HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)

doi:10.3233/FAIA240197

Human-AI Engineering for Adults

André MEYER-VITALI [a,1], and Wico MULDER [b]
[a] *DFKI, Saarbrücken*
[b] *TNO, Groningen*
ORCiD ID: André Meyer-Vitali https://orcid.org/0000-0002-5242-1443, Wico Mulder https://orcid.org/0000-0002-8607-0055

Abstract. The engineering of reliable and trustworthy AI systems needs to mature. While facing unprecedented challenges, there is much to be learned from other engineering disciplines. We focus on the five pillars of (i) Models & Explanations, (ii) Causality & Grounding, (iii) Modularity & Compositionality, (iv) Human Agency & Oversight, and (v) Maturity Models. Based on these pillars, a new AI engineering discipline might emerge, which we aim to support using corresponding methods and tools for 'Trust by Design'. A use case concerning mobility and energy consumption in an urban context is discussed.

Keywords. Software Engineering, Artificial Intelligence, Models, Modules, Trust, Causality, Robustness, Explainability, Agency, Multi-Agent Systems, Smart Cities, Human-Computer Interaction, Context-Aware Pervasive Systems, Maturity Model

1. Introduction

The current wave of Artificial Intelligence (AI) is characterised by Deep Learning [1, 2], Transformers [3, 4] and Large Foundation Models [5]. Whilst the impact of such systems touches almost all veins of our society, it seems that we are reaching the limits of controlled engineering of these large, highly interconnected, AI-based systems.

On the one hand, we see their complexity increase on an individual level, as well as on their connected dependency levels, whilst on the other hand, we see a growing lack of experience on the level of their design and engineering. The complexity of existing AI models is often beyond our understanding, and the methods and processes to ensure safety, reliability, and transparency are lacking. This poses serious risks at the level of trustworthiness, particularly when it comes to critical applications with significant physical, economic, or social impact. The AI systems used in such applications are required – for example by the European AI Act – to have been thoroughly designed, validated and certified according to well-defined criteria.

Recent developments in Generative AI are based on so-called 'Foundation Models', which can appear as Large Language Models (LLM) or as similar multi-modal models of still images, videos and others. The transformer architectures that generate these

[1]Corresponding Author: Dr. André Meyer-Vitali, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Stuhlsatzenhausweg 3, Saarland Informatics Campus D32, 66123 Saarbrücken, Germany; E-mail: andre.meyer-vitali@dfki.de.

models convert huge amounts of text or other media content into statistical models of co-occurrence of tokens (parts of words or other features).

Many are not aware that deep learning does not support a real understanding of problems. At a first glance, these models for generative AI seem to understand human language and creative expression. However, as they are uniquely based on producing probabilistic assemblies of tokens, they do not even understand language itself. There is no grammar involved or any form of semantics. They only reflect high-dimensional statistical correlations.

Great disillusionment set in as problems, such as insufficient internal representation of meaning (interpretability and transparency), susceptibility to changes in the input signal (robustness), lack of transferability to cases not covered by the data (generalisation) and, last but not least, the thirst for big data and processing itself (efficiency, adequacy, sustainability), became apparent. Some of these problems are a direct result of the massive use of deep learning black-box methods that depend solely on data [6].

To increase the grip and understanding of the outcomes of large neural models, new approaches combine data-oriented machine learning with symbolic conclusions and the explicit representation of knowledge [7, 8].

Such types of approaches are being advanced by the term 'Trusted AI'. Trusted AI aims to create a new generation of AI systems that guarantee functionality, allowing use even in critical applications [9]. Developers, domain experts, users, and regulators can rely on performance and reliability even for complex socio-technical systems. Trusted AI is characterised by a high degree of robustness, transparency, fairness, and verifiability, where the functionality of existing systems is in no way compromised, but actually enhanced.

Foundation models are not trustworthy, because they lack any kind of understanding of truth, facts, time, space, concepts, reasons, causes and effects. As they are not consistent, transparent, robust and reliable, it is very risky to trust them in critical applications. Even when they seem to give reasonable answers from time to time, it is impossible to predict when they will fail and start to hallucinate.

We need to stop reinventing the wheel; learning from scratch, but understanding nothing. Instead, we need to use existing knowledge, build on experiments and experience, formulate and validate new hypotheses and theories, in order to gain knowledge and insight at a higher level, and to explain why events happen, predictions are made and decisions or actions are taken. This requires a reinforced attention for engineering processes with an aim to improve scientific progress where one can stand on the shoulders of giants.

This paper therefore sets out five pillars of AI Engineering. Together, they form a supportive framework that fosters 'Trusted AI by Design'. Section 2 discusses the five pillars and section 3 provides a use case in a smart city context.

2. Trusted AI Engineering

There is a dilemma to overcome in building trustworthy AI systems [10, 11, 9]: on the one hand, we expect AI systems to decide autonomously and intelligently on our behalf, which requires agency and delegation; on the other hand, we require them to be predictable, verifiable, safe and accountable. Of course, there are limits to achieving all

these goals and to guarantee correctness under all circumstances and domains. Instead, there is a trade-off to be made between entirely predictable and correct versus plausible and adaptive behaviour. What matters most is that expectations are managed to create validated trust through experience, shared causal models and theory of mind. Therefore, mutual awareness of assumptions, intentions, expectations and capabilities are required to create a dialogue of trust in human-agent collaboration.

A fundamental difference between traditional software and AI systems is that the outcomes are not necessarily deterministic, but probabilistic, and that there may be more than one ”correct answer”. Hence, the goal is shifting from guarantees of correctness towards verifying for plausibility. In the section below we discuss the five pillars of our trustworthy AI engineering framework.

2.1. Models and Explanations

Explicit models[2] of the world or a suitable context in question enable reliable predictions of the behaviour of AI systems, both in the scope of training and outside, because they generalise knowledge beyond the limited and biased scope of the training data. Given a certain context, which can be very narrow or broad, explicit models represent concepts, relationships and rules that are always true in that context. For example, the laws of gravity are applicable to the whole universe. Models can be created by experts or learned from experience and data. Combinations of different types of models are particularly useful and insightful. For example, neuro-symbolic approaches are used to achieve this [13, 14, 15, 16, 17]. In this way, models promote transparency and explainability and, thus, make it possible to render the behaviour of the AI systems understandable and plausible. In simulations, models can enable the understanding – through experiments – of situations that are difficult or impossible to access otherwise. Often, synthetic data can be used to maintain privacy and avoid dangerous conditions.

Because models depend on a given context or domain, it is essential that agents using those models are aware of their competence in the given situation and are able to apply suitable models or adapt to situations gracefully when changing or leaving their scope of competence.

2.2. Causality and Grounding

Causality refers to the ability to identify and predict cause-and-effect relationships, i.e. which effects are the results of which causes and why [18]. An AI system that can understand causal relationships is able to make informed predictions and solve complex problems.

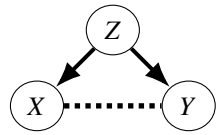


Figure 1.: Correlation vs. Causation

The need to move from correlation to causation is increasingly urgent (see figure 1, where the dotted line indicates correlation and the arrows indicate asymmetric causal relationships). If we want to explain why certain predictions are made or decisions are taken, it is essential to know and act on their causes.

[2]The term ‘model’ is used extensively in the ML community. It is necessary, however, to distinguish between the statistical models of ML and the semantic models of knowledge engineering. Here, we refer to the latter. See also in [12] for a unified taxonomy of AI.

Causal inference is concerned with the resulting effect when a corresponding event (cause) occurs, according to a given causal model. Causal inference asks whether an event indeed causes a certain effect by determining the likelihood that one event was the cause of another. In contrast to statistical correlations, causal relationships are directed and asymmetrical.

Counterfactuals refer to alternative choices that could have been made in the past and the corresponding effects that they might have caused. Therefore, they allow for exploring possibilities to find alternative outcomes according to a causal model, allowing to change policies accordingly in the future.

Causal discovery allows for determining whether a change in one variable (representing a state, action or event) indeed causes a change in another, in order to distinguish between correlated and causal relationships in data and to derive corresponding causal models.

Closely related to causality is understanding the anchoring (grounding) of meanings in the real context. A deep understanding of context and meaning requires not only processing data, but also capturing the real-world phenomena that the data represents, such that predictions, decisions and actions are based on them. Layers of abstractions are fundamental for building rich architectures. Semantic models, such as ontologies, are representations of concepts, their attributes and relationships. They contribute to trustworthy AI systems by explaining and constraining the meaning of those concepts.

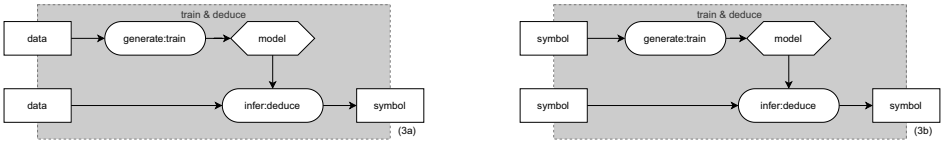
2.3. Modularity and Compositionality

One of the fundamental design principles of (software) engineering is modularity. Modularity guarantees that complex systems are broken down into understandable and manageable parts (functions and features) and assembled into system architectures. This increases the reliability of the individual components and their assemblies as systems of systems. It is much easier to verify smaller components than big monolithic artefacts. The evolution in software engineering from structured to modular and object-oriented programming enabled the design and construction of complex systems. In well-designed systems the transitions between successive components can be controlled and protected, making them explainable such that errors can be detected and repaired effectively. The pre- and post-conditions of each component can be validated and orchestrated in increasingly complex systems of systems.

When designing trustworthy AI systems, there are several important aspects that should be considered to guarantee the characteristics of trustworthy AI. In principle, these aspects apply to all software systems. However, they are of the greatest relevance for complex, intelligent systems for critical applications. AI engineering should make use of the lessons learned from software engineering and apply its engineering principles, such as design patterns and architectures.

An attempt to model design patterns for neuro-symbolic systems is made in [12, 19]. Two examples (see figure 2 show data-driven and knowledge-driven patterns. They are based on a visual language and taxonomy.

An important advantage of modular systems is that compositional patterns of subsystems can be identified and defined, which increases their reliability and documentation through reuse [20].



(a) Data-Driven Pattern (b) Knowledge-Driven Pattern

Figure 2. Modular Design Patterns for Hybrid Learning and Reasoning Systems

It is important to stress that software architectures are not merely static artefacts, but they rely on the interplay between structures and events – the organising principles and the dynamic evolution of complex systems [21]. Neither structure nor events are meaningful on their own, but require and depend on each other. In an extrapolated view, this relationship may be applied to the combination of learning and reasoning. Meaning emerges from a system's structure and its components, when it is operated in a dynamic context of perceiving and acting.

The principle of compositionality also applies to knowledge models and languages [22]: larger constructs are created by joining together smaller units with specific, understandable, and verifiable tasks. Abstract relationships can thus be traced back to their components. These aspects are applied when designing complex systems and should also become a matter of course for AI systems.

2.4. Human Agency and Oversight

Human agency and oversight mean that, in any case, a human being should have the overview, final decision, and responsibility for the actions of an AI system (human empowerment). Even if many tasks are increasingly being transferred to autonomous AI systems (agents), the principle that humans supervise, assess, and approve actions still applies. Keeping in mind the above-mentioned dilemma in building trustworthy AI systems, delegation of tasks needs to be interpretable by both humans and (software) agents – in particular, when humans and agents collaborate as hybrid teams in a symbiotic partnership. It is necessary that suitable task descriptions are handed over to the agents and that they understand and execute them in the relevant context, considering the models, explanations and causal relationships explained above.

When considering the collaboration and competition in hybrid teams of humans and autonomous agents, we consider many-to-many situations where multiple humans and multiple agents form hybrid teams. The purpose of the agents is to empower humans with providing their complementary capabilities, such as fast and precise information exchange and analysis of large data sets. Agents can play many different roles, but the responsibility for decisions remains, in principle, with humans, for example by verifying, validating and approving proposals for decisions. An essential aspect of meaningful collaboration is to make mutual assumptions and expectations explicit, such that they can be used in deliberation and communication. This is a prerequisite for appropriate delegation of tasks and the accurate and concise descriptions of their underlying intentions.

For collaborative decision-making (CDM), it is essential that each human and agent is aware of each others' points of view and has a notion of the others possess points of view that might differ from one's own - which is known as a Theory of Mind (ToM). ToM is defined as the human cognitive ability to perceive and interpret others in terms

of their mental states and it is considered an indispensable requirement of human social life [23, 24, 25, 26, 27, 28]. Rather than reasoning only with one's own beliefs, desires, goals, intentions, emotions, and thoughts, a person or agent with the awareness of others' states of mind can consider different and mindful acts, depending on a perceived context. This ability allows them to more easily understand, predict, and even manipulate the behaviour of others [29].

Trustworthiness in interacting with artificially intelligent systems emerges from experience and as a combination of various properties, such as fairness, robustness, transparency, verification, and accuracy [30]. AI systems are trusted when we have confidence in the decisions that they take, i.e. when we understand why they are made [31], even when we disagree.

2.5. Maturity Models

Maturity of software is commonly denoted by means of Maturity Models[3], which are frameworks that can be used in the process of planning and engineering, as well as in the process of road mapping. An extensive overview of AI maturity models is written by Sadiq et al [32].

In previous work [33] we presented a maturity-model that expressed the level of cooperation of an AI entity (agent) that acts in a human-AI team. The model can be used to reflect on expected capacities, roles and responsibilities of AI entities that act in a Human-AI team. The maturity model classifies the level of cooperation along two dimensions, level of agency and level of communication. It was represented by means of a two-dimensional matrix.

Horizontally, levels of agency range from human-controlled to fully autonomous [34, 35, 36]. Artificial Intelligence is based on the principles of autonomy and agency. Autonomy, the quality or state of being self-governing, is required to avoid purely predictable and reactive behaviour [37]. Whenever an AI entity commits on contributing to a team intention, it has to balance its level of autonomy with required levels of interaction. The model distinguishes four levels of agency with respect to an AI entity acting in a team (Human Trusted, Situational Autonomy, Preferred Autonomy and AI Trusted).

Vertically, levels of communication vary from merely sharing information about alignment of tasks, to forms of interaction in which an agent takes into account the mental state of others. Communication is a prerequisite for deliberation, delegation of tasks and sharing of knowledge within an ecosystem. We distinguish four levels of communication, varying from simple sharing information about coordination of tasks to higher orders of interaction that include the exchange of information about the learning process and each others' mental states. The maturity model distinguishes four levels of communication (Task Alignment, Co-Learning, Mental Modelling and Motivating).

We extended the maturity model with a third axis (see figure 3). The maturity model, named the *AI Interaction Maturity Model*, manifests itself in the form of a three-dimensional cube. the third axis can be used to express the impact of a particular AI team member on the social level, varying from an individual level (silo-ed situation) to organisational, industrial domain-specific and societal level.

[3] Another note on the term 'model', as it is an ambiguous term; where-as in software engineering it may refer to a computational model, theoretical model or architectural model, we use the term model here to refer to an engineering framework, commonly denoted with the term 'maturity model'.

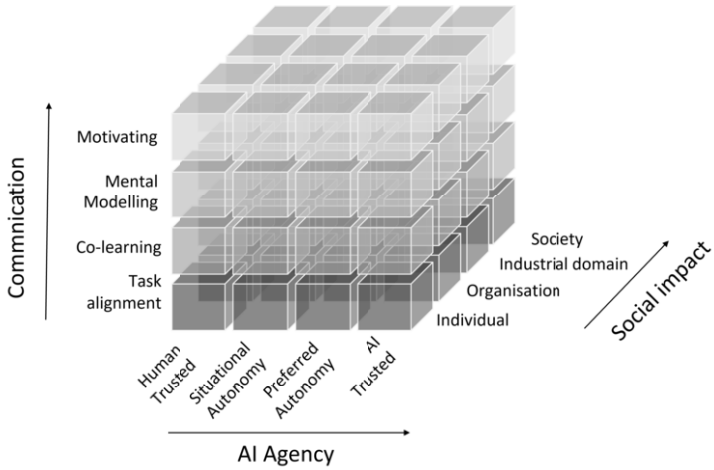


Figure 3. The AI Interaction Maturity Model extends the Collaborative Agent Maturity Model by including a third axis that reflects the impact of an AI entity on the social level.

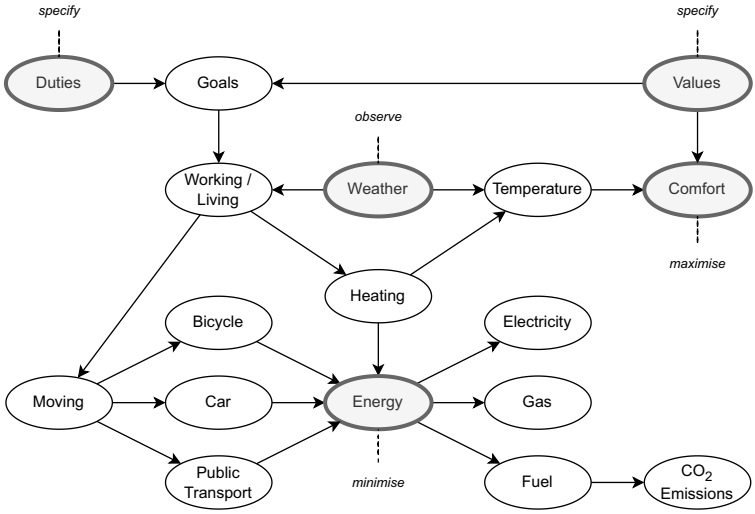


Figure 4. A causal model for living and working in the urban context.

3. Trusted Urban AI Scenario

As an example for applying the above-mentioned AI engineering concepts, we propose a scenario in an urban context.

Urban life has many peculiar characteristics [38, 39, 40, 41, 42, 43]. Some causal relationships in an urban context, focusing on energy consumption and mobility, are shown in figure 4. We are concerned with modelling and understanding human behaviour in an urban context (sustainable smart city), particularly when humans are part of hybrid teams together with agents.

The interactions in an urban environment are diverse, complex and conflicting. Many interests of hybrid actors are related and depend on each other. In the urban context, an overall goal for sustainable use of resources could be the reduction of energy consumption.

Using causal models, adopting and implementing the engineering disciplines and using a maturity model, as explained above, improves our understanding and control of systems that we design and apply in this urban context. The consumption of various types of energy is affected by the need and desire to move about the city and to heat buildings at home and at work (and for leisure, shopping, etc.). As shown in fig. 4, values and duties are the main sources that drive urban behaviour and external factors, such as the weather, influence decision-making. This causal model explains the relationships among several important behavioural aspects, but it is not deterministic. Individual behaviour is influenced by exogenous variables and cooperative behaviour results in complex interactions. A shared goal can be seen and modelled as an effect, that is caused by one or more interventions (actions or events). Consequently, in order to decide and plan which actions to take, it is necessary to understand which actions or events cause the intended effects. For example, your goal can be to arrive at a destination at a given time (work, home, leisure, etc.). By reasoning back which actions are required to get you there, piece by piece, a connected causal path can be constructed to determine the departure time and modes of traffic along the route. Due to shared intentions and causal models, humans and agents can mutually trust each other regarding their actions and outcomes.

The urban context comes with a multidisciplinary stakeholder field, involving a landscape of IT systems varying from traditional components in data sharing platforms to AI-enabled services. A modular approach for both design and realisation of software modules and AI models is crucial to keep the systems at their reuired levels of interoparability and scalability. Roadmapping of AI based systems, inside buildings as well as between various buildings and their interaction with human engineers is facilitated by means of the AIMM model.

4. Conclusions

As the field of Artificial Intelligence is still, and again, facing tremendous and overwhelming changes and progress, there is a strong and quickly growing need for trust in AI systems. The goal of Trust by Design is proposed to be based on the five engineering principles of (i) Models & Explanations, (ii) Causality & Grounding, (iii) Modularity & Compositionality, (iv) Human Agency & Oversight, and (v) a Maturity Model. Our intention is to develop the insights above further into practical methods and tools, based on a design pattern language, to benefit the AI community and its users. The context of energy consumption and mobility in an urban context serves as applied setting setting in various projects, validation of our suggested 5 pillars for controlled engineering and further experimentation in the field of human-AI ecosystems.

Acknowledgements

This research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

1. LeCun Y, Bengio Y, and Hinton G. Deep learning. Nature. 2015 May; 521. Number: 7553 Publisher: Nature Publishing Group:436–44. DOI: 10.1038/nature14539. Available from: https://www.nature.com/articles/nature14539 [Accessed on: 2022 Dec 8]
2. Deng L and Yu D. Deep Learning: Methods and Applications. Foundations and Trends® in Signal Processing. 2014 Jun 29; 7. Publisher: Now Publishers, Inc.:197–387. DOI: 10.1561/2000000039. Available from: https://nowpublishers.com/article/Details/SIG-039 [Accessed on: 2023 Dec 27]
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, and Polosukhin I. Attention is All you Need. *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. Available from: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html [Accessed on: 2023 Jul 26]
4. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, and Polosukhin I. Attention Is All You Need. 2023 Jul 23. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762[cs]. Available from: http://arxiv.org/abs/1706.03762 [Accessed on: 2023 Jul 26]
5. Bommasani R et al. On the Opportunities and Risks of Foundation Models. 2022 Jul 12. DOI: 10.48550/arXiv.2108.07258. arXiv: 2108.07258[cs]. Available from: http://arxiv.org/abs/2108.07258 [Accessed on: 2022 Oct 11]
6. Morocho-Cayamcela ME, Lee H, and Lim W. Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions. IEEE Access. 2019; 7. Conference Name: IEEE Access:137184–206. DOI: 10.1109/ACCESS.2019.2942390. Available from: https://ieeexplore.ieee.org/document/8844682 [Accessed on: 2023 Dec 4]
7. Garcez Ad, Lamb L, and Gabbay D. Neural-Symbolic Cognitive Reasoning. Berlin, Heidelberg: Springer, 2009 Jan 1. DOI: 10.1007/978-3-540-73246-4
8. Garcez Ad and Lamb LC. Neurosymbolic AI: the 3rd wave. Artificial Intelligence Review. 2023 Nov 1; 56:12387–406. DOI: 10.1007/s10462-023-10448-w. Available from: https://doi.org/10.1007/s10462-023-10448-w [Accessed on: 2023 Sep 24]
9. Li B, Qi P, Liu B, Di S, Liu J, Pei J, Yi J, and Zhou B. Trustworthy AI: From Principles to Practices. ACM Computing Surveys. 2023 Jan 16; 55:177:1–177:46. DOI: 10.1145/3555803. Available from: https://dl.acm.org/doi/10.1145/3555803 [Accessed on: 2024 Feb 6]
10. Thiebes S, Lins S, and Sunyaev A. Trustworthy artificial intelligence. Electronic Markets. 2021 Jun 1; 31:447–64. DOI: 10.1007/s12525-020-00441-4. Available from: https://doi.org/10.1007/s12525-020-00441-4 [Accessed on: 2022 Sep 28]
11. Ramchurn SD, Stein S, and Jennings NR. Trustworthy human-AI partnerships. iScience. 2021 Aug 20; 24:102891. DOI: 10.1016/j.isci.2021.102891. Available from: https://www.sciencedirect.com/science/article/pii/S2589004221008592 [Accessed on: 2022 Jun 21]

12. Bekkum M van, Boer M de, Harmelen F van, Meyer-Vitali A, and Teije At. Modular design patterns for hybrid learning and reasoning systems. Applied Intelligence. 2021 Sep 1; 51:6528–46. DOI: 10.1007/s10489-021-02394-3. Available from: https://doi.org/10.1007/s10489-021-02394-3 [Accessed on: 2022 Mar 29]
13. Garcez ASd, Gabbay DM, and Broda KB. Neural-Symbolic Learning System: Foundations and Applications. Berlin, Heidelberg: Springer-Verlag, 2002 Jun. 280 pp.
14. Garcez Ad, Broda KB, and Gabbay DM. Neural-Symbolic Integration: The Road Ahead. *Neural-Symbolic Learning Systems: Foundations and Applications*. Ed. by Garcez Ad, Broda KB, and Gabbay DM. Perspectives in Neural Computing. London: Springer, 2002 :235–52. DOI: 10.1007/978-1-4471-0211-3_9. Available from: https://doi.org/10.1007/978-1-4471-0211-3_9 [Accessed on: 2024 Jan 2]
15. Bader S and Hitzler P. Dimensions of Neural-symbolic Integration - A Structured Survey. version: 1. 2005 Nov 10. DOI: 10.48550/arXiv.cs/0511042. arXiv: cs/0511042. Available from: http://arxiv.org/abs/cs/0511042 [Accessed on: 2024 Feb 5]
16. Lake BM, Ullman TD, Tenenbaum JB, and Gershman SJ. Building machines that learn and think like people. Behavioral and Brain Sciences. 2017; 40. Publisher: Cambridge University Press:e253. DOI: 10.1017/S0140525X16001837. Available from: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/building-machines-that-learn-and-think-like-people/A9535B1D745A0377E16C590E14B94993# [Accessed on: 2022 Oct 3]
17. Yu D, Yang B, Liu D, Wang H, and Pan S. A Survey on Neural-symbolic Learning Systems. arXiv.org. 2021 Nov 10. Available from: https://arxiv.org/abs/2111.08164v3 [Accessed on: 2024 Jan 2]
18. Pearl J and Mackenzie D. The Book of Why: The New Science of Cause and Effect. 1st edition. New York: Basic Books, 2018 May 15. 432 pp.
19. Meyer-Vitali A, Mulder W, and Boer MHT de. Modular Design Patterns for Hybrid Actors. *Cooperative AI Workshop*. Conference on Neural Information Processing Systems. Vol. 2021. NeurIPS. 2021 Dec 14. arXiv: 2109.09331. Available from: http://arxiv.org/abs/2109.09331 [Accessed on: 2021 Nov 17]
20. Gamma E, Helm R, Johnson R, Vlissides J, and Booch G. Design Patterns: Elements of Reusable Object-Oriented Software. 1st edition. Reading, Mass: Addison-Wesley Professional, 1994 Nov 10. 416 pp.
21. Lévi-Strauss C. La pensée sauvage. Google-Books-ID: OoEeAAAAIAAJ. Plon, 1962. 422 pp.
22. Tiddi I, De Boer V, Schlobach S, and Meyer-Vitali A. Knowledge Engineering for Hybrid Intelligence. *Proceedings of the 12th Knowledge Capture Conference 2023*. K-CAP '23. New York, NY, USA: Association for Computing Machinery, 2023 Dec 5:75–82. DOI: 10.1145/3587259.3627541. Available from: https://doi.org/10.1145/3587259.3627541 [Accessed on: 2023 Nov 29]

23. Premack D and Woodruff G. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences. 1978 Dec; 1. Publisher: Cambridge University Press:515–26. DOI: 10.1017/S0140525X00076512. Available from: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/does-the-chimpanzee-have-a-theory-of-mind/1E96B02CD9850016B7C93BC6D2FEF1D0 [Accessed on: 2022 Mar 2]
24. Baron-Cohen S, Leslie AM, and Frith U. Does the autistic child have a "theory of mind" ? Cognition. 1985 Oct 1; 21:37–46. DOI: 10.1016/0010-0277(85)90022-8. Available from: https://www.sciencedirect.com/science/article/pii/0010027785900228 [Accessed on: 2023 May 27]
25. Frith C and Frith U. Theory of mind. Current Biology. 2005 Sep 6; 15. Publisher: Elsevier:R644–R645. DOI: 10.1016/j.cub.2005.08.041. Available from: https://www.cell.com/current-biology/abstract/S0960-9822(05)00960-7 [Accessed on: 2023 May 17]
26. Verbrugge R and Mol L. Learning to Apply Theory of Mind. Journal of Logic, Language and Information. 2008 Oct 1; 17:489–511. DOI: 10.1007/s10849-008-9067-4. Available from: https://doi.org/10.1007/s10849-008-9067-4 [Accessed on: 2022 Mar 2]
27. Byom L and Mutlu B. Theory of mind: mechanisms, methods, and new directions. Frontiers in Human Neuroscience. 2013; 7. Available from: https://www.frontiersin.org/articles/10.3389/fnhum.2013.00413 [Accessed on: 2023 May 27]
28. Buehler MC and Weisswange TH. Theory of Mind based Communication for Human Agent Cooperation. *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. 2020 IEEE International Conference on Human-Machine Systems (ICHMS). 2020 Sep :1–6. DOI: 10.1109/ICHMS49158.2020.9209472
29. Verbrugge R. Testing and Training Theory of Mind for Hybrid Human-agent Environments. *Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 1, Valletta, Malta, February 22-24, 2020*. Ed. by Rocha AP, Steels L, and Herik HJvd. SCITEPRESS, 2020 :11. Available from: https://vimeo.com/396473042
30. Harbers M, Verbrugge R, Sierra C, and Debenham J. The Examination of an Information-Based Approach to Trust. *Coordination, Organizations, Institutions, and Norms in Agent Systems III*. Ed. by Sichman JS, Padget J, Ossowski S, and Noriega P. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008 :71–82. DOI: 10.1007/978-3-540-79003-7_6
31. Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. arXiv:181110154 [cs, stat]. 2019 Sep 21. arXiv: 1811.10154. Available from: http://arxiv.org/abs/1811.10154 [Accessed on: 2022 Mar 9]
32. Sadiq RB, Safie N, Rahman AHA, and Goudarzi S. Artificial intelligence maturity model: a systematic literature review. PeerJ Computer Science. 2021 Aug 25; 7. Publisher: PeerJ Inc.:e661. DOI: 10.7717/peerj-cs.661. Available from: https://peerj.com/articles/cs-661 [Accessed on: 2024 Feb 1]

33. Mulder W and Meyer-Vitali A. A Maturity Model for Collaborative Agents in Human-AI Ecosystems. *Collaborative Networks in Digitalization and Society 5.0*. Ed. by Camarinha-Matos LM, Boucher X, and Ortiz A. IFIP Advances in Information and Communication Technology. Cham: Springer Nature Switzerland, 2023 :328–35. DOI: 10.1007/978-3-031-42622-3_23
34. Bradshaw JM, Hoffman RR, Woods DD, and Johnson M. The Seven Deadly Myths of "Autonomous Systems". IEEE Intelligent Systems. 2013 May; 28. Conference Name: IEEE Intelligent Systems:54–61. DOI: 10.1109/MIS.2013.70. Available from: https://ieeexplore.ieee.org/document/6588858 [Accessed on: 2024 Apr 2]
35. Johnson M, Bradshaw JM, Hoffman RR, Feltovich PJ, and Woods DD. Seven Cardinal Virtues of Human-Machine Teamwork: Examples from the DARPA Robotic Challenge. IEEE Intelligent Systems. 2014 Nov; 29. Conference Name: IEEE Intelligent Systems:74–80. DOI: 10.1109/MIS.2014.100. Available from: https://ieeexplore.ieee.org/document/6982119 [Accessed on: 2024 Apr 2]
36. Musić S and Hirche S. Control sharing in human-robot team interaction. Annual Reviews in Control. 2017 Jan 1; 44:342–54. DOI: 10.1016/j.arcontrol.2017.09.017. Available from: https://www.sciencedirect.com/science/article/pii/S1367578817301153 [Accessed on: 2024 Apr 2]
37. Vecht B van der, Meyer AP, Neef M, Dignum F, and Meyer JJC. Influence-Based Autonomy Levels in Agent Decision-Making. *Coordination, Organizations, Institutions, and Norms in Agent Systems II*. Ed. by Noriega P, Vázquez-Salceda J, Boella G, Boissier O, Dignum V, Fornara N, and Matson E. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007 :322–37. DOI: 10.1007/978-3-540-74459-7_21
38. WBGU – German Advisory Council on Global Change. Humanity on the move: Unlocking the transformative power of cities. Frauke Kraas, Claus Leggewie, Peter Lemke, Ellen Matthies, Dirk Messner, Nebojsa Nakicenovic, Hans Joachim Schellnhuber, Sabine Schlacke, Uwe Schneidewind. Berlin: WBGU, 2016. Available from: https://www.wbgu.de/en/publications/publication/humanity-on-the-move-unlocking-the-transformative-power-of-cities [Accessed on: 2023 Dec 7]
39. Angelidou M, Politis C, Panori A, Bakratsas T, and Fellnhofer K. Emerging smart city, transport and energy trends in urban settings: Results of a pan-European foresight exercise with 120 experts. Technological Forecasting and Social Change. 2022 Oct 1; 183:121915. DOI: 10.1016/j.techfore.2022.121915. Available from: https://www.sciencedirect.com/science/article/pii/S0040162522004371 [Accessed on: 2023 May 15]
40. Oliveira GM, Vidal DG, and Ferraz MP. Urban Lifestyles and Consumption Patterns. *Sustainable Cities and Communities*. Ed. by Leal Filho W, Marisa Azul A, Brandli L, Gökçin Özuyar P, and Wall T. Encyclopedia of the UN Sustainable Development Goals. Cham: Springer International Publishing, 2020 :851–60. DOI: 10.1007/978-3-319-95717-3_54. Available from: https://doi.org/10.1007/978-3-319-95717-3_54 [Accessed on: 2023 Dec 7]
41. Popelka S, Narvaez Zertuche L, and Beroche H. Urban AI Guide. Zenodo, 2023 Mar 8. DOI: 10.5281/zenodo.7708833. Available from: https://zenodo.org/record/7708833 [Accessed on: 2023 Mar 9]

42. Petrikovičová L, Kurilenko V, Akimjak A, Akimjaková B, Majda P, Ďatelinka A, Biryukova Y, Hlad L, Kondrla P, Maryanovich D, Ippolitova L, Roubalová M, and Petrikovič J. Is the Size of the City Important for the Quality of Urban Life? Comparison of a Small and a Large City. Sustainability. 2022 Jan; 14. Number: 23 Publisher: Multidisciplinary Digital Publishing Institute:15589. DOI: 10.3390/su142315589. Available from: https://www.mdpi.com/2071-1050/14/23/15589 [Accessed on: 2023 Dec 7]
43. Hashem IAT, Usmani RSA, Almutairi MS, Ibrahim AO, Zakari A, Alotaibi F, Alhashmi SM, and Chiroma H. Urban Computing for Sustainable Smart Cities: Recent Advances, Taxonomy, and Open Research Challenges. Sustainability. 2023 Jan; 15. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute:3916. DOI: 10.3390/su15053916. Available from: https://www.mdpi.com/2071-1050/15/5/3916 [Accessed on: 2023 Sep 27]