We Make Systems Answer? <i>Shannon Vallor</i> . . . . .	88

### Summary of Breakout Session on “AI in 20 years: 6 Ambitions”

AI Broadens Out . . . . .	88
Knowing <i>How</i> to Use AI vs Knowing <i>About</i> AI . . . . .	89
Greater Professionalisation of the AI Community . . . . .	89
Effective and Balanced AI Regulation . . . . .	89
Standardisation of Responsible AI Design and Development Practices . . . . .	90
A Mature and Collaborative AI Culture . . . . .	90

<b>Participants</b> . . . . .	91
-------------------------------	----

## 3 Overview of Talks

### 3.1 Responsible AI Control

*Nadisha-Marie Aliman (Utrecht University, NL)*

License © Creative Commons BY 4.0 International license  
© Nadisha-Marie Aliman

This talk on “Responsible AI Control” elucidates why when confronted with inconsistent human-level AI/AGI/ASI achievement claims, AI researchers can respond responsibly by rigorously formulating scientific impossibility statements (as has e.g. analogously already been practiced in the Large Hadron Collider Safety case) and developing scientific evaluation frameworks that constrain those achievement claims. For example, related work already introduced diverse AI-related impossibility statements grounded in thermodynamical, biological, cognitive-science-linked and hardware-verification-related explanations. The talk introduces a novel epistemic paradigm termed “cyborgnetic invariance” that entails multiple new impossibility statements. For illustration, a simple new scientific evaluation framework for automated quantity superintelligence achievement claims is discussed. Simply put, the framework extends the tasks of interest for ASI assessment to asymmetrical intelligence/creativity/consciousness levels of civilizations. The cyborgnetic invariance paradigm consists of two postulates: invariance of maximal quantity superintelligence and impossibility of reliable stupidity-based construction. Thereby, asymmetrically measurable intelligence/creativity/consciousness is non-algorithmic (but it involves physical computation). To build an AGI “from scratch” is at least as hard as physically building a new baby universe. To build such a non-controllable but value-alignable creature, humanity would have to at least first become superintelligent in relation to its current self. In the meantime, one can build controllable but non-value-alignable “AI” tools encapsulated in human-centered units of cyborgnetic control loops to deepen critical thinking and broaden human creativity via so-called artificial EM repeaters, EDM miners and EDE generators in order to tackle global risks. The talk ends by stressing that present-day “AI” should not be underestimated either since its use and misuse is currently linked to an “AI”-related epistemic security threat landscape which subsumes multiple novel global/existential risks for a civilization like present-day humanity.

### 3.2 Moral Responsibility for AI Systems

*Sander Beckers (University of Amsterdam, NL)*

License © Creative Commons BY 4.0 International license  
© Sander Beckers

**Main reference** Sander Beckers: “Moral responsibility for AI systems”, in Proc. of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., 2024.

**URL** <https://dl.acm.org/doi/10.5555/3666122.3666312>

As more and more decisions that have a significant ethical dimension are being outsourced to AI systems, it is important to have a definition of moral responsibility that can be applied to AI systems. Moral responsibility for an outcome of an agent who performs some action is commonly taken to involve both a causal condition and an epistemic condition: the action should cause the outcome, and the agent should have been aware – in some form or other – of the possible moral consequences of their action. In this talk I present a formal definition of both conditions within the framework of causal models. I compare my approach to the existing approaches of Braham and van Hees (BvH) and of Halpern and Kleiman-Weiner (HK). I then generalize my definition into a degree of responsibility.

### 3.3 Are We Correct To Ascribe Conversational Agency to LLM-Based Chatbots?

*Jan M. Broersen (Utrecht University, NL)*

**License** © Creative Commons BY 4.0 International license  
© Jan M. Broersen

To trust AIs and give correct assessments of responsibility in situations where they interact with humans, we need to understand their agency. We need to understand if their agency differs from human agency, and if so, what the differences are. For this talk, I will focus on the conversational agency of LLMs.

### 3.4 The Simpson and Bias Amplification Paradoxes

*Yanai Elazar (AI2 – Seattle, US)*

**License** © Creative Commons BY 4.0 International license  
© Yanai Elazar

**Joint work of** Preethi Seshadri, Sameer Singh, Yanai Elazar

**Main reference** Preethi Seshadri, Sameer Singh, Yanai Elazar: “The Bias Amplification Paradox in Text-to-Image Generation”, CoRR, Vol. abs/2308.00755, 2023.

**URL** <https://doi.org/10.48550/ARXIV.2308.00755>

The Simpson’s paradox (and the Sex Bias in Graduate Admission) is a classic example that illustrates the challenges in evaluation data – originating from the real world or AI models. I will introduce Simpson’s paradox briefly and how alternative views of the same data can lead to different conclusions. Then, I will describe our recent work on the Bias Amplification Paradox in the text-to-image models. I argue that bias amplification is highly dependent on the evaluation procedure and sensitive to confounding factors that influence the implications of naive evaluations.

### 3.5 Trustworthy Autonomy

*Michael Fisher (University of Manchester, GB)*

**License** © Creative Commons BY 4.0 International license  
© Michael Fisher

**Joint work of** Dennis, Louise, A., Fisher, Michael

**Main reference** Louise A. Dennis, Michael Fisher: “Verifiable Autonomous Systems: Using Rational Agents to Provide Assurance about Decisions Made by Machines”, Cambridge University Press, 2023.

**URL** <https://doi.org/10.1017/9781108755023>

Autonomous Systems have the ability to make their own decisions and potentially to take their own actions, and to do both without direct human intervention. When we deploy these systems, especially in important or even critical situations, do we know what this use of autonomy will result in? And can we trust it to always work “well”?

I discuss issues around the development of Trustworthy Autonomy, including reliability (does it work?), beneficiality (is it working for our benefit?), and the verification of these both before and after deployment.

This will highlight that not only are there distinct forms of AI, each with different benefits and drawbacks, but that combining these in a heterogeneous way can be beneficial. Such combinations are alternatively termed “hybrid” or “neuro-symbolic” systems.



By utilising a specific hybrid “agent” architecture, where our agents are logical and able to represent and implements concepts such as “belief” and “intention”, we are able to expose the reasons for decisions – i.e: “why did you do that”. Furthermore, we can formally verify this agent decision-making to prove whether the agent, and hence the autonomous system, will never choose to do anything “bad”.

This exposure of decision-making processes also has an impact on the broader issues of these autonomous systems, for example around ethical decision-making and responsibility.

## References

- 1 Bremner, Paul, Dennis, Louise A., Fisher, Michael, Winfield, Alan F.T.: On Proactive, Transparent, and Verifiable Ethical Reasoning for Robots. Proceedings of the IEEE pp. 1–21 (2019). <https://doi.org/10.1109/JPROC.2019.2898267>
- 2 Chatila, Raja, Dignum, Virginia, Fisher, Michael, Giannotti, Fosca, Morik, Katharina, Russell, Stuart, Yeung, Karen. Trustworthy AI. Pages 13–39 of: Braunschweig, Bertrand, and Ghallab, Malik (eds), *Reflections on Artificial Intelligence for Humanity*. Springer, 2021. [https://doi.org/10.1007/978-3-030-69128-8\\_2](https://doi.org/10.1007/978-3-030-69128-8_2)
- 3 Dennis, Louise A., Bentzen, Martin, M., Lindner, Felix, Fisher, Michael: Verifiable Machine Ethics in Changing Contexts. Proceedings of the AAAI Conference on Artificial Intelligence **35**(13), 11470–11478 (May 2021), <https://ojs.aaai.org/index.php/AAAI/article/view/17366>
- 4 Dennis, Louise, A., Fisher, Michael, Slavkovik, Marija, Webster, Matthew, P.: Formal Verification of Ethical Choices in Autonomous Systems. Robotics and Autonomous Systems **77**, 1–14 (2016). <https://doi.org/10.1016/j.robot.2015.11.012>
- 5 Dennis, Louise, A., Fisher, Michael, Webster, Matthew, P., Bordini, Rafael, H.: Model Checking Agent Programming Languages. Automated Software Engineering **19**(1), 5–63 (2012). <https://doi.org/10.1007/S10515-011-0088-X>
- 6 Dennis, Louise, A., Fisher, Michael: Verifiable Autonomous Systems – Using Rational Agents to Provide Assurance about Decisions Made by Machines. Cambridge University Press, 2023. <https://doi.org/10.1017/9781108755023>
- 7 Fisher, Michael, Mascardi, Viviana, Rozier, Kristin Yvonne, Schlingloff, Bernd-Holger, Winikoff, Michael, and Yorke-Smith, Neil. Towards a Framework for Certification of Reliable Autonomous Systems. *Autonomous Agents and MultiAgent Systems*, **35**(1), 2021. <https://doi.org/10.1007/s10458-020-09487-2>
- 8 Koeman, Vincent, Dennis, Louise, Webster, Matthew, P., Fisher, Michael, Hindriks, Koen. The “Why Did You Do That?” Button: Answering Why-Questions for End Users of Robotic Systems. In *Engineering Multi-Agent Systems*, pages 152–172. Springer, 2020. [https://doi.org/10.1007/978-3-030-51417-4\\_8](https://doi.org/10.1007/978-3-030-51417-4_8)

## 3.6 AI Safety and The EU AI Act

Leon Kester (*TNO Netherlands – The Hague, NL*)

License © Creative Commons BY 4.0 International license  
© Leon Kester

Joint work of Nadisha-Marie Aliman, Leon Kester

Main reference Nadisha-Marie Aliman, Leon Kester: “4. Moral programming”, pp. 63–80, Wageningen Academic, 2022.

URL [https://doi.org/10.3920/978-90-8686-922-0\\_4](https://doi.org/10.3920/978-90-8686-922-0_4)

Risk Management aiming at harm minimization and systemic risk mitigation is required for Trustworthy AI compatible with the EU AI Act. Moreover, for a meaningful AI control, there is a need for a rigorous harm model such as e.g. via Augmented Utilitarianism to

safely encapsulate AI systems in a human-centric socio-technological feedback-loop. In this talk, I also explain why one should not overestimate present-day AI since it is linked to a comprehension bottleneck. For instance, as in science, the ethical value alignment among people can include the creation of new unknown better chains of explanations that present-day AI cannot understand. However, taking the case of so-called “deepfake science attacks” as illustration for a systemic risk, I discuss why one should also not underestimate present-day AI. Here, by way of example, it becomes clear why in a risk-aware approach, instead of asking “was this contribution generated by present-day AI or by a human?” a better suited question would be “does this material encode a better new scientific chain of explanations in comparison to the ones that are already available?”. In conclusion, a future risk-aware Trustworthy AI research which is compatible with the EU AI Act should include the AI-aided augmentation of both human critical thinking and human scientific creativity.

### References

- 1 Aliman, Nadisha-Marie and Kester, Leon. 4. Moral Programming. in: *Moral design and technology*, Wageningen Academic, 63–80, 2022.
- 2 Aliman, Nadisha-Marie and Kester, Leon. VR, Deepfakes and epistemic security. 2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), IEEE, 93–98, 2022.

## 3.7 Specification-based Falsification and Repair of DNN Controllers

*Stefan Leue (Universität Konstanz, DE)*

**License** © Creative Commons BY 4.0 International license  
© Stefan Leue

**Joint work of** Fabian Bauer-Marquart, David Boetius, Stefan Leue, Christian Schilling  
**Main reference** Fabian Bauer-Marquart, David Boetius, Stefan Leue, Christian Schilling: “SpecRepair: Counter-Example Guided Safety Repair of Deep Neural Networks”, in Proc. of the Model Checking Software – 28th International Symposium, SPIN 2022, Virtual Event, May 21, 2022, Proceedings, Lecture Notes in Computer Science, Vol. 13255, pp. 79–96, Springer, 2022.  
**URL** [https://doi.org/10.1007/978-3-031-15077-7\\_5](https://doi.org/10.1007/978-3-031-15077-7_5)

I sketch the SpecRepair approach towards specification-based repair of Deep Neural Networks. It implements a counterexample-guided repair approach which includes optimization-based counterexample finding, counterexample-based retraining of the network and finally the certification of the desired property by a complete DNN verifier.

## 3.8 Kaspar Causally Explains

*Mohammad Reza Mousavi (King’s College London, GB)*

**License** © Creative Commons BY 4.0 International license  
© Mohammad Reza Mousavi

The Kaspar robot has been used with great success to work as an education and social mediator with children with autism spectrum disorder. Enabling the robot to automatically generate causal explanations is key to enrich the interaction scenarios for children and promote trust in the robot. In our research, we analysed the human-robot interactions in which causal explanations can contribute substantially to the child’s understanding of Visual Perspective Taking (VPT). The results helped us identify multiple interaction categories

that benefit from causal explanation [3]. Subsequently, we developed a theory of causal explanation to be embedded in Kaspar and built a causal model and an analysis method to calculate causal explanations. We implemented our method to automatically generate causal explanations spoken by Kaspar [2]. We validated our explanations for user satisfaction and brought the robot to a school. The results revealed that children improved their VPT abilities significantly when the robot provided causal explanations [1].

## References

- 1 M. Sarda Gou, G. Lakatos, P. Holthaus, B. Robins, S. Moros, L. Jai Wood, H. Araujo, C. A. E. deGraft-Hanson, M. R. Mousavi, F. Amirabdollahian. Kaspar Explains: The Effect of Causal Explanations on Visual Perspective Taking Skills in Children with Autism Spectrum Disorder. Proceedings of the 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2023), IEEE, 2023.
- 2 H. Araujo, P. Holthaus, M. Sarda Gou, G. Lakatos, G. Galizia, L. Wood, B. Robins, M.R. Mousavi, and F. Amirabdollahian. Kaspar Causally Explains, Proceedings of the 14th International Conference on Social Robotics, Lecture Notes in Computer Science, Springer, 2022.
- 3 M. Sarda Gou, G. Lakatos, P. Holthaus, L. Jai Wood, M.R. Mousavi, B. Robins, and F. Amirabdollahian. Towards understanding causality – a retrospective study of using explanations in interactions between a humanoid robot and autistic children. Proceedings of the 31st IEEE International Conference on Robot & Human Interactive Communication (RO-MAN 2022), IEEE, 2022.

## 3.9 BRIO: A Bias and Risk Assessment Tool

*Giuseppe Primiero (University of Milan, IT)*

**License** © Creative Commons BY 4.0 International license  
© Giuseppe Primiero

**Joint work of** Greta Coraglia, Fabio Aurelio D’Asaro, Francesco Antonio Genco, Davide Giannuzzi, Davide Posillipo, Giuseppe Primiero, Christian Quaggio

**Main reference** Greta Coraglia, Fabio Aurelio D’Asaro, Francesco Antonio Genco, Davide Giannuzzi, Davide Posillipo, Giuseppe Primiero, Christian Quaggio: “BRIOxAlkemy: a Bias Detecting Tool”, in Proc. of the 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AI\*IA 2023), Rome, Italy, November 6, 2023, CEUR Workshop Proceedings, Vol. 3615, pp. 44–60, CEUR-WS.org, 2023.

**URL** <https://ceur-ws.org/Vol-3615/paper4.pdf>

Phenomena of bias by AI systems based on machine learning methods are well known, and largely discussed in the literature. A variety of tools are being developed to assess these undesirable behaviours. In this short talk I present a bias and risk assessment tool [5] developed within the BRIO Research Project (<https://sites.unimi.it/brio/>). The tool is based on various formal logics developed in [1, 2, 3, 4]. The tool works on the I/O data of a ML system remaining agnostic on the model itself. The user can choose one of two distinct modules to evaluate either the difference in behaviour that the model displays on outputs produced by subclasses of inputs, or to evaluate against a desirable output. The type of distance and the threshold for admissible distance from the target distribution can also be selected. The result is a set of all the features and combinations thereof that produce violations with respect to the target distribution. These features can be fed into a risk function which computes an overall value weighting them on parameters such as size of the population and number of features involved, mapping naturally into notions of group and individual fairness.

## References

- 1 F. D’Asaro, G. Primiero, *Probabilistic typed natural deduction for trustworthy computations*, in: Proceedings of the 22nd International Workshop on Trust in Agent Societies (TRUST2021@ AAMAS), 2021.
- 2 Giuseppe Primiero, Fabio Aurelio D’Asaro: Proof-checking Bias in Labeling Methods. BEWARE@AI\*IA 2022: 9-19
- 3 Fabio Aurelio D’Asaro, Giuseppe Primiero: Checking Trustworthiness of Probabilistic Computations in a Typed Natural Deduction System. CoRR abs/2206.12934 (2022)
- 4 Francesco A. Genco, Giuseppe Primiero: A Typed Lambda-Calculus for Establishing Trust in Probabilistic Programs. CoRR abs/2302.00958 (2023)
- 5 Greta Coraglia, Fabio Aurelio D’Asaro, Francesco A. Genco, Davide Giannuzzi, Davide Posillipo, Giuseppe Primiero and Christian Quaggio, *BRIOxAlkemy: A Bias detecting tool*, in Proceedings of the 2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AI\*IA 2023), pp. 44–60. 2024.

### 3.10 Trustworthiness for Medical Diagnostics: What and How?

*Ajitha Rajan (University of Edinburgh, GB)*

License  Creative Commons BY 4.0 International license  
© Ajitha Rajan

The rapidly advancing field of Explainable Artificial Intelligence (XAI) aims to tackle the issue of trust regarding the use of complex black-box deep learning models in real-world applications. Existing post-hoc XAI techniques have recently been shown to have poor performance on medical data, producing unreliable explanations which are infeasible for clinical use. To address this, we propose an ante-hoc approach based on concept bottleneck models that introduces for the first time clinical concepts into the classification pipeline, allowing the user valuable insight into the decision-making process. On a large public dataset of chest X-rays and associated medical reports, we focus on the binary classification task of lung cancer detection. Our approach yields improved classification performance on lung cancer detection when compared to baseline deep learning models ( $F1 > 0.9$ ), while also generating clinically relevant and more reliable explanations than existing techniques. We evaluate our approach against post-hoc image XAI techniques LIME and SHAP, as well as CXR-LLaVA, a recent textual XAI tool that operates in the context of question answering on chest X-rays.

### 3.11 Some Challenges on the Path to Certifying AI-Enabled Autonomy

*Subramanian Ramamoorthy (University of Edinburgh, GB)*

License  Creative Commons BY 4.0 International license  
© Subramanian Ramamoorthy  
URL <https://web.inf.ed.ac.uk/tas>

The increasing use of AI in autonomous systems has made the problem of certifying such systems hard. While the difficulties are associated with broad questions of AI safety, AI-enabled autonomous systems raise certain uniquely challenging questions. This includes the

problem of characterising the dynamic behaviour of adaptive systems in open and human-centred environments. This talk surveys work done within the UKRI Research Node on Trustworthy Autonomous Systems Governance and Regulation (<https://web.inf.ed.ac.uk/tas>), with a focus on the AV certification case study. Within this, we outline results from work on specification gaps [1], scenario generation and sampling with multiple representations [2], [3], and active learning methods for risk-sensitive design [4].

### References

- 1 Abeywickrama DB, Bennaceur A, Chance G, Demiris Y, Kordoni A, Levine M, Moffat L, Moreau L, Mousavi MR, Nuseibeh B, Ramamoorthy S. On specifying for trustworthiness. *Communications of the ACM*. 2023 Dec 21;67(1):98-109.
- 2 Innes C, Ramamoorthy S. Testing rare downstream safety violations via upstream adaptive sampling of perception error models. In *2023 IEEE International Conference on Robotics and Automation (ICRA) 2023* May 29 (pp. 12744-12750).
- 3 Innes C, Ireland A, Lin Y, Ramamoorthy S. Anticipating Accidents through Reasoned Simulation. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems 2023* Jul 11 (pp. 1-11).
- 4 Corso A, Katz S, Innes C, Du X, Ramamoorthy S, Kochenderfer MJ. Risk-driven design of perception systems. *Advances in Neural Information Processing Systems*. 2022 Dec 6;35:9894-906.

## 3.12 A Causal Analysis of Harm

*Joseph Halpern (Cornell University – Ithaca, US)*

License  Creative Commons BY 4.0 International license  
© Joseph Halpern

Joint work of Sander Beckers, Hana Chockler, Joseph Halpern

It has proved notoriously difficult to define harm. Indeed, it has been claimed that the notion of harm is a “Frankensteinian jumble” that should be replaced by other well-behaved notions. On the other hand, harm has become increasingly important as concerns about the potential harms that may be caused by AI systems grow. For example, the European Union’s draft AI act mentions “harm” over 25 times and points out that, given its crucial role, it must be defined carefully.

I start by defining a qualitative notion of harm that uses causal models and is based on a well-known definition of actual causality. The key features of the definition are that it is based on contrastive causation and uses a default utility to which the utility of actual outcomes is compared. I show that our definition is able to handle the problematic examples from the literature. I extend the definition to a quantitative notion of harm, first in the case of a single individual, and then for groups of individuals. I show that the “obvious” way of doing this (just taking the expected harm for an individual and then summing the expected harm over all individuals) can lead to counterintuitive or inappropriate answers, and discuss alternatives, drawing on work from the decision-theory literature.

### References

- 1 Beckers, S., Chockler, H., and Halpern, J.Y. A Causal Analysis of Harm, *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, pp. 2365–2376, 2022. <https://dl.acm.org/doi/abs/10.5555/3600270.3600442>

- 2 Beckers, S., Chockler, H., and Halpern, J.Y. Quantifying harm, Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023), pp. 363–371. 2023. <https://www.ijcai.org/proceedings/2023/0041>

### 3.13 AI Governance and Agential Power: How Can We Make Systems Answer?

*Shannon Vallor (University of Edinburgh, GB)*

**License** © Creative Commons BY 4.0 International license  
© Shannon Vallor

**Joint work of** Shannon Vallor and Bhargavi Ganesh

**Main reference** Shannon Vallor, Bhargavi Ganesh: “Artificial intelligence and the imperative of responsibility: Reconceiving AI governance as social care”. In M. Kiener (ed.), *The Routledge Handbook of Philosophy of Responsibility* (Routledge: NY), p. 395-406, 2023

**URL** <https://doi.org/10.4324/9781003282242>

The accelerating development of artificial intelligence (AI) systems has generated acute and interlinked challenges for social trust, responsibility ascription, and governance. While today’s AI tools lack the type of agency that can bear responsibility, they are deployed in ways that create novel configurations and social appearances of agential power. That is, they allow new things to be done by us, for us, and to us, in ways that do not easily fit our existing practices for governing moral and legal responsibility. This is commonly referred to as the problem of AI “responsibility gaps”.

We confront this challenge by framing normative responsibility for AI actions in a new way: not as a metaphysical fact about agents to be discovered, nor a set of criteria that responsible agents must fully satisfy, but as a set of constructed social practices in the exercise of agential power, that make agential powers answerable for their impact on others’ vulnerabilities and interests. The construction and use of such practices for new or changed agential powers is an essential precondition of social trust.

Drawing from historical examples in steamboat engineering, consumer finance, and environmental governance, we highlight how responsibility gaps have generated the moral and political imperative to construct new forms of responsible agency and governance to balance novel agential powers, of which AI is merely the latest iteration. We conclude with observations about the two general classes of available AI governance strategies, agential obligation and agential constraint, that must be balanced in order to secure public trust in AI technologies that represent new agential powers.

## 4 Summary of Breakout Session on “AI in 20 years: 6 Ambitions”

### 4.1 AI Broadens Out

AI needs to broaden out in two directions; within AI *and* without. Within AI, it needs to be taught that learning AI requires more than just machine learning; other techniques are needed to complement ML and to continue growth and exploration beyond the existing paradigm.

AI must also broaden to incorporate necessary knowledge from other fields: philosophy, law, neuroscience, HCI and design, for example. AI researchers will increasingly need critical thinking skillsets that take them beyond technical work and allow for better evaluation of AI methods and applications.

Suitably broadened, AI itself needs to be a fundamental “core” area of knowledge for university graduates; in the future, understanding how the world works will be impossible without some understanding of AI and its uses. Could AI Studies be a core educational requirement? Something like this has been tried in the Netherlands before, using a broad interdisciplinary model (for example, in Utrecht in the 90s).

In 20 years time, we hope to see first year interdisciplinary courses like “Intro to AI” that all students can take – but how to overcome institutional and disciplinary resistance/tradition? Some countries are very resistant to curriculum change, and this would require retraining of academic staff in universities across disciplines. How can we make this kind of change possible? We can learn from the successes and failures of other interdisciplinary studies created in the last 50 years: Science and Technology Studies, Environmental Studies, Bioethics.

## 4.2 Knowing *How* to Use AI vs Knowing *About* AI

Both are going to be essential. We might see AI trade schools in 20 years, to teach the areas that create new, attractive, well-paying jobs without needing theoretical foundations. Potential career paths include:

- AI User Specialist (domain specific)
- AI Data Quality Officer
- AI Prompt Engineer
- AI Error and Bias Controller
- AI Ombudsperson
- AI UI Specialist

## 4.3 Greater Professionalisation of the AI Community

Professionalisation and accreditation can be mechanisms to prescribe certain educational requirements and also diversify the field into a more balanced set of specializations. We might also consider the “Nuclear Option”: using licensing/certification of AI Professionals for safety-critical industries and applications, in the way that we have seen work in medicine and certain areas of engineering. We do this in medicine and many areas of engineering because they are highly dangerous professions as well as beneficial ones. AI is now *also* a highly dangerous (and beneficial) profession.

## 4.4 Effective and Balanced AI Regulation

Regulation can advance AI further in a number of ways, beyond just making AI safer for people to engage with. It can serve as another incentive for broadening the field of AI – as with privacy regulation, it can require fulfillment of certain roles and create incentives for corporations to invest in more types of AI expertise. Regulators and professional societies might be able to coordinate incentives strategically if not captured by industry. For example, testing and licensing could be an incentive embedded in procurement standards, liability caps, etc. for safety-critical AI development or application. Regulation could help drive the acquisition and normalization of these areas of expertise and more:

- AI Ethics
- AI Law and Policy
- AI Security
- AI Safety
- AI Privacy
- AI Auditor

We note that If no one is willing to take responsibility for an AI system in a high-stakes environment, it arguably should not be deployed in that domain – the burden needs to be on organisations to demonstrate that they have assigned specific and adequate duties to competent, empowered and accountable professional(s).

#### **4.5 Standardisation of Responsible AI Design and Development Practices**

With a more professionalised and well-governed AI ecosystem, we expect to see better ways to standardise the conversion of responsible policy choices into design and engineering choices – right now that falls on AI Developers that aren't trained to formalize values and either aren't doing it or end up doing it poorly.

Professionalisation and standardisation of ethical design principles and processes will also shield individual professionals from being unfairly held personally accountable for unavoidable harms/failures; AI Developers today are disincentivized to make explicit moral choices, for which they will then be personally on the hook if the outcome isn't ideal. No technology can be made risk-free and we need to shield developers from liability or at least cap their liability for making responsible choices that follow best professional practice.

#### **4.6 A Mature and Collaborative AI Culture**

In 20 years we hope to see AI research, learner and practitioner environments that embody openness to interdisciplinarity, effective translation, co-construction and communication of AI knowledge, and intellectual charity. We can start early by moving the learning of AI to earlier phases, before the “hard/soft” skills division (which itself must be challenged and rethought in the next decades), beyond STEM/not STEM, so that AI is marked by a culture of shared intellectual community rather than knowledge hoarding and turf-defending.



## Participants

- Nadisha-Marie Aliman  
Utrecht University – NL
- Emma Beauxis-Aussalet  
VU Amsterdam – NL
- Sander Beckers  
University of Amsterdam – NL
- Vaishak Belle  
University of Edinburgh – GB
- Jan M. Broersen  
Utrecht University – NL
- Georgiana Caltais  
University of Twente –  
Enschede, NL
- Hana Chockler  
King’s College London – GB
- Jens Claßen  
Roskilde University – DK
- Sjur K. Dyrkolbotn  
West. Norway Univ. of Applied  
Sciences – Bergen, NO
- Yanai Elazar  
AI2 – Seattle, US
- Esra Erdem  
Sabanci University –  
Istanbul, TR
- Michael Fisher  
University of Manchester – GB
- Sarah Alice Gaggl  
TU Dresden – DE
- Leilani H. Gilpin  
University of California –  
Santa Cruz, US
- Gregor Goessler  
INRIA – Grenoble, FR
- Joseph Y. Halpern  
Cornell University – Ithaca, US
- Till Hofmann  
RWTH Aachen University – DE
- David Jensen  
University of Massachusetts –  
Amherst, US
- Leon Kester  
TNO Netherlands –  
The Hague, NL
- Ekaterina Komendantskaya  
Heriot-Watt University –  
Edinburgh, GB
- Stefan Leue  
Universität Konstanz – DE
- Joshua Loftus  
London School of Economics and  
Political Science – GB
- Mohammad Reza Mousavi  
King’s College London – GB
- Giuseppe Primiero  
University of Milan – IT
- Ajitha Rajan  
University of Edinburgh – GB
- Subramanian Ramamoorthy  
University of Edinburgh – GB
- Kilian Rückschloß  
LMU München – DE
- Judith Simon  
Universität Hamburg – DE
- Luke Stark  
University of Western Ontario –  
London, CA
- Daniel Susser  
Cornell University – Ithaca, US
- Shannon Vallor  
University of Edinburgh – GB
- Kush R. Varshney  
IBM Research –  
Yorktown Heights, US
- Joost Vennekens  
KU Leuven – BE
- Felix Weitkämper  
LMU München – DE

