





Controllable AI - An Alternative to Trustworthiness in Complex AI Systems?

Peter Kieseberg^{1,2}, Edgar Weippl^{3,4}, A. Min Tjoa^{3,5}, Federico Cabitza^{6,9},
Andrea Campagner⁶, and Andreas Holzinger^{7,8}

¹ Institute of IT Security, St. Pölten University of Applied Sciences,
St. Pölten, Austria

² Josef Ressel Center for Blockchain-Technologies and Security management,
Pölsen, Austria

³ Secure Business Austria, SBA Research, Vienna, Austria

⁴ Research Group Security and Privacy, University of Vienna, Vienna, Austria

⁵ Information Systems Engineering, Vienna University of Technology,
Vienna, Austria

⁶ DISCo, University of Milano-Bicocca, Milan, Italy

⁷ Medical University Graz, Graz, Austria

⁸ Human-Centered AI Lab, University of Natural Resources and Life Sciences
Vienna, Austria

andreas.holzinger@human-centered.ai

⁹ IRCCS Istituto Ortopedico Galeazzi, Milan, Italy

Abstract. The release of ChatGPT to the general public has sparked discussions about the dangers of artificial intelligence (AI) among the public. The European Commission's draft of the AI Act has further fueled these discussions, particularly in relation to the definition of AI and the assignment of risk levels to different technologies. Security concerns in AI systems arise from the need to protect against potential adversaries and to safeguard individuals from AI decisions that may harm their well-being. However, ensuring secure and trustworthy AI systems is challenging, especially with deep learning models that lack explainability. This paper proposes the concept of Controllable AI as an alternative to Trustworthy AI and explores the major differences between the two. The aim is to initiate discussions on securing complex AI systems without sacrificing practical capabilities or transparency. The paper provides an overview of techniques that can be employed to achieve Controllable AI. It discusses the background definitions of explainability, Trustworthy AI, and the AI Act. The principles and techniques of Controllable AI are detailed, including detecting and managing control loss, implementing transparent AI decisions, and addressing intentional bias or backdoors. The paper concludes by discussing the potential applications of Controllable AI and its implications for real-world scenarios.

Keywords: Artificial Intelligence · Digital Transformation · Robustness · Trustworthy AI · Explainability · Explainable AI · Safety · Security · AI risks · AI threats

1 Introduction and Motivation

More than any other subject, Artificial Intelligence (AI) has experienced many ups and downs since its formal introduction as an academic discipline six decades ago. The success of the digital computer [9] along with the remarkable achievements in statistical data-driven machine learning (ML) have rekindled significant interest in digitalization generally and AI specifically. Two key factors have contributed to its practical success: the availability of big data and the growing computational power. Around 2010, a breakthrough occurred with the success of deep learning (DL) algorithms [2] (aka neural networks [7, 18]). This success led to widespread use in all sorts of industrial and everyday applications in virtually every field, literally from agriculture to zoology [15]. This marked the beginning of a new era in AI, often referred to as the second AI spring. A prime example of AI’s capabilities today is OpenAI’s latest natural language technology, GPT-4, which demonstrates the impressive potential of AI while also highlighting its limitations, such as the lack of human common sense [3, 6].

With the release of ChatGPT for the general public, the discussion on the dangers of artificial intelligence has shifted from abstract and rather academic to a discussion required by the general public. In addition, the European Commission issued a draft of the novel *AI Act*, which already generated a lot of discussions, not only with respect to the exact definition of AI in the act, but especially regarding the assignment of risk levels to certain technologies, with the high capabilities of Chat GPT 3.5 being fuel to this discussion. Security as a major concern is seen twofold: On the one hand, security AI systems against potential (typically human) adversaries and thus making them robust and trustworthy. On the other hand, people require protection against AI systems and their decisions, in case these are detrimental to their well-being, a discussion which can be dated back at least until 1941 to Asimov’s three laws of robotics [1] and leading to the definition of Trustworthy AI.

Still, providing secure and trustworthy AI systems is non-trivial when facing modern approaches of machine learning: While rule based systems provide explainability to a certain practical degree, this cannot be said for the deep learning models currently used in a multitude of applications fields ranging from the medical field [17] to smart farming and forestry [12]. Especially when considering reinforcement learning, many approaches like penetration testing [23] become moot, as (i) many testing approaches like input fuzzing might change the underlying model resulting in damage to the tested system while yielding the not-so-astonishing result that a model fed with garbage produces garbage and (ii) the model itself is constantly changing, i.e. the system tested today is different from the system available tomorrow in an unpredictable way from a security perspective.

In this paper we propose the notion of *Controllable Artificial Intelligence* or *Controllable AI* as an alternative to the more classical approach of *Trustworthy AI* and detail the major differences. The main purpose of this editorial paper lies in providing a starting point for discussion on how the new complex AI systems that will definitively get put into service within the next years can be secured,

without either requiring huge improvements in explainability capabilities, nor an unrealistic reduction of the algorithms in use to an explainable and fully transparent selection. Furthermore, we provide an overview on techniques that can be used for achieving Controllable AI.

This paper is organized as follows: Sect. 2 provides an overview on the most important related concepts like trustworthy AI and explainability, as well as some relevant details on the upcoming AI Act. In Sect. 3, we define Controllable AI and compare it to the concept of trustworthiness, while Sect. 4 gives an overview on selected techniques for achieving control. Finally, in Sect. 5 we discuss the approach and its potential in actual application.

2 Background

In this section, we will discuss some background definitions that build the foundation for the notion of controllable AI. It must be noted that sometime definitions differ slightly, we therefore have selected those definitions that we consider to be the most prominent in recent literature, but acknowledging that high level definitions might change depending on authors, time of writing and exact research field.

2.1 The Explainability Problem

The main challenge in the explainability problem is the complexity and opacity of deep learning models used in many AI applications. Deep learning models, such as neural networks, are highly complex, nonlinear and high dimensional and consist of numerous interconnected layers, making it extremely difficult to understand how such a model arrive at their predictions or decisions. Therefore such models are called as black boxes, meaning that it is challenging to trace the reasoning or logic behind their outputs.

Explainability in AI refers to the ability to provide understandable and interpretable explanations for the decisions made by AI systems [4]. It is important for various reasons, including enabling experts but also end-users sometimes to understand and trust AI outputs, ensuring ethical and fair decision-making, identifying and rectifying biases or errors in the models, and facilitating regulatory compliance. Deep learning models learn from vast amounts of data and extract complex patterns and representations, making them highly accurate in many tasks. However, this accuracy often comes at the cost of interpretability. The relationships and features learned by these models are often distributed across multiple layers, making it challenging to provide clear and intuitive explanations for their decisions.

Furthermore, deep learning models are often non-linear and highly parameterized, with billions or even trillions of learnable parameters. This makes it difficult to trace the influence of individual inputs or features on the model's output. As a result, it becomes challenging to provide human-understandable explanations that can be easily interpreted and validated.

Addressing the challenge of explainability in deep learning models requires research and development of new methods and techniques [16]. Various approaches, such as feature importance analysis, gradient-based methods, rule extraction, and model distillation, are being explored to enhance explainability. However, finding a balance between explainability and maintaining high performance and accuracy in deep learning models remains an active area of research and a significant challenge in the field of AI.

To tackle this challenge, it is crucial to seek standardized definitions in both technical standardization and legislation. Standardized definitions provide a shared understanding and a common language for discussing and evaluating AI systems [20]. Efforts are underway to establish consistent definitions and guidelines to ensure the transparency and explainability of AI models, especially in contexts such as regulatory frameworks like the AI Act which goes towards trustworthy AI.

2.2 Trustworthy AI

The most powerful learning methods generally suffer from two fundamental problems: On the one hand, it is difficult to explain why a particular result was obtained (see above), and on the other hand, our best methods are lacking robustness. Even the smallest perturbations in the input data can have dramatic effects on the output, leading to completely different results. In certain non-critical application areas, this may not seem so dramatic. But in critical areas, e.g., medicine, and especially clinical medicine, the issue is trust - and the future trust of clinicians in AI technologies. Explainability and robustness increase reliability and trust in the results [13,14].

Trustworthy AI has been one of the fundamental key concepts for dealing with the problems of AI during the last years. The High-Level Expert Group (HLEG) of the European Union put forth the following seven key requirements for Trustworthy AI [11]: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability. This definition puts great focus on the fact that trustworthiness is not a purely technical issue, but has to consider the socio-technological systems involving AI, therefore requiring an AI to possess three key characteristics throughout its entire life-cycle: (1) Lawfulness, (2) adherence to ethical principles and (3) technological, as well as social, robustness.

The NIST provides a slightly different set of characteristics for trustworthy AI [22], which need to be (1) valid and reliable, (2) safe, (3) secure and resilient, (4) accountable and transparent, (5) explainable and interpretable, (6) privacy-enhanced, and (7) fair with harmful bias managed. While these two sets are quite similar in nature, it puts more focus on the system reliably producing correct results (characteristic 1) and splits safety and security into two characteristics, which we deem useful, as the mindsets behind both approaches are very different. Furthermore, the HLEG-definition focuses far more on human oversight, which can be problematic in many automated decision making processes in e.g.

industrial automation [21], in pervasive health technologies [19], or in clinical applications [24]. Also, key requirement 6 on environmental and societal well-being can be problematic in many application areas, both, in industry, as well as military applications, which, nevertheless, require a high level of trustworthiness.

For the remainder of this work, we will mainly focus on the NIST-characteristics. It must be noted that both publications, while focusing on their definitions, state that they do not claim them to be exhaustive.

2.3 The AI Act

The AI Act that is currently available in draft format [5] focuses on establishing harmonized rules for the development, marketing and use of AI in the EU as its main goal. This includes ensuring that AI systems placed on the EU market are safe, as well as ensuring legal certainty, for investment and innovation in the field of AI, improving governance and effective enforcement and facilitating the development of a single market for legitimate, safe, and trustworthy AI [8]. This also includes the definition of forbidden AI systems like state-run applications for social scoring and dividing AI applications into three distinctive categories based on perceived risk:

- Unacceptable Risk: These systems are prohibited under the AI-Act and include social scoring by any public authority, real-time remote biometric identification in public spaces for law enforcement and behaviour manipulation, amongst others.
- High-Risk: Any AI system that constitutes a safety critical component, or is where the product is protected under a certain range of specific legislation as outlined in Annex II of the act. Furthermore, Annex III provides a taxative enumeration of application fields. Any product containing an AI component that falls under at least one of these fields must be considered as high-risk AI. This includes biometric and medical use cases, but also applications in the field of education amongst others. High-risk AI systems must adhere to several requirements, reflecting fundamentals of trustworthy AI like risk-management, transparency requirements and data management.
- Limited Risk: These systems are subject to additional requirements with respect to transparency and focus on chat bots and deep fakes amongst others. The categorization of the given examples is currently under discussion due to the qualities of Chat GPT [10].
- Minimal Risk: All other system, these are basically unregulated under the AI-Act, but are encouraged to voluntarily follow the requirements for high-risk AI systems as a code of conduct. Systems for spam detection are given as an example for systems of this category.

The definition of what constitutes which category is defined in Annex III of the act and mainly focuses on the application space, like e.g. biometric identification and categorisation of natural persons, management and operation of critical infrastructure or education and vocational than on the technological basis. The

act is currently under scrutiny due to (i) the problematic definition of AI and (ii) because of rating chat bots into as limited risk AI system in the pre-Chat-GPT draft.

3 Principles of Controllable AI

The major idea behind Controllable AI is that certain requirements derived from the definition of trustworthy AI (see Sect. 2) cannot be fulfilled in real live environments. Furthermore, we assume that with the gold digger mindset currently surrounding AI and its application space, developers and companies will not want to utilize explainable but inherently less powerful applications, i.e. we do not agree with the idea that risk management will overrule practical system capabilities due to concerns of trustworthiness in practical system development. This is not only due to the unclear definitions in regulations that leave a wide field for interpretation, but also due to competition with other players and especially between nations.

Furthermore, while typically data is mentioned as an (important) factor for building trustworthy AI, we are of the opinion that the impact of data on these systems is neglected by intrinsically focusing on the system code. Still, in many machine learning applications, the relevant knowledge, as well as many dangers like algorithmic bias, do not lie within the code, but the models, i.e. we need to talk about *data defined software* with a focus shifting from code to data. This is especially important for security considerations, as the code might be perfectly fine, but e.g. backdoors in the model allow for corruption of classification results. This also has an effect on how we need to describe a system life-cycle: While the system might be static from the code side, it might change a lot due to new models being incorporated. This can be especially problematic in cases of reinforcement learning, where the mode changes constantly and even versioning becomes a management nightmare in realistic environments featuring high data volumes. Here, the actors steering the learning process, whether human or also automated, become an additional liability, as they possess a certain influence on the iterative process that shapes the future system.

The principle behind *Controllable AI* is the assumption that no AI-system should be considered trustworthy and that methods need to be put in place that allow to detect malfunction and regain control.

As Controllable AI is a deviation from the definitions of Trustworthy AI, it must be noted that the authors of the two most prominent definitions of Trustworthy AI were very clear about the fact that their principles/characteristics might come into conflict with each other or with the application field in question, thus, even in Trustworthy AI, while the respective principles/characteristics should be followed as good as possible, conflict needs to be resolved. In Controllable AI we, on the other hand, explicitly weaken these principles/characteristics without the advent of an explicit conflict.

Basically, Controllable AI cares about the detection of failure and the application of mechanisms that either allow for rectification, or at least for removal of the AI component:

- *Explainability* is thus relegated from a key requirement/characteristic to a method for achieving control, i.e. we do not assume that we can (or even want to) provide explainability. For example, we do not want to trade in 10% points of detection accuracy for explainability in a cancer detection system. Furthermore, this is very much related to actual, often unexplainable, human behaviour: In the example of driving, human actors can often not explain their decisions that led to certain events like, e.g., overlooking a car, yet we require autonomous driving to be fully explainable.
- The same holds true for *Transparency*, which is a requirement that we typically cannot achieve when dealing with human actors, as these forget things, or take decisions based on intuition.
- Regarding *Security and Resilience*, a fully secured and hardened system might even be a problem in cases where we want to introduce overrides or even emergency backdoors in order to help us remove a system gotten out of hand. So, while we do not tamper with this requirement too much in principle, and the introduction of a backdoor could be defined as a feature, most researchers in IT-Security consider this to be a weakness.
- With respect to *Privacy*, this is very much depending on the actual use-case, but should be considered as best as possible.
- We skip the key requirement of *environmental and societal well-being*, as this (i) is depending on the actual use-case and is highly debatable for applications in e.g. the military sector and also might depend on an ideological point of view. Furthermore, (ii) it does not integrate well with our approach of controlling systems per se.

4 Techniques for Controllable AI

While, of course, many techniques might be used to achieve control over an AI system, we want to focus on the techniques we consider to be either most prominent, interesting, illustrative for the approach or usable for practical applications. Thus, while this list is definitively not comprehensive, it should give a good overview on the key concepts. It must be noted that not all techniques will be applicable in every setting.

4.1 Detecting Control Loss

The first part in order to control a system lies in achieving detection capabilities, whether something went wrong and to what extent. Thus, detection of control loss is a fundamental task.

Providing Explainability: This is certainly one of the most powerful methods. By being able to explain decision making, or even provide a formal model of the system, control loss can be identified straightforward in many cases. Still, as we argued in Sect. 3, this might be impossible to reach for a given set of algorithms and/or data sets, also including methods for reinforcement learning that constantly change their model. Thus, as we have already outlined, we relegated the principle of Explainability to a method for achieving control.

Sanity Checks: In many application fields, while the exact result might be intransparent and hard to control for the human user, certain boundaries can be drawn where violations are simple to detect. Trivial examples include detection of testicular cancer in biological women, but often resort to a deeper understanding of the underlying workings of a (business) process like e.g. traffic in telecommunication networks based on weekdays, events or holidays. Such measures are often already in place in industrial environments when dealing with potentially incorrect sensor information.

Corrective Model with Alternative Data: As an extension of applying sanity checks, which we consider to be rather static, an alternative, corrective model could be trained on a different data set. This set needs to be more or less redundant to the original data, maybe using less data or simpler features, but close enough in order to generate the boundaries for sanity checks. Details, of course, very much depend on the actual use-case and data sets in question, as well as the additional effort introduced.

4.2 Managing Control Loss

In order to (re-)gain control, many different mechanisms can be applied. While the selection and often also the design will largely depend on the actual system in place, we provide an overview on some rather generic approaches that can be used in many different applications.

Divine Rules: Especially in optimization applications, the optimal solution from a mathematical point of view might not be the one aspired due to e.g. ethical reasons. These so-called *divine rules* could be coded into an additional rule-based model that either invokes the reward function in reinforcement learning in order to steer the model away, or overrule a decision made by the AI and trigger a warning.

Training Clearly Defined Non-goals: Defining non-goals and training the model accordingly can be a powerful tool. Thus, these non-goals have to be formulated in the form of training goals and relevant training data needs to be provided in case of trained models. However, stability of these goals needs to be discussed in cases of reinforcement learning or systems introducing an expert in the loop.

Destructive Backdoors: In some selected applications, e.g. in the military sector, it might be important to have means for shutting an AI off completely. While this currently sounds rather like Science Fiction, battlefield automation amongst other applications might require such a technique, especially when self-hardening of the system is also done by the AI. Typically, such a backdoor would be introduced on the logical (code) level, but might also include model components. This measure, of course, directly violates the principles of security from the definitions of Trustworthy AI.

Intentional Bias/Logical Backdoor: In certain cases it might be useful to introduce intentional bias into the trained model in order to steer the decision making, or even make certain results impossible. For example, this could be done in order to introduce positive discrimination into machine learning.

Fail-Safes and Logic Bombs: Apart from backdoors, which constitute a method for arbitrarily taking over control of the system, fail-safes are introduced into the AI beforehand and execute themselves depending on certain events inside the system, e.g. when certain decisions are reached that are incompatible with ethical values. Using logic bombs, these could reset the model or even shut down the entire system.

4.3 Support Measures

This section comprises measures that can help in the detection, as well as the management of control loss and are to some extent even required for controlling a system altogether.

Transparent AI Decisions: Making transparent, which decision was done by an AI and where other processes interfered is a very important technique, very much in vein with the original concept of Trustworthy AI and, to some extent, also required for compliance with the AI-Act. It is a pre-requisite for detecting, as well as managing control loss.

Transparent Data Management: As we have already outlined, in many machine learning based systems, data is as important, if not even more important, for the definition of a system as the code itself - still not a lot of attention has been put on this fact that we have to consider these systems as *data defined software*. Being able to decide, which data had been used at what point in time of the decision making is thus of the utmost importance for exerting control over such a system, as much as being able to explain the algorithm in use. This can be especially challenging in reinforcement learning.

5 Discussion

The notion of Controllable AI presented in this paper offers an alternative approach to addressing the challenges of securing and managing AI systems. By deviating from the strict principles of Trustworthy AI, Controllable AI acknowledges the limitations of achieving complete trustworthiness in real-life environments. Instead, it emphasizes the need for methods that enable the detection of malfunction and the ability to regain control over AI systems.

One of the key observations in Controllable AI is the shift of focus from code-centric approaches to data-centric approaches. While code plays a crucial role, the impact of data on AI systems, including issues like algorithmic bias and model vulnerabilities, cannot be ignored. Controllable AI recognizes the

importance of addressing data-defined software and the continuous evolution of models within the system lifecycle. This recognition highlights the significance of understanding and managing the actors involved in the learning process, as they influence the system's future behavior.

The techniques proposed for achieving Controllable AI provide practical insights into how control loss can be detected and managed. Measures such as sanity checks, alternative data training, transparent AI decisions, and the incorporation of divine rules or corrective models demonstrate potential avenues for ensuring control and mitigating undesired outcomes. However, the applicability of these techniques may vary depending on the specific use case and data sets involved.

It is important to note that the concept of Controllable AI does introduce exceptions to the principles of trustworthiness, particularly in terms of security. Techniques like introducing destructive backdoors or intentional bias raise ethical considerations and potential risks. Striking a balance between control and security while maintaining ethical standards is a critical aspect that needs to be carefully addressed in the development and deployment of Controllable AI systems.

6 Conclusion and Outlook for Future Research

This paper has proposed Controllable AI as an alternative approach to Trustworthy AI, focusing on achieving control and management of AI systems without compromising practical capabilities or transparency. By recognizing the limitations of achieving complete trustworthiness, Controllable AI provides a framework for detecting and managing control loss in AI systems. The techniques discussed offer practical insights into how control can be regained and undesired outcomes can be mitigated.

Future research in the field of Controllable AI should further explore and refine the proposed techniques. Extensive experimentation and case studies across different application domains would help validate the effectiveness of these techniques and identify their limitations. Additionally, ethical considerations associated with exceptions to trustworthiness principles, such as intentional bias or destructive backdoors, require in-depth investigation and guidelines for responsible implementation.

Furthermore, research efforts should focus on developing standardized methodologies and frameworks for assessing and certifying the controllability of AI systems. This would help establish guidelines and best practices for developers, regulators, and end-users, ensuring the safe and responsible deployment of Controllable AI in various domains.

As AI continues to advance and permeate various aspects of society, the discussion on securing AI systems and managing their behavior becomes increasingly crucial. The concept of Controllable AI offers a valuable perspective and opens up new avenues for research and development in this area. By embracing

the idea of control and management in AI systems, we can strive for more practical and accountable AI solutions that cater to the needs and concerns of both developers and end-users.

Acknowledgements. The authors declare that there are no conflict of interests. This work does not raise any ethical issues. Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554.

References

1. Asimov, I.: Three laws of robotics. Asimov, I. Runaround 2 (1941)
2. Bengio, Y., Lecun, Y., Hinton, G.: Deep learning for AI. *Commun. ACM* **64**(7), 58–65 (2021). <https://doi.org/10.1145/3448250>
3. Bubeck, S., et al.: Sparks of artificial general intelligence: early experiments with GPT-4. *arXiv:2303.12712* (2023). <https://doi.org/10.48550/arXiv.2303.12712>
4. Cabitza, F., et al.: Quod erat demonstrandum?-towards a typology of the concept of explanation for the design of explainable AI. *Expert Syst. Appl.* **213**(3), 118888 (2023). <https://doi.org/10.1016/j.eswa.2022.118888>
5. European Commission: Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. European Commission (2021). <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:52021PC0206>. proposal for a Regulation of the European Parliament and of the Council, No. COM/2021/206 final
6. Floridi, L., Chiriatti, M.: GPT-3: its nature, scope, limits, and consequences. *Mind. Mach.* **30**, 681–694 (2020). <https://doi.org/10.1007/s11023-020-09548-1>
7. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**(4), 193–202 (1980). <https://doi.org/10.1007/BF00344251>
8. Hacker, P., Engel, A., Mauer, M.: Regulating ChatGPT and other large generative AI models. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1112–1123 (2023). <https://doi.org/10.1145/3593013.3594067>
9. Hartree, D.R., Newman, M., Wilkes, M.V., Williams, F.C., Wilkinson, J., Booth, A.D.: A discussion on computing machines. *Proc. Royal Soc. London. Ser. A Math. Phys. Sci.* **195**(1042), 265–287 (1948)
10. Helberger, N., Diakopoulos, N.: ChatGPT and the AI act. *Internet Policy Rev.* **12**(1), 1–6 (2023). <https://doi.org/10.14763/2023.1.1682>
11. High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI. Publications Office of the European Union, Luxembourg (2019). <https://doi.org/10.2759/346720>
12. Hoenigsberger, F., et al.: Machine learning and knowledge extraction to support work safety for smart forest operations. In: *Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2022. LNCS, vol. 13480*, pp. 362–375. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-14463-9_23
13. Holzinger, A.: The next frontier: AI we can really trust. In: *Kamp, M., et al. (eds.) ECML PKDD 2021. CCIS, vol. 1524*, pp. 427–440. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-93736-2_33
14. Holzinger, A.: Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Inf. Fusion* **79**(3), 263–278 (2022). <https://doi.org/10.1016/j.inffus.2021.10.007>

15. Holzinger, A., Keiblinger, K., Holub, P., Zatloukal, K., Müller, H.: AI for life: trends in artificial intelligence for biotechnology. *New Biotechnol.* **74**(1), 16–24 (2023). <https://doi.org/10.1016/j.nbt.2023.02.001>
16. Holzinger, A., Saranti, A., Molnar, C., Biececk, P., Samek, W.: Explainable AI methods - a brief overview. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.R., Samek, W. (eds.) *xxAI 2020*. LNCS, vol. 13200, pp. 13–38. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-04083-2_2
17. King, M.R.: The future of AI in medicine: a perspective from a chatbot. *Ann. Biomed. Eng.* **51**(2), 291–295 (2023)
18. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biol.* **5**(4), 115–133 (1943). <https://doi.org/10.1007/BF02459570>
19. Röcker, C., Ziefle, M., Holzinger, A.: From computer innovation to human integration: current trends and challenges for pervasive HealthTechnologies. In: Holzinger, A., Ziefle, M., Röcker, C. (eds.) *Pervasive Health*. HIS, pp. 1–17. Springer, London (2014). https://doi.org/10.1007/978-1-4471-6413-5_1
20. Schneeberger, D., et al.: The tower of babel in explainable artificial intelligence (XAI). In: Holzinger, A., et al. (eds.) *CD-MAKE 2023*, LNCS 14065, pp. 65–81. Springer, Charm (2023). https://doi.org/10.1007/978-3-031-40837-3_5
21. Schwarting, W., Alonso-Mora, J., Rus, D.: Planning and decision-making for autonomous vehicles. *Ann. Rev. Control Robot. Auton. Syst.* **1**, 187–210 (2018). <https://doi.org/10.1146/annurev-control-060117-105157>
22. Tabassi, E.: Artificial intelligence risk management framework (AI RMF 1.0). *NIST AI 100-1* (2023). <https://doi.org/10.6028/NIST.AI.100-1>
23. Tjoa, S., Buttinger, C., Holzinger, K., Kieseberg, P.: Penetration testing artificial intelligence. *ERCIM News* **123**, 36–37 (2020)
24. Yang, Q., Steinfeld, A., Zimmerman, J.: Unremarkable AI: fitting intelligent decision support into critical, clinical decision-making processes. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–11 (2019). <https://doi.org/10.1145/3290605.3300468>

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

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

