Preprint

Moto Kamiura (in press)

The Four Fundamental Components for Intelligibility and Interpretability in AI Ethics.

The American Philosophical Quarterly, Special Issue on "The Ethics of AI".

Editor in chief: Patrick Grim

Guest editor: Luciano Floridi

https://www.press.uillinois.edu/wordpress/call-for-papers-special-issues-of-american-philosophy-quarterly-on-ai/

Accepted 27 February 2024

The Four Fundamental Components for Intelligibility and Interpretability in AI Ethics

Moto Kamiura

Abstract

Intelligibility and interpretability related to artificial intelligence (AI) are crucial for enabling explicability, which is vital for establishing constructive communication and agreement among various stakeholders, including users and designers of AI. It is essential to overcome the challenges of sharing an understanding of the details of the various structures of diverse AI systems, to facilitate effective communication and collaboration. In this paper, we propose four fundamental terms: "I/O," "Constraints," "Objectives," and "Architecture." These terms help mitigate the challenges associated with intelligibility and interpretability in AI by providing appropriate levels of abstraction to describe structure of AI systems generally, thereby facilitating the sharing of understanding among various stakeholders. The relationship between the Objective of AI designers and the Purpose of AI users is linked to the issues of AI alignment.

Key words: intelligibility, interpretability, explicability, level of abstraction, adjoint.

1. Explicability as an Ethical Principle in AI

Numerous organizations have initiated a broad spectrum of programs to set up ethical principles for the

implementation of artificial intelligence (AI) that benefits society. The importance of ethical guidelines concerning AI is widely recognized, and the establishment of these guidelines is to be welcomed. However, there is a potential problem of "principle proliferation" characterized by unnecessary repetition and redundancy, or, confusion and ambiguity among the numerous principles proposed (Floridi and Cowls 2019). To address this problem, Floridi et al. (2018), and, Floridi and Cowls (2019) offer a synthesis of the following six sets of principles produced by various reputable, multi-stakeholder organizations and initiatives: Asilomar AI Principles (2017), the Statement by the European Group on Ethics in Science and New Technologies (2018), the "five overarching principles for an AI code" offered in the UK House of Lords Artificial Intelligence Committee (2018), the Montreal Declaration for a Responsible Development of AI (2017), the IEEE Initiative on Ethics of Autonomous and Intelligent Systems (2017), and the "Tenets" of Partnership on AI (2018). For the yielded 47 principles, they find coherence and overlap between the six sets of principles. They compare the sets of principles with the set of four core principles commonly used in bioethics presented by Beauchamp and Childress (1979); beneficence (doing good), non-maleficence (avoiding harm), respect for autonomy (as a facility for decision-making), and justice (in distributing benefits and harms). Consequently, they find a well-adaptation of the four bioethical principles to the ethical challenges of AI. Moreover, they argue that a new fifth principle is needed in addition: i.e., explicability, understood as a synthesized concept of both in the epistemological sense of intelligibility and in the ethical sense of accountability.

In the field of biomedical ethics, various dedicated examinations are being conducted on the addition of explicability as the new fifth ethical principle. Ursin et al. (2022) conclude that the properties of explicability are

already covered by the four bioethical principles and therefore there is no need for explicability as the fifth principle for biomedical ethics. Conversely, Adams (2023) challenges the critics' premise that explicability cannot be an ethical principle like the classic four because it is explicitly subordinate to them and defends that the five principles including explicability are indeed better than the four when it comes to building an ethical framework for the development and implementation of AI in medicine.

Regardless of the position in these debates on whether explicability should be added as the fifth ethical principle, the following three points seem clear: Firstly, these discussions are about biomedical ethics. Although the starting point of the arguments presented by Floridi et al. (2018) refers to the four bioethical principles, the ultimate goal may be to clarify principles applicable across all areas where AI is used, not limited to the biomedical field. Secondly, the term explicability is not treated as a concept at the same level as explainability or other similar terms, but rather as a higher-order, comprehensive concept related to explanation, either as a principle or something akin to it. Thirdly, and most importantly for our research, there is a consensus that the explicability of AI is important, whether or not it is included as a principle.

2. Intelligibility and interpretability as epistemological sense on functioning of AI

The explicability of AI might be a theme where scientists including the authors of this paper, engineers, and AI developers, are likely to be called upon for more active engagement compared to other themes in AI ethics. It could also be an area where they can actually make significant contributions. This potential for contribution might be

particularly identifiable in aspects related to intelligibility and interpretability, as inferred from the following reasoning.

Floridi et al. (2018, p.700) presented explicability as a concept that synthesizes intelligibility and accountability, where intelligibility means the epistemological sense as an answer to the question "how does it work?" and accountability means the ethical sense as an answer to the question "who is responsible for the way it works?". Morley et al. (2020, p.2155), based on Binns et al. (2018), Cath (2018), and Lipton (2016), presented the arguments regarding explicability as follows: i.e. if a system is explicable (explainable and interpretable) it is inherently more transparent and therefore more accountable in terms of its decision-making properties and the extent to which they include human oversight and are fair, robust and justifiable. Moreover, Ursin et al. (2023) mapped the conceptions that commonly fall under the umbrella term explicability described in various literatures, and distinguished levels of explicability through conceptual analysis. The levels of explicability are associated with levels of opacity, corresponding to four incremental depths of explanation that start with disclosure, proceed through intelligibility and interpretability, and culminate in explainability. They show that these proposed ethical requirements for informed consent are related to the types of hurdles based on Ferretti et al. (2018) and Burrell (2016). Among these, the lack of disclosure might not be much of an issue if there is no intentional attempt to hide the use of AI, or it might be argued that all information systems should now be considered in some way related to AI. Intelligibility is associated with general epistemic opacity on the functioning of AI in general, and interpretability is associated with specific epistemic opacity on the functioning of a specific AI system, respectively. Explainability is associated with

explanatory opacity of the reason why the AI system reached a particular decision. Consequently, intelligibility and interpretability could be related to the epistemological sense on the functioning of AI, and explainability could be related to the ethical sense on accountability or explanatory responsibility.

3. Requirements for our analysis on intelligibility and interpretability

The previous section found that intelligibility and interpretability are challenges in AI ethics to which we, as scientists, can primarily contribute. However, this does not mean they can be separated from explainability or their interdependencies can be ignored. The following three quotes may help further clarify our challenges.

> "Note, however, that ethically speaking transparency and explainability are not necessarily and certainly not only about disclosing the software code. The issue is mainly about explaining *decisions* to people." (Coeckelbergh 2020, p.121)

> "Again, what level of depth is satisfactory in providing details about the conditions of algorithmic (or algorithm-supported) decision making may depend very much on the context and recipient." (Herzog 2022, p220)

> "Part of the difficulty is to get the level of abstraction right (Floridi 2008a; 2008b), …" (Floridi 2019, p.1)

Regarding intelligibility and interpretability, the important aspect is not the mere disclosure of software code, but the provision of a foundation for building better explainability that enables accountability, and the clarification of an appropriate level of abstraction for this purpose.

Let us go back to the ethical guiding question for intelligibility and interpretability in Ursin et al. (2023, p.184): i.e., respectively, "How do AI systems generally work (input, output, training data, parameter, calculation)?" and "How does that specific AI system work (input, output, training data, parameter, calculation)?". While these questions offer significant hints, a detailed analysis is required to establish a more appropriate set of concepts and terminology for intelligibility and interpretability. The requirements for our analysis are threefold: (1) Despite the existence of a very diverse range of current AIs and the further unpredictable development of future AIs, to present a set of concepts and terminology that can "generally" answer their workings, (2) for the correspondence between those concepts and "input, output, training data, parameter, calculation" to be clear, and (3) for the concepts used to describe "AI systems generally" and those used to describe a "specific AI system" to be consistently applied.

4. Four fundamental terms for intelligibility and interpretability

To arrange for intelligibility and interpretability while meeting the requirements extracted in the previous section, I propose the use of a set of four fundamental terms: "I/O," "Constraints," "Objectives," and "Architecture." These terms provide a framework for unified design, analysis, and understanding of systems within system theory and

systems engineering. They are applicable across a wide range of system classes related to AI, including physical systems, control engineering systems, and machine learning systems, aiding in the comprehension of each system's structure. As a result, a terminological foundation is laid for building better explainability that enables accountability, clarifying the appropriate levels of abstraction for intelligibility and interpretability. The term "states" or "internal states," which might be considered as a basic term, is intentionally omitted to avoid the black-boxing of systems as much as possible. The overview of the four terms is as follows, and their correspondence to the terms demonstrated by Ursin et al. (2023) is annotated in brackets [...]:

I/O (Input/Output) refers to the pair of information entered into a system and the results outputted from the system. AI systems receive inputs from the outside, process them, and produce outputs. I/O is the interaction between the system and the external world, and it is important for understanding the system's basic operations and functions. [input, output, training data]

Constraints refer to the conditions that the design and operation of a system must satisfy. This is a generalization of the narrow concept of constraint conditions as a mathematical term. This includes parameters that define internal constraints of the system, constraints related to I/O, physical constraints, and resource constraints. Constraints form the foundation for a system to function feasibly and effectively. [parameter]

Objectives have a dual meaning. In a broad meaning, objectives refer to the ultimate goals or outcomes that the

system aims to achieve, defining what problems it seeks to solve. This serves as a guideline for system design and plays a central role in the system evaluation process. This is determined by system designers and should be distinguished from “Purposes” as motivations of system users. In a narrow meaning, objectives refer to the objective functions that provide criteria for the selection of parameters within the system. The latter can sometimes be considered a mathematical representation of the former, though the relationship between the two may not always be clear, necessitating further consideration. In either interpretation, a system can have one or multiple objectives. In natural sciences, teleological explanations are handled with particular care to avoid. [No corresponding term]

Architecture defines the structure of the entire system by integrally linking the above three elements (I/O, Constraints, Objectives). It outlines how the system’s components or subsystems interact, how information flows, and how these are integrated to achieve the overall objectives. In the simplest cases, this may be represented as an adjoint of I/O and parameters. Concepts on system dynamics such as computation and control are also embedded within a static structure that spans the entire execution time. [calculation]

These four terms together form the overall picture of a system, interrelating with each other. I/O defines the interface between the system and the external world. Constraints limit the scope of its operation. Objectives set the goals that the system aims to achieve. Architecture provides the structure of how these elements are integrated for the entire system to function.

As detailed later, the relationship between I/O and Constraints is abstracted from the adjoint structure of variables and parameters in mathematical functions. Furthermore, Objective is abstract from the role of the objective function. Note that such a perspective is built upon inheriting the concept of dynamic adjoint, a method for modeling and analyzing systems that include observers and observational actions. This concept was developed during the transitional period between the 1990s when complex systems science advanced, and the 2010s when machine learning became prevalent (Gunji and Kamiura 2004; Gunji et al. 2006; Kamiura and Gunji 2006; Kamiura 2013).

To enhance intelligibility and interpretability, it is crucial to have an integrated understanding of how the actual components of a system contribute to its performance and the achievement of its Objectives. These four terms facilitate an understanding of the roles of the system's components. These originate from mathematical models related to machine learning and systems theory, respecting the existence of corresponding mathematical entities, while deliberately adopting names different from technical terms in mathematics and science to align with the levels of abstraction required for intelligibility and interpretability.

5. Relationships of the four fundamental terms

In this section, we analyze in detail how the four fundamental terms introduced in the previous section relate within a system. This analysis corresponds to intelligibility. The consistent premise in this analysis is as follows: all systems involving input and output are modeled and understood through some function $y = f(x; a)$. Here, x represents the input, y the output, a the parameters, and f is a family of functions, all of which possess appropriate

dimensions. The family of functions is a major part of the system's architecture, but not the entirety. Specifying the values of parameters a means selecting a specific function $f_a(x) = f(x; a)$ from the family of functions f. Let us now begin the analysis of the four terms.

5.1. Parameters as Constraints

From the premise mentioned above, we understand why the system's parameters are classified under Constraints. Determining parameters is equated with selecting a specific function from the family of functions defined by the system's architecture, thereby linking parameters to Constraints. While architecture provides the basic framework for the system's structure and function, parameters are the elements that define specific behaviors and performance within that framework.

5.2. Constraints and Objectives

In AI systems, objective functions (i.e., the narrow meaning of Objectives) fulfill the role of determining specific values for parameters as Constraints. While some parameters determined through the trial and error of system developers may still remain, the automation of parameter determination using Objectives is one of the essential characteristics of machine learning systems. The objective function is a function that takes parameters as arguments, and the process of finding the values of parameters that maximize or minimize the value of this objective function is called optimization. Optimization algorithms (i.e. optimizers) may be designed analytically with precision, or as an exploration process in the form of probabilistic mechanical trial and error.

5.3. Objectives and I/O

When practically constructing Objectives and operating an optimizer, the components that become part of those Objectives are the system's I/O. As a very simple example, consider regression analysis, which uses observed data points (pairs of input and output) to estimate the parameters of a certain function (model). The narrow meaning of Objective here is to determine the values of parameters that minimize the sum of squared differences between the outputs predicted by the model and the actual output data. By doing so, a function is obtained that can predict outputs for unknown inputs. Making such predictions is the broad meaning of Objective. The residual sum of squares (RSS) as the objective function consists of data corresponding to I/O. Training data is a collection of pairs of the independent variables (input) and the dependent variables (output) of actual observational data. The teaching data is the predicted values (output) obtained for the input fed into the model. This is a basic method of statistics, but the principles used are fundamentally the same as those operating complex neural networks. However, the relationship between the narrow meaning of Objectives and the broad meaning of Objectives can become more indirect and difficult to understand.

5.4. More general perspective

The above is an example of statistical machine learning, but the idea that the components of Objectives are the system's I/O and that Objectives induce constraints can be applied more generally across a wide variety of systems and fields of application. Regardless of what specifically a system aims to achieve, its outcome depends on inputs to

the system and the generation of outputs under certain conditions. For instance, in system control, the Objective might be to maintain a specific process variable at a control target value. This can be interpreted as the Objective dynamically imposing Constraint based on the state of the system. In feedback control system, the inputs to the system are disturbances and feedback from the controller, with the output being measured values related to the system's state. In PID controller, the parameters are fixed values provided by the designer, and the controller corresponding to the Objective acts directly on I/O, not on parameters. Adaptive control is closer to the concept of regression analysis, performing system identification through parameter estimation. However, the historical development of system control for physical objects and machine learning for data, while not unrelated, are distinct, requiring further interdisciplinary collaboration. Our four terms might also provide new insights into such perspectives.

5.5. Machine learning processes reducing Architecture to adjoint structure and hidden Objectives

There exist systems where Objectives do not appear explicitly (where Objectives are hidden). In the earlier example of regression analysis, after the calculation or the mathematical proof is carried out, only a mathematical structure known as the adjoint remains between I/O and parameters. At first glance, it might seem as though the Objectives have disappeared, but they are merely hidden behind the calculation. The minimization of loss functions in machine learning operates similarly, with the model's parameters being adjusted throughout the training process. In the final model, after training is completed, the loss function (Objective) is optimized and ceases to change, becoming invisible. Thus, the phenomenon of objectives becoming hidden in the design or analysis of a system does

not mean they are gone; rather, it signifies that the necessary optimization for their achievement has been completed. Consequently, Machine learning processes reduce the Architecture of the system to the adjoint structure and hidden Objectives, and the system converges to a function characterized by its adjoint.

The foregoing discussion has elucidated our proposed four fundamental terms: "I/O," "Constraints," "Objectives," and "Architecture," along with their interrelationships. This can be summarized as illustrated in Figure 1 below.

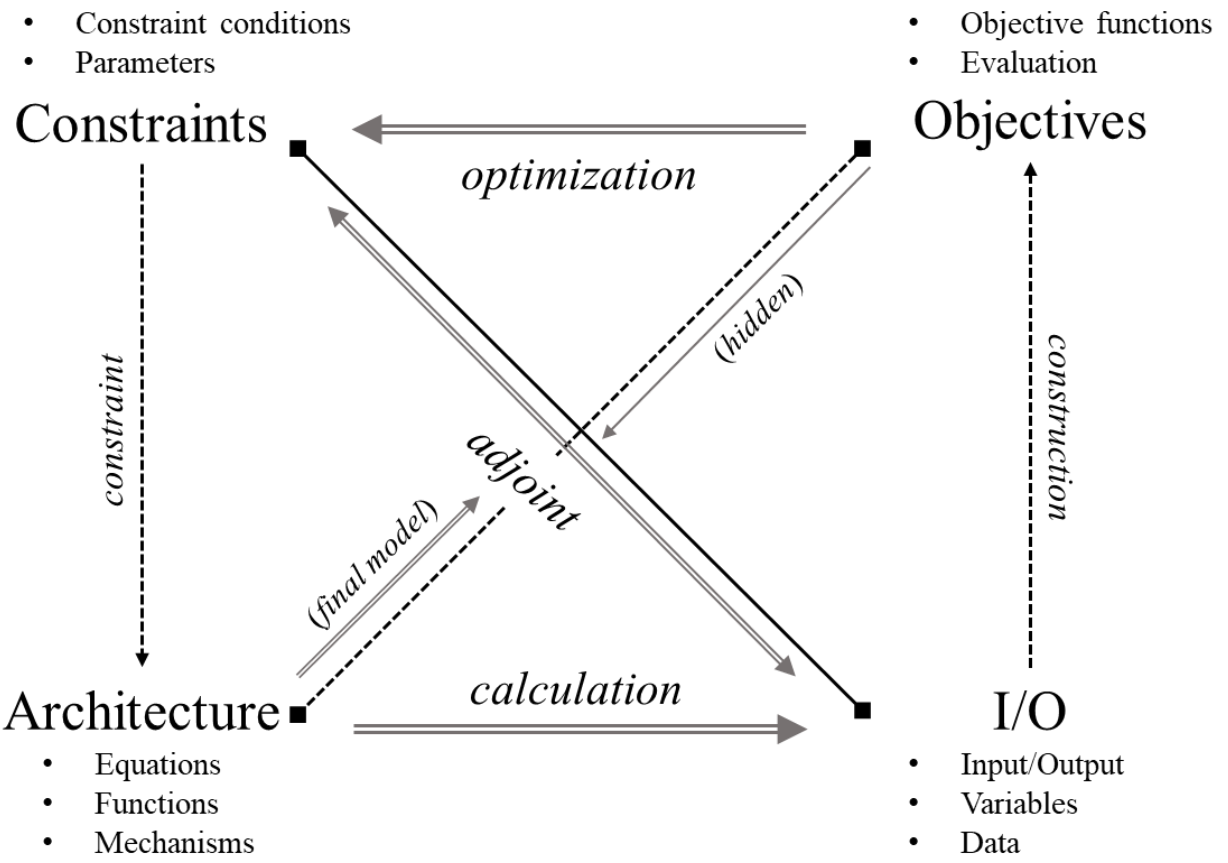

Figure 1. The four fundamental terms and their interrelationships.

5.6. Architecture of deep learning

The description in this section corresponds to interpretability. The interpretability provided by the four fundamental terms and their interrelationships may lower the barrier to technical summaries concerning deep learning: Deep learning, similar to the machine learning approaches mentioned above, adjusts the parameters of each layer in a multi-layered neural network to achieve the desired output in response to the input. [I/O:] Each layer extracts features of the input data at different levels. [Objectives and Constraints:] During the learning process, algorithms such as backpropagation are used to adjust the parameters (weights and biases) of each layer based on the input and the desired output (teaching data). [Architecture:] By repeating this process, the system transforms into a sort of function that makes optimal predictions or classifications for the task at hand.

A trained deep neural network can be considered a "high-dimensional composite function." Essentially, a deep neural network is a composition of functions across multiple layers, where each layer performs some form of mathematical operation on the input (for example, the application of a nonlinear activation function following a linear transformation by weights). Images, texts, and audio signals represent high-dimensional input and output data. Each layer within a deep neural network can be viewed as a function that maps input vectors to another multidimensional space. The composition of these layers ultimately transforms the input data into the desired output format. The structure of this chain of composite functions explains the network's ability to learn and represent complex nonlinear relationships and patterns. In other words, deep learning models can be viewed as highly complex collections of

"high-dimensional composite functions" that sequentially process and transform data through multiple layers (functions) to convert input data into the final output.

6. Alignment of Objectives and Purpose

In this section, we consider the relationship between Objective as an aim and as a function of optimization for AI system designers, and Purpose as an aim of AI system users. When the level of explicability migrates from intelligibility and interpretability to explainability, we must pay attention to the relation between Objective and Purpose. The question for intelligibility and interpretability is "how?", whereas for explainability, it is "why?" (Coeckelbergh 2020; Ursin et al. 2023). The latter question brings us closer to Objective and Purpose compared to the former.

The design of the Objective can be influenced by the Purpose. The proposed narrow meaning of Objective is the technical element that determines the functionalities of the AI system. The broad Objective is linked to the expectations for the functionalities of the AI system as a result of the narrow Objective's operation. On the other hand, the Purpose relates not only to the needs and expectations of the users but also to the external environment and social context surrounding the system. It defines how the system generates value and contributes to users and society, ultimately constraining that AI system.

In conventional industrial products other than AI, Objective and Purpose could be usually aligned. An appropriate

manual for the product may be provided, and the designer's intention might be communicated to the user. While the responsibility for product defects lies with the designer or manufacturer, the responsibility for using a flawless product inappropriately beyond the designer's intention falls on the user. For example, automobiles and gasoline have the potential versatility to be used for various purposes. The mismatch between Objective and Purpose opens up both to the user's positive creativity and to unexpected dangers. By aligning Objective and Purpose, we can use them safely. Such alignment is supported by well-established laws and our common sense. It is also the result of historical development. There was a time lag even between the Watt steam engine and an early experimental steamship. While steam engines have versatility, whether a specific use develops may depend on whether that use is attractive to users and society, and whether it brings significant benefits to designers and manufacturers.

In the case of AI, the users and makers are currently in the phase of trial and error regarding what AI is truly capable of and what it should be tasked with. Even OpenAI, a frontrunner in generative AI, seems not to have a clear grasp of the ultimate goal of artificial general intelligence (AGI) (Altman 2023). Artificial General Intelligence has the positive aspect of being able to respond to the constantly changing and diverse Purposes of various users. However, when this "general" implies a misalignment between Objective and Purpose, or the disappearance of Purpose, it could increase the risk of unforeseen ethical issues. Such challenges are studied as AI alignment and may indeed be formulated as a multi-objective problem (Vamplew et al. 2018). To ensure that the risks associated with AI are managed as adequately as those with existing industrial products, further research and societal consensus are necessary.

7. Conclusion

Intelligibility and interpretability are considered as levels of explanation that answer the "how?" question regarding AI, and are seen as part of the comprehensive concept of explicability. To practically construct intelligibility and interpretability, it is necessary not to disclose AI program codes, but to establish appropriate levels of abstraction for generally explaining the structure of AI. For this issue, we propose four fundamental terms: "I/O," "Constraints," "Objectives," and "Architecture." These terms help mitigate the challenges associated with intelligibility and interpretability in AI by providing appropriate levels of abstraction to describe structure of AI systems generally, thereby facilitating the sharing of understanding among various stakeholders. We are currently in the phase of trial and error regarding what AI is truly capable of and what it should be tasked with. The misalignment between the Objective of AI designers and the Purpose of AI users leads us to the issues of AI alignment.

Acknowledgments

I wish to express my deep gratitude and respect for the thoughtful feedback provided by the two reviewers of the previous version of this article. I would also like to extend my heartfelt thanks to Dr. Kazunori Kondo (Osaka University), Dr. Kohei Nakajima (The University of Tokyo), and Dr. Yuta Nishiyama (Nagaoka University of Technology) for their insightful discussions on this study, although I am solely responsible for all aspects of the final article.

Affiliation of author

Moto Kamiura

Institute for Advanced Research and Education, Doshisha University

1-3, Tatara-Miyakodani, Kyotanabe, Kyoto, JAPAN

mkamiura@mail.doshisha.ac.jp

References

Adams, J. (2023). Defending explicability as a principle for the ethics of artificial intelligence in medicine. *Medicine, Health Care and Philosophy*, Vol.26, pp.615–623. https://doi.org/10.1007/s11019-023-10175-7

Altman, S. (2023). Planning for AGI and beyond. February 24, 2023. Retrieved February 13, 2024 from https://openai.com/blog/planning-for-agi-and-beyond

Asilomar AI Principles. (2017). Retrieved February 13, 2024 from https://futureoflife.org/open-letter/ai-principles/

Beauchamp, T.L. and Childress, J.F. (1979). *Principles of biomedical ethics*. Oxford: Oxford University Press.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. (2018). 'It's reducing a human being to a

percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI conference on human factors in computing systems—CHI'18* (pp. 1–14). https://doi.org/10.1145/3173574.3173951

Burrell, J., (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, Vol.3(1). https://doi.org/10.1177/2053951715622512

Cath, C. (2018). Governing Artificial Intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol.376(2133), 20180080. https://doi.org/10.1098/rsta.2018.0080

Coeckelbergh, M. (2020). *AI Ethics*. Cambridge, MA: The MIT Press.

European Group on Ethics in Science and New Technologies. (2018). Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems. Retrieved February 13, 2024 from https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1

Ferretti A, Schneider M, and Blasimme, A. (2018). Machine learning in medicine: Opening the New Data Protection Black Box. *European Data Protection Law Review*, Vol.4(3), pp.320–332. https://doi.org/10.21552/edpl/2018/3/10

Floridi, L. (2008a). The method of levels of abstraction. *Minds and Machines*, Vol.18(3), pp.303–329. https://psycnet.apa.org/doi/10.1007/s11023-008-9113-7

------ (2008b). Understanding epistemic relevance. *Erkenntnis*, Vol.69(1), pp.69–92. https://doi.org/10.1007/s10670-007-9087-5

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., and Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, Vol.28, pp. 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L. (2019). What the Near Future of Artificial Intelligence Could Be. *Philosophy & Technology*, Vol.32, pp. 1–15. https://doi.org/10.1007/s13347-019-00345-y

Floridi, L. and Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, Vol.1(1). https://doi.org/10.1162/99608f92.8cd550d1

Gunji, Y.-P. and Kamiura, M. (2004). Observational Heterarchy Enhancing Active Coupling. *Physica D*, Vol.198, pp.74-105. https://doi:10.1016/j.physd.2004.08.021

Gunji, Y.-P., Miyoshi, H., Takahashi, T. and Kamiura, M. (2006). Dynamical duality of type- and token-computation as an abstract brain. *Chaos, Solitons & Fractals*, Vol.27, pp.1187-1204. https://doi:10.1016/j.chaos.2005.01.067

Herzog, C. (2022). On the risk of confusing interpretability with explicability. *AI and Ethics*, Vol.2, pp.219–225. https://doi.org/10.1007/s43681-021-00121-9

House of Lords Artificial Intelligence Committee. (2018). AI in the UK: ready, willing and able? Chapter 9, paragraph 417. Retrieved February 13, 2024 from https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10013.htm#_idTextAnchor152

IEEE Initiative on Ethics of Autonomous and Intelligent Systems. (2017). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. Retrieved February 13, 2024 from https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Kamiura, M. and Gunji, Y.-P. (2006). Robust and Ubiquitous On-Off Intermittency in Active Coupling. *Physica D*, Vol.218, pp.122-130. https://doi:10.1016/j.physd.2006.04.006

Kamiura, M. (2013). Implicit Interaction: Mathematics on Local Description of Systems. *Transactions of the Society*

of Instrument and Control Engineers, Vol.49(1), pp.190-196. https://doi.org/10.9746/sicetr.49.190

Lipton, Z. C. (2016). The mythos of model interpretability. http://arxiv.org/abs/1606.03490

Montreal Declaration for a Responsible Development of Artificial Intelligence. (2017). Retrieved February 13, 2024 from https://montrealdeclaration-responsibleai.com/the-declaration/

Morley, J., Floridi, L., Kinsey, L., and Elhalal. A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, Vol.26, pp.2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Partnership on AI. (2018). Retrieved February 13, 2024 from https://partnershiponai.org/about/

Royal Society. (2019). Explainable AI. Retrieved February 13, 2024 from https://royalsociety.org/topics-policy/projects/explainable-ai/

Ursin, F., Timmermann, C., and Steger, F. (2022). Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary? *Bioethics*. *Special Issue: Promises and Challenges of Medical AI*, Vol.36(2), pp.143-153. https://doi.org/10.1111/bioe.12918

Ursin, F., Lindner, F., Ropinski, T., Salloch, S., and Timmermann, C. (2023) Levels of explicability for medical artificial intelligence: What do we normatively need and what can we technically reach? *Ethik in der Medizin*. Vol.35, pp.173–199. https://doi.org/10.1007/s00481-023-00761-x

Vamplew, P., Dazeley, R., Foale, C., Firmin, S., and Mummery, J. (2018). Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, Vol.20, pp.27–40. https://doi.org/10.1007/s10676-017-9440-6