Artificial Intelligence, Context and Emotion.
Dreyfus' Critique of Symbolic AI from an Ethical Perspective

Master's Thesis

Helena Winiger

helenalow97@zedat.fu-berlin.de
5390427

Institute of Philosophy
Department of Philosophy and Humanities
Freie Universität Berlin

Supervisors:
PD Dr Werner Kogge
Prof Dr Christoph Benzmüller

July 18, 2022

To Flo, Rahaf and Philomena.

I would like to thank Prof Dr Kogge and Prof Dr Benzmüller very sincerely.
With your guidance, I was able to find a small consensus between apparent opponents -
the symbolic and the sub-symbolic approach - even though other tensions arose along
the way: rationalist and phenomenological undertakings, mind and brain,
analogies and anticipations. I hope I was able to maintain my position
as a compromise of the approaches despite these tensions.

Table of contents

List of Figures

Introduction

The phenomenon of intelligence still raises fascinating questions. Hence studying it means entering unknown depths and constantly giving rise to new disciplines. Perhaps the attraction of the phenomenon stems precisely from the fact that numerous mysteries still seem unsolved: We are left with the phenomenon itself. Nevertheless, in the research field of AI, efforts are being made to thoroughly assess and even replicate the phenomenon by technical means. Historically and methodologically, two different approaches are opposed to one another, reflecting the dichotomy of mind and brain: The initially dominant symbolic approach to AI attempts to capture the phenomenon of the mind by logically representing world knowledge, ontologically relating it, and automatically reasoning on its basis. In contrast, today's dominant sub-symbolic AI intends to bionically mimic the human brain with artificial neural networks that sense the world and learn from it through data.

Dreyfus' work „What Computers Still Can't Do" criticizes symbolic AI from a phenomenological perspective: Its underlying computationalist assumptions are misleading, as intelligence is embodied, context-sensitive, and emotionally anchored. Perceiving symbolic AI as a realization of the assumptions in practice, Dreyfus rates it as a lost cause early on and advocates the sub-symbolic approach. Although he seems to have been right in his prediction of the dominance of sub-symbolic AI, the symbolic approach is championed in this thesis. Symbolic AI today brings forth promising techniques that enable logical reasoning and ethical governance in hybrid systems: Symbolic techniques are not only transparent themselves but can also introduce transparency into sub-symbolic black box systems by committing AI actions to reasons based on ethical values.

After an introduction to the symbolic and sub-symbolic AI approaches, this thesis reconstructs Dreyfus' arguments against symbolic AI, which became known as GOFAI during his time. On this ground, the strengths and weaknesses of both approaches are discussed from today's point of view. It is shown that despite the phenomenological strengths of sub-symbolic AI and remaining questions about the axiomatic nature of symbolic AI, symbolic AI must regain focus because of its unique potential to bring ethical standards to AI systems. From today's perspective, an essential dimension of AI criticism is not only "What Computers Still Can't Do" but also "what computers shouldn't do"[1]. Dreyfus' critique should therefore be extended to include the dimension of AI ethics. An approach on three levels – the ethical,

[1] van der Meulen, S. & Bruinsma, M., 2019. Man as 'aggregate of data'. What computers shouldn't do. *AI & SOCIETY, 34*, p. 343.

technical, and legal – is proposed, which aims at enabling the ethical development and deployment of AI technologies holistically. This requires the integration of symbolic AI in a hybrid setup.

Literature Review

The thesis builds on Dreyfus' GOFAI critique 'What Computers Still Can't Do'[2]. Since Dreyfus relies strongly on them, Heidegger[3] and Merleau-Ponty[4] were also included, whereas Heidegger shaped the understanding of contextual situatedness and Merleau-Ponty the role of intentionality. Hereby, the emphasis on the role of emotion for intelligence was given in advance by the reflections of Asma and Gabriel[5] and by the interpretation of Dreyfus' work given by Dreyfus' scholar Haugeland interviewed for the documentary 'Being in the world'[6]. As with this work, Coeckelbergh[7] highlights that Dreyfus neglected a social and ethical perspective and consequently relates Dreyfus to specific ethics.

The defense and advocacy of symbolic AI from an ethical perspective is based on Benzmüller's theories of ethico-legal governance and applications of ethico-legal 'governors'[8,9]. To open and illustrate a philosophical perspective on an interdisciplinary field, Kogge's[10] guidelines on interdisciplinary collaboration were followed, and the scientific type of interpretative adequation was methodologically adopted. Concerning ethico-legal governance, and thus the explanatory works of Boden[11], Buckner[12], Marcus[13], and Sun[14], a confrontation between the symbolic and the sub-symbolic approach was undertaken.

---

[2] Dreyfus, H., 1992. *What Computers Still Can't Do. A Critique of Artificial Reason.* Cambridge: The MIT Press.
[3] Heidegger, M., 1993. *Sein und Zeit.* 17. Edition. Tübingen: Max Niemeyer Verlag.
[4] Merleau-Ponty, M., 2005. *Phenomenology of Perception.* London: Routledge.
[5] Asma, S. & Gabriel, R., 2019. *The Emotional Mind. The Affective Roots of Culture and Cognition.* Cambridge MA: Harvard University Press.
[6] *Being in the World.* 2010. [Movie] Direction: Tao Ruspoli. US: Canavesio, Giancarlo; Redlich, Christopher.
[7] Coeckelbergh, M., 2019. Skillful coping with and through technologies. *AI & SOCIETY, 34*, p. 269–287.
[8] Benzmüller, C. & Lomfeld, B., 2020a. Reasonable Machines: A Research Manifesto. In: U. Schmid, F. Klügl & D. Wolter, Edt. *Advances in Artificial Intelligence, 43rd German Conference on AI.* Berlin: Springer, pp. 251-258.
[9] Benzmüller, C., Parent, X. & van der Torre, L., 2020b. Designing Normative Theories for Ethical and Legal Reasoning: LogiKEy Framework, Methodology, and Tool Support. *Arxiv*, pp. 1-50.
[10] Kogge, W., 2022a. *Einführung in die Wissenschaften. Wissenschaftstypen – Deutungskämpfe – Interdisziplinäre Kooperation.* Bielefeld: Transcript.
[11] Boden, M. A., 2014. GOFAI. In: K. Frankish & W. M. Ramsey, Edt. *The Cambridge Handbook of Artificial Intelligence.* Cambridge: Cambridge University Press, pp. 89-107.
[12] Buckner, C., 2019. Deep learning: A philosophical introduction. *Philosophy Compass, 14,* pp. 1-19.
[13] Marcus, G., 2018. Deep Learning: A Critical Appraisal. *Arxiv*, pp. 1-27.
[14] Sun, R., 2014. Connectionism and neural networks. In: K. Frankish & W. Ramsey, Edt. *The Cambridge Handbook of Artificial Intelligence.* Cambridge: Cambridge University Press, pp. 109-127.

The phrase "what computers should not do" was derived from van der Meulen and Bruinsma[15] , and it serves as a leitmotif within the interpretative adequation. However, the original authors use it primarily to refer to the increasing liquification of human identity and dissolution of authenticity in an expanding virtual world in which emerging technologies abolish boundaries and, above all, human beings are quantified by bits. The phrase "what computers should not do" is used here in a new context because it also indicates the ethical dimension by which Dreyfus' critique must be expanded from today's perspective. Moreover, it can be understood as a metaphor for how AI criticism grew out of the supposedly constructivist symbolic-sub-symbolic tension and took on a social view that formed a whole subfield of AI: AI ethics. In finding an ethical position within AI ethics, Grunwald's[16] and Misselhorn's[17] works served as an orientation in technology ethics. The differentiation of the concrete ethical values was thus achieved with the help of Floridi et al.[18], Jobin et al.,[19] and Hagendorff.[20] Several regulations and reports of the European Commission in the field of AI were considered in order to illuminate external possibilities for governance[21,22]. An outlook on standardization processes was initiated by Lorenz.[23]

Furthermore, books were consulted which addressed the juxtaposition of symbolic and sub-symbolic AI: Domingos' 'The Master Algorithm'[24] strongly advocates sub-symbolic AI, as it describes the vision of a singular super-intelligent algorithm capable of learning all world knowledge. Marcus and Ernest's 'Rebooting AI'[25] stresses that symbolic AI should not be neglected in research and development from an ethical point of view. Russell and Norvig's

[15] van der Meulen & Bruinsma, 2019, pp. 343-354.
[16] Grunwald, A., 2011. Technikethik. In: M. Düwell, C. Hübenthal & M. Werner, Edt. *Handbuch Ethik.* Stuttgart / Weimar: J.B. Metzler, pp. 283-287.
[17] Misselhorn, C., 2019. Maschinenethik und Philosophie. In: O. Bendel, Edt. *Handbuch Maschinenethik.* Wiesbaden: Springer, pp. 33-56.
[18] Floridi, L. et al., 2018. AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines, 28*, pp. 689-707.
[19] Jobin, A., Ienca, M. & Vayena, E., 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, pp. 389–399.
[20] Hagendorff, T., 2022. A Virtue-Based Framework to Support Putting AI Ethics into Practice. *Philosophy & Technology, 35*, pp. 1-24.
[21] European Commission, 2021a. Regulation of the European Parliament and of the Council. Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. *COM(2021) 206 final*, pp. 1-107.
[22] European Commission, 2021b. Trustworthy Autonomous Vehicles. Assessment criteria for trustworthy AI in the autonomous driving domain. *JRC Science for Policy Report*, pp. 1-72.
[23] Lorenz, P., 2021. *AI Standardization and Foreign Policy. How European Foreign Policy Makers can engage with Technical AI Standardization,* Berlin: Stiftung Neue Verantwortung.
[24] Domingos, P., 2017. *The Master Algorithm. How the Quest for the ultimate Learning Machine will remake our World.* London: Penguin Books.
[25] Marcus, G. & Davis, E., 2019. *Rebooting AI. Building Artificial Intelligence We Can Trust.* New York: Penguin.

'Artificial Intelligence A Modern Approach'[26] was used to get an overview and to trace specific approaches in more detail, whereas Russell's 'Human Compatible'[27] opened up various views on risks posed by AI technologies. Lastly, Wooldridge's 'The Road to Conscious Machines'[28] helped for a historical lens on AI.

Main part

Overall, the work is methodologically oriented toward the scientific type of *interpretative adequation* as identified and characterized by Kogge:[29] Particularly, it is shown that Dreyfus' critique of 'What Computers Still Can't Do' must be interpretatively expanded to include an ethical dimension called "What computers shouldn't do", which makes symbolic AI a necessity despite Dreyfus' critique of GOFAI. In this thesis, GOFAI is in contrast to symbolic AI in general referring only to the early phase of symbolic AI research and development.[30]

In order to engage with Dreyfus' work, the thesis will show how his critique is positioned concerning the aspects of context and emotion in a reconstructive manner: What premises does Dreyfus imply for the phenomenon of intelligence by identifying missing assumptions of GOFAI research that prevent essential preconditions for intelligence? Only then can the further dimension be opened up through 'interpretative adequation', which describes a broad adaptation of the critique in terms of a holistic ethical approach to the research, development, and deployment of AI on three levels: the ethical, technical and legal level. It is more precisely assumed that Dreyfus' critique must be extended from today's perspective to embrace the concept of hybrid AI – as symbolic AI enables techniques that are necessary from an ethical point of view.

[26] Russell, S. & Norvig, P., 2021. *Artificial Intelligence. A Modern Approach.* 4th Edition Edt. Hoboken: Pearson.
[27] Russell, S., 2020. *Human Compatible. AI and the Problem of Control.* New York: Penguin Random House.
[28] Wooldridge, M., 2021. *The Road to Conscious Machines. The Story of AI.* Dublin: Pelican Books.
[29] Cf. Kogge, 2022a, p. 175 f.
[30] Although Boden explains that GOFAI refers to symbolic AI in general, it will in this thesis be differentiated between the notions of GOFAI and symbolic AI. To make time-bound aspects of Dreyfus critique of symbolic AI evident, which was first published in 1972 and ultimately revised in 1992, GOFAI will solely denote symbolic AI research and development between the 1950s and the 1980s because, in these years, symbolic AI was dominating AI research and development. In contrast, ‚symbolic AI' will refer to the *approach in general*, including new developments in the field. By the time Dreyfus' scholar Haugeland introduced the term, the symbolic AI approach was already experiencing an AI winter and has been dominated by sub-symbolic approaches in many capabilities ever since. However, nowadays symbolic AI systems bear new potential which seems to stand in contrast with the attribute "old-fashioned". Cf. also Boden, 2014, p. 89.

1. Background: Symbolic and sub-symbolic AI

1.1. Human and Artificial Intelligence

Human Intelligence is a phenomenon that opens up many questions in different research fields. It is fundamental to human existence, thus, it is also central to philosophy. In the philosophy of mind, intelligence is often referred to[31] as the "general ability required for complex cognitive tasks like language processing, analogical reasoning, mathematical and logical reasoning, creative reasoning […], theoretical and practical problem-solving, playing chess, etc."[32]. Important components of intelligence are not only reasoning but also "the features of rationality, effectiveness, and flexibility"[33]. As a conceptual part of intelligence, the notion of cognition is strongly influenced by cognitive science. The term focuses on the skills of "language […], problem solving […], attention, memory […], and perception"[34]. Cognitive processes take place in the nervous system. From a cognitive science view, they are understood as similar or even analogous to a computer as they can be described as involving cognitive outputs as a symbolic and rule-based result of distinct inputs, i.e., mainly physical stimuli.[35] The computationalist position[36] intensifies this understanding of cognition by holding that the *brain is actually a computer*.[37] This view is criticized by philosophers such as Hubert Dreyfus, who hold embodiment and emotion as constitutive to intelligence[38] and supported by neuroscientific perspectives accessing the evolutionary role of emotion in intelligence:[39] The computer metaphor of the mind is missing cogency in terms of embodiment, as the software of

[31] However, searching for a broader definition opens up a field of competing concepts stemming from different research fields and traditions each of which is described by a characterizing metaphor. According to Sternberg, choosing the right metaphor for the context in question requires locating the scientific context: the field and direction of research. Intelligence is, therefore, here understood with regard to the epistemological and the computational metaphor. On the one hand, the epistemological metaphor contributes Piaget's understanding of equilibration which enables information acquisition by the cognitive processes of assimilation and accommodation. Thus, intelligence is based on "periods of development, starting with the sensorimotor period and ending with the formal-operational period". See: Sternberg, 2020, p. 10. On the other hand, the computational metaphor contributes to an understanding of intelligence according to which intelligence, especially though – cognition –, is compared to a computer whereas the hardware is similar to the brain and the software similar to the mind. The software is responsible for cognitive processes of reasoning in the mind.
See for further information: Sternberg, R., 2020. The Concept of Intelligence. In: R. Sternberg, Edt. *The Cambridge Handbook of Intelligence.* Cambridge: Cambridge University Press, pp. 3-17.
[32] Rakova, M., 2006. *Philosophy of Mind A-Z.* Edinburgh: Edinburgh University Press, p. 86.
[33] Rakova, 2006, p. 86.
[34] Shapiro, L. & Spaulding, S., 2021. *Embodied Cognition.* [Online].
[35] Shapiro & Spaulding, 2021.
[36] The computationalist position refers to the defense of the *computational theory of mind* (CTM).
See for further information: Rescorla, M., 2020. *The Computational Theory of Mind.* [Online].
[37] Cf. Rescorla, 2020.
[38] Cf. Dreyfus, 1992.
[39] Cf., for instance, Asma & Gabriel, 2019.

a computer can hardly be described as being embodied in hardware. Thus, the metaphor doesn't include any understanding or regard for emotion. Human intelligence is still bearing many questions that need to be answered to grounding a definition of the phenomenon.

The computationalist position, nevertheless, seems to become realized in the attempt to reconstruct intelligence in order to understand it: the research field of *Artificial Intelligence (AI)* is aiming "not just [to] understand[…] but also [to] build[…] intelligent entities – machines that can compute how to act effectively and safely in a wide variety of novel situations"[40]. Within AI, different aims are set: *Strong AI* refers to AI systems with the same or a higher amount of understanding as humans do and even acquire consciousness. However, the possibility of strong AI is highly questionable and hence controversial. To aim for an arguably equally distant yet mitigated goal within AI research, *Artificial General Intelligence (AGI)* refers to systems with at least the same level of cognitive understanding as humans, including the skills of conversation in natural language, reasoning, problem-solving, and environmental perception. Still, it does not presuppose consciousness. In contrast, *weak AI* refers to weakly intelligent capabilities without understanding, as we have already touched on in the decades of deep learning.[41] Subdisciplines in AI are "ranging from the general (learning, reasoning, perception, and so on) to the specific, such as playing chess, proving mathematical theorems, writing poetry, driving a car, or diagnosing diseases"[42]. The specifics show that the universality of AI possibly relates it to a broad range of human purposes.[43]

AI is therefore not to be understood as a unified technology but as a research field characterized by its interdisciplinarity: Engaged in it are computer scientists as well as cognitive scientists, psychologists, linguists, philosophers, logicians, and mathematicians.[44] Philosophy is perceived as foundational to AI as philosophers in the field of AI are trying to give answers to questions concerning the foundations and representational scope of various logics, the

[40] Russell & Norvig, 2021, p. 19.
[41] Cf. Wooldridge, 2021, pp. 38-41.
[42] Russell & Norvig, 2021, p. 19.
[43] Russell and Norvig explain in more detail that "[h]istorically, researchers have pursued several different versions of AI […]: some consider intelligence to be a property of internal thought processes and reasoning, while others focus on intelligent behavior, an external characterization. […] The methods used are necessarily different: the pursuit of human-like intelligence must be in part an empirical science related to psychology, involving observations and hypotheses about actual human behavior and thought processes; a rationalist approach, on the other hand, involves a combination of mathematics and engineering, and connects to statistics, control theory, and economics. The various groups have both disparaged and helped each other".
See: Russell & Norvig, 2021, p. 19 f.
[44] Cf. O'Regan, G., 2021. History of Artificial Intelligence. In: G. O'Regan, Edt. *A Brief History of Computing.* Cham: Springer, pp. 295.

formalizability of common sense, the foundations and forms of knowledge in general, the mind-brain dichotomy and, increasingly at the moment, the ethics of AI.[45]

1.2. History of AI: the birth of both approaches

The history of AI[46] goes back to logics, languages, and concepts developed by mathematicians and philosophers for centuries. However, usually, the time when AI was born is referred to take its form within the second half of the 20th century: Gödel's and Turing's achievements in computation[47] deeply inspired thinkers from associated fields, and the vision of machine intelligence arose together with a reflection on materialist concepts of human intelligence being analogous to a computer as a physical symbol system. In 1943, McCulloch and Pitts published an article comprising the very first artificial neural network (NN), which was stated to be capable of computing and learning. However, the two pioneers seemed to be ahead of their time. It was yet after the 'Dartmouth summer', a workshop organized by McCarthy at Dartmouth College in 1956 in which AI was defined as a research field[48] when the first so-called AI summer began: During the time following the workshop, results in computation and automation seemed promising and the field grew rapidly which reached a lot of attention. Drawn by optimism, computers were claimed to be "Electronic Super-Brains" that were "[f]aster than Einstein"[49]. The phenomenon of intelligence was hoped to be solved and replicated within one decade.[50]

---

[45] Cf. Russell & Norvig, 2021, pp. 24-26. A central role is attributed to Aristotle: "In the Nicomachean Ethics, […] Aristotle further elaborates on th[e] topic [that actions are justified by a logical connection between goals and knowledge of the action's outcome], suggesting an algorithm[.] […] Aristotle's algorithm was implemented 2300 years later by Newell and Simon in their General Problem Solver program." See: Russell & Norvig, 2021, p. 25.

[46] For the History of AI, it is followed the illustration of Russell & Norvig, 2021, pp. 35-45.

[47] These achievements can be described as: „Gödel […] showed that there exists an effective procedure to prove any true statement in the first-order logic of Frege and Russell, but that first-order logic could not capture the principle of mathematical induction needed to characterize the natural numbers. In 1931, Gödel showed that limits on deduction do exist. His incompleteness theorem showed that in any formal theory as strong as Peano arithmetic (the elementary theory of natural numbers), there are necessarily true statements that have no proof within the theory. This fundamental result can also be interpreted as showing that some functions on the integers cannot be represented by an algorithm – that is, they cannot be computed. This motivated Alan Turing […] to try to characterize exactly which functions are computable[…]. The Church-Turing thesis proposes to identify the general notion of computability with functions computed by a Turing machine […]. Turing also showed that there were some functions that no Turing machine can compute. For example, no machine can tell in general whether a given program will return an answer on a given input or run forever". See: Russell & Norvig, 2021, p. 27.

[48] Cf. McCarthy, J., Minsky, M. L., Rochester, N. & Shannon, C. E., 2006. A Proposal for the Dartmouth Summer Research Project on AI, August 31, 1955. *AI Magazine, 27*, pp. 12-14.

[49] Russell & Norvig, 2021, p. 27.

[50] Herbert Simon made this popular promise in 1957: "[T]he simplest way I can summarize is to say that there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until – in the visible future [of ten years] – the range of problems they can handle will be coextensive with the range to which the human mind has been applied". See: Simon, 1957. In: Russell & Norvig, 2021, p. 39.

One could say that the history of AI is characterized by two temporal seasons – AI summers as golden ages and AI winters as times of disillusionment – and thus, two different approaches – symbolic and sub-symbolic AI. Characteristic of the first AI summer was that intelligence was thought to be *symbolic* – therefore, machines were built to represent knowledge logically and to reason on this basis. However, Dreyfus criticized the symbolic approach early on for being limited in terms of its applicability to the real world and standing in the tradition of rationalist philosophers[51] whose legacy was fundamentally questioned by new realizations within philosophy's continuous discourse – such as phenomenology.[52]

Soon, Dreyfus' critique appeared to be right: The initial AI euphoria ended and turned into a wave of scepticism that led to the first AI winter in 1974. Though, with the breakthrough invention of symbolic expert systems, the next AI summer set in soon after, in 1980, and lasted for another seven years. Problem-solving was thought to be equal to cognition, and expert systems were developed to do just that. In sum, the symbolic approach dominated AI research and development from the 1950s to the 1980s.[53] The second AI winter in 1987 ultimately limited the hopes for the symbolic AI approach to success. Symbolic AI proved to be very limited as the real world exceeded the capacities of logical representation by far. Symbolic AI systems showed some success in well-set and defined domains but were everything other than flexible when taken out of the specific setting. Therefore, this AI winter came with intense disappointment for the symbolic faction and seemed to silence the symbolic approach broadly. Symbolic AI, therefore, gained the nickname "Good Old-Fashioned AI"[54] – *GOFAI*.

However, it eventually brought forth the rediscovery and advancement of McCulloch and Pitt's neural network and therefore gave rise to the third AI summer in 1994 – whole

[51] In more detail, he elaborates: "Since the Greeks invented logic and geometry, the idea that all reasoning might be reduced to some kind of calculation—so that all arguments could be settled once and for all—has fascinated most of the Western tradition's rigorous thinkers. Socrates was the first to give voice to this vision. The story of artificial intelligence might well begin around 450 B.C. when (according to Plato) Socrates demands of Euthyphro, a fellow Athenian who, in the name of piety, is about to turn in his own father for murder[.] […] Socrates is asking Euthyphro for what modern computer theorists would call an 'effective procedure,' 'a set of rules which tells us, from moment to moment, precisely how to behave.' Plato generalized this demand for moral certainty into an epistemological demand. According to Plato, all knowledge must be stateable in explicit definitions which anyone could apply […]. The belief that such a total formalization of knowledge must be possible soon came to dominate Western thought. It already expressed a basic moral and intellectual demand, and the success of physical science seemed to imply to sixteenth-century philosophers, as it still seems to suggest to thinkers such as Minsky, that the demand could be satisfied. […] Leibniz, the inventor of the binary system, dedicated himself to working out the necessary unambiguous formal language. […] Like a modern computer theorist announcing a program about to be written, Leibniz claims: 'I have invented an elegant artifice by virtue of which certain relations may be represented and fixed numerically and which may thus then be further determined in numerical calculation'" See: Dreyfus, 1992, pp. 67-69.

[52] Cf. Dreyfus, H., 1974. Artificial Intelligence. *The Annals of the American Academy of Political and Social Science, 412*, pp. 21-33.

[53] Cf. Boden, 2014, p. 89.

[54] Haugeland, J., 1989. *Artificial Intelligence. The very Idea.* Cambridge MA: MIT Press.

decades to come in which the *sub-symbolic* approach showed promising successes. In fact, the approach seems to have dominated AI research ever since: From that moment on, the novel focus was primarily set on neural networks as they enabled probabilistic modeling and, with it, a method to approach the ambiguous phenomena of the real world. The symbolic approach to AI took a back seat in funding because sub-symbolic AI techniques were steadily progressing with the advent of big data. At present, we are still experiencing this ongoing AI summer, which yielded new hopes as deep learning advanced in the last decade.[55]

To better understand the tension of the historical symbolic-sub-symbolic distinction, the central techniques and methodological characteristics of both approaches are briefly illustrated in the following. In order to analyze commonalities and fundamental differences between the approaches in the further course of the thesis and to discuss strengths, weaknesses, and promising visions later on, this preliminary understanding is necessary. Therefore, let us take a closer look at two putative opposites: the symbolic approach to artificial reasoning and the sub-symbolic approach to artificial learning.

1.3. Symbolic AI: reasoning

Symbolic AI refers to the initially dominating approach to AI, which is based on "programmed instructions operating on formal symbolic representations"[56]. Symbols of formal programming languages are, in this matter, representative of particular semantics and logically structured according to defined rules: Symbolic "computation involves the construction and transformation of symbolic data structures"[57]. Hence, symbols that are formal and relational by nature are structurally preprocessed, and resulting data structures are subsequently manipulated. Defined symbolic knowledge is stored in a knowledge base that enables computation through logical inferences. Therefore, it is often described as knowledge engineering. The symbolic approach to AI can be framed by the notion of *reasoning* as it attempts to reason on the basis of formal logical axioms, defined rules, mainly deductive logical inferences, and a knowledge base that is ideally representing the entire corpus of human common sense. The capability of representation is central to the approach as common sense is here modeled representationally. Various levels of abstraction of the representations are possible.[58]

[55] Cf. Russell & Norvig, 2021, pp. 35-45.
[56] Boden, 2014, p. 89.
[57] Boden, 2014, p. 90.
[58] Cf. Russell & Norvig, 2021, pp. 37-41.

The central techniques of symbolic AI are planning and heuristic search. A symbolic AI quest in terms of heuristic search is representationally mapped as a feasible region: "a set of possibilities (defined by a finite set of generative rules), within which the solution lies – and within which it must be found"[59]. This can be exemplified by the various valid possibilities to make a move in a game of chess which are identified by search. Because often the number of possibilities is too high to quickly consider all possibilities, heuristics can serve as a way to find an appropriate solution. A symbolic AI task regarding the technique of planning is usually hierarchically structured as a prioritization of different goals. The program then attempts to minimize the conditions varying between the actual state of the task and the ideal state of the task that was defined before as the task's highest-rated goal. To achieve the ideal state, the program has the means of defined heuristics to initially draft the appropriate hierarchy (goal-setting), access possible varying conditions, and decide how to operate in order to realize differently rated goals.[60]

Expert systems are a typical application of symbolic AI and are often deployed in health care for the purpose of medical diagnosis. The expert system consists of an (expert) knowledge base and an inference engine, logically deriving decisions or actions from defined facts and 'if-then rules'. As the name already indicates, the systems enable reasoning with expert knowledge of a certain domain, for instance, medicine, and rely on conditional expressions.[61] In a simplified example, for which no doctor would need a machine but which serves to illustrate the method, the system could handle a medical case in the following way: *If* the patient is a child *and* fever is reported *and* a rash of red dots is visible *then* the diagnosis is chickenpox *and* the recommendation is the clipping of the patient's fingernails *and* the prescription of anti-fever medication.

The two most important programming languages in classical symbolic AI, respectively GOFAI, were the high-level[62] languages Lisp and Prolog.[63,64] Lisp is a functional programming language based on the lambda-calculus and was adopted for approximating machine representation and reasoning. It is still used for specific applications today.[65] Prolog is a logic

[59] Boden, 2014, p. 90.
[60] Boden, 2014, p. 90 f.
[61] Boden, 2014, p. 91 f.
[62] 'High-level' refers to a high amount of possibilities to allow abstractions from the underlying machine code in a programming language. This means, that the syntax of the source code is approximated to either mathematics or natural language – in this way, it is easier for the programmer to understand and use the language in question. Cf. Butterfield, A. & Szymanski, J., 2018. *A Dictionary of Electronics and Electrical Engineering,* Oxford: Oxford University Press.
[63] Cf. Wooldridge, 2021, pp. 49; 112-114.
[64] Cf. Ertel, W., 2017. *Introduction to Artificial Intelligence.* Berlin: Springer, p. 81.
[65] Cf. Russell & Norvig, 2021, p. 37.

programming language used for symbol manipulation purposes. For instance, it was used for the development of expert systems. Likewise, it is still taught and used today.[66]

In terms of the symbolic approach to AI, intelligence can be understood as the ability to find the best solution to reason with the before-defined knowledge and 'act' or decide on the basis of this flexible motion within the knowledge base. This comprehension was based on the physical symbol system hypothesis, stating that human or machine reasoning is achieved by the manipulation of symbolic data structures.[67] Therefore, from a philosophical point of view, the symbolic approach is often compared to the human mind in the mind-brain dichotomy as the aim to create intelligent systems is here based on the representations of formal contents, which are considered to be underlying human thoughts.[68] The mental processes simulated by symbolic approaches are therefore often compared to conscious top-down processes: considerations, reflections, abstraction, problem-solving, and regulative thoughts.[69]

1.4. Sub-symbolic AI: learning

Sub-symbolic AI is often understood as the opposite approach to GOFAI: In analogy to the human brain, instead of the mind, (artificial) neural networks are constructed which are able to learn from data. The notion of *learning* is therefore central to the approach:

> "An agent is learning if it improves its performance after making observations about the world. […] When the agent is a computer, we call it machine learning [ML]: a computer observes some data, builds a model based on the data, and uses the model as both a hypothesis about the world and a piece of software that can solve problems"[70].

The approach is also referred to as PDP connectionism (whereas PDP is short for parallel distributed processing) because it relies on the networks' interconnection nodes which process information through parallel and distributed computations. Those computations are achieved by the activation of the closely linked nodes through gradual changes in the nodes' weights. This enables the sub-symbolic system to adapt to new situations, which are mirrored in the data on which the network is trained (whereas the training data can be labeled or unlabeled). In this way, the neural network acquires a learning behavior in which intelligent features such as pattern recognition are shown. Although McCulloch and Pitts drafted the first neural network

[66] Cf. Russell & Norvig, 2021, p. 312 f.
[67] Cf. Russell & Norvig, 2021, p. 37.
[68] Cf. Wooldridge, 2021, p. 42.
[69] Cf. Boden, 2014, p. 92 f.
[70] Russell & Norvig, 2021, p. 669.

in 1943,[71] the approach was not popular in AI research and development until the 1980s. It continuously gained importance with the emergence of big data and according computing capacities. Central methods of the sub-symbolic approach are deep learning (DL) which is a specific advancement of machine learning, thus, machine perception, natural language processing (NLP), and context-specific memory.[72,73] Neural networks can find "solutions to a wide range of classification and decision problems"[74], whereas the solutions can almost be said to be creative as the network induces solutions straight from the data.

Deep learning comprises multiple layers that enable a high processing intensity – an extensive depth in the data processing. The successes achieved by various deep learning applications in spite of a short history of the technique[75] strongly impressed the public and the scientific community. Deep learning has enabled various applications such as predicting protein structures or autonomous driving and thereby disproved some of AI scepticism. A typical and promising deep learning approach is convolutional neural networks. New or exceptional properties of it are, aside from its processing depth, also a variety of different nodes and activation functions. Thus, the layers and connections are organized in a local manner – which means that only the output layer is fully-connected in terms of its nodes, whereas the other layers' nodes are solely connected with those being either in their closer environment or receiving inputs at the same time. In this way, neural networks can run more efficiently.[76]

Various programming languages are used in the field of sub-symbolic AI. Very prominent, however, are the high-level languages Python and Java. Python is syntactically intuitive and flexible in its access to libraries, whereas Java is a general-purpose language specially used for online environments such as a cloud.[77,78]

Intelligence can be understood here as the ability to learn from given training data and to find the best solution to transfer it to new situations on which basis a decision or action is made. The sub-symbolic approach is often compared to the human brain in the mind-brain dichotomy as the aim to create intelligent systems is here based on architectural features of the human brain such as neurons and synapses.[79] The brain processes simulated by the sub-

[71] McCulloch, W. & Pitts, W., 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics, 5*, pp. 115-133.
[72] Cf. Sun, 2014, pp. 108-110.
[73] Luger, G., 2021. *Knowing our World. An Artificial Intelligence Perspective.* Cham: Springer.
[74] Buckner, 2019, p. 2.
[75] The breakthrough of deep learning was achieved by Geoffrey Hinton in 2006, with his publication of "deep belief nets". See for further information: Hinton, G., Osindero, S. & Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. *Neural Computation, 18, 7,* p. 1527–1554.
[76] Cf. Buckner, 2019, pp. 2-8.
[77] Cf. Ertel, 2017, p. 175.
[78] Cf. Ciesla, R., 2021. *Programming Basics.* Helsinki: Apress, p. 13 f.
[79] Cf. Wooldridge, 2021, p. 44.

symbolic approach can therefore be compared to mainly subconscious bottom-up processes: perceptions, data collections, pattern recognition, observations, and learning in general.[80]

In order to maintain an overview of what is about to follow, Figure 1 shows a simplified depiction of the opposing approaches. The main concepts of symbolic AI are heuristic search and planning, while expert systems rely on both techniques and thus introduce new methods. The approaches in sub-symbolic AI build on top of each other and can therefore be represented as a simple hierarchy: Deep learning is currently the most promising technique in the realm of neural networks. Neural networks are, in turn, based on the concept of machine learning, which is considered a key approach in sub-symbolic AI.

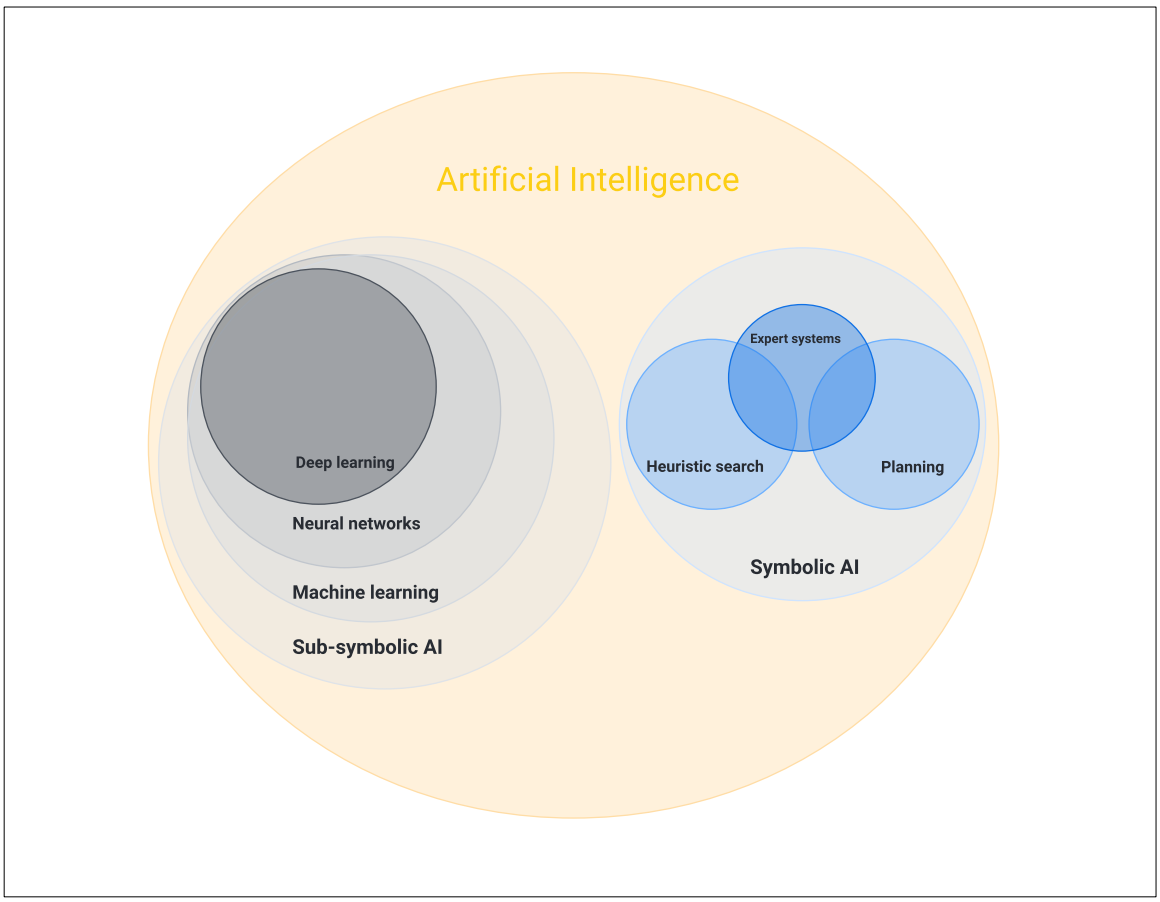

Figure 1: Simplified AI overview as conceptualized in the symbolic-sub-symbolic distinction (own illustration)[81]

To base the thesis' position, which advocates symbolic AI from an ethical perspective, on solid ground, 'deeper' philosophical challenges have to be reflected upon first. Dreyfus, a strong

[80] Cf. Boden, 2014, p. 92.
[81] The overview is drawn on the basis of a synthesis of Lawson et al. and Boden. Cf. Lawson, C. et al., 2021. Machine learning for metabolic engineering: A review. *Metabolic Engineering, 63*, p. 36, and Boden, 2014, pp. 90-92.

sceptic of symbolic AI introducing a phenomenological position into the philosophy of technology, illustrated not only the weaknesses of GOFAI but also its underlying *computationalist position* in 'What Computers Still Can't Do'. According to him, the computationalist position is comprising assumptions that are misleading from a philosophical perspective. As he considered symbolic AI, however, to be a realization of these assumptions in practice, he marked it as a hopeless case early on. In the following, Dreyfus' arguments against GOFAI are taken closer into focus as he revealed fundamental philosophical questions through the identification of these implicit assumptions.

2. "What Computers Still Can't Do": Dreyfus' phenomenological perspective on symbolic AI

Dreyfus stands in the tradition of a phenomenological movement within the philosophy of technology that is, based on Husserl, Heidegger, and Merleau-Ponty, critically opposing the Cartesian dualism that divides the material from the mental world. Kogge delineates the confrontation of two opposing, however, yet related philosophical traditions in the 20th century.[82,83] Both traditions originated from an intensive examination of mathematical foundations, which was to be conducted philosophically. The Cartesian dualism opens up a dichotomy of 'two ontological spheres'[84], whereas (material) objects and their relations in the real world are symbolically represented in (mental) knowledge. On the one hand, this dualistic view was passed on by rationalist schools of thought, including the logicians, who considered the symbol to be a universal form of explicitly representing and conveying knowledge about the world. World knowledge was assumed to be accessible by means of mathematics and thus postulated in logical rationale as the basis of existence and science.

Phenomenologists, on the other hand, attempted to overcome the Cartesian division: Phenomenologists hold that in the act of apprehension, the subject is not internally grasping and mathematically comprehending objects of an external sphere. But rather is consciousness and hence, the reality we can access by it, constituted as a unity. Within, not only the subject but also all objects of consciousness are situated, to which the subject intentionally relates. This largely dissolves Cartesian dualism and, beyond, demands new notions, conceptions, and terminologies for the description and philosophical investigation of the knowing subject in the

[82] Kogge, W., 2017. *Experimentelle Begriffsforschung. Philosophische Interventionen am Beispiel von Code, Information und Skript in der Molekularbiologie.* Weilerswist: Velbrück Wissenschaft, pp. 28-33.
[83] The term was here translated and is originally to be found in Kogge, W., 2016. Verkörperung – Embodiment – Körperwissen. Eine historisch-systematische Kartierung. *Paragrana, 25, 1*, pp. 33-46.
[84] Kogge, 2016, p. 37.

world. This understanding of the phenomenal constitution is thus expanded by elaborating on the role of bodily or incorporated knowledge (that is, "Körperwissen"[85]), which is implicit by nature. The mind is not isolated and thus does not centrally and abstractly perform reasoning by mentally representing logical relations of the world. Rather, all abstraction is grounded in the intuitive knowledge acquired by the body itself – as we as humans are continuously bodily perceiving and experiencing the world we are in. Abstract knowledge must therefore be imagined as constituted by the body.

While the rationalist and computationalist understanding of intelligence encompass explicit representations as the foundation of rational thinking – hence, reasoning –, the phenomenological understanding of intelligence zeroes in on implicit knowledge that is deeply anchored in the experience of our bodily engagement with the world. Because implicit knowledge is acquired pre-conceptually, it precedes symbolic representations as well as reasoning with these representations. Hence, the phenomenologist position reintegrates the constitution of embodiment as the mediating form of knowledge and the perceiving subject that acquires knowledge through its social and cultural practices.

As the focus of phenomenological examination is on the nature of perception originating from the body and its engagement with the world, the comprehension of embodiment bridges the Cartesian gap between body and mind, respectively also between brain and mind. The dissolution of abstract thoughts from their embodied origin is the object of critique: Phenomenologists such as Dreyfus emphasize that the object of analysis is always perceived as deeply bound up with its environment, that is, embedded in its situated context, including the perceiving subject. To simply remove the contextual and subjective dimension in the activity of analysis is hence inadequate to scientifically access or describe the perceived world. The phenomenological approach to body, mind, and technology influenced not only the research field of AI but also conceptions in psychology, cognitive science, and neuroscience.[86]

The phenomenological critique gives rise to the demand for a new approach to body and mind – a methodologically holistic approach to the phenomenon of embodied intelligence that stands in contrast to the Cartesian dualistic comprehension and thus in contrast to the computationalist position. In this manner, Dreyfus strongly criticizes the computationalist position with regard to AI. It is focused on his phenomenological critique of GOFAI and the underlying computationalist position in the following: What are his phenomenological arguments for a dismissal of symbolic AI and, in turn, the tendential advocacy of sub-symbolic

[85] Kogge, 2016, p. 42.
[86] Cf. Kogge, 2016, pp. 33-46.

AI? With regard to Heidegger[87,88] and Merleau-Ponty,[89] on whom Dreyfus bases his position, his arguments are reconstructed and focused on the following key aspects: AI in general, AI and context, as well as AI and emotion.

2.1. AI in general

Why the research field of AI was, according to Dreyfus, generally about to lead to disappointing results, substantiates by misleading assumptions underlying the computationalist position[90] and, with it, AI in general. From a phenomenological perspective, Dreyfus criticizes GOFAI as an applied rationalist conception stemming from "the Cartesian idea that all understanding consists in forming and using appropriate symbolic representations"[91]. The conception's further tradition relies on Kantian "rules for relating such [representations]"[92] and Frege's formalization and manipulation of the rules in question, which were received by the computationalist position adopting a rule-based "computer model of the mind"[93]. Not only does Dreyfus hold the notion of understanding in Descartes' rationalist tradition but also its further comprehension in the computationalist position arising with early AI research and being continued in AI development, to be misleading as it is "staking everything on man's ability to formalize his behavior[.] to bypass brain and body, and arrive, all the more surely, at the essence of rationality".[94]

---

[87] Heidegger, 1993.
[88] Dreyfus, H., 1991. *Being-in-the-world: a commentary on Heidegger's Being and time, Division I.* Cambridge MA: MIT Press.
[89] Merleau-Ponty, 2005.
[90] Most often, Dreyfus uses the notion of "Cognitive Simulation (CS)" for the position relying on the computer metaphor and thus underlying GOFAI. For instance, see Dreyfus, 1992, p. 85. The cognitive psychologist Newell and the social scientist Simon laid the foundations of the position and its following tradition in research. Their physical symbol system hypothesis was therefore often subject to Dreyfus' critique. However, 'computationalism', the 'computationalist position' respectively, is here chosen to translate the position denoted with CS by Dreyfus, to the current terminology of research in cognitive science – because CS is hardly referred to in current research. Computationalism, however, is nowadays usually used to refer to the tradition in cognitive science that is based on the computer metaphor. It includes not only Newell and Simon's early hypothesis but also its resulting tradition of research. Thus, GOFAI is described as a variant of computationalism. It is therefore assumed to be a suitable translation for the purpose of this thesis. Cf. for the latter: Miłkowski, M., 2018. From Computer Metaphor to Computational Modeling: The Evolution of Computationalism. *Minds and Machines, 28,* p. 517.
[91] Dreyfus, 1992, p. xi.
[92] Dreyfus, 1992, p. xi.
[93] Dreyfus, 1992, p. 156.
[94] Dreyfus, 1992, p. 77f.

2.1.1. GOFAI as a realization of computationalism

The computationalist position, which holds that the human mind is not only analogous to but actually *is* itself a computer, a so-called "symbol-manipulating device"[95], is criticized as a false assumption.[96] It, in turn, comprises four further misleading assumptions (the biological, psychological, epistemological, and ontological assumptions), which are examined in the further course of chapter two. Therefore, Dreyfus' critique of GOFAI must not be understood on a primarily technical level but rather on an epistemological level – as he views AI as the realization of philosophical schools of thought.

The critique as a whole can thus be assigned to the tradition of criticizing symbolisms as a supposed universal method of seeking the truth: Kogge describes, for example, that symbolic representation does not correspond to truth per se but rather serves as a mere medium of representational orders capturing phenomena. Symbolic representations are, therefore, to be understood more as instruments of systematic acquisition of knowledge than as an exact compression of truth into formal structures. At the same time, symbolic representations can be the result of the conduct of these instrumental practices.[97] This, in turn, sums up the core of Dreyfus' critique: While a symbolic representation functions as a medium and information can certainly be derived and understood from it – this does *not* mean that, conversely, the automated manipulation of such symbolic representations can functionally generate truth or even intelligence.

Computationalists deem human intelligence to be the activity of reasoning consisting of computational information processing and consequently building objective representations of the world, which are then propositionally related.[98] However, the aim of formalizing the body of represented common sense knowledge into distinct facts, relating and automating those appropriately to create machine reasoning on the level of human intelligence is, according to Dreyfus, nothing more than a "rationalist vision"[99] of computationalists whose realization in AI was about to founder. Dreyfus emphasizes not only the bodily roots but also the permanent *embodiment* of human intelligence as being prior to any mental representation: Human intelligence would not mysteriously emerge from bodily roots and thereafter exist as an abstract and rule-based entity of the brain. Therefore, the mind could not be viewed as a phenomenon detached from the body but rather as a phenomenon constituted by the body – continually being

[95] Dreyfus, 1992, p. 155.
[96] Dreyfus states: "This assumption, that human and mechanical information processing ultimately involve the same elementary processes, is sometimes made naïvely explicit". See: Dreyfus, 1992, p. 155.
[97] Cf. Kogge, 2022a, p. 67 f.
[98] Cf. Dreyfus, 1992, p. xvii.
[99] Dreyfus, 1992, p. xvii.

embodied. Therefore, the phenomenon of intelligence could only be accessed and explored by preliminary considering its embodiment: "The human world with its recognizable objects is organized by human beings using their embodied capacities to satisfy their embodied needs. There is no reason to suppose that a world organized in terms of the body should be accessible by other means"[100]. This access for exploration, however, is determined by the form of what is to be explored: Because the human world is experienced through bodily phenomena, it is incumbent to also scientifically access it on the basis of phenomenology.[101]

2.1.2. The psychological assumption

The argument evolves in one of the four further assumptions yielded by the computationalist position: The '*psychological assumption*' identified by Dreyfus[102] holds that "the mind can be viewed as a device operating on bits of information according to formal rules […] [encompassing] a third-person process in which the involvement of the 'processor' plays no essential role"[103]. Herein, it is assumed that the human mind is processing precepted information distinctly as abstract sets of facts in the same way as a computer processes informational bits, and thus, that reasoning can be thought to be objective rather than subjective. Dreyfus first questions whether a computer and the human mind share the same formal information processing mechanisms: Because psychology is not equal to biology, the mind can not be regarded as equal to the brain and, thus, can not be described through the reduction to physical and chemical entities. Hence, although human cognition is processing physical information within the laws of nature, the phenomenal experience of the mind can not simply be reduced to a physical information processing entity.

Rather, Dreyfus questions the existence of an identifiable information processing entity that operates on purely syntactical rules and hence formalizes perceived items according to those rules in order to be further computed. Information in a computational way differs fundamentally from information as a matter of human perception – as the latter is always semantically anchored in meaningful coherences within the world of the perceiving subject: "[T]here are no facts with built-in significance and no fixed human forms of life which one could ever hope to program"[104]. Only because human perception is meaningful to the subject in question, relevance and significance could be attributed to specific items perceived. Only on the basis of the distinguishment between relevance and irrelevance as well as significance and

[100] Dreyfus, 1974, p. 32 f.
[101] Cf. Dreyfus, 1992, p. 181.
[102] Cf. Dreyfus, 1992, pp. 163-188.
[103] Dreyfus, 1992, p. 156.
[104] Dreyfus, 1992, p. 290.

insignificance enabled by subjective meaning is the perceiving subject able to focus on segments within holistically – not supposedly analytically – experienced situations.

Nevertheless, the computationalist position unfoundedly assumes that information as a matter of human perception is structured and processed on the basis of heuristic rules similar to a program. Dreyfus questions the existence of thought-determining rules as the mind might not rely on rules at all but rather "may well arrive at its thoughts and perceptions by responding to 'fields,' 'force,' 'configurations,' and so on, as, in fact, we seem to do so insofar as our thinking is open to phenomenological description"[105]. Dreyfus concludes that even though "no independent, empirical evidence exists for the psychological assumption"[106], for computationalists and AI researchers "the psychological assumption seems not to be an empirical hypothesis that can be supported or disconfirmed, but some sort of philosophical axiom whose truth is assured a priori"[107].

However, as the psychological assumption is taken a priori as the foundation of an explanation on the psychological level, no phenomenological or physiological explanation could be adduced to mitigate its axiomatic nature.[108] Thus, the explanations based on the psychological assumption could not bridge the gap between (cognitive) psychology and phenomenology as they could not psychologically assess phenomenal experiences: Seeing light and listening to music differ inherently from passively receiving "sensory inputs"[109]. Dreyfus holds phenomenology necessary to meet the requirements of a plausible explanation because it is the meaning in the world that humans as perceiving subjects attribute to the perceived situations to make sense of them: Meaning enables the subject to explanatory access phenomenal experiences and to render them intelligently. However, meanings of the world could neither be accessed psychologically nor represented computationally – they could be seized only phenomenologically, in taking into account the bodily roots and bodily grounding of human intelligence, which were not correlating with heuristic rules representable by a computer program.

We could visualize this, for example, by imagining the view of a foggy Berlin in early autumn. According to Dreyfus, the situation can only be assessed phenomenologically: A psychological explanation suffices at most to explain an individual reaction of the observer to

[105] Dreyfus, 1992, p. 166.
[106] Dreyfus, 1992, p. 174.
[107] Dreyfus, 1992, p. 174.
[108] Dreyfus states: "The confusion can best be brought to light by bearing firmly in mind the neurophysiological and phenomenological levels of description and then trying to locate the psychological level somewhere between these two. [There does not seem to be] […] place for [a] information-processing level" (Dreyfus, 1992, p. 179 f).
[109] Dreyfus, 1992, p. 183.

the situation – perhaps a rejection of the fog when absolute clarity of vision is preferred due to a neuroticism trait. A physical explanation, in turn, is at most sufficient to describe the occurrence of the weather and the observer's physiological reaction to the weather, such as processes on the skin, which absorbs the fog and moisture, or the cold feet. Solely from a phenomenological perspective, however, seems it possible to describe why the holistically experienced situation – for example, the sun shrouded in a veil of mist, the crunching of fallen leaves under chilly feet, and the faint melancholy of another past summer – can trigger associations, thoughts, processing of experiences, desires, hopes, and ideas.

In contrast, the computationalist position, its inherent psychological assumption, thus in its tradition, the GOFAI systems, were not only disregarding the embodiment of intelligence but also critical dimensions coming with it, which can be best understood as situatedness as illustrated by Heidegger and intentionality as illustrated by Merleau-Ponty. Why AI in general, GOFAI at Dreyfus' time, is generally based on misleading philosophical positions is now further examined with regard to context and emotion: Dreyfus emphasizes that because it is embodied, human intelligence has the core capacities to conceive and experience situated context, as well as to emotionally engage with the world, which governs cognition in an intentional way. Because of their bodily roots and their phenomenological shape, both of these dimensions of embodiment are described as "preconceptual"[110] and therefore non-representable. As GOFAI, however, relies strongly on representational methods, Dreyfus deems embodiment and hence intelligence to be beyond the reach of understanding or reconstruction by GOFAI systems.

2.2. AI and context

2.2.1. Heidegger's being-in-the-world

As a dimension of embodiment, human context-sensitivity is examined below with regard to the situatedness of intelligence. Dreyfus epistemologically argues against the possibility of representing situational context with symbolic AI on the basis of Heidegger's comprehension of *being-in-the-world*. Heidegger drafts being-in-the-world phenomenologically as the phenomenal and entirely holistic constitution of being in existence: "Diese Seinsbestimmungen

[110] Dreyfus, 1992, p. xii.

des Daseins müssen […] auf dem Grunde der Seinsverfassung […] verstanden werden, die wir das In-der-Welt-sein nennen"[111].[112]

In arguing against GOFAI, Dreyfus highlights that the holistic experience of being-in-the-world dissolves the supposed distinction between subjectivity and objectivity and rather brings forth human intelligence involved and situated in the constitution of the world. Thus, the involvement in a holistic experience is basic to human intentionality, which can be understood as the meaningful but non-representational relatedness of conscious beings in the world. As such, it may indeed refer to certain objects. However, they cannot be understood as outside of the holistic experience but rather as situated within: "When we are at home in the world, the meaningful objects embedded in their context of references among which we live are not a model of the world stored in our mind or brain; they are the world itself"[113].

In that sense, cognitivist theories and the computationalist position wrongly claim symbols to be able to formally contain and therefore represent objects of the real world: As soon as an abstraction of meaning is semantically embedded in formal syntax, for example, the syntax of a computer, the holistic unity is distorted in the sense that the *context* of the represented object is entirely missed. Thus, the object itself is distorted because its meaning is continually *shaped by* situational contexts.

Transferring these insights to the methodology of GOFAI means that GOFAI's foundational method, which is representation, can in itself be viewed as contradictory as it is always applied to from human perception abstracted objects syntactically structured as logical relations (but not real-world relations) and thus, semantically denoted with meanings that are distorted by its very isolation from their meaning-ness: their context. Heidegger's phenomenological understanding of being-in-the-world counters the idea of formalizing and representing phenomenal experience and, with it, the phenomenon of human intelligence.[114]

---

[111] Heidegger, 1993, p. 53.

[112] In more detail, Heidegger relates: "Dasein ist Seiendes, das sich seinem Sein verstehend zu diesem Sein verhält. Damit ist der formale Begriff von Existenz angezeigt. Dasein existiert. Dasein ist ferner Seiendes, das je ich selbst bin. Zum existierenden Dasein gehört die Jemeinigkeit als Bedingung der Möglichkeit von Eigentlichkeit und Uneigentlichkeit. Dasein existiert je in einem dieser Modi, bzw. in der modalen Indifferenz ihrer. Diese Seinsbestimmungen des Daseins müssen nun aber a priori auf dem Grunde der Seinsverfassung gesehen und verstanden werden, die wir das In-der-Welt-sein nennen. Der rechte Ansatz der Analytik des Daseins besteht in der Auslegung dieser Verfassung. Der zusammengesetzte Ausdruck ‚In-der-Welt-sein' zeigt schon in seiner Prägung an, daß mit ihm ein einheitliches Phänomen gemeint ist. Dieser primäre Befund muß im Ganzen gesehen werden. Die Unauflösbarkeit in zusammenstückbare Bestände schließt nicht eine Mehrfältigkeit konstitutiver Strukturmomente dieser Verfassung aus". See: Heidegger, 1993, p. 53.

[113] Dreyfus, 1992, p. 265 f.

[114] Cf. Dreyfus, 1991, p. 5.

2.2.2. The epistemological assumption

Dreyfus further evolves this position as he identifies a false '*epistemological assumption*'[115] underlying GOFAI research and development: It states that all forms of knowledge are formalizable and logically representable on which basis a computer can further reason with the knowledge. Therefore, even if the plausible explicability of human behavior by representations and formal rules as an axiom taken up in the psychological assumption is unjustified, the epistemological assumption considers that human intelligence is still formalizable and reproducible with AI.

The distinction from the psychological assumption is to be stressed: While, according to Dreyfus, mainly computationalists are subject to the psychological assumption considering intelligence to be *emerging from* mental representations and formal rules – mainly GOFAI researchers are subject to the epistemological assumption in considering human intelligence to be formalizable by representations and rules which are then serving as some general mechanisms of intelligence ready to be replicated by the computer. Even if Dreyfus rates the epistemological less controvertible than the psychological assumption, it was still leading to wrong conclusions about human and artificial intelligence. Because although the epistemological assumption does not entail the requirement of AI to base on the exact same rules that supposedly underly human intelligence, it still disregards the important role of being-in-the-world for human intelligence as it does not consider the context of abstracted objects.[116]

However, according to Dreyfus, there is no empirical reason to assume that the human mind and behavior are of a symbolic nature. As a capability of human intelligence, natural language might, for instance, therefore not be reducible to symbolic representation and manipulation which is most obvious when trying to formalize the linguist category of pragmatics in a unified theory. To find a "formal theory of pragmatics, one would [first] have to have a theory of all human knowledge, but this may well be impossible"[117]. Second, exceptions to linguistic rules are common in natural language, which refutes the possibility of a consistent formal theory of natural language competence as well as the possibility of GOFAI systems being able to adopt it for natural language performance. Exceptions to linguistic rules might be recognized by a human audience and still understood: In being aware of common sense knowledge as well as the specific situated context of the speaker and the audience, the audience is able to attribute meaning to the utterance in interpreting it accordingly.

[115] Cf. Dreyfus, 1992, pp. 189-205.
[116] Cf. Dreyfus, 1992, p. 190 f.
[117] Dreyfus, 1992, p. 198.

However, the exception from a rule leaves a GOFAI system fragile as it neither owns sufficient common sense capabilities nor context-sensitivity that would allow for creative interpretations of the linguistic exception. To cope with the exception, it could make up new rules for the exceptions in question. However, this would lead to an infinite regress of rules, as arbitrary exceptions always seem possible to be found. Therefore, even if theories of natural language competence might be formalizable to a certain degree, Dreyfus deems it unlikely that GOFAI-based machine intelligence will be capable of actually performing communication in natural language. Dreyfus refers to Wittgenstein in holding the GOFAI approach inadequate for natural language processing as a field of intelligence because being bound to linguistic rules would counter natural language: "In general we don't use language according to strict rules – it hasn't been taught us by means of strict rules either"[118].

Thus, individual situational contexts are, according to Dreyfus, neither formalizable to rules nor are they reproducible as they are characterized by the individuality of events. That means that universal rules for context-sensitivity can not be found and applied to GOFAI systems. In order to access individual situated context, GOFAI systems would require some consolidation of the singularity of situated context and universal laws:

> "[T]he machine must use its formalism to cope with real-life situations as they occur. It must deal with phenomena which belong to the situational world of human beings as if these phenomena belonged to the objective formal universe of science"[119].

One could say, for instance, that the situated context of a conference might be formalizable as it does bear a certain regularity of specific features of conferences in general – such as their scientific background, presentational speeches, more than one speaker being present to contribute, and a broad international audience. However, many features are individually constituted by the singularity that comes with the nature of events determined by their specific location and, thus, its time: It is in turn bound to a particular background, such as the exact reason for the conference, which might be the foundation of a new research field, in a certain time, which might be immediately after a scientific breakthrough. Thus, even if the type of the event is regularly, such as annually hosted, and seems therefore not to be characterized by its singularity – it still is singular in practice, for example, in terms of the people attending on the specific date, their languages, dialects, interests, speeches, and conversations, as well as the behavioral atmosphere surrounding the event as subjectively perceived by its participants.

[118] Wittgenstein, L., 1960. *The Blue and Brown Books.* Oxford: Basil Blackwell, p. 25. As quoted by Dreyfus: Dreyfus, 1992, p. 203.
[119] Dreyfus, 1992, p. 201.

What constitutes a situated context and its access goes, therefore, beyond the explanatory level of similarities and regularities logically deduced from different contexts:

> "If there could be an autonomous theory of performance, it would have to be an entirely new kind of theory, a theory for a local context which described this context entirely in universal yet nonphysical terms"[120].

It is, according to Dreyfus, the enduring subjective involvement in the individual situated context in particular, the way of being-in-the-world, that brings forth intelligent human behavior. This involvement in practice contradicts a rule-governed and therefore formalizable behavior which can be taught to and adopted by AI systems theory-wise, that means without regard to the information being semantically constituted by its situated context: It, therefore, refutes the epistemological assumption.[121]

2.2.3. The ontological assumption

Dreyfus further illustrates GOFAI's weaknesses in terms of situated context through the identification of the '*ontological assumption*':[122] It holds that intelligence is based on the perceiving and processing of sequenced distinct and definite elements. As data, these elements are taken to be the subject to computation in GOFAI systems. However, Dreyfus claims this assumption to be misleading because when it comes to human intelligence, reality is perceived holistically in a phenomenological way. In contrast, it does not have to but can be analyzed through distinct facts. A representation of these facts as disjointed products of analysis would neither equal reality nor intelligence because the world itself does not actually consist of self-contained and formalizable elements. Thus, as shown in the psychological and epistemological assumption, human intelligence is not enabled through some rule-based mechanism representing these formal elements.[123] Rules can be considered analytical products of the perceived world, which is consequently abstracted and sectioned to single items, constructively related, and represented in a resulting synthesis. However, these rules can also be considered as distorted from the veritable consistency of the world, which may well differ from the cognitive edifice of ideas logically relating and representing the situations perceived.

The ontological problem is intensified when supposing a perceived abstracted element as actually semantically representing its situated role and meaning in the world. Assuming that a distinct element serving as an informational bit contains all relevant information when it is de

[120] Dreyfus, 1992, p. 202.
[121] Cf. Dreyfus, 1992, pp. 198-205.
[122] Cf. Dreyfus, 1992, pp. 206-224.
[123] Cf. Dreyfus, 1992, p. 206.

facto taken out of its context by the very abstraction to its logical form and its addition into a data structure for the computational purpose is "creat[ing] a problem by determining the way all questions concerning giving information to computers must be raised"[124]: It is, in particular, the context which makes it possible to attribute the relevant meaning to the abstracted element in question. However, on the one hand, the process of analysis does not allow for a holistic outcome. And on the other hand, only data structured as informational bits can be processed by GOFAI systems. Hence, the context can not be transduced to the computer as no capacity for holistic and, less so, phenomenal perception is taught to the systems in the attempt to approximate actual human intelligence.[125]

By trying to solve the problem of context representation, AI researchers tried to objectify the context of the abstracted elements as well, leading to a problem of 'micro-worlds'[126]. Referring to Heidegger, Dreyfus describes a further problematic aspect arising from the "ceteris paribus condition"[127] of rules: Rules, by their logical nature, contain the implicit demand of 'all other conditions being equal' in order to be valid. To explain an object in question by rules – for instance, a specific form of human behavior – its context might be represented, however, under the explicit prerequisite of 'all other conditions being equal'. Even if the contextual conditions are not clear or yet to be explored, they have to be taken as fixed to be represented. This means that the behavioral context is not represented as modifiable and therefore not represented as open to mirror contextual flexibility. Although situated contexts are essentially characterized by continuous situational changes as well as the adaptability of human behavior to its environment or situated context on the basis of implicit knowledge, relevance has to be given explicitly and statically to specific features to be objectified in advance. Therefore, the rule-based GOFAI approach only allows the representation of abstracted elements symbolized by single objects whose context underlies rules of axiomatic grounds. However, this contradicts the openness and situatedness of the real world: It leads to a representation of objects in 'micro-worlds' where intelligence is framed by limited conditions of further fixed contextual objects. Any possible transfer of a 'micro-world' to the real world requires a contextual norm that must be met in order for the system to formally adapt to relevant analogous aspects. However, as the real world is not preorganized by norms and thus relevance

[124] Dreyfus, 1992, p. 208.
[125] Dreyfus describes: "Stated in a neutral way the problem is this: as we have seen, in order to understand an utterance, structure a problem, or recognize a pattern, a computer must select and interpret its data in terms of a context. But how are we to impart this context itself to the computer?" See: Dreyfus, 1992, p. 208.
[126] Dreyfus, 1992, p. 58.
[127] Dreyfus, 1992, p. 57.

can not be given universally to individual situations, a logic-based transfer is contested by Dreyfus.[128]

While the real world is open to infinite possible influences and situational changes, a 'micro-world' is synthetically constructed by and constantly bound to a limited amount of abstract elements and their defined relations.[129] GOFAI machine reasoning can be notable within such an isolated 'micro-world'[130]. However, it can not be considered machine understanding as it has no comprehension "of what a world is"[131]. Also, partially understanding specific domains of the world must be questioned as 'micro-worlds' are not specific forms of the real world. According to Dreyfus, a "world is an organized body of objects, purposes, skills, and practices in terms of which human activities have meaning or make sense"[132]. Such a world can be considered a local "sub-world"[133] of human involvement as, for instance, different cultures or scientific fields. As such, it must be taken into account in a holistic understanding, as a specific formation of the real world and not as an isolated context.[134] Dreyfus claims that GOFAI's reasoning successes are limited to 'micro-worlds' and not generalizable because the task of representing common sense bears not only challenges for AI research but also to Philosophy in general.[135]

Thus, in a broader picture, the representation of objects and contexts in 'micro-worlds' leads to a "regress of contexts"[136]. The objectification of contexts[137] requires selecting relevant features of the context constituting its explicability and representability. However, selecting relevant features and adding them consistently in a representation reduces the representation's explicability: Whereas the explanandum is usually an object to be explained, it is explained by its context, which relates to it as the explanans. Nevertheless, to be explained, the represented context's broader context has to be taken into account, whereas, in being represented in turn – the context of the explanandum is taken as a further explanandum. The "ceteris paribus condition" here prevents an actual explanans from being found, which is usually anchored in the explanandum's context. Therefore, in the case of representing real-world contexts, the

[128] Cf. Dreyfus, 1992, pp. 56-58.
[129] Cf. Dreyfus, 1992, p. 10.
[130] Cf. Dreyfus, 1992, p. 27 f.
[131] Dreyfus, 1992, p. 14.
[132] Dreyfus, 1992, p. 14.
[133] Dreyfus, 1992, p. 14.
[134] Cf. Dreyfus, 1992, p. 13 f.
[135] Cf. Dreyfus, 1992, p. 26 f.
[136] Dreyfus, 1992, p. 289.
[137] Cf. Dreyfus, 1992, p. 56.

explanatory method might lead to an infinite "regress of contexts", in which a sufficient explanation can not be found.[138]

How do we, however, acquire an objective understanding of intelligence in a situated context that differs from the GOFAI approach in including an understanding of being-in-the-world? As already seized on above, it is, according to Dreyfus, the *embodiment* that bridges the gap between symbolic representation and subjective meaning within situations that allows human intelligence to be context-sensitive, attribute relevance, and thoroughly understand the world and its various formations. The embodied constitution of human intelligence needs to be examined by phenomenological practice:

> "Instead of modeling intelligence as a passive receiving of context-free facts into a structure of already stored data, Husserl[, the father of phenomenology,] thinks of intelligence as a context-determined, goal-directed activity—as a search for anticipated facts"[139].

Herein, the focus lies upon the situated subject's meaningful engagement with the world – in forms originating from cultural and social practices in contrast to formal rules. Based on this rationale, Merleau-Ponty highlights the body as the ground for responsiveness to and action within the meaningful situation.[140] According to Dreyfus, the Husserlian notion of intelligence advises AI researchers to make "a step forward in AI techniques from a passive model of information processing to one which tries to take account of the context of the interactions between a knower and his world"[141]. In order to understand human intelligence and possibly create machine intelligence, the embodiment of human intelligence must be investigated: "intelligence requires understanding, and understanding requires giving a computer the background of common sense that adult human beings have by virtue of having bodies, interacting skillfully with the material world, and being trained into a culture"[142].[143]

Referring to Heidegger, Dreyfus illustrates that intelligent human access to context, that is, setting a focus in a situation of infinite possible characteristics, is pre-conceptually organized by meaning that we continually experience in being-in-the-world. It is the holistic nature of the world that determines the "meaningful patterns and regions"[144] according to which we orientate ourselves and move through the world – which, however, stand in contrast to logics of various abstraction degrees that breach a holistic situation in order to be applied.[145] Given the subject-

[138] Cf. Dreyfus, 1992, pp. 288-290.
[139] Dreyfus, 1992, p. 34.
[140] Cf. Dreyfus, 1974, pp. 26-36.
[141] Dreyfus, 1992, p. 35.
[142] Dreyfus, 1992, p. 3.
[143] Cf. Dreyfus, 1992, p. 36.
[144] Dreyfus, 1992, p. 274.
[145] Cf. Dreyfus, 1992, p. 274.

object resolution within *being-in-the-world*, building rule-based intelligence with objectified contexts would, to say the least, bypass the subject and its meaning – or it simply would not be possible as this is only half the battle: When it comes to human intelligence, there is no object perceived without being contextually interpreted by an embodied subject in terms of meaning.[146]

> ”This use of paradigms and context, rather than class definitions, allows our recognition of patterns to be open-textured in a way which is impossible to for recognition based on a specific list of traits. […] For further help we must turn to the existential phenomenologists and, in particular, to Merleau-Ponty who postulates that it is the body which confers the meaning discovered by Husserl. […] Moreover, as Merleau-Ponty has pointed out, the body is able to respond as a whole to its environment”[147].

2.3. AI and emotion

2.3.1. Merleau-Ponty’s intentionality

As another important dimension of embodiment, the focus in this chapter is on emotion.[148] In Dreyfus’ comprehension of human intelligence, human beings attribute relevance to specific aspects of a situation by meanings they experience in situations – whereas meaning itself is strongly influenced or even constituted by emotional experience. The differentiation between significant and insignificant aspects of a situation, hence the focus set in the situation, is, for instance, brought forth by needs that can be emotionally composed. Dreyfus illustrates this by the example of falling in love: “In such a creative discovery the world reveals a new order of significance which is neither simply discovered nor arbitrarily chosen”[149]. What the GOFAI approach is therefore missing is certainly not the artificial capacity to fall in love. However, it is not only lacking the recognizability of and sensitivity to the context of distinct represented facts but also an emotional capacity to guide the attribution of relevance and thus to learn, that is, to acquire skills.[150] Learning requires the cultivation of implicit and incorporated knowledge obtained by “shared practices which seem to be picked up in everyday interactions not as facts and beliefs but as bodily skills for coping with the world”[151].

---

[146] Cf. Dreyfus, 1992, pp. 288-290.
[147] Dreyfus, 1974, p. 26.
[148] As a further conceptual element of intelligence, *emotion* is here pragmatically understood as “phenomenal states as […] [e.g.,] fear, grief, gratitude, guilt, happiness” whereas six emotions are categorized as basic emotions (anger, disgust, fear, happiness, sadness, surprise) and are therefore constitutive to other emotions. Emotions differ from bodily sensations and moods in being with some exceptions generally „object-directed (emotional intentionality)”. Thus, emotions are said to be “involuntary in character […][,] crucial to decision-making […][,] maintaining adequate belief formation […] and regulating social relations (moral emotions or emotions of self-consciousness)”. Quotation from Rakova, 2006, p. 55.
[149] Dreyfus, 1992, p. 277.
[150] Cf. Dreyfus, 1992, pp. xlv, 45.
[151] Dreyfus, 1992, p. 47.

Dreyfus understands emotion in a phenomenological sense on the basis of Merleau-Ponty's notion of *intentionality*: Intentionality is to be understood as the human beings' relatedness in the world, not in terms of physical or logical relations to abstract objects but in terms of a "unity of the world, [which is 'lived' as ready-made or already there] before being posited by knowledge in a specific act of identification"[152]. Therefore, intentionality anchors human consciousness in the world pre-conceptually: The phenomenological notion of intentionality describes consciousness as pre-conceptually and bodily experienced by the human being, prior to single elements being mentally analyzed and formally represented. Understanding does, therefore, not refer to reasoning and judgment based on representation but to getting aware of „the total intention […] sum[ming] up some unique manner of behavior towards others, towards Nature, time and death"[153] that is regarded the structuring principle of perceiving the world. Dreyfus, therefore, presumes that the attempt to model intelligence by replacing human intentionality with defined and stored representations must fail:

> "The problem precisely was that this know-how, along with all the interests, feelings, motivations, and bodily capacities that go to make a human being, would have had to be conveyed to the computer as knowledge — as a huge and complex belief system—and making our inarticulate, preconceptual background understanding of what it is like to be a human being explicit in a symbolic representation seemed to me a hopeless task"[154].

Merleau-Ponty views phenomenology as the collapse of two extremes: "extreme subjectivism and extreme objectivism in its notion of the world or of rationality:"[155] Therefore, the Cartesian dualism of body and mind, which is criticized by Dreyfus, is eroding not only in Heidegger's understanding of being-in-the-world but also in Merleau-Ponty's phenomenological understanding of intentionality and rationality. The concepts are no longer ingrained in the ontological tension of bringing together objective philosophical and mathematical relations with the meaningful subjective experience of human beings.[156] Human emotion and memory are considered by Merleau-Ponty to anchor the human being in its holistic existence, in being-in-the-world. Emotions are responsible for upholding the deep human involvement in lived situations and directing intentions on the basis of emotive phenomena that can be considered

[152] Merleau-Ponty, 2005, p. xix.
[153] Merleau-Ponty, 2005, p. xx.
[154] Dreyfus, 1992, p. xi f.
[155] Merleau-Ponty, 2005, p. xxii.
[156] Merleau-Ponty points out to the dissolution as follows: "As for consciousness, it has to be conceived, no longer as a constituting consciousness and, as it were, a pure being-for-itself, but as a perceptual consciousness, as the subject of a pattern of behaviour, as being-in-the-world or existence, for only thus can another appear at the top of his phenomenal body, and be endowed with a sort of 'locality'. Under these conditions the antinomies of objective thought vanish. Through phenomenological reflection I discover vision, not as a 'thinking about seeing', to use Descartes' expression, but as a gaze at grips with a visible world, and that is why for me there can be another's gaze; that expressive instrument called a face can carry an existence, as my own existence is carried by my body, that knowledge-acquiring apparatus". See: Merleau-Ponty, 2005, p. 409.

motivational modes of being.[157] Therefore, Merleau-Ponty's understanding of meaning is said to strongly incorporate emotions as they facilitate human involvement in the world and the intersubjective experiences shared by human beings, which can not be reduced to biological or physical explanations.[158]

2.3.2. Emotion underlying cognition

However, Dreyfus' notion of emotion as an embodiment dimension is even intensifying the role of emotion in human perception and intelligence. Here, the role of emotion is understood as fundamental to intelligence which differs essentially from the computationalist conception of intelligence in the research field of AI: In the traditional computationalist position, emotions are not involved at all, as human intelligence is equal to computational information processing whereas informational entities are considered primarily syntactical. That means that the way of processing would not change if the information to be processed was semantically differentiated. Dreyfus notes that although some AI researchers attempt to integrate approaches to emotion in their research, they overlook the fundamental role of emotion: "[E]motions and concerns accompany and guide our cognitive behavior"[159].

Dreyfus proposes a lower and a higher level of human intelligent activities – not in terms of the evaluation of intelligence – but in terms of subconscious activities, emotion, and intuition underlying a higher level of conscious reflection: "[T]o put it phenomenologically, what if the 'higher,' determinate, logical, and detached forms of intelligence are necessarily derived from and guided by global and involved 'lower' forms?"[160] To explicitly derive what may be implied by this description, the lower level of intelligence Dreyfus could refer to "[in]determinate, [non-]logical, and […] [embodied], global and involved forms" actively constituting the higher level. However, because the lower level is still difficult to scientifically access, explore or describe, according to Dreyfus, also GOFAI intelligence would only mirror the higher level of intelligence: „It turns out that it is the sort of intelligence which we share with animals, such as pattern recognition (along with the use of language, which may indeed be uniquely human) that has resisted machine simulation"[161]. Dreyfus holds embodiment and its dimensions to be the presupposition for a machine reconstruction of components of the lower level. However, most importantly, embodiment would, in turn, presuppose the neural connections underlying human

[157] Cf. Merleau-Ponty, 2005, p. 99.
[158] Tone, R., Levin, K. & Køppe, S., 2018. Affective Incarnations: Maurice Merleau-Ponty's Challenge to Bodily Theories of Emotion. *Journal of Theoretical and Philosophical Psychology, 38, 4*, pp. 207-213.
[159] Dreyfus, 1992, p. 276.
[160] Dreyfus, 1992, p. 237.
[161] Dreyfus, 1992, p. 237.

brain activities and, therefore, fundamentally differ from heuristic attempts to robotic forms of embodiment. Dreyfus deems the nervous system to be responsible for activities necessary for "indeterminate, global anticipation"[162] which was underlying intelligence on the lower level, such as pattern recognition and communication. Therefore, the lower level of intelligence could not be described or rebuilt by rules and representations.[163]

2.3.3. The biological assumption

The important role of neurobiological mind-brain activities for intelligence is further evolved in Dreyfus' '*biological assumption*',[164] which predates the psychological argument against the computationalist position and the possibility of symbolic AI. However, it will conclude the argumentative reconstruction of the thesis because it is not only central to the following discussion section but also relative to emotion theories. As in the psychological assumption, Dreyfus emphasizes that the brain might work in a completely dissimilar way than a computer does and hence refutes the computational position once again. The GOFAI researchers assumed that information was centrally processed by sequenced firings of neurons: A singular neural impulse "was taken to be the unit of information in the brain corresponding to the bit of information in a computer"[165]. Although the "all or nothing"-firing of neurons might remind of a binary digit, Dreyfus doubts that the human neural impulse model indicates any digital or even symbolic processing entity. Instead, he assumes a decentralized and distributed way of information processing in the brain. He states that a causal relation between distinct symbol elements and associated distinct neural impulses can not be neurobiologically proven. Rather, as symbolic processing cannot be proven, the neurobiological basis of human intelligence differs pivotally from the architectural approach to intelligence in GOFAI:[166]

> "In fact, the difference between the 'strongly interactive' nature of brain organization and the noninteractive character of machine organization suggests that insofar as arguments from biology are relevant, the evidence is against the possibility of using digital computers to produce intelligence"[167].

The importance of refuting the biological assumption becomes particularly clear when one considers the context-bound nature of emotion: The philosopher Stephen Asma and the psychologist Rami Gabriel argue for an extended understanding of emotion as emotion is

[162] Dreyfus, 1992, p. 237.
[163] Cf. Dreyfus, 1992, p. 236 f.
[164] Cf. Dreyfus, 1992, pp. 159-162.
[165] Dreyfus, 1992, p. 159.
[166] However, Dreyfus already recognizes the potential of early neural networks as implying a relevant analogy to the actual functionality of the human brain. Though, it contradicts the GOFAI approach insofar as neural networks' constitution is neither rule-based nor representational but instead sub-symbolic. This analogy apprehended by Dreyfus is evaluated in more detail in the following chapter.
[167] Dreyfus, 1992, p. 162.

distributed not only all over the body but also all over the environment, the situated context respectively, of the feeling person. Emotions are directly interlinked with situations and therefore shaped by social and cultural experiences, which results in our "'world,' or umwelt, […] [being] intrinsically emotional"[168]. Based on this view, it becomes evident what separates GOFAI from human intelligence: While GOFAI attempts to centrally and abstractly formalize and fix the meaning of objects detached from their context, human intelligence incessantly interacts with its situated context, which is imbued with emotions and therefore filled with meaning.

Not only was the biological assumption later proven to be wrong as predicted by Dreyfus,[169] but also do Asma and Gabriel empirically support Dreyfus' hypothesis of the lower level of intelligence constituting a higher level. However, the lower level is essentially comprised of human emotion, which leads them to the formulation of a theory confronting the computationalist position: "The Emotional Theory of Mind"[170]. This emerging understanding of emotion as being constitutive to intelligence is central to a holistic concept of embodiment: As Dreyfus already stated, emotion is not only accompanying but also governing cognitions – that means the mind constantly involves emotions in its activities. Referring to Merleau-Ponty and Dreyfus, Asma and Gabriel attribute an even more important role to emotions: they enable human beings to acquire intelligence.[171] In interaction with cognitions, emotions are here understood as essential components of a mental feedback loop of learning: While emotions constitute learning bottom-up, cognitions regulate learning top-down.[172,173] Without this emotional experience of the world, learning and, hence, intelligence is deemed impossible.

Learning here, however, does not only refer to bodily practices such as tying one's shoes but also to the acquisition of common sense knowledge which is considered deeply bound to

[168] Asma & Gabriel, 2019, p. 6.
[169] Dreyfus explains: "In surveying the four assumptions underlying the optimistic interpretation of results in AI we have observed a recurrent pattern: In each case the assumption was taken to be self-evident—an axiom seldom articulated and never called into question. In fact, the assumption turned out to be only one alternative hypothesis, and a questionable one at that. The biological assumption that the brain must function like a digital computer no longer fits the evidence. The others lead to conceptual difficulties" See: Dreyfus, 1992, p. 225.
[170] Asma & Gabriel, 2019.
[171] Cf. Asma & Gabriel, 2019, pp. 31-34.
[172] Cf. Asma & Gabriel, 2019, pp. 7-10.
[173] Asma and Gabriel depict the relation of emotion and mind as: "Affective science can demonstrate the surprising relevance of feelings to perception, thinking, decision-making, and social behavior. The mind is saturated with feelings. Almost every perception and thought is valenced or emotionally weighted with some attraction or repulsion quality. Moreover, those feelings, sculpted in the encounter between neuroplasticity and ecological setting, provide the true semantic contours of the mind. Meaning is foundationally a product of embodiment, our relation to the immediate environment, and the emotional cues of social interaction—not abstract correspondence between sign and referent. The challenge then is to unpack this embodiment. How do emotions like care, rage, lust, and even playfulness create a successful social world for mammals, an information-rich niche for human learning, and a somatic marking system for higher-level ideational salience?" See: Asma & Gabriel, 2019, p. 3 f.

the bodily and emotional experience of the world.[174] When Dreyfus, later on, approaches intelligence in terms of skill acquisition, the role of emotion becomes yet clearer: Emotion goes hand in hand with learning which is in general considered to be non-representational.[175,176] In the process of learning skills, humans are becoming increasingly "emotionally involved in [their] tasks"[177]. Learning progress is, in fact, enabled because humans are emotively rewarded with the elation of successes or are undergoing the contrition of mistakes. Emotions are hence guiding cognitions insofar as they are prompting further efforts, motivations, and involvements in specific tasks as well as the initial engagement with different subjects. The unity of emotional experiences subsequently further guides the progress to manifest as skillful or intelligent behavior.[178] As such, it is accessible to us not as conscious facts or rules but rather as intuitive directedness in being-in-the-world on the basis of implicit and incorporated knowledge.[179]

What becomes apparent, then, is that phenomenology not only raises fundamental questions in relation to epistemology but also in relation to emergent technologies such as AI. Human intelligence's embodied, situated, and emotional nature is a prerequisite for how the world is perceived and scientifically accessed by individuals involved in the world. Therefore, research concepts should not ignore the questions raised by phenomenology but rather seek methods to address them. These questions must be taken into account in order not to uphold a concept of intelligence that is based on false assumptions and that, subsequently, is technologically realized in AI. Dreyfus' identification of the misleading assumptions united in the computationalist position showed that phenomenology can be considered a key determinant in revealing important components of human intelligence. Therefore, the following juxtaposition is based on a comprehensive impression of Dreyfus' phenomenological critique: Having outlined these fundamental philosophical challenges to AI, and nevertheless returning to a defense of symbolic AI from an ethical perspective, symbolic and sub-symbolic AI is confronted below. Why did Dreyfus advocate sub-symbolic AI? And how must the debate be continued from today's perspective?

[174] Cf. Dreyfus, 1992, p. xx f.
[175] Cf. Dreyfus, H., 2002. Intelligence without representation – Merleau-Ponty's critique of mental representation. The relevance of phenomenology to scientific explanation. *Phenomenology and the Cognitive Sciences, 1*, pp. 367-383.
[176] Dreyfus, H. & Dreyfus, S., 2004. The Ethical Implications of the Five-Stage Skill-Acquisition Model. *Bulletin of Science, Technology & Society, 24, 3*, pp. 251-264.
[177] Dreyfus, 2002, p. 370.
[178] Cf. Dreyfus, 2002, p. 370 f.
[179] Cf. Dreyfus, 1992, p. xix f.

2.4. Symbolic and sub-symbolic AI in confrontation

2.4.1. Dreyfus' phenomenological advocacy of sub-symbolic AI

Dreyfus already recognizes the potential of early sub-symbolic AI approaches to seize phenomena of human intelligence: According to him, the computationalist assumptions are mainly held and realized by the symbolic approach to AI because the sub-symbolic approach comprises features that are, from a phenomenological perspective, promising for reaching capabilities that supposedly lay beyond those of GOFAI systems. Hence, what are these features exactly, and how does he evaluate them?

Although Dreyfus is still sceptical about the possibility of strong AI in general, [180] he deems sub-symbolic approaches to be auspicious in terms of approximating the embodiment dimensions that GOFAI excluded. Because the approach aims at a bionic simulation of the human brain, sub-symbolic AI is aligned with an analogy to the brain and hence differs essentially from physical symbol systems: The approach attempts to "creat[e] artificial intelligence by modeling the brain's learning power rather than the mind's symbolic representation of the world"[181]. He advocates a methodological substitution of reasoning by learning as the corpus of common sense is rich in ambiguous meanings which align with social, cultural, and time-bound developments. To formally represent this corpus of common sense in a static way on which basis inferences are being enabled is, according to Dreyfus, an unfeasible undertaking. He deems the objective to access the world's various structures, matters, patterns, and meanings by means of learning more promising than to achieve flexible, intelligent features or mechanisms on the basis of representation and automated reasoning.[182] Beyond, he doubts that representation plays an essential role or even any role in human intelligence. Hence, by circumventing the idea of representation, sub-symbolic approaches would actually come closer to human intelligence.

> "Thus we can say that so far neural-network research has tended to substantiate the belief that coping does not require the abstraction of a theory of the skill domain. This is bad news for rationalism but fives networks a great advantage over GOFAI"[183].

The sub-symbolic approach approximates an integration of implicit and incorporated knowledge because neural networks do not reason based on stored representations and rules but instead learn through processing large amounts of data. Thereby, the neural network is able to utilize the data to assess new situations. This machine learning capacity by data processing

[180] Cf. Dreyfus, 1974, p. 32 f.
[181] Dreyfus, 1992, p. xiv.
[182] Cf. Dreyfus, 1992, p. xiv; xlv.
[183] Dreyfus, 1992, p. xxxv f.

resembles human experience: Since we access different situated contexts in a form that is shaped by experiences from prior contexts, the way we experience present phenomena is constituted by earlier forms of experience.[184]

Dreyfus links this constitutive relation of the neural networks' training processes to a potential condition of the networks to access phenomenal experience bringing forth intelligence: "This would […] make [a sophisticated neural network] a perfect candidate for the neural basis of the phenomenon Merleau-Ponty calls the intentional arc"[185]. The intentional arc is here understood as the specific focus set by prior experiences with which human beings enter a new situation. As such, it is not actively stored in memory or consciously present in the situation. Rather it is a pre-conceptual and non-representational form of implicit knowledge: "The idea of an intentional arc is meant to capture the idea that all past experience is projected back into the world. The best representation of the world is thus the world itself"[186]. Hence, Dreyfus favors the sub-symbolic AI in terms of the experience-bound composition of neural networks.[187]

According to Dreyfus, sub-symbolic AI resembles not only human experience better than GOFAI but also other conditions of embodiment such as "skillful coping"[188] in respect of various contexts. Especially the case of reinforcement learning would approximate human intelligence because it is oriented toward a functional state of maximal reward or satisfaction. The strategy to reach that state is acquired by the artificial agent itself in processing different environmental scenarios and evaluating past outputs for optimizing future action. Its learning technique is hence less dependent on human supervision. Dreyfus regards this as similar to human motivation, which aims to satisfy context-specific needs. While emotion motivates the learning progress of humans through rewards, reinforcement learning would be a potential first step to mirroring this orientation toward functionally equal rewards.

However, embodiment goes far beyond an objective reward function. Because – when respect is paid to being-in-the-world – the subject is equally as present in cognitive processes as the world itself is, intelligence must be understood holistically so that intelligent qualities are driven by the constitutive sources of embodiment such as personal needs and emotions. Although, according to Dreyfus, sub-symbolic are preferred to GOFAI approaches, such an objective reward function would still be tied to Cartesian dualism because it widely excludes

[184] Cf. Dreyfus, 1992, p. 265 f.
[185] Dreyfus, H., 1996. *The Current Relevance of Merleau-Ponty's Phenomenology of Embodiment, Archive: The Electronic Journal of Analytic Philosophy, 4, 1996.* [Online].
[186] Dreyfus, 2002, p. 373.
[187] Cf. Dreyfus, 2002, p. 374.
[188] Dreyfus, 1992, p. xli.

the role of the perceiving subject. As such, it would lead to an automatic exclusion of meanings that emerge from the perceiving subject's involvement in the world and prevent the attribution of relevance that is enabled by the subjective and qualitative experience of these meanings.[189]

> "Our needs, desires, and emotions provide us directly with a sense of the appropriateness of our behavior. If these needs, desires, and emotions in turn depend on the abilities and vulnerabilities of a biological body socialized into a culture, even reinforcement-learning devices still have a very long way to go"[190].

Thus, Dreyfus also observes that the sub-symbolic approach is missing the capacities of generalization for which the understanding of common sense might be necessary: „No one has any idea how to get a network or any other mechanism to generalize in the way that would be required for human-like intelligence"[191]. He doubts that without generalization, similarities can be recognized and relevance determined appropriately in order to achieve an intelligent adaption to new or even unforeseen situations. While humans determine relevant aspects by the attribution of meaning according to the generalization of situational similarities, the sub-symbolic approach would be very limited in imitating these capabilities. In sum, Dreyfus assumes that certain sub-symbolic successes might seem promising but that it might still be possible that the sub-symbolic approach founders in a similar way as GOFAI.[192] Although sub-symbolic approaches are closer to the human embodiment of intelligence than GOFAI approaches, a proper understanding of these dimensions is still hampered by the computationalist view that underlies not only conceptions of human intelligence but also approaches to artificial intelligence:

> "All these uniquely human capacities [of embodiment] provide a 'richness' or a 'thickness' to our way of being-in-the-world and thus seem to play an essential role in situatedness, which in turn underlies all intelligent behavior. There is no reason to suppose that moods, mattering, and embodied skills can be captured in any formal web of belief […]. [Yet], all AI workers and cognitive psychologists are committed, more or less lucidly, to the view that such noncognitive aspects of the mind can simply be ignored"[193].

Dreyfus' critique and his remaining scepticism about both approaches were partially disproved by the times: "The present difficulties in game playing, language translation, problem solving, and pattern recognition"[194] now belong to the past, as, for example, sub-symbolic systems as DeepMind's AlphaGo beat the world champion of the ancient board game Go in 2016,[195] various online tools such as Google Translate are capable of translating texts into more

[189] Cf. Dreyfus, 1992, pp. xl-xlv.
[190] Dreyfus, 1992, p. xlv.
[191] Dreyfus, 1992, p. xlii.
[192] Cf. Dreyfus, 1992, pp. xxxiii-xxxix.
[193] Dreyfus, 1992, p. 53.
[194] Dreyfus, 1992, p. 226.
[195] Spiegel Online, 2016. *Software schlägt Go-Genie mit 4 zu 1.* [Online].

languages than individuals are capable of,[196] and symbolic approaches to problem-solving have also made significant advances.[197]

Dreyfus' has rightly increased the focus on sub-symbolic AI from a phenomenological point of view as it can be assumed that it comes closest to a form of incorporated knowledge within its processing capacities of vast amounts of data which are in part sensory. However, his rejection of symbolic AI does not seem justified, as strongly cognitive attributes of human intelligence are better simulated by symbolic AI.[198] The symbolic approach to AI is still promising as it has progressed since GOFAI criticism. Most importantly, symbolic techniques are essential to transparent and reason-based approaches in AI research and development.

To understand the tension of the symbolic-sub-symbolic distinction, which is still perceptible in AI, the strengths and weaknesses of both approaches are discussed in the following from a contemporary point of view. Why was symbolic AI thought to be defeated, and why is sub-symbolic AI dominating present research and development? Are the approaches friends or enemies or are we allowed to rely on both when it comes to hybrid AI? It is shown in the following section that the strengths and weaknesses of symbolic and sub-symbolic AI can nowadays be regarded as complementary. A hybrid approach to AI research and development is advocated in regard to an ethical dimension: "What computers shouldn't do".

2.4.2. Assessment of symbolic and sub-symbolic AI from today's view

As indicated before, it is shown that even if symbolic and sub-symbolic approaches are often seized to be opponents due to their different key concepts, milestones, and schools of thought, they are actually complementary in theory: Although Dreyfus was convinced about the failure of GOFAI early on, it is argued here, that for a substantial contemporary vision of AI, we are requiring both *reasoning* and *learning* in hybrid interaction to complement strengths and balance weaknesses. In this thesis 'hybrid AI', hence, denotes a system that combines symbolic and sub-symbolic components.

As previously stated, symbolic AI is referred to as the approach of *reasoning*. It is hence particularly strong in the *representation of hierarchical and sequential structures*.[199] This is exemplified by the expert systems that have been quite advanced already in the 1980s. With the hierarchical and sequenced representations, a narrow reconstruction of mechanical steps supposedly enabling logical reasoning was achieved and thus organized similarly to human

[196] Cf. Perez, S., 2022. *Google Translate adds 24 new languages, including its first indigenous languages of the Americas.* [Online].
[197] Zhang, D. et al., 2021. *The AI Index 2021 Annual Report*, pp. 72-74. [Online].
[198] Cf. Boden, 2014, pp. 96-100.
[199] Cf. Boden, 2014, p. 93.

expert knowledge. It made expert systems a powerful tool, broadly deployed by numerous corporations in the United States. The hope was to grow a knowledge base able to represent human common sense entirely. In that way, the expert system could flexibly change the expert domain and be truly intelligent.

However, Dreyfus was right that GOFAI systems turned out to be *fragile* in application to the real world: The represented knowledge must be continuously maintained, which means updated to developments in the world. Because the representations are defined, the adaption to semantic changes in common sense is a fine-grained task that, in contrast to machine learning approaches, often requires human intervention. These updates are required if semantical changes in natural language occur. For instance, a word could gain a new meaning by language trends or if new meanings arise from events that are described and discussed by the public. Furthermore, if the required definitions for GOFAI were not sufficiently exact, missing or contradicting each other, expert systems and GOFAI systems in general turned out to be very fragile as the systems were not able to adapt to uncertain conditions by learning techniques. These problems are still relevant today in symbolic AI but mitigated. For instance, expert systems can be combined with learning approaches.[200]

Thus, with symbolic AI, *various formal logics can be represented*. The method of problem-solving can therefore rely on different representational foundations such as mathematical or philosophical logic, predicate logics with different quantification degrees, for instance, first-order logic (FOL) and higher-order logic (HOL), and even deontic logic when it comes to validation of permitted or prohibited actions.[201] Another strength of the symbolic AI approach is that it is very *precise*, as accurate definitions are the premise for its logic-based techniques. Therefore, problem-solving has a promising foundation in symbolic AI: Problems can be illustrated in great detail, and the solutions found are outlined very exactly.[202] An auspicious example is automated theorem proving (ATP), whereby formalized arguments can be formulated as problems, and the validity of conclusions subsequently verified or falsified on formal means.[203]

However, a weakness is *dealing with uncertainty and ambiguity*. Even though our world can, to a certain degree, be logically accessed and described, there is no universal logic underlying the real-world composition, events, or beings. The real world is rather often ambiguous and uncertain. You can think about natural language, for example, where the

[200] Cf. Russell & Norvig, 2021, p. 40-42.
[201] Cf. Russell & Norvig, 2021, pp. 272-274, and thus cf. Benzmüller & Lomfeld, 2020a, p. 4.
[202] Cf. Boden, 2014, p. 93.
[203] Cf. Russell & Norvig, 2021, p. 240.

meanings of symbols are often multiple, flexible and arbitrary. GOFAI systems were too fragile when it came to applications to new situations which were not or even could not be defined and therefore introduced uncertain factors to the system. Depending on the scope of fragmentary or contradictory data, uncertainty led to a meaningless result or to a whole error in the system.

Thus, the GOFAI systems were confronted with the so-called '*frame problem*' which stated that the application domain did not represent the real world in an appropriate way: As criticized by Dreyfus, therefore, intelligent actions were limited to specific domains or 'micro-worlds', which were framed by beforehand definition. Because the exact changes of aspects in a certain context induced by action could not be predicted by the system and thus, could not be learned without exact prior definition, the 'micro-world' results were not sufficiently transferable to real-world scenarios. The problem of the transferability deficit of 'micro-worlds' to the real world intensifies in relation to a philosophy of science dimension pointed out by Dreyfus: „if this phenomenological description of human intelligence is correct, there are in principle reasons why artificial intelligence can never be completely realized. […] [T]here are in the last analysis no fixed facts […] [s]ince human beings produce facts, the facts themselves are changed by conceptual revolutions"[204]. Dreyfus, therefore, views facts as constructivist, flexible, and textually open to social and cultural transformations. Since the meaning of facts must be determined and thus textually closed in order to be representable, the ontological problem of 'micro-worlds' could not be adequately solved by GOFAI.

A strength of symbolic AI, however, is its *transparency*.[205] Dreyfus accused GOFAI of being a realization of the rationalist school of thought and emphasized the importance of investigating phenomenological issues brought forth by it. Therefore, his advocacy of sub-symbolic AI is associated with exploratory accessing human intelligence by observing, collecting data on, and describing phenomena. Emerging analogies are then responsible for understanding intelligence from within the correlating phenomenon. Zigon points out, for example, that Dreyfus' phenomenological appeal is partly realized in sub-symbolic AI:

> „Although the contemporary successes of data-centric machine learning are largely a matter of engineering advances coupled with the accumulation and storage of massive amounts of data, it is clear that these new technologies satisfy in some ways the phenomenological approach to intelligence that Dreyfus argued for"[206].

[204] Dreyfus, 1992, p. 282.
[205] Cf. Bolander, T., 2019. What do we loose when machines take the decisions?. *Journal of Management and Governance, 23,* p. 866.
[206] Zigon, J., 2019. Can Machines Be Ethical? On the Necessity of Relational Ethics and Empathic Attunement for Data-Centric Technologies. *Social Research, 86, 4* , p. 1004.

However, realizing these phenomena exclusively out of observation, such as in sub-symbolic AI, conclusively averts the rationalist stance to set causal, symbolic, and explanatory relations. In contrast to solely theoretical domains, these relations are necessary, in practice, to accomplish transparency through explanation. Since AI applications are now widely used, transparency is needed from a social and practical perspective: Through symbolic representations, rules and conclusions, it becomes understandable which problems are subject to which causes and how those causes must be prevented to solve problems. The further course of the thesis will specifically address the ethical dimension of such a disregard for transparency and concrete explanatory relations by AI systems.

Although symbolic AI is favorable in terms of its transparent constitution, a weakness is, according to Boden, its dependence on the principle of a *central execution entity*.[207] The concept of a knowledge base where actions are executed from a central inference engine is criticized as it differs from natural intelligence: Animal or human intelligence was shown to be mainly constituted as rather decentralized and distributed over the whole nervous system and did not involve the principle of central control by distinct inference rules executed by a singular entity. This was early on criticized by Dreyfus as outlined in the biological assumption, which was reconstructed in chapter two.

How does the sub-symbolic approach, referred to as the approach of *learning,* deal with the weaknesses shown above? How is intelligence sub-symbolically grasped and modeled? Taking a closer look at contemporary sub-symbolic AI, it is strong in modeling intelligence in a *decentralized and distributed* way. The parallel computations in the neural network differ from symbolic AI's concept of a central knowledge base and inference engine constituting all intelligent execution and insofar comes closer to natural intelligence: "Intelligent, purposeful problem-solving behavior can be found in parts of all living things: single cells and tissues, individual neurons and networks of neurons, viruses, ribosomes and RNA fragments, down to motor proteins and molecular networks"[208]. Although animal or human nervous systems still differ significantly from neural networks, the principle of central executive control is avoided.[209]

The *representation of hierarchical and sequential structures*, however, is, in turn, a weakness of the sub-symbolic approach: Any "systematic[ally]"[210] structured order, such as semantics organized in a hierarchical way, can hardly be represented. The weakness arises from

[207] Cf. Boden, 2014, pp. 93-96.
[208] Yuste, R. & Levin, M., 2021. *New Clues about the Origins of Biological Intelligence.* [Online].
[209] Cf. Boden, 2014, pp. 93-96.
[210] Marcus, 2018, p. 9.

the sub-symbolic architecture that tends to relate features "that are themselves 'flat' or nonhierarchical"[211] and hence are perceived on a singular level in contrast to hierarchical representations on multiple levels. It negatively affects other capabilities for which a hierarchical structuring is necessary, such as planning.[212]

However, the sub-symbolic approach is not as fragile as the symbolic: This strength is sometimes described as *graceful degradation*, "that is, being able to avoid catastrophic breakdowns in the face of errors in processing or in input"[213]. Roughly speaking, there is no point of absolute failure in sub-symbolic systems that leads to meaningless or no results once it is reached – as it was the case with GOFAI. Rather, simplified, one might say: the better the input data, the better the result, and conversely, as the input data gets progressively worse, the performance of the sub-symbolic system 'degrades' progressively. Hence, it can be thought to do so 'gracefully'.

Another weakness of sub-symbolic AI systems is the *lack of representability of various logics:* "They have no obvious ways of performing logical inferences, and they are also still a long way from integrating abstract knowledge, such as information about what objects are, what they are for, and how they are typically used"[214]. This results in difficulties in representing common sense, meaning, and thus, causal relations. Therefore, sub-symbolic AI is also less precise than symbolic AI: Obviously, there is a great difference between proceeding to action by defined means and proceeding to action by probabilistic means.[215]

Because sub-symbolic systems don't rely on appropriate and exact definitions in order to generate results that proceed to AI action on the basis of logic, their *action is less dependent on their design* and more on the data they are trained on or the amount of computing power they can access. This is of advantage because, first, they are able to learn from new data and recognize patterns themselves, they are more flexible to adapt to situations unseen by the programmer in the process of designing the system. Second, it might seem easier to maintain or control these external factors of data and compute than inner factors, for instance, GOFAI's logical representations of real-world relations brought forth by our common sense understanding.[216]

---

[211] Marcus, 2018, p. 10.
[212] Cf. Marcus, 2018, p. 9 f as well as cf. Boden, 2014, p. 95 f.
[213] Sun, 2014, p. 109.
[214] Marcus, 2018, p. 23.
[215] Cf. Marcus, 2018, p. 12 as well as cf. Boden, 2014, p. 95 f.
[216] Cf. Boden, 2014, pp. 94-96. Thus, cf. Huizing, A., Veenman, C., Neerincx, M. & Dijk, J., 2021. Hybrid AI: The Way Forward in AI by Developing Four Dimensions. In: F. Heintz, M. Milano & B. O'Sullivan, Edt. *Trustworthy AI – Integrating Learning, Optimization and Reasoning. Revised Selected Papers from the First International TAILOR Workshop.* 2020: Springer, pp. 73.

Still, this might be only a matter of perspective. The dependence on external factors comes with new problems: One could, for instance, lament the scarcity of labeled data or an *overreliance on data* in general. While humans are able not only to reason on the basis but also to learn on the basis of definitions – besides acquiring implicit knowledge –, sub-symbolic techniques are usually not able to do so. Instead, they need a vast amount of training data to develop intelligent features.[217] Often, deep learning neural networks already rely on significantly more training examples than humans do.[218] As Buckner puts it, "[sc]eptics […] wonder whether deep neural networks will ever be able to learn from smaller, more human-like amounts of experience"[219].

A strength of sub-symbolic AI, however, is its capacity to process incomplete data or to *deal with uncertainty and ambiguity* in general. Where GOFAI systems were fundamentally challenged, and contemporary symbolic systems still have problems with data that is either fragmentary or contradictory, sub-symbolic systems still come to feasible results. This is possible because the training data has 'taught' them many different examples that introduce volatility into the data, in which random distortions are already included but outweighed on average: "noise tolerance and pattern completion, both of which are problematic for GOFAI, result 'naturally' from the design of PDP networks"[220].[221]

This training data serves best if it is labeled and thus qualitative to represent the real world unbiased. However, not every situation can be captured data-wise as some are unforeseeable, and thus, not every dataset is labeled: "[t]he *scarcity of labelled training data* [emphasis added] is a big challenge for machine learning which restricts the applications in which AI can be deployed effectively and safely"[222]. The need to data-wise represent the world as it is and as it evolves can be considered a difficulty parallel to aiming to symbolically define the common sense in terms of an, ideally, universal logic.

Beyond, to assure that the action in question is rightful and does not harm individuals or groups, the training data has to represent our world in a fair way. However, this is easier said than done, and in the relatively short history of deep learning, there have been many incidents of *biased training data*. For example, Amazon used a tool for recruiting talent that was highly

[217] Cf. Marcus, 2018, p. 6 f.
[218] Buckner gives an example: "AlphaGo's networks were trained on over 160,000 stored Go games recorded from human grandmaster play and then further learned by playing millions of games against iteratively stronger versions of itself (over 100 million matches in total); its human opponent Lee, by contrast, could not have played more than 50,000 matches in his entire life". See: Buckner, 2019, p. 12 f.
[219] Buckner, 2019, p. 13.
[220] Boden, 2014, p. 95.
[221] Cf. also: Huizing, et al., 2021, p. 73.
[222] Huizing, et al., 2021, p. 73.

biased: It learned primarily to recruit candidates that were male, which means being male was pivotal for the decision of being hired.[223] Bias can have different causes: On the one hand, the data itself could be bad and show bias that originated from its collection or pre-processing. On the other hand, the model could cause distortion and hence be a reason for bias. Another possibility, however, is that the data simply reveals real-world issues that are socially unfair: For instance, social groups that are underrepresented in a certain area in the real world – are likely to be underrepresented in the data as well. In this way, real-world bias gets perpetuated through automation.[224] In the example of Amazon, the data and the model might have been actually realistic in strongly underrepresenting women due to the underrepresentation of women in jobs related to STEM subjects.

What is still and most importantly missing in contemporary publicly deployed sub-symbolic systems is a capacity for understanding for which reasoning can be considered crucial. Understanding here does, first, refer to a comprehension of the seemingly infinite meanings to be found in our world. Second, it refers to a self-referential machine understanding of the systems concerning their AI actions or decisions.[225] However, not only do AI systems often lack understanding of their actions but also are developers and users often left with many questions: *Sub-symbolic systems are often said to be black boxes* because the systems' functioning is opaque – that means it is often too complex or too data-intensive to be entirely understood by humans.[226] This is of serious concern from an ethical perspective: The resulting deficiency of transparency comes with the risk of harming humans and violating human rights through non-reflected or implicit causes for AI action, such as bias in the training data.[227]

In sum, it still seems like GOFAI's weaknesses have outweighed in the history of AI as the sub-symbolic approach dominated AI research and development since the 1990s. GOFAI systems represented 'micro-worlds' where it was possible to develop domain-specific symbolic AI applications rather than dealing with real-world problems. The approach thus came to its limits in terms of handling uncertainty, such as the ambiguity of natural language. The symbolic systems were highly fragile as it was impossible to converge to the integration of learning concepts to counterbalance their fragility. Instead, sub-symbolic systems are flexible in dealing with new situations as long as they are captured qualitatively sufficient data-wise. The sub-symbolic approach, thus, is decisively stronger in dealing with uncertainty and ambiguity and

[223] Cf. Hamilton, I., 2018. *Why it's totally unsurprising that Amazon's recruitment AI was biased against women.* [Online].
[224] Cf. Johnson, G., 2021. Algorithmic bias: on the implicit biases of social technology. *Synthese, 198*, p. 9948 f.
[225] Cf. Wooldridge, 2021, p. 303 f.
[226] A more precise differentiation of black boxes is given in the third chapter.
[227] Cf. Huizing, et al., 2021, p. 74 f.

not endangered by input data that is out of the norm or contradictory because the concept of learning is central to the approach. However, even if the sub-symbolic approach is momentarily dominant, it can not be said that symbolic AI failed. Boden does a reality check on the "myth of GOFAI failure"[228] and comes to a different result than 'mainstream narratives' tell us. She states it to be wrong that a general rejection of symbolic AI occurred as it was, for instance, early predicted and assumed to ultimately have happened in the late 1980s by Dreyfus. Boden stresses that important applications did 'live on', such as expert systems. Thus, seeing the bigger picture, symbolic techniques are still central to AI, for instance, planning and heuristic search, but are easily overlooked. The unjust narrative of failure is, however, strong because symbolic techniques outgrew the research field of AI and are nowadays located in the field of general computer science.[229] Hence, in part they are either invisible or unidentifiable:

> "Invisibility is only one reason why GOFAI's successes go largely unsung. Another is unidentifiability. Many aspects of AI have been so successful that people (including other computer scientists) think of them merely as part of mainstream computer science. These include computing techniques that are now taken for granted […]. Their roots in AI are forgotten. In this, AI is comparable to philosophy. It bravely asks the unanswered, almost unaskable, questions – but when it finds a reliable way of answering them, they are relabeled as questions for 'respectable' science"[230].

Thus, symbolic AI applications were decisively optimized since the time of GOFAI: Dreyfus argued that GOFAI had almost completely ground to a halt in the 1980s and that symbolic approaches to AI were thus effectively disproved in their success. For example, he claimed in the 1990s that the symbolic AI project CYC by Lenat, whom Dreyfus called the "last heir of GOFAI"[231], would have failed by early 2000. However, the CYC project continues today – still based on the original logical programming language Lisp – and is indeed making slow but steady progress.[232] Although symbolic systems today still tend to be fragile due to their dependence on precise definition, representations can be probabilistically modeled, contradictions avoided by more options for action introduced, and learning techniques integrated. For example, theorem provers are combined with learning techniques in the case of the DeepHOL system.[233] Thus, according to Boden, the frame problem has been significantly mitigated nowadays.[234] In sum, despite the weaknesses of symbolic AI, the approach is still

[228] Boden, 2014, p. 100.
[229] Cf. Boden, 2014, pp. 100-102.
[230] Boden, 2014, pp. 100-102.
[231] Dreyfus, 1992, p. xxxiii.
[232] Cf. Hutson, M., 2022. *Can Computers Learn Common Sense?.* [Online].
[233] Bansal, K. et al., 2019. HOList: An Environment for Machine Learning of Higher-Order Theorem Proving. *Proceedings of the 36th International Conference on Machine Learning*, pp. 454-463. As depicted by Russell & Norvig, 2021, p. 327.
[234] Cf. Boden, 2014, pp. 94-102.

highly relevant in AI research and development today and thus combined with other, including sub-symbolic approaches, as it will be shown in the following section.

The "myth of GOFAI failure" thus needs to be questioned when it comes to the remaining weaknesses of sub-symbolic AI. Buckner remarks that the narratives accompanying sub-symbolic AI are exaggerated as well – sub-symbolic AI is claimed to "either […] [evoke the next] 'AI winter' or […] soon usher us into a singularity of exponentially-increasing levels of intelligence"[235]. However, it must be stressed that some weaknesses must be analyzed with serious attention from an ethical perspective, as they have direct ethically problematic implications. For example, the overreliance on data in combination with various biases existing in the real world or in the pre-processed data set is ethically problematic. Without adequate training data, the systems can put humans at risk either psychologically – by being underrepresented or discriminated – or physically – by being endangered by systems that steer hardware applications such as transportation. In the automotive sector, for example, a minor modification of traffic signs can lead so-called 'advanced driver assist systems' to estimate the permissible speed significantly higher and hence cause severe car accidents.[236]

Because sub-symbolic applications are already broadly deployed in public, this urgently begs the question of whether they are reflected sufficiently from an ethical perspective. Definitely, however, it seems not to be done sufficiently when it comes to black boxes in sub-symbolic systems: To say the least, it can not be thoroughly reflected upon a matter which can not be examined because it is left in the dark. To understand how a black box system works, we may need to shed some light on its construction. Often, however, this is not possible because the complexity is too high to get an insight into the causes of action, hence rendering them transparent: "[D]eep learning systems have millions or even billions of parameters, identifiable to their developers not in terms of […] human interpretable labels […] but only in terms of their geography within a complex network"[237].

After reflecting on the weaknesses of both approaches, one might ask if there is a compromise to counterbalance them. Can we, for example, combine learning power and transparency? Perhaps it was already implicitly noticeable that the strengths and weaknesses of the symbolic and sub-symbolic approaches discussed are, in fact, complementary in theory – although the two approaches are often seen as opponents due to their different key concepts, milestones, and schools of thought. Even the GOFAI critic Dreyfus remarks that the cleavage between logic and phenomenology, between conceptualization and holism, and potentially so

[235] Buckner, 2019, p. 11.
[236] Cf. Huizing, et al., 2021, p. 73.
[237] Marcus, 2018, p. 10 f.

between symbolic and sub-symbolic AI should be bridged: "The conceptualists can't give an account of how we are absorbed in the world, while the phenomenologists can't account for what makes it possible for us to step back and observe it."[238] For a substantive vision of AI, hence, we seem to need both *reasoning* and *learning* to interact. Only then can the strengths of both sub-symbolic AI, such as its learning power, and symbolic AI, such as its transparency in sophisticated reasoning techniques, come into play. In the next chapter, it is argued that the latter needs to be strongly regarded from an ethical standpoint leading to the general plea for hybrid systems.

2.4.3. Hybrid AI: can we rely on both approaches?

As already indicated, approaches of symbolic and sub-symbolic AI are often perceived as opposing as their history is marked by different milestones. Both approaches seem to be oppositional because the sub-symbolic approach arose when the symbolic approach seemed to fall silent. However, it seems conceptually useful to view the approaches to be complementary instead: With the aim of creating strongly or generally intelligent systems, the AI achievements made so far are considered to be only "components of intelligence"[239], whose integrative composition is yet to be accomplished. Hybrid AI systems instead, which combine symbolic and sub-symbolic methods, are consequently based on the capacities of both logical reasoning and machine learning. In this way, specific strengths of both approaches can be complemented, and specific weaknesses counterbalanced. For example, machine perception, memory, and planning could be combined – which seems to be necessary to approximate a complete view of human intelligent capabilities.[240]

Quite a few both symbolic and sub-symbolic AI researchers assume that "high-level cognition"[241], respectively explicit reasoning skills, are likely better represented by symbolic techniques and hence require a hybrid system to be based on a symbolic architecture. In contrast, other exponents of the sub-symbolic faction hold that sub-symbolic activity "underlie[s] all aspects of human cognition […] [and that] reasoning and problem solving often arise from insight or intuition, or directly from perception"[242]. They are therefore advocating a sub-symbolic architecture that lays the foundation for "high-level cognition", either emerging sub-symbolically or being granularly integrated by symbolic means. What can be stated in general, however, is that the vision of hybrid systems seems not only encouraging in theory.

[238] Dreyfus, 2007, p. 364.
[239] Wooldridge, 2021, p. 303.
[240] Cf. Boden, 2014, pp. 96-100.
[241] Sun, 2014, p. 118.
[242] Sun, 2014, p. 118.

But also do hybrid applications already deployed in practice "tend to be more expressive, more powerful, often more efficient, and thus more useful"[243] because they include not only explicit knowledge, which is symbolic but also implicit knowledge, which is sub-symbolic.[244]

As described above, symbolic AI is unjustly publicly conceived as coming with more disadvantages than advantages, and it is central to the aim of true machine *understanding:* To achieve human-level understanding by AI, "planning, and internal representation, is […] essential"[245] and both capabilities are central concepts of the symbolic approach.[246] Sub-symbolic systems without the capacities of logical representation and inference are able to learn but not capable of making sense of what the learned contents actually mean. In contrast, hybrid systems that strongly rely on symbolic techniques, for example, IBM's logical neural networks, "are capable of greater understandability, tolerance to incomplete knowledge, and full logical expressivity"[247].

Most importantly, however, the symbolic approach is transparent through and through and hence doesn't come with ethical challenges just arising directly from its architecture.[248] Thus, symbolic techniques can increase transparency in sub-symbolic systems, for instance, by proving the correctness of actions, detecting errors, or communicating the causes for sub-symbolic AI action.[249]

> "If we succeed in a deep integration of symbolic and connectionist [that is sub-symbolic] approaches, we might have a hope to get future AI systems that can both learn, reason about others, use language to explain themselves in human-comprehensible ways, and engage in dialogues with humans about their reasoning and decisions. That would make algorithmic decision making much more trustworthy and have a much larger general potential"[250].

Research and development in hybrid AI are increasingly being carried out to integrate symbolic strengths such as transparency and sub-symbolic strengths as ambiguity tolerance. Figure 2 by the German Standardization Institute (DIN & DKE) reconstructs the different AI approaches historically and regarding the degree of intelligence achieved. The initial phase of AI was symbolically dominated and included heuristic systems and knowledge-based systems. The AI summer still ongoing is characterized by the sub-symbolic approach and learning systems. Hybrid systems, which according to the Association for the Advancement of Artificial

[243] Sun, 2014, p. 119.
[244] Cf. Sun, 2014, p. 118 f.
[245] Boden, 2014, p. 96.
[246] Cf. Boden, 2014, p. 100.
[247] IBM, 2020. *Getting AI to reason: using neuro-symbolic AI for knowledge-based question answering.* [Online]. This research stream goes back to 2011, were IBM Watson's defeating the champions of 'Jeopardy!' – a quizshow that can be considered to be played 'vice versa' as the participants need to find the right question to single answers.
[248] Cf. Bolander, 2019, pp. 854-856; 866.
[249] Cf. Huizing, et al., 2021, p. 74 f.
[250] Bolander, 2019, p. 866.

Intelligence AI (AAAI), will dominate the next twenty years of AI research and development "currently exhibit the highest degree of intelligence, robustness, transparency and adaptability"[251]. Symbolic modules of automated reasoning serve to introduce transparency to the hybrid system and thus verify the results of neural networks through different underlying logics, such as propositional logic and first-order logic.[252] The symbolic approach is therefore strongly advocated from an ethical perspective in the further course of this thesis – it needs to be brought into focus in order to build hybrid systems that are actually based on an ethical foundation and the goal of transparency to verify corresponding ethical values.

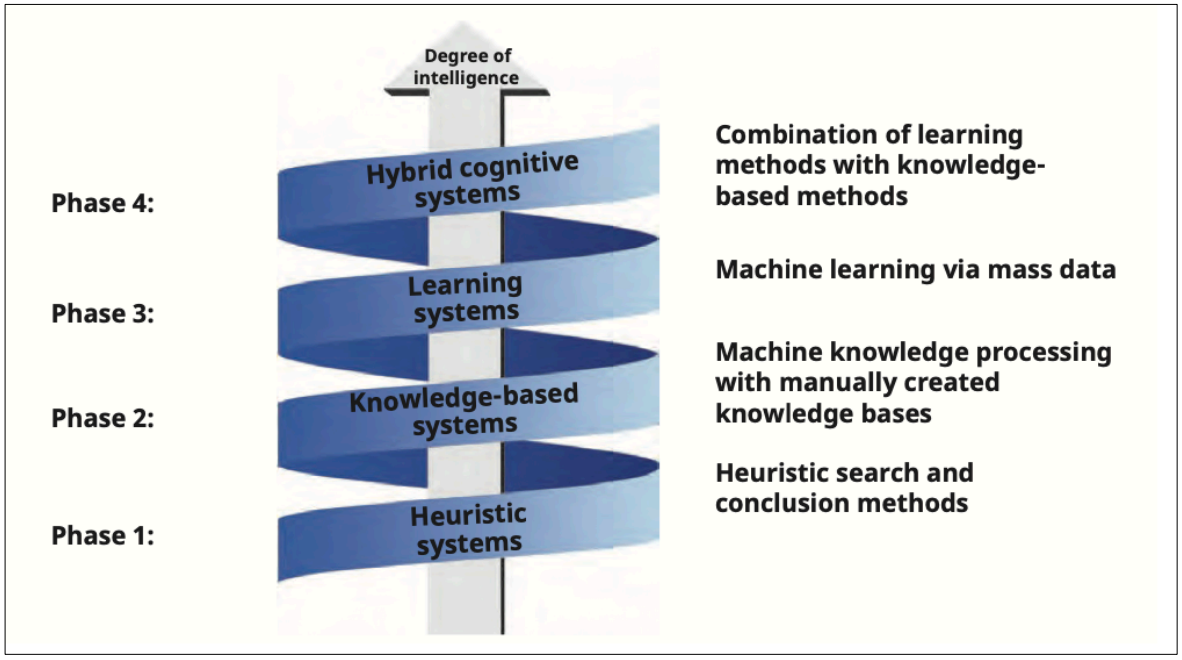

Figure 2: „The four phases of AI" by DIN & DKE[253]

A promising example of such a hybrid system is the 'Neuro-Symbolic Concept Learner', which was introduced by IBM, MIT, and DeepMind in 2019.[254] The innovation combines learning and representational techniques in "a neuro-symbolic reasoning module that executes [symbolic] programs on [a] latent scene representation"[255]. The application achieves a certain

[251] Deutsches Institut für Normung (DIN) & Deutsche Kommission Elektrotechnik Elektronik Informationstechnik (DKE), 2020. *German Standardization Roadmap on Artificial Intelligence*, p.12. [Online]. Here, the abbreviations (i.e., DIN & DKE) are preferred to the full names, as they are fairly long.
[252] Cf. DIN & DKE, 2020, p. 12; 88.
[253] DIN & DKE, 2020, p. 12.
[254] According to Confalonieri et al., the term neuro-symbolic AI or 'neural-symbolic AI' is used synonymously to the term 'hybrid AI'.
Cf. Confalonieri, R., Coba, L., Wagner, B. & Besold, T., 2020. A historical perspective of explainable Artificial Intelligence. *WIREs Data Mining and Knowledge Discovery, 11,* p. 4.
[255] Mao, J. et al., 2019. The Neuro-Symbolic Concept Learner: Interpreting scenes, words, and sentences from natural supervision. *Published as a conference paper at ICLR 2019* , p. 1.

degree of conceptual interpretation by a generalization capacity.[256] Thinking further, advanced capacities of generalization and interpretation might be a first step towards machine understanding.

3. "What computers shouldn't do": An ethical perspective on symbolic AI based on Dreyfus and current AI Ethics

Ethical viewpoints are hardly present in Dreyfus' AI critique. While it is not entirely clear why an ethical dimension was not included, it is likely that the possibility of strong AI seemed more in keeping with the times from an epistemological perspective than ethical concerns.[257] However, can Dreyfus be criticized for something he had not declared as his goal: the integration of a supposed ethical perspective? Surely not. Nevertheless, it has become apparent in recent decades that, on the one hand, merely turning away from symbolic AI and toward sub-symbolic AI poses enormous ethical challenges. Thus, it has recently been shown that symbolic approaches can be a first step towards solving these ethical challenges – and are therefore useful and necessary. Therefore, it seems that Dreyfus' critique in this regard seems incomplete from today's perspective – and needs to be interpretatively expanded to "What computers shouldn't do"[258], namely AI Ethics.

So, what is it, "[t]hat computers shouldn't do"? In the view defended here, AI systems should generally not make decisions that negatively affect or violate human rights, that is, they should not act unethically in a broader sense. To meet this need, it is necessary to understand on what grounds AI systems make decisions and why certain AI actions result from the decision-making processes. However, opaque sub-symbolic AI systems are already used in vulnerable areas today, for instance, in law enforcement in the United States or in autonomous driving – they are, therefore, already affecting individuals' human rights. Because AI systems that raise such ethical concerns often lack transparency, they are, as already mentioned, referred

[256] Cf. Mao, et al., 2019, pp. 7-9.

[257] The following statement was made by Hubert Dreyfus in relation to the barely represented dimensions of ethics and power in his and his brother Stuart Dreyfus' book 'Mind over Machine': „Our line constantly was, we talk about what we know about and at least one thing is clear: if you don't understand what expertise is and what tacit knowledge, you can't even discuss the social, political, labor movement issues, etc. intelligently. You spend your time worrying about what to if computers and expert systems come along and replace experts and workers when that is not the real problem, because they can't. So we didn't talk about the social issues". See: Dreyfus, H. & Dreyfus, S., 1990. *Sustaining Non-Rationalized Practices: Body-Mind, Power, and Situational Ethics. An Interview with Hubert and Stuart Dreyfus* [Interview] 1990, p. 71.
Therefore, it is assumed here that Dreyfus did not consider the field of AI to have grown enough in expertise and techniques to discuss ethical or social dimensions. However, the state of the art in the field of AI has changed fundamentally since then – ethical and social issues now arise directly from the application of AI systems in our daily lives.

[258] van der Meulen & Bruinsma, 2019, p. 343.

to as *black boxes*. Rudin distinguishes between two forms of black boxes:[259] First, *proprietary* black boxes are AI applications protected by trade secrets. Here, we can not understand the functioning of the system because it is protected by intellectual property rights. Second, *complex* black boxes prevent us from understanding the functioning of the system because the AI application in question is too complex or comes to conclusions seemingly arbitrary. In sum, sub-symbolic AI systems applied as black boxes prevent compliance with fundamental ethical values or rights.

Because AI is a dual-use technology, its deployment can be socially useful or, conversely, not useful or even harmful to society. While an opaque system used in public can be accused of being a black box or even harming individuals or groups, instead, the developers and operators must be held accountable for developing or using (non-)transparent systems. Without transparency, it cannot be proven whether the AI system itself is being used in a socially useful or harmful way. Therefore, responsible AI research and development should always begin in adherence to transparency, which is considered the most important ethical value in this thesis.

A rejection of symbolic AI must be contested from an ethical perspective: First, unlike sub-symbolic AI, symbolic AI can be considered central to the approach of "ethics by design"[260] because it is based on rules that are transparent. Second, symbolic AI can constitute value-based ethical reasoning as it is central to the approach of *ethico-legal governance.*[261] The term ethico-legal governance refers to normative, both ethical and legal, theories formalized and automated by symbolic techniques. It can be viewed as a future key element for compliance with ethical values in opaque sub-symbolic AI systems, respectively black box systems: Using symbolic applications, for instance, automated theorem provers, the black box system's compliance with formalized ethical or (and) legal values is proved or disproved – hence, controlled. Ethico-legal governance can be considered an effort to coordinate a human rights-based approach to AI research and development at the technical level by symbolic means.[262]

It is argued in the following that symbolic AI layers or modules should be examined as a potential standardized element of future hybrid systems. As indicated before, hybrid systems seem to be particularly valuable from an ethical perspective. Dreyfus was right concerning the strengths of sub-symbolic AI: Sub-symbolic techniques are dominant in current AI systems as

[259] Cf. Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, pp. 206-208.
[260] Ekmekci, P. & Arda, B., 2020. Bioethical Inquiries About Artificial Intelligence. In: P. Ekmekci & B. Arda, Edt. *Artificial Intelligence and Bioethics.* Cham: Springer Nature Switzerland, p. 62.
[261] The term ethico-legal governance is coined by Benzmüller, such as in Benzmüller & Lomfeld, 2020a.
[262] Cf. Benzmüller, et al., 2020b, pp. 1-6.

their functional strengths have outweighed those of symbolic AI, at least in a broad intersectoral application. However, unconditional use of sub-symbolic AI systems is ethically highly questionable: Black box systems are already in operation today and make non-transparent decisions in vulnerable areas of society, such as transportation. The decisions are non-transparent insofar as the process of decision-making remains hidden – its explanation or interpretation is impossible either for secrecy or complexity reasons.

Therefore, it is now illustrated within the approach of interpretative adequation[263] how attempts can be made to ensure that computers don't do "what [they] shouldn't do": Required is an ethical approach to AI research, development, and deployment that refers to ethical, technical and legal evaluations. Hence, it is fair to say that two forms of AI governance should be exercised for the adherence to ethical values: At the technical level, ethico-legal governance should be achieved through symbolic techniques, and at the legal level, governance should be implemented through appropriate standards and policies. Kogge identifies the interdisciplinary notion of governance as a form of steering and hence governing economic and political processes, which may also be coordinated by governmental institutions, but are more often characterized by self-organizing system dynamics.[264] In AI research and development, this could be an important concept: External regulation is to some extent necessary, but to ensure governance, the systems of the future should integrate a system-internal component of governance mechanisms. This is the case, for example, in ethico-legal governance, in which verifying and regulative mechanisms are achieved.

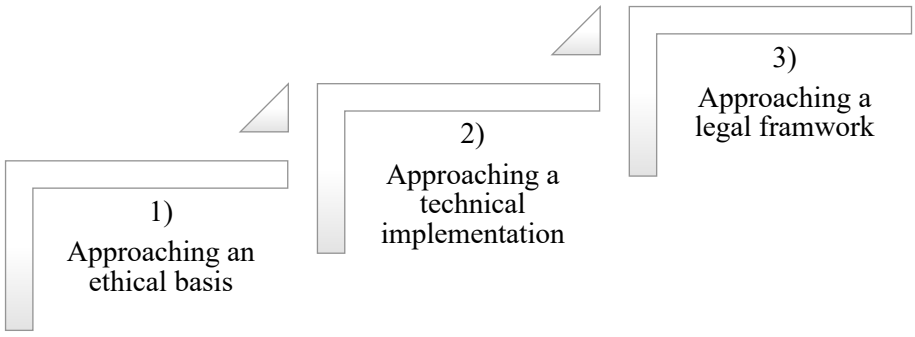

Figure 3: Approaching a solution to the black box problem (own illustration)

[263] Cf. Kogge, 2022a, p. 176.
[264] Cf. Kogge, W., 2022b. *Governance. Organon terminology toolbox.* [Online].

An approach to a solution for the problem of black box systems is described hereinafter on three levels (illustrated in Figure 3): Firstly, an *ethical basis* needs to be explored and defined as a foundation. Secondly, a *technical implementation* needs to be developed by symbolic means, and thirdly, AI systems that pose a high risk to individuals or society need to be bound into a *legal framework*. In this concept, ethical values are aimed to be implemented internally (at the technological level) by specific symbolic solutions and standards, and thus externally (at the legal level) through binding standards and laws. A symbolic solution for the implementation of an ethical value, illustrated beyond by the example of transparency, connects the first two levels, that is, the ethical and the technical level. Thus, technical AI standardization bridges the gap between the technological and the legal level as it might be thought of as legally binding in future applications that bear a high risk.

3.1. Values of AI Ethics: approaching an ethical basis

AI technologies can be considered as technologies that bear a technology conflict with moral implications in terms of technical means, concepts of future as well as concepts of humanity and society.[265] In order to establish a constitutive ethical understanding of AI that does justice to an attempt to resolve these conflicts, the definition of values with a high potential for universality is taken as a starting point – not only to identify and define ethical goals but also to examine the potential to formalize them in legal or technical norms and rules.[266]

Methodologically, considering Dreyfus' critique, the high importance of context-sensitivity and ambiguity-tolerance in context-dependent meanings needs to be taken into account. It might be difficult to abstract values that are universal sufficiently to be formalized in order for them to be valid in all possible contexts. The issue is discussed, and a solution is differentiated in more detail in the further course of this chapter. As an objective for the identification of values, however, Umbrello and van de Poel suggest an approach of "value-sensitive design […] consisting of four iterative basic steps: contextual analysis, value identification, translation of values into design requirements, and prototyping"[267]. Herein, every value constitutes normative requirements for the specific design of AI systems, as shown in Figure 4.

---

[265] Cf. Grunwald, 2011, p. 284.
[266] Cf. IEEE Global 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.* [Online].
[267] Umbrello, S. & van de Poel, I., 2021. Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, p. 294.

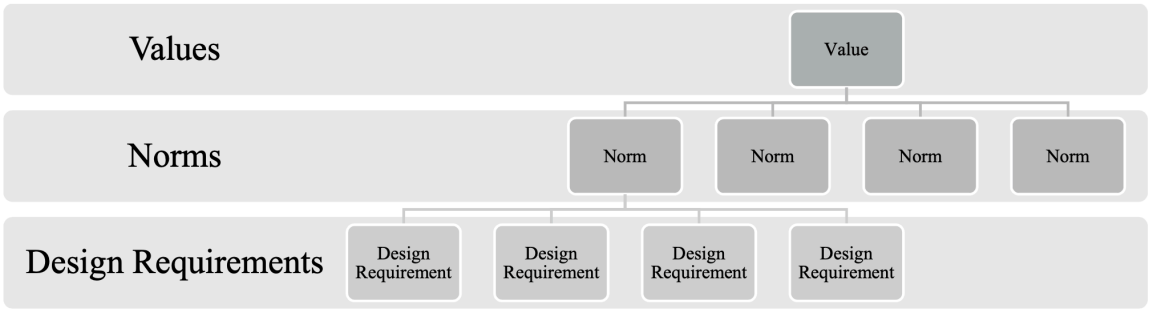

Figure 4: "Values hierarchy" by Umbrello & van de Poel[268]

According to IEEE's "First Global Ontological Standard for Ethically Driven Robotics and Automation Systems"[269] which can be considered a cornerstone of the ethical AI standardization field that is to grow tremendously in the upcoming years,[270] values proposed can be either deontological principles, consequentialist values, or ethical virtues formulated as values. They, however, need to be formalizable to concrete norms in formal logic in order to be adaptable to the architecture of an AI system.

However, in order to avoid a conflict of ethical aims, we should have a closer look at which normative ethics can be considered as an ethical foundation: Even though different ethical approaches are in the discourse within the ethics of technology, the focus here is on teleological normative approaches – in contrast to descriptive ethics, which remain too weak for a development perspective and, moreover, in contrast to ethics that are not primarily oriented towards an ethical goal ('*telos'*). To include Dreyfus in the conception of an ethical foundation, we will now examine which ethics he perceives to be coherent with his phenomenological standpoints. Although *Dreyfus did not explicitly link AI to ethics* and excluded it from his critique of GOFAI, he proposed a model of skill acquisition with his brother Stuart that descriptively includes the *acquisition of ethical skills*. Two objectives, therefore, determine the approach to an ethical basis: a reflection of his phenomenological standpoints as well as an orientation towards the perspective of technical implementation.

[268] Umbrello & van de Poel, 2021, p. 293.
[269] IEEE, 2021. The First Global Ontological Standard for Ethically Driven Robotics and Automation Systems. *IEEE ROBOTICS & AUTOMATION MAGAZINE*, p. 124.
[270] Cf. Lorenz, 2021, p. 5 f.

3.1.1. Finding ethics with Dreyfus

By the time when AI had further evolved, and the sub-symbolic summer had already arisen, Dreyfus argued for the "ethics of situated involvement"[271], which can be understood as a relativistic type of virtue ethics.[272] According to these ethics, the learning process of skills in general, including ethical skills, begins with the learning of abstract rules on the basis of independent facts. Subsequently, it gradually develops into a situated, embodied, and above all emotionally guided further development of the skills, which only implicitly observes the initial rules and develops new rules which are a result of experience-bound "spontaneous ethical response[s]"[273] to different contexts. Emotions herein initiate a reflection of learning outcomes in terms of different contexts and enable new skill developments until the acquisition reaches "its telos in involved intuitive expertise"[274], which is ultimately rewarded by satisfactory emotions.

Casacuberta and Guersenzvaig highlight that transferred to AI, this means that both symbolic and sub-symbolic approaches are necessary to achieve such ethical expertise:[275] Abstract rules and logical inferences serve as the symbolic foundation of ethical skill acquisition as well as a method for verification and validation, whereas sub-symbolic models are necessary to achieve a further intuitive development through learning behaviors.

> "If we carefully consider Dreyfus' legacy, we can realize that we are still far away from an artificial system able to take fair decisions. […] [A] pure [symbolic] system […] is not enough. […] A pure machine learning approach would not work either. […] To have an AI able to become an ethical expert we will need: Ethical declarative concepts which the system can present to justify higher order decisions […] [and] [a] pattern recognition system based on some machine learning paradigm that can capture common pre-reflective ethical judgments that are the basis of ethical expertise"[276].

Some hybrid AI control system of an autonomous vehicle could, for example, firstly be prescribed by the rule not to exceed the maximum speed. The system would initially be symbolically bound to comply with this instruction and act in accordance with it. In the sub-symbolical learning progress of the system, however, variations and slight modifications might be possible, which seem to be necessary for specific contexts. If, for instance, the vehicle recognizes that adhering to the speed limit in the context of a prematurely closing barrier in front of a railway crossing is risky, it learns to exceed the speed limit for a short time in order to still pass the railway crossing safely.

---

[271] Dreyfus & Dreyfus, 2004, p. 251.
[272] Cf. Coeckelbergh, 2019, p. 280 f.
[273] Dreyfus & Dreyfus, 2004, p. 254.
[274] Dreyfus & Dreyfus, 2004, p. 262.
[275] Cf. Casacuberta, D. & Guersenzvaig, A., 2019. Using Dreyfus' legacy to understand justice in algorithm-based processes. *AI & SOCIETY, 34,* p. 316 f.
[276] Casacuberta, D. & Guersenzvaig, 2019, p. 316 f.

Although this comprehension might contribute well to an advocacy of hybrid systems, as indicated before, one fundamental component is, however, not only missing but also seemingly out of reach: According to Dreyfus, "ethical judgments are grounded in basic human emotions"[277]. In the example described above, this could mean that the autonomous vehicle was only able to decide ethically if it actually understood the emotional dimension of harming its passengers, which is likely to be a painful mixture of grief and guilt. In order to guarantee ethical learning, "we should [hence] consider the possibility […] [of] hav[ing] some sort of artificial emotion repertory implemented"[278]. However, since AI systems are not embodied in a manner comparable to humans and do not have qualitative emotions in any case, a situated form of ethics would clearly yet be pointless.[279] Moreover, the field of AI ethics, which attempts to identify values, formalize, and possibly implement them in rule systems, can still be considered very young. Not only do eligible rules first have to be designed in the abstract for AI systems to begin acquiring ethical reasoning – but an emotionally guided further development of ethical skills also seems out of the question in the near future. The sub-symbolic climb to ethical expertise described by Dreyfus hence still seems a long way off. What we can do now, however, is focus on concrete ethical values and related rules in terms of symbolic AI to ensure an ethical foundation and symbolic governance techniques as the first step towards ethical AI research and development.

Thus, to underestimate the role of symbolic AI would mean staying with mainly perceptive inputs and arbitrary outputs without any capacity for reflection or cognitive regulation as, in simplified terms, sub-symbolic AI is often compared to lower, perceptive, levels, whereas symbolic AI is often compared to higher cognitive levels.[280] In this sense, one could argue that it is not only emotions but also higher-order cognitions that govern ethics – which in this view can be considered to be regulatorily influenced by reflection and classification, generalization and prioritization, goal setting, standardization, and rule-setting, as well as the regulation of bodily perceptions. Although the role of sub-symbolic AI for ethical expertise is definitely important, it should not be overestimated. For instance, today's promising

[277] Casacuberta & Guersenzvaig, 2019, p. 316 f.
[278] Casacuberta & Guersenzvaig, 2019, p. 316 f.
[279] First attempts of implementing synthetic emotions in AI systems have actually been suggested in the research field of Affective Computing. For instance, Arkin and Ulam suggested the implementation of an "ethical adaptor capable of using a moral affective function". See: IEEE, 2009. An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions. *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09)*, p. 381.
However, as these attempts seem very experimental and could (likely will) bring forth a range of new risks in the application, they should not serve as an ethical foundation in AI ethics until a transparent, consistent, and validated corpus of results is established.
[280] Cf. Boden, 2014, p. 96 as well as cf. Asma & Gabriel, 2019, pp. 8-10.

deep learning neural networks still lack in the fields of "meaning or emotional significance of a scene, i.e., the human ability for generalization, conceptual learning, [and] selective attention"[281] which can be considered crucial for ethical decisions.

As the scope of this work is limited, it remains open how a sub-symbolic further development of the rule-based ethical foundations, that is, the potential increase in expertise with the means of learning, could be achieved. However, as the foundation of ethics is not only from the standpoint of this thesis but also from Dreyfus' point of view, considered to be rule-based and hence symbolic, it will be focused on symbolic means to elaborate a concept of ethical AI research and development. With the aim of incorporating Dreyfus' "ethics of situated involvement", this thesis will consider virtue ethics as an apt approach to normative ethics. However, the foundation of virtue ethics will be concretized in a stronger normative manner in terms of technically ensuring transparency, as Coeckelbergh highlights that Dreyfus' account of virtue ethics is "mainly descriptive and aimed at understanding the kind of knowledge involved in skilled coping, [although] it has normative implications"[282].

Besides virtue ethics, other main theories underlying AI ethics are, for example, deontological ethics and consequentialism.[283] Why might virtue ethics have been preferred by Dreyfus, and why is it regarded as adequate here? Deontological ethics focuses on universal normative principles that individuals need to comply with in order to act ethically. However, deontological principles can be considered problematic for AI ethics because the strict adherence to principles might lead to contradictions in contexts that are fundamentally different or even bear a conflict of different principles. For instance, a principle that leads to ethical action in the majority of cases could be harmful in a singular specific case. To circumvent this risk brought forth by the totality of principles, which is limiting the flexibility to adapt to different contexts in ethical actions, teleological ethics are here preferred to deontological ethics. Teleological ethics aim at an ethical goal[284] and include consequentialism as well as virtue ethics. However, consequentialism, which is only concentrated on ethical consequences, might, in turn, be also problematic in AI ethics, as it has not to be an ethics by design approach per se. The actual reason for an action in question might move into the background as only the

[281] Evers, K., Farisco, M. & Salles, A., 2022. On the Contribution of Neuroethics to the Ethics and Regulation of Artificial Intelligence. *Neuroethics, 15, 4* , p. 7.
[282] Coeckelbergh, 2019, p. 280.
[283] Cf. Russell & Norvig, 2021, p. 26.
[284] The ancient Greek notion *telos* means orientation toward an end, goal, or completion. However, Aristotle coined the term in describing that every natural being has its specific biological telos which determines its life form. Humans, for example, have their telos in well-being as a form of holistic happiness (*eudaimonia*): Herein, the individual and the social well-being are unified in the moral good. Aristotle devised virtue ethics inspired by Plato. Most modern approaches to virtue ethics are still strongly influenced by Aristotle. Cf. Grunwald, 2011, pp. 61 f; 69-75.

outcome, the consequence, needs to be regarded as ethical. For instance, an AI system that is acting on the basis of strong bias and still leads seemingly passable decisions could be investigated too late because the decisions seem to be tolerable on average when instead it should be focused on the reason for decisions.

Virtue ethics, in contrast, holistically takes into account the acting subject and asks how virtues can be cultivated that are oriented towards ethical actions. Virtues hereby are moral positions or qualities aimed at the well-being of society that govern actions and thoughts. In this regard, what is ethically examined is neither the principle underlying an action nor solely the consequences of an action but rather the acting subject within society: Virtue ethics is essentially concerned with the *how* of the morally right or good, not primarily with the *what*.[285] For instance, a virtue ethical approach to AI would intend to develop systems that themselves mirror the ethical virtues of the developers, operators, and users. The systems in question would come to ethical decisions because they are bound to the virtue ethics approach in their design and deployment. Virtue ethics for AI is here preferred because either is the discovery of universal and consistent principles a challenge for itself or the exact prediction of consequences often hardly possible.[286] Additionally, it goes hand in hand with Dreyfus' emphasis on context-sensitivity and a holistic account of the acting subject: As illustrated in chapter two, he questions the existence of universal principles underlying human decision-making as well as the delineation of consequences from a subject acting 'in-the-world'. Rather, the unlimited depth of a situation constitutes the subject's world in which it acts embodied and emotionally: "[Dreyfus'] [e]thics seems to require […] a practical wisdom which can respond intuitively and appropriately to specific situations"[287].[288]

3.1.2. Values of international AI Ethics

AI ethics is considered an interdisciplinary field of applied ethics, namely a subclass of machine ethics[289] intersecting with and influenced by values from bioethics such as justice and non-

[285] Translation of a quote by Hillerbrand and Poznic: "Eine Tugendethik befasst sich originär mit dem Wie des moralisch Richtigen oder Guten, nicht vordringlich mit dem Was". See: Hillerbrand, R. & Poznic, M., 2021. Tugendethik. In: A. Grunwald & R. Hillerbrand, Edt. *Handbuch Technikethik.* Berlin: J.B. Metzler, p. 166.
[286] Cf. Hillerbrand & Poznic, 2021, pp. 165-167.
[287] Coeckelbergh, 2019, p. 280.
[288] The comprehension of practical wisdom is central to virtue ethics: The Aristotelian notion of phrónēsis, ancient Greek for prudence, denotes the anchoring of intelligent behavior of the individual in social practices where virtues are cultivated because of moral interactions. The Aristotelian notion of intelligence here goes beyond intelligence in a mathematical-analytical understanding as it implies ethical reasoning. Cf. Hagendorff, 2022, p. 9.
Specifically, Aristotle's method of accessing truth, which could also be described as intelligence, includes: Craft (techné), knowledge (episteme), prudence (phrónēsis), wisdom (sophia) and intellect (nous). The attainment of intelligence through practice is only given in phrónēsis and techné. Cf. Kogge, 2022a, p. 22 f.
[289] Cf. Misselhorn, 2019, p. 34.

maleficence.[290] AI ethics is underpinned by the different ethical theories delineated above, such as deontological ethics and virtue ethics.[291] Hence, various ethical values have emerged in the field during the last years, which need to be drawn upon to concretize an ethical approach to AI research and development. Although values of AI ethics are flourishing, Jobin et al. suggest a delineation of values to transparency, justice, non-maleficence, responsibility, and privacy as they internationally reached a wide consensus. However, it must be emphasized that the discourse on values of AI ethics is geographically very limited, as mainly Western and wealthier states are involved in the debate. To arrive at a discourse in a global setting, it should remain open to new participating states.[292] To further elaborate a concentration on virtue ethics, Hagendorff limits and assembles Jobin et al.'s values into four 'AI virtues', namely *justice, honesty, responsibility, and care*. Transparency is hereby subsumed under the virtue of honesty, whereas non-maleficence and privacy contribute to the virtue of care:

Justice in AI research and development is depicted as a fair and hence non-discriminatory inclusion and representation of all individuals and groups, for example, in terms of preventing any form of algorithmic bias. Thus, human oversight (that means human governance of and intervention in automated decision-making) and a right to remedy automated decision-making by legal means are included in the notion of justice. *Honesty* refers to the transparent organization of AI research and development, for example, the disclosure of the funding in research institutes and companies, as well as to a transparent design of AI systems in terms of an explainable and interpretable functionality. For instance, open-source development is advocated, which stands in contrast to non-disclosure through proprietary black box systems. Developers and operators of AI systems bear a great *responsibility* to society, and they need to embrace it with high awareness and sincerity. Thus, it comes with legal obligations as it needs to be clear who is held liable and accountable for AI systems: "Diffusions of responsibility in complex technological as well as social networks can cause individuals to detach themselves from moral obligations"[293]. Finally, *care* denotes the pivotal attitude to developing AI systems to serve the society and, therefore, to dedicate them to social well-being. The virtue of care, therefore, implies the dedication to non-maleficence and safety as well as the preservation of privacy, among other central democratic values. For example, the harming of individuals and groups through the deployment of AI systems that are bearing a high risk for the well-being of humans explicitly contrasts with the virtue of care.[294]

[290] Cf. Floridi, et al., 2018, p. 696.
[291] Cf. Misselhorn, 2019, p. 48 f.
[292] Cf. Jobin, et al., 2019, p. 391.
[293] Hagendorff, 2022, p. 7.
[294] Cf. Hagendorff, 2022, pp. 4-14 as well as cf. Jobin, et al., 2019, pp. 391-395.

As already mentioned, a virtue-ethical grounding prevents the situational conflict of several ethical values since the body of virtues can be perceived as a holistic attitude in AI research and development. A virtue-ethical approach in application consequently refers to all virtues at the same time. An ideal autonomous vehicle, for example, mirrors all the virtues described by Hagendorff if it is developed responsibly: It would be just and 'honest' in design by making decisions free of bias and explaining why the decisions are made. It would additionally serve the common good ('care') by equalizing the mobility levels of all social groups, including the mobility of elderly and disabled people. In this sense, a virtue-ethical approach as a foundation to AI research and development is endorsed.

However, Hagendorff limits his concretization to a perspective on the 'humans behind the technology' and excludes an approach to technically implementing concrete virtues. This seems to be reasonable due to the issue of formalization highlighted by Dreyfus: How could a universal method suffice to ensure ethical actions in every possible context? Thus, could an AI technique ever be mature enough to illustrate the context-sensitivity necessary to recognize situational aspects that determine ethical relevance in every possible context without having recourse to human meaning and emotion? If, for example, a virtue ethics approach is firmly anchored in the theory of research and development, it could be expected to automatically strengthen the awareness of developers' ethical responsibility so that innovations are ethically designed in practice. However, this hope is already practically limited. It seems like the field of AI ethics would not have emerged as firmly as it did without current ethical challenges in the application of AI systems: The black box problem is an example where the virtues of the developer could be laudable, and still, the system is able to be socially harmful. In case the black box is due to the technological complexity of the system, the expert itself is often in the dark about the causes of actions as they are non-transparent.[295] Hence, it is argued here that *transparency* should be considered the most foundational ethical value as it is the only one that ensures insight into compliance with the others. This is in accordance with the findings of Jobin et al., which show that transparency is perceived as being of the highest priority on average.[296]

Although a virtue-ethical foundation of AI research and development should be preferred to a deontological foundation of AI ethics, a normative concretization seems still necessary with regard to transparency as a stronger normative value than the virtue of honesty: In order to prevent social harm in a binding way and to be able to verify compliance with all

[295] Cf. Marcus, 2018, p. 10 f.
[296] Cf. Jobin, et al., 2019, p. 391.

virtues, we need above all the means to sufficiently understand why ethical virtues may be violated in certain contexts.

> "Virtue ethics does not come without shortcomings. […] Ultimately, trustworthy AI will be the result of both strands, ethics as well as law. Both strands interact and inspire each other. However, especially virtue ethics with its focus on individual dispositions is perhaps less apt to inspire systemic changes or legal norms than principlism"[297].

Therefore, the fundamental value of transparency needs to go beyond a virtue-ethical interpretation since clear normative specifications and obligations should be made in order to technically grant compliance. A deontological understanding of the value of transparency within the framework of the virtue of honesty needs to be advocated in order to translate the theoretical approach into practice: Transparency must be understood as an ethical principle that, when adhered to in AI action, enables the realization of other ethical approaches. Even if it is taken as a deontological principle such as '*all AI actions shall be rendered transparent*', transparency seems to avoid possible conflicts with other values – as it seems to be located one argumentative level below other ethical values or virtues. For instance, if an automated vehicle is in a dangerous situation for its passengers, it does not make sense to determine whether it is more urgent to communicate the risk transparently or to preserve the virtue of care. Rather, it seems that by being transparent, such as communicating the risk, the virtue of care can be upheld by, for example, demanding human oversight or intervention.

The understanding in this thesis, hence, goes beyond an interpretation of transparency as the AI virtue of honesty. It is argued here that transparency requires a technical implementation: We need a concrete shift from *causes of AI actions* to *reasons for AI actions* that guarantee transparency in the application. Ensuring reasons for AI actions should be achieved by technical means: it has to be the "machine architecture [that is] reasonable"[298] in itself. The first step toward a realization of this goal is already achieved by symbolic techniques and will be illustrated in the following.

3.2. Reasons for AI Actions: approaching a technical implementation

Approaching the technical implementation is hence based on the foundation of virtue ethics: This means that Hagendorff's AI virtues must be adhered to in AI research and development in order to have virtuous AI systems as a future goal. Beyond that, however, the concrete value of transparency is conceived as more fundamental as it allows not only to reveal causes of AI

[297] Hagendorff, 2022, p. 15; 18.
[298] Benzmüller & Lomfeld, 2020a, p. 252 f.

action but also to derive reasons for AI actions – therefore, it requires the exploration of the possibilities of a concrete technical implementation: For AI systems to operate trustworthily, they need a "reasonable […] machine architecture"[299] ensured by symbolic techniques such as ethico-legal governance. As shown before, this necessity seems to contradict the nature of sub-symbolic AI, as its complexity often leads to a black box in which actions are enabled by (seemingly arbitrary) causes rather than reasons. Therefore, symbolic AI needs to be used to technically integrate reasoning architectures, layers, or modules that outweigh the opacity of sub-symbolic applications by increasing transparency within a hybrid setup. A two-sided framework of reasons suggested by Benzmüller and Lomfeld is therefore introduced to show a concrete approach to such a technical implementation. The approach is "[a]llowing various kinds of reasons, […] advanc[ing] normative pluralism […] [in possibly] integrat[ing] different (machine-)ethical traditions: deontological, consequentialist and virtue ethics"[300]. It is therefore assumed that the grounding in virtue ethics defended above, including a stronger normative or even deontological interpretation of the value of transparency, can serve as a sound ethical foundation.

The framework of value-based reasons can be thought of as a hermeneutic feedback loop where the system is "able to give and take reasons for their decisions to act"[301]: Firstly, AI systems need a priori reasons for taking actions – that means that AI actions should only be allowed if they are justified by reasons consistent with the virtues or values of AI ethics – and, secondly, actions generated by AI systems should be justified a posteriori by the communication of value-based reasons in order to regulate and adapt future actions.[302] The framework might be enabled through the integration of an upper level of ethico-legal governance through symbolic techniques, backing the sub-symbolic layer of the system as a second layer (illustrated in Figure 5).

As indicated before, ethico-legal governance can be considered a core element of a technical approach to AI ethics, especially in increasing transparency. To that end, the symbolic layer or module of an AI system facilitates machine reasoning on the basis of formally defined ethical (or legal) values that are represented in a hierarchical structure of norms, such as in ethico-legal ontologies.[303] This can, for example, be achieved by the application of interactive and automated theorem provers in an architectural framework of higher-order logic (HOL), such as in Benzmüller et al.'s normative reasoning framework LogiKEy. Hereby, several

[299] Benzmüller & Lomfeld, 2020a, p. 254.
[300] Benzmüller & Lomfeld, 2020a, p. 254.
[301] Benzmüller & Lomfeld, 2020a, p. 251.
[302] Cf. Benzmüller & Lomfeld, 2020a, p. 253 as well as cf. Misselhorn, 2019, p. 41 f.
[303] Cf. Benzmüller & Lomfeld, 2020a, pp. 251-253.

deontic logics and their combinations are used to represent normative modalities of AI action, such as the permission, prohibition, or obligation of certain actions. For this purpose, they are semantically embedded in higher-order logic that is strongly expressive. The interactive proof assistant Isabelle in the environment of HOL (therefore, it is mostly known as Isabelle/HOL), which is usually often used for software verification, is in this context used to prove compliance with formal normative requirements of AI systems. To achieve the proof or disproof (that is, the formal verification) of action regarding the underlying normative requirements, that is, the respective formalized ethico-legal theory, various tools such as automated theorem provers can be consulted within Isabelle/HOL. It hence enables the critical assessment of the decision-making process against set ethical standards *before* the AI system takes action. As a consequence, the sub-symbolic system can be governed by the integrated symbolic module: By the twofold commitment to rational reasons for AI actions, not only by verification of reasons for AI action (a priori reasons) but also by communication of reasons for AI action (a posteriori reasons), actions are not only controlled but also made transparent. Such a framework is hence suggested as a solution for the black box problem of sub-symbolic systems as it enables actions that are ethically governed on the basis of reasons.[304]

In view of the virtue-ethical foundation previously identified, however, it must be critically examined whether virtues can be sufficiently abstracted as ones that can apply independently of context and that can therefore be considered general enough regarding deontic modalities.[305] While virtue ethics is useful in developing a holistic approach to AI research and development, it is less likely to allow conclusions to be drawn about concrete norms. For example, if one were to formalize the AI virtue of care, it would only obtain semantic validity if the context is known in advance, which, as Dreyfus has already pointed out, still seems contradictory because contexts are always individually situated. The virtue of care could indeed be translated to deontic statements, for instance, in the sense of 'it is obligatory to act benevolently towards individuals and groups', 'an action is permissible if it serves the social good', or 'an action is prohibited if it is maleficent'. However, in diverse situations, concepts such as the social good can mean something very different: For example, if a child does not normally go to school, this is unlikely to serve the social good. However, if a child, for instance, does not go to school because of a climate demonstration on Friday but rather makes its first experiences as a political human being and consequently engages with the idea of social or ecological responsibility, it can probably be said that this serves the social good. Overall, virtue

[304] Cf. Benzmüller, et al., 2020b, pp. 1-40.
[305] Cf. Benzmüller, et al., 2020b, p. 33 f.

ethics, therefore, seems to lie outside the realm of concrete formalization by ethico-legal domain theories. However, if we look at the ethical value of transparency, which is interpreted from a more normative, deontological perspective here, a different view emerges. Not only are there very few contexts in which transparency could cause harm, but transparency is also constitutive for the fulfillment of other ethical values and virtues. Moreover, transparency can be implemented concretely through Benzmüller and Lomfeld's approach of the two-sided framework of reasons.

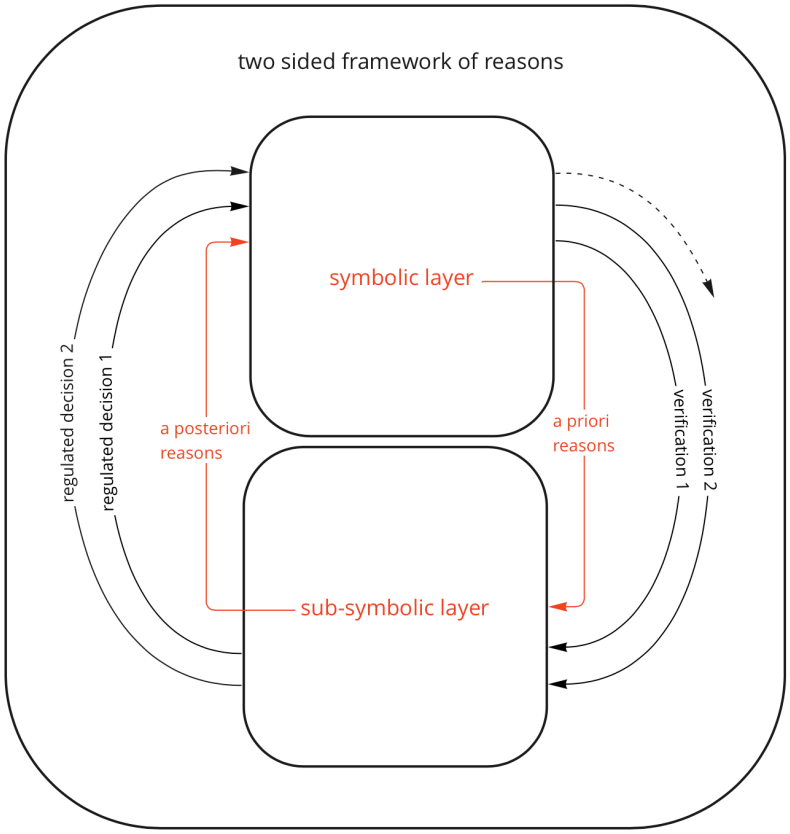

Figure 5: Two sided framework of reasons by Benzmüller and Lomfeld (own illustration)

Figure 5 illustrates the approach in a simplified setup of a hybrid system combining sub-symbolic techniques with symbolic techniques that serve the purpose of ethico-legal governance. The hermeneutic interaction of both layers is illustrated, whereas all decisions on the sub-symbolic layer must hence comply with reasons defined a priori on the symbolic layer. Thus, the sub-symbolic layer would transmit the reasons for the decisions a posteriori to the symbolic layer. Subsequently, a symbolic verification of the communicated a posteriori reasons enables the regulation of subsequent decisions on the basis of logics such as deontic logic or even a logic combination. After the decision-making process has been governed by symbolic

techniques at the symbolic level by the verification and regulation of decisions on the grounds of value-based reasons, an AI action can be conducted. AI action undergoing the ethico-legally governing of decision-making shows a higher degree of transparency because it is based on reasons that comply with ethical values or virtues. Within the hermeneutic structure connecting the sub-symbolic and symbolic layer, learning might be possible in the future as the neural network at the sub-symbolic layer acquires behavior that adapts to given reasons and increasingly aligns outcomes with required reasons.[306]

As the hybrid setup seems quite abstract here, an example might help. One could consider the example of the control system of an automated vehicle in the process of overtaking. In order to execute the planned action on the basis of a reasonable decision-making process, both sub-symbolic and symbolic levels are necessary: Sensory environmental data are decisive for scene understanding and hence the sub-symbolic initiation of an overtaking maneuver. Previously symbolically defined ethical norms would be assessed initially, for example, the operation must not endanger persons on the road under any circumstances. If this norm can be met, the decision is made to overtake. The reasons for the decision, for example, that acceleration is sensible and that the required distance to other participants in the road traffic will be maintained in the overtaking process, is in turn communicated to the symbolic layer. Through the subsequent symbolic verification loop, action can then be realized despite the continuous scene motion and the vehicle overtakes.

In summary, the logic-based techniques of the symbolic approaches, in contrast to the sub-symbolic approaches, are suitable for a "reasonable […] machine architecture" and enable ethico-legal governance. Their integration into future hybrid systems should therefore be pursued with the aim of standardization.

3.3. Trustworthy AI: approaching a legal framework

Since AI applications are already being used very ambivalently today, it has become apparent that they not only have the potential to increase societal well-being but can also be used for harmful purposes. Therefore, AI technologies are considered dual-use technologies.[307] Hence, the belief here is that they should be "trustworthy"[308], which is not only based on ethical values but also framed by law to legally ensure their beneficial use. Because AI comes with many ethical challenges, these must be taken into consideration in the context of human rights but

[306] Cf. Benzmüller & Lomfeld, 2020a, pp. 252-256.
[307] Cf. Lorenz, 2021, p. 20.
[308] European Commission, 2021a.

also of regional and national legal frameworks: The areas of application of AI technologies must be embedded in concrete legislation in the future – to prevent any harmful use. Approaching a legal framework is, for instance, already being attempted at the EU level as the European Commission presented a legislative proposal for regulating the use of AI applications in 2021: the *AI Act*. It defines specific legal requirements for AI technologies in use that aim at achieving 'trustworthy AI'.[309,310] The AI Act is the first legal framework at a regional level that attempts to address ethical challenges through legislation. It has stimulated discussions internationally to make similar normative requirements legally binding.[311]

The risk posed by AI systems is assessed within the AI Act at three stages: Firstly, AI technologies that entail an "unacceptable high risk" are banned from the outset, for instance, systems aiming at social scoring. Secondly, applications that entail a "high risk" are regulatory bound to legal processes of adherence to various standards, including transparency, obligatory documentation and disclosure, specific communications to the user, and other requirements. Thirdly, AI technologies with no noticeable risk remain mainly unregulated. AI systems and components of AI systems listed in Annexes II and III are identified by the EU as systems bearing a high risk.[312] These include but are not limited to systems applied for the purpose of transportation, extraction and exploitation of various natural resources, of biometric identification, medical procedures, as well as procedures in the areas of employment, private and public services, law enforcement, migration, and legal administration.[313] It is anticipated here that the AI Act provides a solid basis for legally addressing current and future risks originating from the deployment of AI systems: It is promising for governing research, development, and deployment on the legal level.

[309] European Commission, 2021a. The term 'trustworthy AI' originates from the Assessment List for Trustworthy AI (ALTAI), which was elaborated by an expert group and published in 2020. It presupposes ethical principles in development that must be adhered to in order to guarantee ethical AI technologies and thus their ethical deployment. However, since we have, firstly – based on Dreyfus' phenomenological standpoints – chosen a foundation of AI virtues rather than principles and, secondly, taken into account the possibility of the technical implementation of the concrete normative value of transparency to ensure further adherence to ethics in development, we will not discuss the underlying principles in more detail here. Rather we will take a closer look at the risk-based approach to legal requirements for AI technologies that has emerged from these considerations. For further information, See: HLEG on AI & European Commission, 2020. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self Assessment.* [Online].

[310] Caution is advised here when it comes to related terms: A distinction must be made between 'trustworthy AI' and other terms intended to increase trust in AI systems: While 'trustworthy AI' refers to the concrete legal understanding in the European Commission's proposal, 'trusted AI' as discussed within IEEE (Zhang, et al., 2021, p. 55313), for instance, refers to concrete techniques or technical requirements for achieving trusted AI systems. It is hence located on the technical rather than the legal level.

[311] Future of Life Institute, 2022. *The AI Act.* [Online].

[312] Cf. European Commission, 2021a, pp. 38-58.

[313] Cf. European Commission, 2021a, *Annexes*, pp. 2-5.

What should still be focused on in the future, though, is the prevention of the protection of black box systems by trade secrets in high-risk systems. Although the AI Act empowers authorities to demand transparency in confidence,[314] it would still be desirable to discuss problematic systems as widely as possible. Because while open-source software or even the publications of patents give insight into the functioning of AI technologies, with only the possibility of a black box due to model complexity remaining, the functioning of AI systems protected by trade secrets is not even accessible to a broader evaluation. Hence, these systems can contain a double black box model – referring to secrecy plus complexity. For this reason, proprietary, non-transparent AI systems are increasingly questioned from a societal perspective:[315] While open-source solutions are usually supposed to foster scientific collaboration as well as ethical revision, and intellectual property rights are generally intended to augment social welfare, trade secrets can limit both. To prevent harm to society, both types of black boxes – technical and proprietary ones – used in the context of high-risk areas, for instance, in autonomous driving, should be addressed accordingly by means of ethico-legal governance and the obligation to comply with strict disclosure standards. The danger posed by black boxes in systems that bear a high risk when applied needs to be tackled by competent authorities – by demanding a high degree of transparency through means such as verification, documentation, and communication.

In the following, autonomous driving is discussed with respect to the three stages of ethical AI research and development proposed above – an ethical basis and a technical implementation within the proposed framework of the Commission's AI Act. What problems arise here from an ethical perspective? How can they be addressed on a technical level? Thus, how does the example behave from the legal perspective of trustworthy AI? The need for symbolic AI in the specific case of autonomous driving is examined in light of these current and future ethical challenges.

3.4. Looking closer at autonomous driving as an example

To discuss the thesis' theoretical stances application to a real-world example, we will take a provisional and, therefore, humble look at autonomous driving as a use case. It is primarily intended to serve the purpose of illustration rather than generalization. Why is it considered as an example? First, the deployment of autonomous cars is, in practice, an ethical one: The

[314] Cf. European Commission, 2021a, pp. 11; 81.
[315] Cf. Moore, T. R., 2017. *Trade Secrets and Algorithms as Barriers to Social Justice.* [Online]. Thus, cf. Pasquale, F., 2017. *Secret Algorithms Threaten the Rule of Law.* [Online].

motivation for autonomous cars is sound from an ethical perspective as it aims to improve the mobility of disabled and elderly people. In addition, autonomous driving could reduce the number of traffic accidents: While human drivers make mistakes by losing concentration when answering the phone or driving home tired from work, an ideal autonomous vehicle would always be focused, alert, and aware of the situation. The European Commission assumes that such human error is responsible for 94 percent of road accidents.[316]

Furthermore, the often opposing approaches of AI reasoning and AI learning are conflated in the field of autonomous driving as logic-based reasoning, such as planning, is essential for research and development, and the vehicles largely rely on sub-symbolic methods. Hence, autonomous vehicles are hybrid systems already – relying on both symbolic and sub-symbolic techniques.[317] Symbolic techniques are, for example, used for the purposes of error detection, motion control, and the formal verification of systems, amongst others.[318] However, there seems to be still room for improvement in using symbolic AI in a modularized, goal-oriented way to safeguard compliance with ethical values, such as transparency.

Although here, it is referred to 'autonomous driving' in general, this does not have to be understood literally but rather as a denotation of the research field of aiming at high degrees of automation in vehicles used for transportation.[319] From an ethical perspective, the practical

---

[316] Cf. European Commission, 2018. On the road to automated mobility: An EU strategy for mobility of the future. *COM(2018) 283 final*, p. 1.

[317] Cf. Veres, S., Molnar, L., Lincoln, N. & Morice, C., 2011. Autonomous Vehicle Control Systems. A Review of Decision Making. *Journal of Systems and Control Engineering 225, 2*, p. 158. Thus, cf. Giancola, M., Bringsjord, S., Govindarajulu, N. & Licato, J., 2020. Adjudication of Symbolic & Connectionist Arguments in Autonomous-Driving AI. *EPiC Series in Computing, 72*, pp. 29–32.

[318] Cf. Fraunhofer IEM, 2022. *Hybride KI-Methoden für das Testen von elektrischen und elektronischen Systemen.* [Online]. Thus, cf. Rizaldi, A., Immler, F., Schürmann, B. & Althoff, M., 2018. A Formally Verified Motion Planner for Autonomous Vehicles. In: S. Lahiri & C. Wang, Edt. *Automated Technology for Verification and Analysis. 16th International Symposium Proceedings.* Cham: Springer, p. 75.

[319] In most of the literature, the levels of autonomous driving are categorized according to the criteria defined by the Society of Automotive Engineers (SAE). Hereby, six levels – Level 0 to Level 5 five – denote ascendingly higher stages of automation. Level 0 consequently starts without any automation, that is, all driving is done by humans themselves. While stage 1 denotes the presence of a driver assistance system for limited motion control, Level 2 already involves partial automation of driving – the vehicle is able to maintain its lane or brake in several situations. Level 3 comprises conditional driving automation, where the automated vehicle is already able to overtake, brake, and accelerate without human intervention. As long as the human driver is not explicitly asked to intervene, paying attention to the driving is not necessary. Level 4 already denotes high driving automation in which all driving operations are taken over by the system. The degree of autonomy is already sufficiently high to allow the human driver to sleep while driving. Lastly, Level 5 designates full driving automation, which means full autonomy of the vehicle in all possible situations: Not only does the system no longer require human intervention, but such intervention is also no longer possible – since the human driver has now become a passenger entirely. In most cases, a level of automation of Level 2 is achieved nowadays, whereas in rare cases, individual components of level 3 are included. However in Germany, for instance, only partially automated vehicles (Level 2) can be licensed. A significant amount of research is needed to ultimately bridge the gap to Level 3. Although autonomous vehicles are hence still more vision than reality, the term 'autonomous driving' is often used for research and development in the whole area and summarizes the developments in this field better than individual distinctions based on the level of automation.
Cf. European Commission, 2021b, p. 14 as well as cf. Fraunhofer IKS, 2022. *Autonomes Fahren.* [Online].

implementation of autonomous driving as a vision for society faces many challenges. In the first driving lesson, we are explained how important responsibility is for driving out on the roads. We often experience directly how risky it can be not to drive carefully and how rapidly a car accident can happen. If the AI virtues of justice, honesty, responsibility, and care introduced above are not realized accordingly in all forms of driving and, above all, if the normative value of transparency is neglected in autonomous driving, there is a risk of individuals being physically harmed when vehicles are tested or used in public accordingly. The use of partially automated vehicles has already led to severe accidents in several cases. In the US, for instance, the frequency of accidents with causal involvement of partially automated systems was investigated – it turned out that 400 such accidents occurred in a ten-month period, among which few were fatal.[320] Moreover, at the moment, compliance with the ethical basis must still be questioned insofar as the black box problem also arises in the context of autonomous vehicles.[321]

To realize defined ethical values on the level of technical implementation, it is useful to look at suggestions that target concrete solutions for ethics by design. An approach to governance on the technical level of autonomous vehicles such as ethico-legal governance might in the future be regulated by standards and policies such as the AI Act. Specifically, we will explore how a technical implementation of the normative value of transparency, which is understood as constitutive for the adherence to further ethical values or AI virtues, could look for autonomous vehicles in the context of trustworthy AI. The use of symbolic techniques for technically increasing transparency in the design of autonomous vehicles, in contrast to black box models, is hence examined briefly. It could go along with regulating AI action by symbolic verification of sub-symbolic decision-making processes as suggested by Benzmüller and Lomfeld's two sided framework of reasons as well as Benzmüller et al.'s symbolic approach to the verification of AI action by the use of automated theorem provers.

Autonomous vehicles are a subtype of "autonomous ground robots"[322]. As such, they are subject to many urgent ethical questions, such as the controllability of integrated AI systems in order to ensure the safety of participants in road traffic. As already mentioned, however, in order to verify safety, sub-symbolic techniques are not sufficient as they often constitute a black box in autonomous vehicles consisting of "hard-to-interpret models that are difficult to debug

[320] Cf. Krisher, T., 2022. *US report: Nearly 400 crashes of automated tech vehicles.* [Online].
[321] Cf. Kriebitz, A., Max, R. & Lütge, C., 2022. The German Act on Autonomous Driving: Why Ethics Still Matters. *Philosophy & Technology, 35*, p. 3.
[322] Mitsch, S., Ghorbal, K., Vogelbacher, D. & Platzer, A., 2017. Formal Verification of Obstacle Avoidance and Navigation of Ground Robots. *International Journal of Robotics, 36, 12*, p. 1.

and challenging to maintain"[323]. Therefore, the integration of symbolic techniques in the form of a layer or standard-suitable module for autonomous vehicles could serve the purpose of formally verifying decision-making processes such as navigation and control. Symbolic verification is necessary because the system is tested to a broad but limited amount of scenarios, whereas it is confronted with further and possibly unforeseen situations in application to the real world.[324] AI action based on sub-symbolic techniques can be verified by theorem provers, including satisfiability theories for various logics.[325] Mitsch et al. as well as Rizaldi et al.[326] suggest the deployment of automated theorem provers for the verification of actions generated by autonomous systems such as autonomous vehicles. Hereby it can be formally proven that all actions carried out are correct and oriented toward specific goals on the basis of beforehand defined logical axioms. Rizaldi et al.'s approach relies on the proof assistant Isabelle/HOL that was depicted in chapter 3.2. Theorem provers are in the approach applied to prove the correctness of motion planning. Environmental that is contextual, changes are modeled through temporal logics such as Linear Temporal Logic (LTL). However, it is to be stated that the authors perceive their framework as still preliminary and advise future work in the field as the application of Isabelle is promising in the context of autonomous vehicles.[327]

In addition, on a legal level, autonomous vehicles should be included in specific frameworks addressing the risks posed by the incorporation of AI systems. Because autonomous vehicles are strongly relying on various complex AI systems in combination for achieving capabilities of "localization, scene understanding, planning, control, and user interaction"[328], the European Commission suggested including autonomous vehicles in the notion of trustworthy AI as defined by the proposed AI Act legislation.[329] As such, autonomous vehicles are considered high-risk applications as "their adoption involves addressing significant technical, political and societal challenges"[330]. Hence, "all the ethical principles, key requirements and assessment criteria of a trustworthy AI […] must[.] necessarily be applied to the specific context of A[utonomous] V[ehicle]s"[331]. Thus, it is an aim of the European Commission to internationally promote technical AI standardization as well as policies that

[323] Marcus & Davis, 2019, p. 183.
[324] Cf. Marcus & Davis, 2019, pp. 183-193 as well as cf. Mitsch, et al., 2017, pp. 2-4. Caution is advised when Mitsch et al. refer to hybrid dynamical system models as they have to be distinguished from our understanding of hybrid systems here, as hybrid dynamical systems denote the capability of combining discrete states and continuous motion in the behavior of dynamical system.
[325] Cf. DIN & DKE, 2020, p. 88.
[326] Cf. Rizaldi, et al., 2018, p. 76.
[327] Cf. Rizaldi, et al., 2018, pp. 75-88.
[328] European Commission, 2021b, p. 2.
[329] Cf. European Commission, 2021a.
[330] European Commission, 2021b, p. 2.
[331] European Commission, 2021b, p. 5.

shall ensure trustworthy autonomous vehicles.[332] All in all, the objective of autonomous driving lends itself to an ethical perspective. Therefore, the need for internal governance, such as ethico-legal governance, and external governance through standardization and legislation should be further tackled to ensure ethical development and deployment in a holistic manner by addressing the dual-use nature of AI ethically, technically, and legally.

4. Outlook: Standardized ethico-legal governance in hybrid systems

The findings of the preceding chapters are condensed into two positions presented as an outlook. Firstly, the concept of a standardized module of symbolic AI that is compatible with sub-symbolic AI technologies and guarantees transparency is proposed for the case of a black box system – as it can be considered ethically challenging in the context of high-risk systems. Secondly, hybrid systems are generally advocated, as symbolic AI is considered necessary for ethico-legal governance.

4.1. Proposal of the concept of a standardized ethico-legal governance

At large, legislation according to the European AI Act seems to be a sensible proceeding on the legal level. However, since laws are very general and have to be decided in different national or regional legal frameworks for broad validity, technical standardization offers a possibility for ensuring compliance with ethical values in an impactful way. Therefore, a future concept of modularized ethico-legal governance is proposed for the black box problem of sub-symbolic systems. With the integration of a standardized symbolic AI module, transparency can be increased through the verification of the compliance with value-based reasons for AI actions and the technical communication of decision-making processes, as the approach is rule-based. For instance, automated theorem provers can be applied to sub-symbolic AI systems in the form of "ethico-legal governors" [333] as suggested by Benzmüller and Lomfeld. Applied modularly, symbolic governance can contribute reasoning processes in the form of ethical and legal normative theories to the sub-symbolic AI system in question. Therefore, as soon as sub-symbolic AI is used as a proprietary or complex black-box model in the context of high-risk AI systems, concrete technical solutions should be explored, adopted, and standardized for international validity. In the case of black-box AI systems in high-risk areas of operation, such

[332] Cf. European Commission, 2021b, pp. 2-5.
[333] Benzmüller & Lomfeld, 2020a.

a standard could even be thought of as legally binding. An example of such an AI standard is the IEEE Standard 7007, that "contains a set of ontologies that represents norms and ethical principles […], data privacy and protection […], transparency and accountability, and ethical violation management"[334]. It provides the opportunity to integrate different ethics, for example, values of virtue ethics or deontological ethics formalized as norms. The standard is expected to make waves in approaching ethical AI research and development on the technical level, hence through ethics by design.[335]

The concept of a standardized module of symbolic AI is thus assumed to be consistent with Russell's proposal to go beyond regulatory disclosure standards and create standardized designs for the verification of AI systems and also to impart the findings in education and research.[336] Although this would certainly take some time to gain acceptance in Silicon Valley, according to Russell, such transparency and safety standards are comparable to the standards required in the development of pharmaceuticals. The concept of an internationally standardized module of ethico-legal governance, for instance, in the form of a theorem prover as a tool of symbolic reasoning, is therefore advocated. This means it would inevitably extend the sub-symbolic AI system to a hybrid AI system which is argued for beyond.

4.2. Advocacy of hybrid systems

As Dreyfus pointed out, sub-symbolic AI is indeed promising to simulate or at least imitate human perception and perception-based cognition. However, Dreyfus largely excluded an ethical dimension from his assessment of symbolic AI. As shown in the preceding chapters of the thesis, hybrid systems that incorporate both sub-symbolic and symbolic approaches can be considered sensible for performance reasons as well as ethical reasons.

> "Despite the diversity that exists in the research on hybrid […] models, there is a clear unifying theme: […] The various methods, models, and architectures proposed manifest the common belief that connectionist [i.e., sub-symbolic] and symbolic methods can be usefully integrated, and that such integration may lead to advances in the understanding of cognition and intelligence".[337]

Given the ethical challenges associated with sub-symbolic AI systems, however, sub-symbolic approaches should aim for broad compatibility with symbolic approaches to ensure transparency and further adherence to ethical values. Hence, a general endorsement of such hybrid systems appears reasonable from an ethical perspective: The integration of symbolic

[334] IEEE, 2021, p. 121.
[335] IEEE, 2021, pp. 121-124.
[336] Cf. Russell, 2020, p. 252.
[337] Sun, 2014, p. 123.

techniques, layers, or even standardized modules, which are central to the compliance of AI systems with ethical values, should be given significant consideration in all approaches to AI. This is also pointedly outlined by Marcus and Davis:

> "The real risk is not superintelligence, it is idiots savants with power, such as autonomous weapons could target people, with no values to constrain them, or AI-driven newsfeeds, lacking superintelligence, prioritize short-term sales without evaluating their impact on long-term values"[338].

According to them, solely the solution of hybrid systems as the set goal of AI research and development, which are by means of symbolic techniques "equipped with common sense, […] and powerful tools for reasoning"[339] can open a "way out of this mess"[340]. Because even if sub-symbolic AI, such as deep learning, seems indispensable nowadays – only symbolic AI is rule-based and therefore able to ensure reasons rather than causes for AI actions.

Conclusion

Dreyfus' phenomenological lens on AI, his elaboration of situated context, and emphasis on the emotional roots of human cognition have opposed and unmasked the computationalist assumptions realized in GOFAI. His findings have already influenced the field of AI and should be further considered in practice to not perpetuate a potentially misleading conception of intelligence. As predicted by him, sub-symbolic AI has shown promising successes in this ongoing AI summer. Nevertheless, symbolic AI was championed in the thesis as it should be integrated into the development of hybrid systems to enable reasoning processes that counterbalance sub-symbolic weaknesses such as the black box problem. Beyond, it was made apparent that Dreyfus' critique needs to be interpretively expanded from today's perspective on AI technologies: It needs to include the ethical dimension of "what computers should not do". It was argued that AI ethics in practice must refer to an ethical, technical and legal level in order to comprehensively target ethically developed and deployed AI systems that are transparent instead of opaque. In the future, the approach of hybrid AI should be broadly adopted, which enables an embedding of ethico-legal governance methods within the framework of AI standards and policies. Sub-symbolic and symbolic AI methods should be brought together in a hermeneutic design in order to develop AI technologies responsibly and ethically far-sighted.

[338] Marcus & Davis, 2019, p. 199.
[339] Marcus & Davis, 2019, p. 199.
[340] Marcus & Davis, 2019, p. 199.

Limitations

Finally, the limitations and weaknesses of the work are briefly pointed out to enable findings to be classified accordingly and to draw attention to possible areas of future work as an incentive for further research. The following aspects were excluded due to the limited scope of this thesis and the yet open texture of the topic discussed.

Firstly, the widespread implementation of symbolic techniques might be challenging from an economic and, consequently, an ecological perspective. One might contest that symbolic verification procedures might require an enormous amount of computing power in complex scenarios such as autonomous driving. The 'hunger' for compute in the verification procedures of neural networks in autonomous driving might therefore be regarded as a disadvantage of the approach.[341] However, the objective of verified autonomous vehicles was prioritized over the problem of computing power here, as it was argued that an ethical development of autonomous systems requires, above all, transparency for further adherence to ethical values. Future research should explore the possibilities of developing low-power ways of verification.

With regard to autonomous driving, the example given is fairly narrow, as I am, openly speaking, not an expert in the field. However, it was my aim to show a concrete field of application of the previously discussed findings on all three levels (the ethical, technical, and legal level) that is not only very relevant but also likely contributing to the social good from today's perspective. Thus, often in the discussion of ethical risks of autonomous vehicles, the moral machine dilemma is discussed as brought up by the moral machine experiment.[342] However, it was not addressed here because the scope of this thesis is limited, and the experiment itself seems ethically questionable. While it is unlikely that such scenarios actually occur in road traffic, in which the vehicle is exclusively confronted with the binary constellation of harming two different parties, the public is being sensitized to it as if it were a common scenario in reality.[343] Overall, it seems more important to bring other issues into the focus of public awareness, such as the need for standards incorporating ethical values through symbolic techniques.

In the design of hybrid systems, according to Sun,[344] it is also necessary to methodologically explore which characteristics can be generalized in order to adopt either

[341] Cf. DIN & DKE, 2020, p. 88.
[342] Bonnefon, J.-F.et al., 2018. The Moral Machine experiment. *Nature, 563*, S. 59–64.
[343] Cf. also: European Commission, 2021b, p. 47.
[344] Cf. Sun, 2014, p. 122 f.

symbolic or sub-symbolic approaches for the architecture of the system. Overall, it should be studied how the broadest possible compatibility of the approaches can be accomplished: It seems necessary to turn away from specific use cases and instead turn towards the attempt to draft unified concepts that can fill the gaps left by the heterogeneity of hybrid systems in the future. Therefore, interfaces between symbolic and sub-symbolic AI should be investigated in depth.

Buckner[345] also advocates a thorough distinction to be made between terminologies of symbolic and sub-symbolic AI, but more specifically, reasoning and learning techniques. Often, terms are not precisely delineated in discussions and literature, making results more difficult to achieve and leading to conceptual misunderstandings. Buckner, therefore, advises initiating a differentiation on the philosophical level of discourse, which sets out with the distinction based on schools of thought, as Dreyfus did, for instance, with his identification of the assumptions held by computationalism – based on rationalist axioms and realized in GOFAI.

[345] Cf. Buckner, 2019, p. 12.

References

Asma, S. & Gabriel, R., 2019. *The Emotional Mind. The Affective Roots of Culture and Cognition.* Cambridge MA: Harvard University Press.

Bansal, K. et al., 2019. HOList: An Environment for Machine Learning of Higher-Order Theorem Proving. *Proceedings of the 36th International Conference on Machine Learning*, pp. 454-463.

Being in the World. 2010. [Film] Regie: Tao Ruspoli. US: Canavesio, Giancarlo; Redlich, Christopher.

Benzmüller, C. & Lomfeld, B., 2020a. Reasonable Machines: A Research Manifesto. In: U. Schmid, F. Klügl & D. Wolter, Edt. *Advances in Artificial Intelligence, 43rd German Conference on AI.* Berlin: Springer, pp. 251-258.

Benzmüller, C., Parent, X. & van der Torre, L., 2020b. Designing Normative Theories for Ethical and Legal Reasoning: LogiKEy Framework, Methodology, and Tool Support. *Arxiv*, pp. 1-50.

Boden, M. A., 2014. GOFAI. In: K. Frankish & W. M. Ramsey, Edt. *The Cambridge Handbook of Artificial Intelligence.* Cambridge: Cambridge University Press, pp. 89-107.

Bolander, T., 2019. What do we loose when machines take the decisions?. *Journal of Management and Governance, 23,* pp. 849–867.

Bonnefon, J.-F.et al., 2018. The Moral Machine experiment. *Nature, 563*, pp. 59–64.

Bosch, 2021. *Introduction to knowledge-infused learning for autonomous driving.* [Online] Available at: https://www.bosch.com/stories/knowledge-infused-learning-for-autonomous-driving/

Buckner, C., 2019. Deep learning: A philosophical introduction. *Philosophy Compass, 14,* pp. 1-19.

Butterfield, A. & Szymanski, J., 2018. *A Dictionary of Electronics and Electrical Engineering ,* Oxford: Oxford University Press.

Casacuberta, D. & Guersenzvaig, A., 2019. Using Dreyfus' legacy to understand justice in algorithm-based processes. *AI & SOCIETY, 34, p*p. 313–319.

Ciesla, R., 2021. *Programming Basics.* Helsinki: Apress.

Coeckelbergh, M., 2019. Skillful coping with and through technologies. *AI & SOCIETY, 34*, pp. 269–287.

Confalonieri, R., Coba, L., Wagner, B. & Besold, T., 2020. A historical perspective of explainable Artificial Intelligence. *WIREs Data Mining and Knowledge Discovery, 11,* pp. 1-21.

Deutsches Institut für Normung (DIN) & Deutsche Kommission Elektrotechnik Elektronik Informationstechnik (DKE), 2020. *German Standardization Roadmap on Artificial Intelligence.* [Online] Available at: https://www.din.de/resource/blob/772610/e96c34dd6b12900ea75b460538805349/normungsroadmap-en-data.pdf

Domingos, P., 2017. *The Master Algorithm. How the Quest for the ultimate Learning Machine will remake our World.* London: Penguin Books.

Dreyfus, H., 1974. Artificial Intelligence. *The Annals of the American Academy of Political and Social Science, 412*, pp. 21-33.

Dreyfus, H., 1991. *Being-in-the-world: a commentary on Heidegger's Being and time, Division I.* Cambridge MA: MIT Press.

Dreyfus, H., 1992. *What Computers Still Can't Do. A Critique of Artificial Reason.* Cambridge: The MIT Press.

Dreyfus, H., 1996. *The Current Relevance of Merleau-Ponty's Phenomenology of Embodiment, Archive: The Electronic Journal of Analytic Philosophy, 4, 1996.* [Online] Available at: https://ejap.louisiana.edu/EJAP/1996.spring/dreyfus.1996.spring.html

Dreyfus, H., 2002. Intelligence without representation – Merleau-Ponty's critique of mental representation. The relevance of phenomenology to scientific explanation. *Phenomenology and the Cognitive Sciences, 1*, pp. 367-383.

Dreyfus, H., 2007. The Return of the Myth of the Mental. *Inquiry, 50, 4*, pp. 352-365.

Dreyfus, H. & Dreyfus, S., 1990. *Sustaining Non-Rationalized Practices: Body-Mind, Power, and Situational Ethics. An Interview with Hubert and Stuart Dreyfus* [Interview] 1990.

Dreyfus, H. & Dreyfus, S., 2004. The Ethical Implications of the Five-Stage Skill-Acquisition Model. *Bulletin of Science, Technology & Society, 24, 3*, pp. 251-264.
Ekmekci, P. & Arda, B., 2020. Bioethical Inquiries About Artificial Intelligence. In: P. Ekmekci & B. Arda, Edt. *Artificial Intelligence and Bioethics.* Cham: Springer Nature Switzerland, pp. 41-78.
Ertel, W., 2017. *Introduction to Artificial Intelligence.* Berlin: Springer.
European Commission, 2018. On the road to automated mobility: An EU strategy for mobility of the future. *COM(2018) 283 final*, pp. 1-17.
European Commission, 2021a. Regulation of the European Parliament and of the Council. Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. *COM(2021) 206 final*, pp. 1-107.
European Commission, 2021b. Trustworthy Autonomous Vehicles. Assessment criteria for trustworthy AI in the autonomous driving domain. *JRC Science for Policy Report*, pp. 1-72.
Evers, K., Farisco, M. & Salles, A., 2022. On the Contribution of Neuroethics to the Ethics and Regulation of Artificial Intelligence. *Neuroethics, 15, 4* , pp. 1-12.
Floridi, L. et al., 2018. AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines, 28*, pp. 689-707.
Fraunhofer IEM, 2022. *Hybride KI-Methoden für das Testen von elektrischen und elektronischen Systemen.* [Online] Available at: https://www.iem.fraunhofer.de/de/referenzen/industrieprojekte/hybride-ki-methoden.html
Fraunhofer IKS, 2022. *Autonomes Fahren.* [Online] Available at: https://www.iks.fraunhofer.de/de/themen/autonomes-fahren.html
Future of Life Institute, 2022. *The AI Act.* [Online] Available at: https://artificialintelligenceact.eu
Giancola, M., Bringsjord, S., Govindarajulu, N. & Licato, J., 2020. Adjudication of Symbolic & Connectionist Arguments in Autonomous-Driving AI. *EPiC Series in Computing, 72*, pp. 28–33.
Grunwald, A., 2011. Technikethik. In: M. Düwell, C. Hübenthal & M. Werner, Edt. *Handbuch Ethik.* Stuttgart / Weimar: J.B. Metzler, pp. 283-287.
Hagendorff, T., 2022. A Virtue-Based Framework to Support Putting AI Ethics into Practice. *Philosophy & Technology, 35*, pp. 1-24.
Hamilton, I., 2018. *Why it's totally unsurprising that Amazon's recruitment AI was biased against women.* [Online] Available at: https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10?r=US&IR=T
Haugeland, J., 1989. *Artificial Intelligence. The very Idea.* Cambridge MA: MIT Press.
Heidegger, M., 1993. *Sein und Zeit.* 17. Edition. Tübingen: Max Niemeyer Verlag.
Heil, R., 2021. Künstliche Intelligenz/Maschinelles Lernen. In: A. Grunwald & R. Hillerbrand, Edt.. *Handbuch Technikethik.* Berlin: J.B. Metzler, pp. 424-428.
Hillerbrand, R. & Poznic, M., 2021. Tugendethik. In: A. Grunwald & R. Hillerbrand, Edt. *Handbuch Technikethik.* Berlin: J.B. Metzler, pp. 165-170.
Hinton, G., Osindero, S. & Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. *Neural Computation, 18, 7,* pp. 1527–1554.
HLEG on AI & European Commission, 2020. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self Assessment.* [Online] Available at: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment
Huizing, A., Veenman, C., Neerincx, M. & Dijk, J., 2021. Hybrid AI: The Way Forward in AI by Developing Four Dimensions. In: F. Heintz, M. Milano & B. O'Sullivan, Edt. *Trustworthy AI – Integrating Learning, Optimization and Reasoning. Revised Selected Papers from the First International TAILOR Workshop.* 2020: Springer, pp. 71-76.
Hutson, M., 2022. *Can Computers Learn Common Sense?.* [Online] Available at: https://www.newyorker.com/tech/annals-of-technology/can-computers-learn-common-sense
IBM, 2020. *Getting AI to reason: using neuro-symbolic AI for knowledge-based question answering.* [Online] Available at: https://research.ibm.com/blog/ai-neurosymbolic-common-sense
IEEE Global 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. [Online]
Available at: http://standards. ieee.org/develop/indconn/ec/autonomous_ systems.html
IEEE, 2009. An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions. *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09)*, pp. 381-387.
IEEE, 2021. The First Global Ontological Standard for Ethically Driven Robotics and Automation Systems. *IEEE ROBOTICS & AUTOMATION MAGAZINE*, pp. 120-124.
Jobin, A., Ienca, M. & Vayena, E., 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, pp. 389–399.
Johnson, G., 2021. Algorithmic bias: on the implicit biases of social technology. *Synthese, 198*, pp. 9941–9961.
Kogge, W., 2016. Verkörperung – Embodiment – Körperwissen. Eine historisch-systematische Kartierung. *Paragrana, 25, 1*, pp. 33-48.
Kogge, W., 2017. *Experimentelle Begriffsforschung. Philosophische Interventionen am Beispiel von Code, Information und Skript in der Molekularbiologie.* Weilerswist: Velbrück Wissenschaft.
Kogge, W., 2022a. *Einführung in die Wissenschaften. Wissenschaftstypen – Deutungskämpfe – Interdisziplinäre Kooperation.* Bielefeld: Transcript.
Kogge, W., 2022b. *Governance. Organon terminology toolbox.* [Online] Available at: https://gkorganon.userpage.fu-berlin.de/2021/05/31/governance/
Kriebitz, A., Max, R. & Lütge, C., 2022. The German Act on Autonomous Driving: Why Ethics Still Matters. *Philosophy & Technology, 35*, pp. 1-13.
Krisher, T., 2022. *US report: Nearly 400 crashes of automated tech vehicles.* [Online] Available at: https://www.seattletimes.com/business/us-report-273-teslas-with-automated-driving-systems-crashed/
Lawson, C. et al., 2021. Machine learning for metabolic engineering: A review. *Metabolic Engineering, 63*, pp. 34-60.
Lorenz, P., 2021. *AI Standardization and Foreign Policy. How European Foreign Policy Makers can engage with Technical AI Standardization,* Berlin: Stiftung Neue Verantwortung.
Luger, G., 2021. *Knowing our World. An Artificial Intelligence Perspective.* Cham: Springer.
Mao, J. et al., 2019. The Neuro-Symbolic Concept Learner: Interpreting scenes, words, and sentences from natural supervision. *Published as a conference paper at ICLR 2019* , pp. 1-28.
Marcus, G., 2018. Deep Learning: A Critical Appraisal. *Arxiv*, pp. 1-27.
Marcus, G., 2020. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. *Preprint: Arxiv*, pp. 1-59.
Marcus, G. & Davis, E., 2019. *Rebooting AI. Building Artificial Intelligence We Can Trust.* New York: Penguin.
McCarthy, J., Minsky, M. L., Rochester, N. & Shannon, C. E., 2006. A Proposal for the Dartmouth Summer Research Project on AI, August 31, 1955. *AI Magazine, 27*, pp. 12-14.
McCulloch, W. & Pitts, W., 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics, 5*, pp. 115-133.
Merleau-Ponty, M., 2005. *Phenomenology of Perception.* London: Routledge.
Miłkowski, M., 2018. From Computer Metaphor to Computational Modeling: The Evolution of Computationalism. *Minds and Machines, 28,* pp. 515–541.
Misselhorn, C., 2019. Maschinenethik und Philosophie. In: O. Bendel, Edt. *Handbuch Maschinenethik.* Wiesbaden: Springer, pp. 33-56.
Mitsch, S., Ghorbal, K., Vogelbacher, D. & Platzer, A., 2017. Formal Verification of Obstacle Avoidance and Navigation of Ground Robots. *International Journal of Robotics, 36, 12*, pp. 1-36.
Moore, T. R., 2017. *Trade Secrets and Algorithms as Barriers to Social Justice.* [Online] Available at: https://cdt.org/wp-content/uploads/2017/08/2017-07-31-Trade-Secret-Algorithms-as-Barriers-to-Social-Justice.pdf
O'Regan, G., 2021. History of Artificial Intelligence. In: G. O'Regan, Edt. *A Brief History of Computing.* Cham: Springer, pp. 295-320.
Pasquale, F., 2017. *Secret Algorithms Threaten the Rule of Law.* [Online] Available at:

https://www.technologyreview.com/2017/06/01/151447/secret-algorithms-threaten-the-rule-of-law/
Perez, S., 2022. *Google Translate adds 24 new languages, including its first indigenous languages of the Americas.* [Online] Available at: https://techcrunch.com/2022/05/11/google-translate-adds-24-new-languages-including-its-first-indigenous-languages-of-the-americas/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAANQX5uRf_gcuKp9SHPs9W4hrSn1f8FPAFYKcQ4mtsCo
Rakova, M., 2006. *Philosophy of Mind A-Z.* Edinburgh: Edinburgh University Press.
Rescorla, M., 2020. *The Computational Theory of Mind.* [Online] Available at: https://plato.stanford.edu/entries/computational-mind/
Rizaldi, A., Immler, F., Schürmann, B. & Althoff, M., 2018. A Formally Verified Motion Planner for Autonomous Vehicles. In: S. Lahiri & C. Wang, Edt. *Automated Technology for Verification and Analysis. 16th International Symposium Proceedings.* Cham: Springer, pp. 75-90.
Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, pp. 206-215.
Russell, S., 2020. *Human Compatible. AI and the Problem of Control.* New York: Penguin Random House.
Russell, S. & Norvig, P., 2021. *Artificial Intelligence. A Modern Approach.* 4th Edition. Hoboken: Pearson.
Shapiro, L. & Spaulding, S., 2021. *Embodied Cognition.* [Online] Available at: https://plato.stanford.edu/entries/embodied-cognition/#FoilInspForEmboCogn
Spiegel Online, 2016. *Software schlägt Go-Genie mit 4 zu 1.* [Online] Available at: https://www.spiegel.de/netzwelt/gadgets/alphago-besiegt-lee-sedol-mit-4-zu-1-a-1082388.html
Sternberg, R., 2020. The Concept of Intelligence. In: R. Sternberg, Edt. *The Cambridge Handbook of Intelligence.* Cambridge: Cambridge University Press, pp. 3-17.
Sun, R., 2014. Connectionism and neural networks. In: K. Frankish & W. Ramsey, Edt. *The Cambridge Handbook of Artificial Intelligence.* Cambridge: Cambridge University Press, pp. 109-127.
Tone, R., Levin, K. & Køppe, S., 2018. Affective Incarnations: Maurice Merleau-Ponty's Challenge to Bodily Theories of Emotion. *Journal of Theoretical and Philosophical Psychology, 38, 4*, pp. 205-218.
Umbrello, S. & van de Poel, I., 2021. Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, pp. 283–296.
van der Meulen, S. & Bruinsma, M., 2019. Man as 'aggregate of data'. What computers shouldn't do. *AI & SOCIETY, 34*, pp. 343–354.
Veres, S., Molnar, L., Lincoln, N. & Morice, C., 2011. Autonomous Vehicle Control Systems. A Review of Decision Making. *Journal of Systems and Control Engineering 225, 2*, pp. 155–195.
Wittgenstein, L., 1960. *The Blue and Brown Books.* Oxford: Basil Blackwell.
Wooldridge, M., 2021. *The Road to Conscious Machines. The Story of AI.* Dublin: Pelican Books.
Yuste, R. & Levin, M., 2021. *New Clues about the Origins of Biological Intelligence.* [Online] Available at: https://www.scientificamerican.com/article/new-clues-about-the-origins-of-biological-intelligence/
Zhang, D. et al., 2021. *The AI Index 2021 Annual Report.* [Online] Available at: https://aiindex.stanford.edu/ai-index-report-2021/
Zhang, T., Qin, Y. & Li, Q., 2021. Trusted Artificial Intelligence: Technique Requirements and Best Practices. *IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 1458-1462.
Zigon, J., 2019. Can Machines Be Ethical? On the Necessity of Relational Ethics and Empathic Attunement for Data-Centric Technologies. *Social Research, 86, 4* , pp. 1001-1022.