AI Engineering for Trust by Design

André Meyer-Vitali[a]
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI),
Stuhlsatzenhausweg 3, Saarland Informatics Campus D32, 66123 Saarbrucken, Germany

Keywords: Software Engineering, Artificial Intelligence, Causality, Trust, Robustness, Explainability.

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 four pillars of (i) Models & Explanations, (ii) Causality & Grounding, (iii) Modularity & Compositionality, and (iv) Human Agency & Oversight. Based on these pillars, a new AI engineering disciple could emerge, which we aim to support using corresponding methods and tools for “Trust by Design”.

1 INTRODUCTION

The current wave of Artificial Intelligence (AI) has emerged as a leading technology in the digital transformation, changing the economy, society, and our lives, while attracting massive investment worldwide. The past decade has been characterised by Deep Learning (LeCun et al., 2015; Deng and Yu, 2014), Transformers (Vaswani et al., 2017; Vaswani et al., 2023) and Large ”Foundation” Models Machine learning methods have transformed AI from a niche science to a socially relevant “mega-technology,” especially in the fields of image and video analysis, as well as in text and language processing. This new technology is made possible primarily by the latest graphics processors and the availability of vast amounts of data from social media and similar sources.

However, we are reaching the limits of control over these large, highly interconnected, AI-based systems. The complexity of existing AI models is often beyond our understanding, and the methods and processes to ensure safety, reliability, and transparency are lacking. We must overcome these novel and serious limitations or face an inevitable dwindling public and consumer acceptance of AI and dramatic losses in business opportunities and markets. This is clearly visible already in the automotive sector‘s broad retreat from highly automated driving. AI-based technology is also a key enabler in other economic sectors – including healthcare, mobility, energy, and the digital industry itself. All of these markets depend on complex and highly connected AI systems designed to support people in decision making and situational analysis.

Despite all the successes, many are not aware that deep learning does not support a real understanding of the problem, but only reflects complex statistical relationships. 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 itself (efficiency, adequacy, sustainability) became apparent.

Recently, however, a new overall approach to solving these problems is 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. Developers, domain experts, users, and regulators can rely on performance and reliability even for complex sociotechnical 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.

2 MOTIVATION

Current machine learning systems perform quite well and reliably in the context of their training data sets. To be useful, however, they also need to predict, clas-

[a] https://orcid.org/0000-0002-5242-1443

Meyer-Vitali, A.
AI Engineering for Trust by Design.
DOI: 10.5220/0012622400003645

In *Proceedings of the 12th International Conference on Model-Based Software and Systems Engineering (MODELSWARD 2024)*, pages 357-364
ISBN: 978-989-758-682-8; ISSN: 2184-4348

sify, decide and act in situations that they were not explicitly trained for. Therefore, they are evaluated with test data sets that should not overlap with the training data set. The measured level of ”generalisation” is an indication for how well they can perform in general (at least with respect to the test data, which is a limiting factor, indeed). Even a good level of such generalisation is not sufficient, however, because the systems are not able to distinguish between ”normal” and ”abnormal” situations.

Robustness is the ability of software systems to react appropriately to abnormal conditions (Meyer, 1997). For this purpose, it is a necessity to recognise situations or contexts where (implicit) assumptions do no longer hold. Without making those assumptions explicit, it is impossible to detect the edge of competence and to adapt accordingly. Explicit assumptions or (world) models include rules, norms or laws. These models include physical and natural laws (thermodynamics, electromagnetism, gravity, quantum mechanics, etc.), legal rules, socio-cultural norms, medical models (anatomy, mechanistic models of disease transmission, etc.), and others – that are always true, independent of the training data. Adaptation to changes in or of the context may include changes of rules. The use of model-based software engineering allows to exclude impossible or to invalidate highly improbable options and to enforce or guide learning and adaptation towards the most plausible and realistic outcomes. Many combinations of methods using knowledge-based reasoning models and data-driven learning components are possible (van Bekkum et al., 2021) and contribute to mutual system-level enhancements.

3 BUILDING TRUSTWORTHY AI SYSTEMS FOR THE FUTURE

Some of the current problems related to a lack of trust in AI systems are a direct result of the massive use of black-box methods that depend solely on data (Morocho-Cayamcela et al., 2019). Instead, the new AI generation has its foundation built on hybrid AI systems (also known as *neuro-symbolic* or *neuro-explicit*). These hybrids do not rely solely on data-driven approaches but on the full range of AI technologies (“All of AI”), which includes symbolic AI methods, search, reasoning, planning, and other operations. “Trust by Design” is achieved through the combination of Machine Learning with symbolic conclusions and the explicit representation of knowledge in hybrid AI systems. Knowledge no longer needs to be machine learned when it is represented by semantic and other explicit models, which can also guide the learning process in a direction that improves generalisation, robustness, and interpretability. This hybrid approach is also known as the third wave of AI (Garcez et al., 2009; Garcez and Lamb, 2023). The requirements are particularly strict when it comes to applications with significant physical, economic, or social risk. The AI systems used in such applications are required – for example by the European AI Act – to have been validated and certified.

With respect to the recent excitement about generative AI, a few critical considerations need to be highlighted. Generative AI is based on so-called ”Foundation Models”, which can appear as Large Language Models (LLM) or as similar models of still images or videos. The transformer architectures that generate these models convert huge amounts or text or other media content into statistical models of co-occurrence of tokens (parts of words or other features). The resulting models can then be used to generate text, images and video as predictions of probabilistic patterns of adjacency in the model’s huge space. For text, it is also possible to extract summaries and to conduct dialogues in natural language. 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 even language itself. There is no grammar involved or any form of semantics. Foundational 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.

3.1 Trusted AI Engineering

There is a dilemma to overcome in building trustworthy AI systems (Thiebes et al., 2021; Ramchurn et al., 2021): 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.

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.

A fundamental difference between tradition 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.

Very importantly, an autonomous AI system should be aware of its level of competence and self-confidence of its results. The area of competence is also known as the operational design domain (ODD), but a system may perform well beyond its designed or trained expertise by generalising also out of domain (OOD).

The following four pillars of AI Engineering are proposed as a framework for creating Trusted AI by Design. Each of these pillars will be described in more detail in the sections below, with a special focus on causality.

Models & Explanations. Reliable predictions about system behaviour for insightful and plausible explanations and simulations with generalised models from knowledge and training.

Causality & Grounding. Identification and predictions of cause-effect relationships for informed predictions and anchoring of meaning in real-world context and phenomena.

Modularity & Compositionality. Design of complex systems broken down into comprehensible and manageable parts (functions and features), reliably composed in system architectures.

Human Agency & Oversight. Overview, final decision and responsibility by humans for actions of AI systems, also when delegating tasks to autonomous agents in collaborative teams.

3.2 Models and Explanations

Explicit models[1] of the world or a suitable context in question enable reliable predictions of the behaviour of AI systems, both in the scope of training data 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 (Garcez et al., 2002b; Garcez et al., 2002a; Bader and Hitzler, 2005; Lake et al., 2017; Yu et al., 2021). 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. Privacy is thus maintained, as is the avoidance of 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. Each context includes a corresponding bias. Often, bias is attempted to be removed. However, agents need to be aware of their bias and to apply it thoughtfully, because bias is a measure of information, when bias-awareness exists.

3.3 Causality and Grounding

The need to move from correlation to causation is becoming more and more evident. If we want to explain why certain predictions are made or decisions are taken, it is essential to know their causes (Pearl et al., 2016; Pearl and Mackenzie, 2018). 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. An AI system that can understand causal relationships is able to make informed predictions and solve complex problems. Counterfactual inference can be performed in a wider scope of domain than given by the training data alone, because answers can be found to questions that involve hypotheses about changes in the past (”what would have happened if...?”), which give reasons for alternative outcomes in the future.

Causal Models. Structural causal models (SCM) and Structural Equation Models (SEM) (Pearl, 2010; Pearl et al., 2016) provide a concise method for modelling and analysing causal relationships as graphs (SCM) and sets of equations (SEM),

[1] 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 (van Bekkum et al., 2021) for a unified taxonomy of AI.

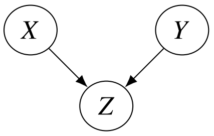
as shown in figure 1. SCMs support the human understanding and explanations, while SEMs are more suitable for representing causality in combination with logical expressions to define the specific functions that relate the variables (F). Variables represent events, processes, states or objects.

Causal Inference. Causal inference is typically concerned with the resulting effect when a corresponding event (cause) occurs, according to a given causal model, such that the respective dependency can be verified. 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 asymmetrical (Price, 1992; Kutach, 2013; Ismael, 2023), i.e. that there is a directed relationship from a cause to an effect, rather than a spurious co-occurrence of events.

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 that exist in imagined worlds – in contrast to what actually happened – by intervening in the value of specific variables and, hence, find alternative outcomes according to the same model as usual. Counterfactuals can represent situations that may not be practical for observation due to practical or ethical reasons, which enables the causal exploration of a wider scope of domain.

Causal Discovery. Even when causal models are not known in advance, 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 models. Approaches to make the distinction are interventions, random control trials and counterfactual reasoning (Eberhardt, 2017; Zhu et al., 2020; Schölkopf et al., 2021; Schölkopf and von Kügelgen, 2022). The use of known causal models can improve (language) understanding and causal discovery can bring understanding from data to a higher level, i.e., formulate new hypotheses and insights that transcend the previous body of knowledge (for example, in a similar way as discovering the laws of thermodynamics or electromagnetism).

Causal Machine Learning. Causally-informed Machine Learning (CML) uses causal models to influence and direct ML methods for improved predictability through reuse of domain knowledge, as well as explainability and robustness through interventions and counterfactuals (Vlontzos et al., 2023; Kyono and van der Schaar, 2019; Zhang et al., 2020; Rawal et al., 2023).



(a) Structural Causal Model (SCM).

$$U = \{X, Y\}$$
$$V = \{Z\}$$
$$F = \{f_Z : Z = 2X + 3Y\}$$

(b) Structural Equation Model (SEM).

Figure 1: Causal Models.

Shared causal models increase trust among team members, because they help to explain to each other *why* certain actions are to be taken (Janssen et al., 2022). Delegation without reason or motivation is not trustworthy (unless the authority or reputation of the delegator is very high). This enables users to better understand the rationale and have greater confidence in others making a fair decision. Causality can also be seen as an enabler (or even a requirement) for explainable artificial intelligence (Carloni et al., 2023).

There are several important aspects by which causality can improve the trustworthiness of AI systems. Besides precision and accuracy, which are fundamental to trustworthiness in AI, they are (Greifeneder, 2021; Ganguly et al., 2023; Bartling et al., 2018; Yap and Tomlinson, 2016):

Transparency & Interpretability. The reasoning behind decisions is explainable and easily understood by humans. Causal models provide the reasons for predictions and causal explanations help to build a correct mental model of the problem.

Reproducibility. The ability to repeat experiments and get the same results increases the trustworthiness and accuracy of scientific output.

Fairness. Causal AI can remove bias, because it understands how variables are interconnected and dependent on each other. Understanding causal relationships between sensitive input variables (such as gender or race) and predicted outcomes is important for assessing biased behaviour. Counterfactual fairness is achieved when the output is identical for each sensitive input variable.

Robustness. Causal models can avoid the brittleness of most machine learning systems, due to spurious correlations. They can handle data that is not independent and identically distributed (IID) or out of distribution (OOD), because they can discern between relevant and irrelevant data and variables (Sherman and Shpitser, 2019; Zhang et al., 2023).

Privacy. The robustness of causal models helps in preventing privacy attacks, because weaknesses of trained models cannot easily be exploited, for example in federated learning.

Safety & Accountability (Auditing). Regulations for safe-guarding AI systems for use in critical applications and domains demand impact assessment (IA) to prevent from algorithmic and data-driven harm by finding potential negative effects before (large-scale) deployment. Causal models that represent dependencies between system design and impact can be used to assess and mitigate corresponding risks by identifying which system elements are responsible for undesired effects.

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. This applies also to large language models, so that statements are not only made based on statistical probabilities, but in the knowledge of the concepts, contexts, phenomena, and semantic and causal relationships grounded in reality (Searle, 1980; Harnad, 1990). Whether this knowledge necessarily requires physical interaction of the agent with its environment remains a subject of debate (Gärdenfors, 2019). Harnad argues for the need for sensations to induce and stimulate representations via distal objects – things that exist in the environment and emit signals that can be perceived by means of a medium and means of perception. Not all concepts are physical though, which is a strength of abstract thinking, namely that more abstract concepts can be formed, represented and communicated from less abstract ones. Also, perception is guided by intellect and constrained by the available means of perception.

Layers of abstractions are fundamental for building rich architectures in software engineering and AI systems are no exception. Semantic models, such as ontologies (Fensel et al., 2001; Antoniou and Van Harmelen, 2004), are representations of concepts, their attributes and relationships, and, therefore, contribute to trustworthy AI systems by explaining and constraining the meaning of those concepts. The difference and close interaction between perception and reasoning on various levels of abstractions is documented as *System 1* and *System 2* in (Kahneman, 2011).

3.4 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 reliably 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 effectively. The pre- and post-conditions of each component can be validated and orchestrated in increasingly complex systems of systems.

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 (Gamma et al., 1994; van Bekkum et al., 2021).

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 (Lévi-Strauss, 1962). 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 (Tiddi et al., 2023): 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.

3.5 Human Agency and Oversight

Human agency and oversight mean that in any case a human 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.

For collaborative decision-making (CDM), it is essential that each human and agent is aware of each others' points of view and understands that others possess mental states 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, such as beliefs, desires, goals, intentions and emotions, and it is considered an indispensable requirement of human social life (Premack and Woodruff, 1978; Baron-Cohen et al., 1985; Frith and Frith, 2005; Verbrugge and Mol, 2008; Byom and Mutlu, 2013; Buehler and Weisswange, 2020). Rather than reasoning only with one's own beliefs, desires, 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 (Verbrugge, 2020).

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.

Instead of relying on AI systems to take over human activities, as some have predicted, it is better to focus on how humans and machines can complement each other's strengths (Marcus, 2022). For example, radiologists are still needed to interpret MRI images (Chan and Siegel, 2019), but they will have to collaborate with AI systems and those systems need to support the human collaborators by providing insight into their decision-making process. Therefore, a new approach of hybrid or neuro-symbolic AI is necessary for creating trustworthiness (Marcus and Davis, 2019).

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 (Harbers et al., 2008). AI systems are trusted when we have confidence in the decisions that they take, i.e. when we understand why they are made (Rudin, 2019), even when we disagree.

In a community with trustworthy interactions, it is crucial to establish and enforce social norms (Haynes et al., 2017; Emelin et al., 2020; Jiang et al., 2021; Savarimuthu et al., 2008; Haynes et al., 2017). Such norms can be of generic nature or valid only within certain communities or teams and specify transparently what is expected behaviour, what is allowed or forbidden and which are the consequences in case of violations. In addition, knowledge and intentions, but also norms, can change and need to be adapted in due course. Otherwise, such systems and interactions cannot be trusted any longer.

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 four engineering principles of (i) Models & Explanations, (ii) Causality & Grounding, (iii) Modularity & Compositionality, and (iv) Human Agency & Oversight. Our intention is to develop the insights above further into practical methods and tools to benefit the AI community and its users. The Boxology in (van Bekkum et al., 2021) provides a stepping stone to further develop trustworthy AI engineering methods, based on neuro-symbolic and causal AI.

ACKNOWLEDGEMENTS

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

REFERENCES

Antoniou, G. and Van Harmelen, F. (2004). *A semantic web primer*. MIT press.

Bader, S. and Hitzler, P. (2005). Dimensions of Neural-symbolic Integration - A Structured Survey.

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37–46.

Bartling, B., Fehr, E., Huffman, D., and Netzer, N. (2018). The Causal Effect of Trust.

Buehler, M. C. and Weisswange, T. H. (2020). Theory of Mind based Communication for Human Agent Cooperation. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pages 1–6.

Byom, L. and Mutlu, B. (2013). Theory of mind: mechanisms, methods, and new directions. *Frontiers in Human Neuroscience*, 7.

Carloni, G., Berti, A., and Colantonio, S. (2023). The role of causality in explainable artificial intelligence. arXiv:2309.09901 [cs].

Chan, S. and Siegel, E. L. (2019). Will machine learning end the viability of radiology as a thriving medical specialty? *The British Journal of Radiology*, 92(1094):20180416.

Deng, L. and Yu, D. (2014). Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387. Publisher: Now Publishers, Inc.

Eberhardt, F. (2017). Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3(2):81–91.

Emelin, D., Bras, R. L., Hwang, J. D., Forbes, M., and Choi, Y. (2020). Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences. *arXiv:2012.15738 [cs]*. arXiv: 2012.15738.

Fensel, D., van Harmelen, F., Horrocks, I., McGuinness, D., and Patel-Schneider, P. (2001). OIL: an ontology infrastructure for the Semantic Web. *IEEE Intelligent Systems*, 16(2):38–45. Conference Name: IEEE Intelligent Systems.

Frith, C. and Frith, U. (2005). Theory of mind. *Current Biology*, 15(17):R644–R645. Publisher: Elsevier.

Gamma, E., Helm, R., Johnson, R., Vlissides, J., and Booch, G. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley Professional, Reading, Mass, 1st edition edition.

Ganguly, N., Fazlija, D., Badar, M., Fisichella, M., Sikdar, S., Schrader, J., Wallat, J., Rudra, K., Koubarakis, M., Patro, G. K., Amri, W. Z. E., and Nejdl, W. (2023). A Review of the Role of Causality in Developing Trustworthy AI Systems. arXiv:2302.06975 [cs].

Garcez, A. d., Broda, K. B., and Gabbay, D. M. (2002a). Neural-Symbolic Integration: The Road Ahead. In Garcez, A. d., Broda, K. B., and Gabbay, D. M., editors, *Neural-Symbolic Learning Systems: Foundations and Applications*, Perspectives in Neural Computing, pages 235–252. Springer, London.

Garcez, A. d., Lamb, L., and Gabbay, D. (2009). *Neural-Symbolic Cognitive Reasoning*. Springer, Berlin, Heidelberg.

Garcez, A. d. and Lamb, L. C. (2023). Neurosymbolic AI: the 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406.

Garcez, A. S. d., Gabbay, D. M., and Broda, K. B. (2002b). *Neural-Symbolic Learning System: Foundations and Applications*. Springer-Verlag, Berlin, Heidelberg.

Gärdenfors, P. (2019). From Sensations to Concepts: a Proposal for Two Learning Processes. *Review of Philosophy and Psychology*, 10(3):441–464.

Greifeneder, B. (2021). Three Ways A Causal Approach Can Improve Trust In AI. Section: Innovation.

Harbers, M., Verbrugge, R., Sierra, C., and Debenham, J. (2008). The Examination of an Information-Based Approach to Trust. In Sichman, J. S., Padget, J., Ossowski, S., and Noriega, P., editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems III*, Lecture Notes in Computer Science, pages 71–82, Berlin, Heidelberg. Springer.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346.

Haynes, C., Luck, M., McBurney, P., Mahmoud, S., Vítek, T., and Miles, S. (2017). Engineering the emergence of norms: a review. *The Knowledge Engineering Review*, 32. Publisher: Cambridge University Press.

Ismael, J. (2023). Reflections on the asymmetry of causation. *Interface Focus*, 13(3):20220081. Publisher: Royal Society.

Janssen, S., Sharpanskykh, A., and Mohammadi Ziabari, S. S. (2022). Using Causal Discovery to Design Agent-Based Models. In Van Dam, K. H. and Verstaevel, N., editors, *Multi-Agent-Based Simulation XXII*, Lecture Notes in Computer Science, pages 15–28, Cham. Springer International Publishing.

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M., and Choi, Y. (2021). Delphi: Towards Machine Ethics and Norms. *arXiv:2110.07574 [cs]*. arXiv: 2110.07574.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 1st edition edition.

Kutach, D. (2013). Causal Asymmetry. In Kutach, D., editor, *Causation and its Basis in Fundamental Physics*, page 0. Oxford University Press.

Kyono, T. and van der Schaar, M. (2019). Improving Model Robustness Using Causal Knowledge. arXiv:1911.12441 [cs, stat].

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253. Publisher: Cambridge University Press.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. Number: 7553 Publisher: Nature Publishing Group.

Lévi-Strauss, C. (1962). *La pensée sauvage*. Plon. Google-Books-ID: OoEeAAAAIAAJ.

Marcus, G. (2022). Deep Learning Is Hitting a Wall.

Marcus, G. and Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Vintage.

Meyer, B. (1997). *Object-Oriented Software Construction*. Prentice Hall, Upper Saddle River, NJ, 2 edition.

Morocho-Cayamcela, M. E., Lee, H., and Lim, W. (2019). Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions. *IEEE Access*, 7:137184–137206. Conference Name: IEEE Access.

Pearl, J. (2010). An Introduction to Causal Inference. *The International Journal of Biostatistics*, 6(2). Publisher: De Gruyter.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. John Wiley & Sons. Google-Books-ID: I0V2CwAAQBAJ.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 1st edition edition.

Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526. Publisher: Cambridge University Press.

Price, H. (1992). Agency and Causal Asymmetry. *Mind*, 101(403):501–520. Publisher: [Oxford University Press, Mind Association].

Ramchurn, S. D., Stein, S., and Jennings, N. R. (2021). Trustworthy human-AI partnerships. *iScience*, 24(8):102891.

Rawal, A., Raglin, A., Sadler, B. M., and Rawat, D. B. (2023). Explainability and causality for robust, fair, and trustworthy artificial reasoning. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications V*, volume 12538, pages 493–500. SPIE.

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *arXiv:1811.10154 [cs, stat]*. arXiv: 1811.10154.

Savarimuthu, B. T. R., Cranefield, S., Purvis, M., and Purvis, M. (2008). Role Model Based Mechanism for Norm Emergence in Artificial Agent Societies. In Sichman, J. S., Padget, J., Ossowski, S., and Noriega, P., editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems III*, Lecture Notes in Computer Science, pages 203–217, Berlin, Heidelberg. Springer.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634. Conference Name: Proceedings of the IEEE.

Schölkopf, B. and von Kügelgen, J. (2022). From Statistical to Causal Learning. arXiv:2204.00607 [cs, stat].

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424. Publisher: Cambridge University Press.

Sherman, E. and Shpitser, I. (2019). Identification and Estimation of Causal Effects from Dependent Data. arXiv:1902.01443 [stat].

Thiebes, S., Lins, S., and Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, 31(2):447–464.

Tiddi, I., De Boer, V., Schlobach, S., and Meyer-Vitali, A. (2023). Knowledge Engineering for Hybrid Intelligence. In *Proceedings of the 12th Knowledge Capture Conference 2023*, K-CAP ’23, pages 75–82, New York, NY, USA. Association for Computing Machinery.

van Bekkum, M., de Boer, M., van Harmelen, F., Meyer-Vitali, A., and Teije, A. t. (2021). Modular design patterns for hybrid learning and reasoning systems. *Applied Intelligence*, 51(9):6528–6546.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention Is All You Need. arXiv:1706.03762 [cs].

Verbrugge, R. (2020). Testing and Training Theory of Mind for Hybrid Human-agent Environments. In Rocha, A. P., Steels, L., and Herik, H. J. v. d., editors, *Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 1, Valletta, Malta, February 22-24, 2020*, page 11. SCITEPRESS.

Verbrugge, R. and Mol, L. (2008). Learning to Apply Theory of Mind. *Journal of Logic, Language and Information*, 17(4):489–511.

Vlontzos, A., Kainz, B., and Gilligan-Lee, C. M. (2023). Estimating categorical counterfactuals via deep twin networks. *Nature Machine Intelligence*, 5(2):159–168.

Yap, J. Y. and Tomlinson, A. (2016). A Causality-Based Model for Describing the Trustworthiness of a Computing Device. In Yung, M., Zhang, J., and Yang, Z., editors, *Trusted Systems*, Lecture Notes in Computer Science, pages 130–149, Cham. Springer International Publishing.

Yu, D., Yang, B., Liu, D., Wang, H., and Pan, S. (2021). A Survey on Neural-symbolic Learning Systems.

Zhang, C., Mohan, K., and Pearl, J. (2023). Causal Inference under Interference and Model Uncertainty. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, pages 371–385. PMLR. ISSN: 2640-3498.

Zhang, C., Zhang, K., and Li, Y. (2020). A Causal View on Robustness of Neural Networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 289–301. Curran Associates, Inc.

Zhu, S., Ng, I., and Chen, Z. (2020). Causal Discovery with Reinforcement Learning. In *International Conference on Learning Representations*, Online.