Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI)

Alexis Fritz, Wiebke Brandt, Henner Gimpel and Sarah Bayer

Philosophical and sociological approaches in technology have increasingly shifted toward describing AI (artificial intelligence) systems as '(moral) agents,' while also attributing 'agency' to them. It is only in this way – so their principal argument goes – that the effects of technological components in a complex human-computer interaction can be understood sufficiently in phenomenological-descriptive and ethical-normative respects. By contrast, this article aims to demonstrate that an explanatory model only achieves a descriptively and normatively satisfactory result if the concepts of '(moral) agent' and '(moral) agency' are exclusively related to human agents. Initially, the division between symbolic and sub-symbolic AI, the black box character of (deep) machine learning, and the complex relationship network in the provision and application of machine learning are outlined. Next, the ontological and action-theoretical basic assumptions of an 'agency' attribution regarding both the current teleology-naturalism debate and the explanatory model of actor network theory are examined. On this basis, the technical-philosophical approaches of Luciano Floridi, Deborah G. Johnson, and Peter-Paul Verbeek will all be critically discussed. Despite their different approaches, they tend to fully integrate computational behavior into their concept of '(moral) agency.' By contrast, this essay recommends distinguishing conceptually between the different entities, causalities, and relationships in a human-computer interaction, arguing that this is the only way to do justice to both human responsibility and the moral significance and causality of computational behavior.

Introduction: Exemplary harmful outcomes

Artifacts have played a substantial role in human activity since the first Paleolithic hand axes came into use. However, the emergence of an (ethical) discussion about which roles can be attributed to the people and artifacts involved in an action is only a consequence of the increasing penetration of artifacts carrying 'artificial intelligence' (AI) into our everyday lives.

Let us consider three examples of the potentially harmful effect of sophisticated machine learning approaches:

1) Google's search engine shows ads for high-paying executive jobs to men, but not so much to women.[1] Google's photo tagging service incorrectly labeled photos showing African-American people as showing 'gorillas.'[2] Even years after being alerted to this racist behavior, Google did not fix the machine learning approach itself, instead simply removing the word 'gorilla' from the set of possible labels.[3]
2) Amazon developed a machine learning system designed to analyze the résumés of job applicants and rate them with respect to their technical skills. The system was shown to be sexist in how it distinguished between applicants: 'It penalized résumés that included the word 'women's,' as in 'women's chess club captain.' And it downgraded graduates of two all-women's colleges.'[4] Amazon eventually shut down the system after failing to fully prevent discrimination.
3) In pretrial, parole, and sentencing decisions in the U.S., machine learning algorithms frequently predict a criminal defendant's likelihood of committing a future crime. The calculation of these so-called 'recidivism scores' is made by commercial providers that do not disclose the workings of their models. It was demonstrated for a widely used criminal risk assessment tool that used 137 features concerning an individual that the model performs no better than a simple logistic regression using just two features: age and the defendant's total number of previous convictions.[5] Yet, the seemingly more sophisticated 137-feature black box is being used in practice and has been accused of having a racial bias.[6,7]

We do not suggest that Google, Amazon, or the providers of criminal risk assessment tools are sexist, racist, or discriminatory by purpose in any other way. These examples merely illustrate that even well-intentioned initiatives using subsymbolic AI black boxes can lead to harmful outcomes. These systems may do very well with respect to some performance measures but may have inductive biases which are hard to detect and hard to fix. Overall, applications of AI, and especially subsymbolic machine learning-based

[1] Cf. Julia Carpenter, 'Google's Algorithm Shows Prestigious Job Ads to Men, But Not to Women. Here's Why That Should Worry You', *The Washington Post* (July 6, 2015), online at https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/ (accessed 2019-11-10).
[2] Cf. Alex Hern, 'Google's Solution to Accidental Algorithmic Racism: Ban Gorillas', *The Guardian* (January 12, 2018), online at https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people (accessed 2019-11-10).
[3] Cf. ibid.
[4] Reuters, 'Amazon Ditched AI Recruiting Tool that Favored Men for Technical Jobs', *The Guardian* (October 11, 2018), online at https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine (accessed 2019-11-10).
[5] Cf. Julia Dressel and Hany Farid, 'The Accuracy, Fairness, and Limits of Predicting Recidivism', *Science Advances* 4:1 (2018).
[6] Cf. Anthony W. Flores, Kristin Bechtel and Christopher T. Lowenkamp, 'False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks'', *Federal Probation Journal* 80:2 (2016), pp. 38-46.
[7] Cf. Sam Corbett-Davies et al., 'A Computer Program Used for Bail and Sentencing Decisions was Labeled Biased Against Blacks. It's Actually Not That Clear', *The Washington Post* (October 17, 2016), online at www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas (accessed 2019-11-10).

systems, are part of complex socio-technical systems. There is no doubt that AI systems have moral impact, but do they act and reason morally?[8]

The question of whether it is possible to create ethically acting machines represents an ongoing discussion.[9,10] Additionally, the dominant approaches of technical philosophy and sociology currently emphasize the moral significance of AI systems, and have moved towards calling them '(moral) agents' and attributing them 'agency.' The principal argument of this approach is that it allows us to describe both the moral effect of an action's technological components and the complex network of human-computer interaction in a sufficiently descriptive and ethical manner. It is therefore crucial to elucidate the semantics of 'agency' and 'moral agency,' as well as their connection to the concept of responsibility, in order to provide more clarity in settings involving hybrid human-computer intelligence. The central issue is whether we can better grasp the descriptive and normative dimensions of AI and especially subsymbolic machine-learning-based systems with the help of the 'agency' attribution.

In the first part of this research, we provide basic information on symbolic and subsymbolic AI, the black box character of (deep) machine learning, and the complex relationship networks in the supply and application of machine learning.

The second part elaborates ontological and action-theoretical basic assumptions of agency attribution regarding the current teleology-naturalism debate, as well as an explanatory model of Actor-Network Theory (ANT).

Thirdly, three technical philosophical models describing computer systems as '(moral) agents' are critically analyzed with regard to whether an extended agency attribution really illuminates the descriptive and ethical-normative structure of human-computer interaction, or whether it obscures this.

Background on artificial intelligence

AI describes a computer 'system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation.'[11] Different levels of AI include narrow AI (below human-level intelligence, outperforming humans in specific domains but not being potent in other domains), general AI (human-level intelligence across many domains), and artificial super intelligence (above human-level intelligence). Contemporary AI systems show narrow AI (also known as weak AI).

Early computer programs solved tasks that can logically be described with a set of rules and are therefore easy for computers but require prolonged effort for people. A branch of AI still follows this route: computers are equipped with a formal representation of knowledge about the world and the rules of logical reasoning. Thus, they deductively generate new insights. This type of AI is *symbolic AI* because it builds on explicit symbolic

[8] Cf. ibid.
[9] Cf. Michael Anderson and Susan Leigh Anderson, 'Machine Ethics. Creating an Ethical Intelligent Agent', *AI Magazine* 28:4 (2007), pp. 15-26.
[10] Cf. Gordana Dodig Crnkovic and Baran Çürüklü, 'Robots: Ethical by Design', *Ethics and Information Technology* 14:1 (2012), pp. 61-71.
[11] Andreas Kaplan and Michael Haenlein, 'Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence', *Business Horizons* 62:1 (2019), pp. 15-25.

programming and inference algorithms. IBM's chess computer Deep Blue defeating the chess world champion Gary Kasparov in 1997 is an example of a symbolic (narrow) AI system. The other type of AI is *subsymbolic AI* using machine learning. The challenge for today's computer programs is to solve tasks that for humans are hard to describe formally, as they are more intuitive; for example, speech recognition, face recognition, or emotions.[12] Machine learning aims to build computers that automatically improve through experience.[13] A computer program learns from experience with respect to a class of tasks and a specific performance measure, if its performance on tasks of that class improves with experience.[14] However, this focus on experience might lead to an inductive bias if training data is not representative of the data and situations a machine learning model will face after training. Within AI, 'machine learning has emerged as the method of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications.'[15] Contemporary voice assistants, such as Amazon's Alexa, Apple's Siri, and Microsoft's Cortana, leverage such subsymbolic (narrow) AI.

Symbolic AI is easier to debug, easier to explain, and easier to control than subsymbolic AI, as symbolic programming lends itself to human inspection. Subsymbolic AI requires less upfront knowledge, builds on learning from data more successfully and shows better performance than symbolic AI in many domains, especially on perceptual tasks.

Deep learning is a form of machine learning that has gained popularity in recent years due to advances in (big) data availability, (cloud-based) massive computing power, algorithms, and openly available libraries for using these algorithms. In this context, the 'depth' refers to the number of layers in the network's structure; for example, in an artificial neural network (ANN). In the training phase, the strength of the connections (an analogy to brain synapses) between different nodes (an analogy to brain neurons) in the network is identified and learned. The more nodes and connections a network has, the better the network can acquire structural descriptions of the domain (if sufficient training data is available). Some of the largest artificial neural networks have millions of nodes and billions of connections.

Black box character of (deep) machine learning

Machine learning models, especially deep ANN, are frequently perceived as a black box.[16] Once such a model is then trained, and calculating the output based on a given input is rather simple. In principle, all the weights and functions to apply can be inspected manually. However, the sheer number of nodes and connections in a deep ANN, as well as the non-linearity of the calculations, make it practically very difficult, if not impossible, to fully understand the model's behavior for all but the most trivial examples. It is even more difficult to ex-ante predict the outcome of the statistical learning process. Thus, many people effectively perceive deep learning as a black box.

[12] Cf. ibid., p. 15.
[13] Cf. Stuart J. Russell and Peter Norvig, *Artificial Intelligence. A Modern Approach* (Boston: Pearson, 2016).
[14] Cf. Tom M. Mitchell, *Machine Learning* (Boston, Mass.: WBC/McGraw-Hill, 1997).
[15] Russell and Norvig, *Artificial Intelligence*, p. 255.
[16] Cf. Davide Castelvecchi, 'Can we open the black box of AI?', *Nature* 538:7623 (2016), pp. 20-23.

Over recent years, applications of AI became more sophisticated in terms of high-impact and high-risk tasks, such as autonomous driving or medical diagnosis. This has led to an increasing need for explanations.[17] At the same time, this rising complexity has made it more difficult to get insights and to understand and trust the system's functions – not just for users, but also for the programmers of those algorithms.[18] A logical model, like a decision-tree with statements involving 'and,' 'if-then,' etc., is comprehensible for the user. The larger the decision tree, the longer it takes, but humans are able to work through this process. Understanding deep learning models with millions or even billions of connections can be compared to understanding human predictions: we might anticipate what the system predicts, based on prior experience with the system, but we will never be completely sure if our assumption about the system's operating principles is correct.

This lack of transparency stands at the core of the discussion about the accountability and responsibility of humans regarding AI systems: can the user trust a prediction or be responsible for a decision made by a system that she or he cannot understand? To solve this issue, the research stream of *explainable AI* discusses two main options: white box and black box approaches. White box approaches aim at transparency, for instance, by displaying verbally or graphically the 'information contained in the knowledge base,' or via explaining the evidence, such as displaying the symptoms and test results that indicate the existence of a disease.[19] As the operating principles of linear models or decision trees are easier to understand, those models still dominate in many application areas.[20] Nevertheless, complex machine-learning models are in the fast lane and should offer explanations of their predictions to users. Due to the rising complexity of such systems, we cannot expect users to understand how the models work.[21]

Taking the example of an ANN, black box approaches focus on, for example, visualizing the input-output relationship, thus showing which input is most responsible for reaching a certain output.[22,23] These approaches help users and programmers shed light on the black box, but they do not reveal the whole complex functions of the ANN. Therefore, such approaches make AI 'more of a grey than a black box.'[24] Still, these highly performant black and grey box machine learning systems pose challenges in terms of agency, especially as these artifacts are part of complex systems involving multiple actors.

[17] Cf. Jichen Zhu et al., 'Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation', *IEEE Conference on Computational Intelligence and Games* (2018), pp. 1-8.
[18] Cf. Mitchell, *Machine Learning*.
[19] Cf. Carmen Lacave and Francisco J. Díez, 'A Review of Explanation Methods for Bayesian Networks', *The Knowledge Engineering Review* 17:2 (2002), pp. 107-127.
[20] Cf. Grégoire Montavon, Wojciech Samek and Klaus-Robert Müller, 'Methods for Interpreting and Understanding Deep Neural Networks', *Digital Signal Processing* 73 (2018), pp. 1-15.
[21] Cf. Or Biran and Kathleen McKeown, 'Human-Centric Justification of Machine Learning Predictions', *Proceedings of International Joint Conferences on Artificial Intelligence* (2017), pp. 1461-1467.
[22] Cf. Zhu et al., 'Explainable AI for Designers'.
[23] Cf. Ruth C. Fong and Andrea Vedaldi, 'Interpretable Explanations of Black Boxes by Meaningful Perturbation', *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 3429-3437.
[24] Zhu et al., 'Explainable AI for Designers'.

Complex relationship networks in the supply and application of machine learning

Figure 1 is a stylized picture of the value chain from algorithm development, all the way through to the human being affected by a decision. It is an abstract depiction of the processes behind the examples given above. By showing the different types of human actors involved, it can thereby illustrate the complex interplay between different human actors and artifacts.

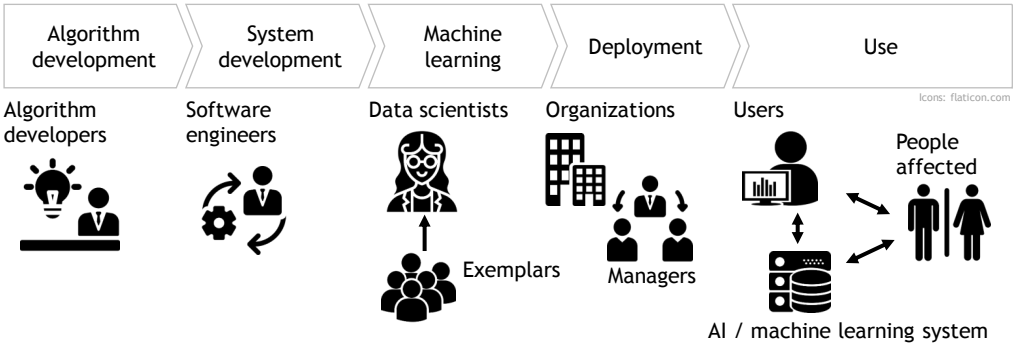

Figure 1: Stylized value chain from algorithm development to use of machine learning systems

Algorithm development conceives general-purpose machine learning algorithms. System development embeds these algorithms in a software system, typically for a specific purpose like criminal risk assessment or personnel decisions. The system is trained on the basis of data that originates from it (e.g., prior decisions by humans like evaluating résumés or sentencing criminals). Organizations like a court system or a company – or, more specifically, managers within an organization – then decide to use the system. Finally, individual users (like a clerk in the personnel department or a judge) interact with the machine learning-based system to obtain information and make decisions that affect others, like applicants or defendants.

If this overall socio-technical system harms people, who is responsible? There are eight candidates: (1) the technical AI system, despite it being an artifact; (2) the users obliged to use a system they do not understand; (3) the managers who neither understand the black box nor make individual decisions; (4) the organization; (5) the data scientists, despite the fact they do not make decisions concerning individual persons; (6) the people providing the training data, oftentimes unknowingly; (7) the software engineers, despite their inability to foresee the system's behavior after learning; and (8) the algorithm developers who created the multi-purpose black boxes in the first place. Is any single candidate responsible, several of them (each to a certain degree), is the overall socio-technical system responsible without individual responsibility, or are none of them responsible?

Pre-assumptions of agency attribution based on action theory

Asking what an actor or an action is and how it can be explained leads to a branched discussion of very different approaches to action theory. This makes it clear that agency attribution depends on several ontological and action-theoretical basic assumptions. Whoever uses concepts of action must not shy away from reflecting on these fundamental implications. Only against this background can different positions and their possible conclusions be adequately understood and discussed.

The teleology-naturalism debate concerns whether we can adequately describe and understand human actions and natural events by the same language and at the same level. Actor-Network Theory seeks to overcome the distinction between humans and non-humans by describing an actor as the symmetrical interplay between social, technical, and natural entities.

The teleology-naturalism debate in action theory

In order to determine the ways in which an action differs from a natural event, it is instructive to take a closer look at how we talk about it. We usually explain actions through the intentions of the person doing them ('She opened the window to air the room'), thus attributing the mental capacity to have goals, make decisions, etc. In contrast, we consider a natural event as the (provisional) end of a causal chain, and name the previous chain links as an explanation for its taking place ('The window opened because a gust of wind blew against it').[25] Obviously, we distinguish between a 'mental' language, which refers to actions, and a 'physical'[26] language, which refers to natural events.[27] As long as both are applied only in their respective fields, there is no problem. However, it is questionable whether the same event can be expressed in both languages: is the window opening perhaps also due to certain neuronal states that triggered the woman's arm movement? Is such a physical description perhaps even more accurate than referring to mental states and abilities?

How do these different descriptions of the same event relate to each other? Are both of them legitimate perspectives that are able to coexist, or do they exclude each other so that at least one of them must be wrong? As a third option, one language might be translatable into the other.[28]

This is exactly the basic assumption of the naturalistic approach: anything expressed in mental language can be translated into physical language without any loss of meaning. Ultimately, there is no ontological difference between actions and natural events.[29] Accordingly, actions are subject to the same causal laws as natural events. Therefore, they can, in theory, be retrospectively deduced from a certain set of necessary and sufficient conditions, as well as predicted for the future if those very conditions are fulfilled (deductive-nomological explanatory scheme) – even if an accurate prediction is practically difficult to realize due to the complex interplay of numerous internal and external conditional factors.[30] In order to avoid this problem, a simpler action pattern is declared the object of investigation: the so-called 'basic action,' which consists of only a

[25] Cf. Edmund Runggaldier, *Was sind Handlungen? Eine philosophische Auseinandersetzung mit dem Naturalismus*. Stuttgart: W. Kohlhammer, 1996, p. 17, 106; Christoph Horn and Guido Löhrer, 'Einleitung: Die Wiederentdeckung teleologischer Handlungserklärungen', in *Gründe und Zwecke. Texte zur aktuellen Handlungstheorie*, edited by Christoph Horn and Guido Löhrer (Berlin: Suhrkamp, 2010), pp. 7-45, at p. 8.

[26] Cf. Runggaldier, *Was sind Handlungen?*, p. 18.

[27] Cf. Runggaldier, *Was sind Handlungen?*, p. 106.

[28] Cf. Scott R. Sehon, 'Abweichende Kausalketten und die Irreduzibilität telologischer Erklärungen', in *Gründe und Zwecke. Texte zur aktuellen Handlungstheorie*, edited by Christoph Horn and Guido Löhrer (Berlin: Suhrkamp, 2010), pp. 85-111, at p. 87; Horn, 'Einleitung', pp. 15f.

[29] Cf. Runggaldier, *Was sind Handlungen?*, pp. 15, 24-26.

[30] Cf. ibid., pp. 26, 106f, 110; Josef Quitterer, 'Basishandlungen und die Naturalisierung von Handlungserklärungen', in *Soziologische Handlungstheorie. Einheit oder Vielfalt*, edited by Andreas Balog and Manfred Gabriel (Opladen: Westdeutscher Verlag, 1998), pp. 105-122, at pp. 106f.

simple body movement (e.g. bending a finger).[31] If one regards the different levels of an action as an 'action tree,' then this 'basic action' represents the lowest, most basal level, which cannot be further explained by other partial actions. You get to higher levels by asking 'why?': he bent his finger to pull the trigger of a weapon, to fire a bullet at a person, to kill that person, etc. By contrast, you reach a lower level by asking 'how?': he killed him by shooting at him, by using the trigger, by bending the finger, etc. At this point, where you cannot break down the question of 'how?' any further, you have reached the lowest level.[32] Regardless of whether you consider these levels to describe the same action or many different actions,[33] both positions agree that the 'basic action' is the main, essential action on which further analysis has to concentrate.

The teleological approach contrasts with the naturalistic approach, and its followers criticize the orientation towards 'basic actions': in order to do justice to the nature of an action, it cannot be reduced to a body movement. On the contrary, the higher levels of the action tree are to be examined, where the actor's intentions, systems of rules and signs, the situational context with possibly involved third parties, etc. are situated.[34] Certain actions (e.g. greeting, betting, lecturing) are not dependent on a certain movement of the body, and therefore cannot be reduced to it.[35] But even actions whose correlation to body movements is evident, such as firing a weapon, are principally comprehensible only against the background of their circumstances and references: not the bending of the finger, but the intention to kill, the connection with the victim, etc., which constitute the action.[36] The reference to lower levels of action can be misleading, and even be used to deliberately conceal the essence of the action: 'I have only...'[37]

Teleologists agree that intentions are the criterion that distinguishes an action from a natural event.[38] In contrast to the naturalistic translation thesis, they insist that mental language cannot be reduced to physical language, since intentions cannot be equated with the links of a causal chain.[39]

Not only is it practically impossible to completely determine all the causal conditions for an action taking place, but this is also theoretically opposed by the conviction that a human being is fundamentally free in his decision to act.[40]

[31] Cf. Quitterer, 'Basishandlungen', pp. 107f.
[32] Cf. Runggaldier, *Was sind Handlungen?*, pp. 46-48; Quitterer, 'Basishandlungen', pp. 115f; Georg Kamp, 'Basishandlungen', in *Handbuch Handlungstheorie. Grundlagen, Kontexte, Perspektiven*, edited by Michael Kühler and Markus Rüther (Stuttgart: J. B. Metzler Verlag, 2016), pp. 69-77, at pp. 69f.
[33] According to the 'unifiers'/'minimizers' bending the finger and killing the victim represent a single action; from the point of view of the 'multipliers'/'maximizers' these are numerically different actions (cf. Runggaldier, *Was sind Handlungen?*, pp. 50f; Quitterer, 'Basishandlungen', pp. 116f; Christian Budnik, 'Handlungsindividuation', in *Handbuch Handlungstheorie. Grundlagen, Kontexte, Perspektiven*, edited by Michael Kühler and Markus Rüther (Stuttgart: J. B. Metzler Verlag, 2016), pp. 60-68, at p. 60).
[34] Cf. Runggaldier, *Was sind Handlungen?*, pp. 55, 59, 62; Quitterer, 'Basishandlungen', p. 106.
[35] Cf. Runggaldier, *Was sind Handlungen?*, pp. 65f.
[36] Cf. ibid., p. 62; Quitterer, 'Basishandlungen', pp. 118f.
[37] Cf. Runggaldier, *Was sind Handlungen?*, p. 62f.
[38] Cf. Friedo Ricken, *Allgemeine Ethik* (Stuttgart: W. Kohlhammer, 2013 [1983])., pp. 103f; Horn, 'Einleitung', p. 9; Sehon, 'Abweichende Kausalketten', p. 85; Runggaldier, *Was sind Handlungen?*, pp. 12, 68; Donald Davidson, 'Handlungen, Gründe und Ursachen', in *Gründe und Zwecke. Texte zur aktuellen Handlungstheorie*, edited by Christoph Horn and Guido Löhrer (Berlin: Suhrkamp, 2010), pp. 46-69, at p. 48.
[39] Cf. Runggaldier, *Was sind Handlungen?*, p. 76; Horn, 'Einleitung', p. 8; Sehon, 'Abweichende Kausalketten', p. 110.
[40] Cf. Runggaldier, *Was sind Handlungen?,* pp. 110-113.

Donald Davidson, a representative of a moderate naturalism, takes this objection seriously and does not claim any principal predictability of human action. In the case of a broken windowpane, it can be stated afterwards, without any doubt, that a certain stone caused its breaking. However, to move from such a causal analysis to a prognosis about how hard one has to throw a stone against a window to break it in the future is something completely different.[41] For actions, it applies analogously that individual, concrete actions can be explained causally and, in these individual cases, be translated into physical language. However, there are no laws either in the mental realm or between the mental and the physical sphere according to which predictions about future actions can be made. The name of this position, 'anomalous monism,' derives from the negation of such overarching laws.

Teleologists reply that such a concept devalues the mental side, since it is causally effective only insofar as it can be translated into physical terms.[42] Again, the intentionality of the actor is reduced.

Instead of searching for mental or physical events within the actor that have produced his action, one should simply accept the actor himself as the origin of his action ('agent-causality').[43]

The concept of 'agency' in Actor-Network Theory (ANT)

Both naturalistic and teleological theories of action require a distinct separation between the subject and the object of an action. ANT criticizes this basic assumption. It opposes mechanistic, quasi-automatic explanations of actions, as well as models of understanding that presuppose the intention, autonomy, or consciousness of the human actor. But how are the terms 'action' and 'agency' to be understood if there is no subject-object difference, no primary principle, or no modern concept of the subject?

ANT is a challenging alternative to traditional theories of action, and has become one of the classic approaches of technical sociology.[44] Bruno Latour, Michel Callon, and John Law founded this theory in the 1980s and continue to develop it further to this day. Despite the diversity and complexity of the concepts within this family of ANTs, some key aspects shall be briefly highlighted.[45]

ANT does not ask why an actor acts in this way and not differently. Rather, it describes how an actor is transformed into an agent through the interplay of social, technical, and natural entities. The surprising thing is not so much that action always refers to others, but that non-humans are not simply passive objects of human action. Instead, they act themselves in a heterogeneous network.[46]

[41] Cf. Davidson, 'Handlungen, Gründe und Ursachen', pp. 63f.
[42] Cf. Runggaldier, *Was sind Handlungen?,* pp. 122-127, 132; Quitterer, 'Basishandlungen', pp. 109; 112-114.
[43] Cf. Runggaldier, *Was sind Handlungen?*, pp. 144-147.
[44] Cf. Roger Häußling, *Techniksoziologie. Eine Einführung* (Opladen, Toronto: Verlag Barbara Budrich, 2019), pp. 240-252.
[45] A differentiated introduction to ANT in German is offered by Andréa Belliger and David J. Krieger, 'Einführung in die Akteur-Netzwerk-Theorie', in *ANThology. Ein einführendes Handbuch zur Akteur-Netzwerk-Theorie*, edited by Andréa Belliger and David J. Krieger (Bielefeld: transcript, 2006), pp. 13-50; Ingo Schulz-Schaeffer, *Sozialtheorie der Technik* (Frankfurt am Main: Campus-Verlag, 2000).
[46] Cf. Bruno Latour, 'Social Theory and the Study of Computerized Work Sites', in *Information Technology and Changes in Organizational Work*, edited by W. J. Orlinokowsky and Geoff Walsham (London: Chapman and Hall, 1996), pp. 295-307, at pp. 303ff.

This basic assumption is formulated by ANT as the general principle of symmetry, which claims a radically equal treatment of humans and non-humans. Social, technical, and natural factors are equal and depend on each other.[47] In order to clarify the concept that not only humans are capable of acting, ANT replaces the 'actor' with an 'actant.' An actant is generally someone or something with the ability to act and to exercise activity.[48] Both human and non-human actants begin to create heterogeneous networks by themselves. They do not precede their networking but are produced by the networking process. The results of such networking are hybrids (i.e. hybrid forms of the social, the technical, and the natural).[49]

Actants transform into actors when a role and interests are assigned to them in the process of building networks (figuration).[50] The successive and different steps of the network-building process are summarized under the term 'translation.' This is 'the continuous attempt to integrate actors into a network by 'translating' them into roles and interests.'[51] Translations create the 'identities, characteristics, competences, qualifications, behaviors, institutions, organizations and structures necessary to build a network of relatively stable, irreversible processes and procedures.'[52] A 'network' is not an external social reality, but a theoretical term for a concept that 'is traced by those translations in the scholars' accounts.'[53] Statements about actants and actors are always moments in the process of network building or translation.

Latour exemplified his ANT by closing a door.[54] He understands this process as a network in which both human (= the user) and technical (= the door) actants are involved. If you regularly forget to close the door, this can quickly become a problem. This problem can then be solved, for instance, by introducing a sign, hiring a porter, or implementing a door-closing mechanism. If, for instance, a door-closing mechanism is installed, the new technical actant changes the characteristics and behavior of the existing network. For example, people have to adapt to the speed of the closing door.

While humans determine technical behavior, technical artifacts can also lead to human behavioral changes. In ANT, there is no clearly assignable making and being made; instead, there is only the network of actants (e.g. texts, people, animals, architectures, machines, or money).[55]

[47] Cf. Bruno Latour, *Wir sind nie modern gewesen. Versuch einer symmetrischen Anthropologie* (Berlin: Akad.-Verl., 1995), pp. 125ff.
[48] Cf. Madeleine Akrich and Bruno Latour, 'A Summary of a Convenient Vocabulary for the Semiotics of Human and Nonhuman Assemblies', in *Shaping Technology/ Building Society. Studies in Sociotechnical Change*, edited by Wiebe E. Bijker and John Law (Cambridge, Mass.: The MIT Press, 1992), pp. 259-264, at p. 259.
[49] Cf. Latour, *Wir sind nie modern gewesen*, pp. 7f.
[50] Cf. Michel Callon, 'Einige Elemente einer Soziologie der Übersetzung: Die Domestikation der Kammmuscheln und der Fischer der S. Brieuc-Bucht', in *ANThology. Ein einführendes Handbuch zur Akteur-Netzwerk-Theorie*, edited by Andréa Belliger and David J. Krieger (Bielefeld: transcript, 2006), pp. 135-174, at pp. 146f; Bruno Latour, *Reassembling the Social. An Introduction to Actor-Network-Theory* (Oxford: Oxford University Press, 2007) p. 53.
[51] Belliger and Krieger, 'Einführung in die Akteur-Netzwerk-Theorie', p. 39 (translated by authors).
[52] ibid., p. 39 (translated by authors).
[53] Latour, *Reassembling the Social*, p.108.
[54] Cf. Jim Johnson, 'Mixing Humans and Nonhumans Together: The Sociology of a Door-Closer', *Social Problems* 35:3 (1988), pp. 298-310.
[55] Cf. Michel Callon, 'Techno-ökonomische Netzwerke und Irreversibilität', in *ANThology. Ein einführendes Handbuch zur Akteur-Netzwerk-Theorie*, edited by Andréa Belliger and David J. Krieger (Bielefeld: transcript, 2006), pp. 309-342, at p. 313.

This sometimes results in controversial, even irritating formulations in Latour's writing. Thus, a clumsy hotel key chain acts more morally than its human user. Due to its size, it forces the guest to hand in the key at the reception desk before leaving the hotel.[56] When asked whether a person or a weapon was responsible for killing a person, Latour replied: 'It is neither people nor guns that kill. Responsibility for action must be shared among the various actants.'[57] It is a hybrid that cannot be reduced to a technical or human actant. Agency emerges from a connection of actants in the network: 'Action is a property of associated entities.'[58] Action and agency are always distributed among different entities. According to the sociologist M. Wieser, the notion of the agency in terms of non-human things must 'not be understood as animism or as the naive intentionality of things, but as the power of things, highlighting their resistance.'[59] 'Agency' is not a substance, but a process.[60] In this sense, non-humans also possess the ability to act, for which the English term 'Agency' or 'Material Agency' has prevailed in technical sociology.[61]

Three technical-philosophical approaches

It turned out that 'agent' or 'agency' are multifaceted concepts in the field of action theory. Their semantics and language practice depend on controversial and sometimes contradictory basic assumptions. The following technical-philosophical approaches are not identical with any of the action-theoretical directions discussed above. Nevertheless, the basic concerns, the course, or the focus of the following technical-philosophical approaches can each be traced back to one of the previously discussed theories of action.

The following approaches aim to describe and ethically evaluate the complex human-computer interaction appropriately and descriptively with the help of the terms '(moral) agent' or 'agency.'

The original problem and the basic concern of the three systemic models coincide. Nevertheless, Floridi's, Johnson's and Verbeek's answers compete with each other, and thus cannot be sensibly combined. To put it simply, we can describe Floridi's model as 'techno-centric,' Johnson's as 'anthropocentric,' and Verbeek's as 'constructivist.'

L. Floridi: Artificial Agency

According to Floridi, the so-called standard ethics (i.e. deontological – like discourse-theoretical and contractualistic – or teleological – like virtue-ethical or consequentialist

[56] Cf. Bruno Latour, 'Technology is Society Made Durable', in *A Sociology of Monsters? Essays on Power, Technology and Domination*, edited by John Law (London/ New York: Routledge, 1991), pp. 103-131.
[57] Bruno Latour, *Pandora's Hope. Essays on the Reality of Science Studies* (Cambridge, Mass.: Harvard Univ. Press, 1999), p. 180.
[58] Ibid., p. 182.
[59] Matthias Wieser, *Das Netzwerk von Bruno Latour. Die Akteur-Netzwerk-Theorie zwischen Science & Technology Studies und poststrukturalistischer Soziologie* (Bielefeld: transcript, 2012), p. 182 (translated by authors).
[60] Cf. ibid., pp. 184f.
[61] Cf. Latour, *Reassembling the Social*, p. 45; Werner Rammert, *Technik – Handeln – Wissen. Zu einer pragmatistischen Technik- und Sozialtheorie* (Wiesbaden: Springer VS, 2016 [2007]), p. 14; Wieser, *Das Netzwerk von Bruno Latour*, pp. 175-184.

ethics) are hopelessly overwhelmed by the challenges of human-computer interaction.[62] The first reason for this is that in conventional philosophy, only human beings (and thus no AI), are considered 'moral agents.' Thus, the human actor is burdened by a disproportionally great responsibility.[63] Secondly, actions are judged on the basis of the actor's intentions:[64] it is morally relevant whether a person is injured intentionally or unintentionally. However, this focus on intentions does not help us where AI is used. In fact, the impact of a self-learning computer system can never be overlooked completely and therefore cannot be answered for by the designer or user. It is for this reason that Floridi suggests that we broaden the concept of 'moral agency' and refrain from judging intentions.[65]

Starting from the question who or what a 'moral agent' is, Floridi argues that definitions must be looked at in their particular context:[66] A car mechanic looks at a car from a different point of view than an ethicist. To refer to these different points of view, Floridi uses the technical term 'level of abstraction.' At different levels of abstraction, different observables are relevant. For example, an ethicist delights in low pollutant emission, while a car mechanic is pleased by an unbroken V-belt.[67]

In order to define 'agent' properly, Floridi suggests a higher level of abstraction than is usually adopted. Candidates for 'agents' should no longer be examined for intentionality or other mental abilities; instead, they should be observed from a more distant perspective, appearing only vaguely as 'systems.' To be called 'agents,' systems have to be interactive, autonomous, and adaptive.[68]

According to Floridi, whether, for example, a computer program checking CVs is considered an 'agent' depends on the granularity of the level of abstraction employed: if only the incoming CVs and their outgoing evaluation are regarded as 'observables,' but the algorithm itself is hidden, the recruitment program appears interactive, autonomous, and adaptive, consequently, as an 'agent': 'interactive,' because it begins to work in reaction to an external input; 'autonomous,' because it arranges the many applications automatically – as in a black box –; and 'adaptive,' because it learns on the basis of the data records.[69]

From 'agent' to 'moral agent' takes only a small step: for Floridi, all 'agents' whose actions have morally qualifiable consequences are 'moral agents.'[70] Consequently, the recruitment program is not only an 'agent,' but also a 'moral agent,' because its selection is sexually discriminatory.

[62] Cf. Luciano Floridi and Jeff W. Sanders, 'Artificial Evil and the Foundation of Computer Ethics', *Ethics and Information Technology* 3 (2001), pp. 55-66, at pp. 57, 64f.
[63] Cf. Luciano Floridi and Jeff W. Sanders, 'On the Morality of Artificial Agents', *Minds and Machines* 14 (2004), pp. 349-379, at pp. 350f.
[64] Cf. Luciano Floridi, 'Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2016), Issue 2083, at p. 4.
[65] Cf. ibid., p. 3f.
[66] Cf. Floridi and Sanders, 'On the Morality of Artificial Agents', pp. 352f.
[67] Cf. Luciano Floridi, 'Levels of Abstraction and the Turing Test', *Kybernetes* 39 (2010), pp. 423-440, at p. 426; Floridi and Sanders, 'On the Morality of Artificial Agents', p. 354.
[68] Cf. Floridi and Sanders, 'On the Morality of Artificial Agents', pp. 357f.; Floridi, 'Levels of Abstraction', p. 432.
[69] Cf. Floridi and Sanders, 'On the Morality of Artificial Agents', p. 362; Floridi, 'Levels of Abstraction', p. 432.
[70] Cf. Floridi and Sanders, 'On the Morality of Artificial Agents', p. 364.

However, the program is not morally responsible for its consequences, as responsibility requires intention,[71] but intention does not matter at the level of abstraction chosen for 'agency.' According to Floridi, 'moral agents' without intentions are not morally responsible for their actions but accountable.[72] If artificial 'moral agents' cause damage – by analogy with sanctions on people – they can be modified, disconnected from the data network, or completely deleted or destroyed.[73]

Floridi finally concludes that his understanding of 'moral agency' and 'accountability' sufficiently clarifies the ethical questions of human-computer interaction: 'The great advantage is a better grasp of the moral discourse in non-human contexts.'[74]

This positive self-evaluation of Floridi has to be questioned:

First, the AI debate is – according to Floridi – about attributing responsibility. If we stick to this assumption, we cannot see how the existence of non-responsible 'moral agents' can help in the search for a culprit.

Second, Floridi's reference to non-human 'moral' sources of good and evil of all kinds is nothing new in itself: a serious illness, a large avalanche, a chainsaw, a rabid dog, or falling roof tiles can all cause human suffering. However, despite the damage, we would never speak of a 'moral' avalanche, chainsaw, disease, dog, or tile.

By calling computer systems 'moral,' we can neither describe their mode of action better (causality), nor come closer to resolving moral issues (evaluation of an action or attribution of responsibility).

It can perhaps be said that the novelty of Floridi's approach lies not so much in qualifying the impacts of computer systems as 'moral' but in perceiving them as 'agents' at a certain level of abstraction. However, would that take us any further descriptively or normatively? This raises three thoughts: first, the necessity of making computer systems 'accountable' (i.e. that they have to be reprogrammed or even switched off if deficient) may be realized without there being any need of calling them 'moral agents.' While we may call our computer names when it does not do what we want it to, we do not do so because we seriously believe it will somehow impress our computer. Second, not all links in a causal chain need to be called 'moral agents' in order to become the object of ethical thought. Even in the standard ethics scolded by Floridi, a moral evaluation of an action or the attribution of responsibility is only possible after a precise and sufficient description of the causal connections. Third, it must also be criticized that if something goes wrong, at the level of abstraction favored by Floridi, the question of responsibility can no longer be posed for AI as a 'moral agent,' since Floridi abstracts from human intention, and computer systems are accountable but not morally responsible. In this way, ethically questionable incentive structures emerge, where the responsible party can be excused prematurely.

Thus, the impression is reinforced that the term 'moral agents' in Floridi's explanatory model contributes nothing toward gaining a better descriptive and normative understanding of human-computer interaction. It can thus be dismissed without consequences, since 'moral agent' or 'moral agency' is an empty concept if separated from responsibility.

[71] Cf. ibid., p. 365.
[72] Cf. ibid., pp. 351, 376.
[73] Cf. ibid., pp. 372f.
[74] Ibid., p. 376.

D. G. Johnson: Triadic agency

Deborah Johnson struggles to find a happy medium between two extremes: one position undermines human responsibility to the extent that computer systems are referred to as 'moral agents,' and Johnson explicitly criticizes Floridi's approach. Representatives of the other position, on the other hand, misjudge the moral quality of machine behavior since they regard technology as extra-moral.

In the course of a larger searching movement, Johnson developed the so-called 'Triadic agency' model. According to Johnson, a state is caused neither by man nor by the computer system alone, but by a differentiated interaction. Basically, 'agency' means a 'capability to act.' Johnson distinguishes between three forms of agency:

(1) 'causal agency': things have a causal effect;[75]

(2) 'intentional agency': people act intentionally; their intention causes the action;[76]

(3) 'triadic agency': these forms of 'agency' relate to each other and are more than the sum of their individual parts. When people cooperate with computer systems, then:

a. the user wants to achieve a certain goal – in our case the Amazon HR department wants an efficient and effective personnel selection –and delegates this task to the designers;

b. the designer project team creates the recruitment program;

c. with the help of this program the initial goal is achieved.[77]

In the 'triadic agency' model, responsibility is attributed only to those who are able to act intentionally. Since AI has no intention, it bears no responsibility for its causal effectiveness. Only humans can be 'moral agents' due to their intentional capacity. People therefore remain responsible, even if they delegate increasingly complex tasks to AI. In the search for the responsible person(s), it has to be asked in the direction of the designer or user until a person (or a group of persons) is found. However, an answer to the question of how much responsibility each person bears cannot be found without also considering the technological component.

By differentiating between three modes of action, Johnson first succeeds in maintaining the ontological difference between man and machine in terms of action theory. This differentiation is not essentialist, since it does not refer to fixed descriptive characteristics, but to certain abilities. Secondly, although only human beings can be responsible, their responsibility can only be clarified if all components of action are considered. Because of the descriptive and normative significance of machine behavior, Johnson does not want to renounce the agency attribution.

However, Johnson's inclusive use of the term 'agency' gives rise to misunderstandings and side scenes, since one term refers to human beings, computer systems, and human-computer interaction. Johnson strives to name the difference and interrelationship between man and computer systems, but she shrinks from taking the final step and continues to call computer systems 'agents.' Unlike Floridi's use of the term, Johnson's 'agency' is not meaningless but misleading. It would have been more

[75] Cf. Markus Schlosser, 'Agency', *The Stanford Encyclopedia of Philosophy* (2015), online at https://plato.stanford.edu/entries/agency/ (accessed 2019-11-15).

[76] Cf. ibid.

[77] Cf. Deborah G. Johnson and Mario Verdicchio, 'AI, Agency and Responsibility: The VW Fraud Case and Beyond', *AI & SOCIETY* (2018), online at https://doi.org/10.1007/s00146-017-0781-9 (accessed 2019-11-15), p. 4.

beneficial to use different terms such as 'factor,' 'cause,' or 'actor' in order to emphasize the specific descriptive and normative contribution of computer systems.

P.-P. Verbeek: Hybrid agency

Peter-Paul Verbeek's 'mediation theory' is based on Don Ihde's postphenomenological approach and Bruno Latour's 'actor-network theory.'[78] Verbeek emphasizes the joint causality of man and technology. Hence, technology actively mediates between human beings and their environment.[79] It does so on two levels: hermeneutically, by influencing human perception of the world, and pragmatically, in partaking in human action.[80]

Returning to our example of a recruitment program, the question of how the human resources department perceives the applicants – as deficient or positive – is decisively mediated by technology (hermeneutical mediation), and the final recruitment decision is pragmatically mediated. It is neither determined by, nor can it be made completely independently of, technology.

Consequently, according to Verbeek, moral decisions and actions are joint products of human beings and technology;[81] morality is 'hybrid,' and 'moral agency' is a mixture ('composite moral agency').[82] No thing or living being possesses 'moral agency' by itself. Rather, 'moral agency' results from complex technical-human interaction; it does not form the basis for an action but emerges from it.[83]

Verbeek goes so far as to describe even the actors themselves as the result of interaction.[84] Nevertheless, Verbeek's theorem of a hybrid 'moral agency' does not mean that people cannot bear responsibility. In particular, designers of computer systems bear great responsibility because technology shapes the way of being in the world, and thus the human being himself. Verbeek shows the ethical dimensions with sentences such as 'Designers materialize morality'[85] and 'Designing technology is designing human beings.'[86]

Against this background, we would like to ask whether Verbeek's 'moral agency' attribution helps us to understand human-computer interaction better both descriptively and ethical-normatively. The strength of Verbeek's postphenomenological-constructivist mediation theory undoubtedly lies in the fact that it acknowledges the complexity of human-computer interaction. Verbeek's approach is particularly successful in reflecting

[78] Cf. Peter-Paul Verbeek. 'Materializing Morality. Design Ethics and Technological Mediation', *Science, Technology, & Human Values* 31 (2006), pp. 361-380, at pp. 362f.; Peter-Paul Verbeek, *Moralizing Technology. Understanding and Designing the Morality of Things* (Chicago: Univ. of Chicago Press, 2011) pp. 33, 45-47, 52.
[79] Cf. Verbeek, 'Materializing Morality', p. 364; Peter-Paul Verbeek, 'Some Misunderstandings About the Moral Significance of Technology', in *The Moral Status of Technical Artefacts*, edited by Peter Kroes and Peter-Paul Verbeek (Dordrecht: Springer, 2014), pp. 75-88, at pp. 77f.
[80] Cf. Verbeek, 'Materializing Morality', pp. 364, 368.
[81] Cf. Verbeek, 'Some Misunderstandings', p. 78.
[82] Ibid., pp. 77f.
[83] Cf. ibid., pp. 75, 80; Peter-Paul Verbeek, 'Designing the Morality of Things: The Ethics of Behaviour-Guiding Technology', in *Designing in Ethics*, edited by Jeroen van den Hoven, Seumas Miller and Thomas Pogge (New York: Cambridge Univ. Press, 2017), pp. 78-94, at p. 84.
[84] Cf. Peter-Paul Verbeek, 'Beyond Interaction: A Short Introduction to Mediation Theory', *Interactions* 22 (2015), pp. 26-31, at p. 28.
[85] Verbeek, 'Beyond Interaction', p. 31 (cf. Verbeek, 'Materializing Morality', pp. 361, 369, 379; Verbeek, 'Designing the Morality of Things', p. 88).
[86] Verbeek, 'Beyond Interaction', p. 28.

reality. If we accept that technology creates reality in terms of its interplay with human beings, and if this awareness replaces both obsession with, as well as forgetfulness about, technology, then much is gained for the debate about the responsible use of technology in both a descriptive and normative sense. This is true even if mediation is not a specific characteristic of technology alone.

However, with regard to Verbeek's understanding of 'moral agency,' there are important inquiries to make:

Unlike Floridi, Verbeek considers intentionality and freedom as part of the term 'moral agency,' albeit in a mediated, hybrid form. However, intentionality and freedom do not constitute 'moral agency'. Instead, and much like 'moral agency' itself, this only results from a complex human-computer interaction.

The strength of the postphenomenological-constructivist view of reality turns into a weakness as soon as we want to attribute agency or responsibility to individual, concrete entities. In Verbeek's mediation theory, 'moral agency,' intention, freedom, and thus responsibility can no longer be attributed to individuals, since they always emerge from an overall structure. Ultimately, in Verbeek's theory of mediation, the individual and his actions cannot be conceived without technical influences or mediation. Human beings and computer systems are 'actants' – only as a mixture are they also 'agents.'

Verbeek's two concerns – reconstructing the understanding of human-computer interaction and attributing moral responsibility – could also be fulfilled if the human actors remained 'moral agents.' For the realization that human capacity to act is always mediated is nothing new from a philosophical point of view. However, in order to avoid a circular conclusion in the attribution of 'moral agency' and moral responsibility, the freedom of human actors must be regarded as taking precedence. This is because interaction does not have its origin in itself but is a consequence of the human ability to reflect, decide, and act freely.

Conclusion

This study has revealed the opportunities and risks of applying the concept of 'moral agency' to human-computer interaction. Ultimately, the risks of agency attribution to computational behavior are disproportionate to the benefits of such language practice.

From a descriptive and ethical-normative point of view, this practice proves to be both unnecessary and risky. Floridi's use of 'moral agents' for computer systems is redundant. Exclusive features for human or social contexts (e.g. 'intentionality' or 'responsibility'), which should be preserved, come out of sight.

Verbeek offers a comprehensive and promising understanding of human-computer interaction. However, his 'moral agent' attribution is circular or leads to an infinite regression, thus making it objectionable. This is illustrated by the fact that it is difficult to identify a specific human capacity or actor for responsibility.

Johnson's results are consistent in view of their ontological and action-theoretical premises. She also conceptually differentiates the contribution of each component and is thus able to provide an almost accurate understanding of human-computer interaction. However, the 'agency' attribution gives rise to misunderstandings. At the same time, there is a serious risk that the extensive use of 'moral agents' undermines the question of responsibility.

Consequently, an appropriate differentiation between humans and computers should also be conceptually discernible. In this way, human-computer interaction can not only be described more precisely but the ethical-normative structure can also be elaborated more clearly.

Alexis Fritz, Wiebke Brandt, Henner Gimpel and Sarah Bayer
alexis.fritz@ku.de, wiebke.brandt@ku.de,
henner.gimpel@fit.fraunhofer.de, sarah.bayer@fim-rc.de

Bibliography

Akrich, Madeleine and Bruno Latour. 'A Summary of a Convenient Vocabulary for the Semiotics of Human and Nonhuman Assemblies', in *Shaping Technology/ Building Society. Studies in Sociotechnical Change*, edited by Wiebe E. Bijker and John Law. Cambridge, Mass.: The MIT Press, 1992, pp. 259-264.

Anderson, Michael and Susan Leigh Anderson. 'Machine Ethics. Creating an Ethical Intelligent Agent', *AI Magazine* 28:4 (2007), pp. 15-26.

Belliger, Andreá and David J. Krieger. 'Einführung in die Akteur-Netzwerk-Theorie', in *ANThology. Ein einführendes Handbuch zur Akteur-Netzwerk-Theorie*, edited by Andréa Belliger and David J. Krieger. Bielefeld: transcript, 2006, pp. 13-50.

Biran, Or and Kathleen McKeown. 'Human-Centric Justification of Machine Learning Predictions', *Proceedings of International Joint Conferences on Artificial Intelligence* (2017), pp. 1461-1467.

Budnik, Christian. 'Handlungsindividuation', in *Handbuch Handlungstheorie. Grundlagen, Kontexte, Perspektiven*, edited by Michael Kühler and Markus Rüther. Stuttgart: J. B. Metzler Verlag, 2016, pp. 60-68.

Callon, Michel. 'Einige Elemente einer Soziologie der Übersetzung: Die Domestikation der Kammmuscheln und der Fischer der S. Brieuc-Bucht', in *ANThology. Ein einführendes Handbuch zur Akteur-Netzwerk-Theorie*, edited by Andréa Belliger and David J. Krieger. Bielefeld: transcript, 2006, pp. 135-174.

[original English version:

Callon, Michel. 'Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St. Brieuc Bay', in *Power, Action and Belief: A New Sociology of Knowledge?*, edited by John Law. London: Routledge, 1986, pp. 196-233.]

Callon, Michel. 'Techno-ökonomische Netzwerke und Irreversibilität', in *ANThology. Ein einführendes Handbuch zur Akteur-Netzwerk-Theorie*, edited by Andréa Belliger and David J. Krieger. Bielefeld: transcript, 2006, pp. 309-342.

[original English version:

Callon, Michel. 'Techno-Economic Networks and Irreversibility', in *A Sociology of Monsters? Essays on Power, Technology and Domination,* edited by John Law. London/ New York: Routledge, 1991, pp. 132-161.]

Carpenter, Julia. 'Google's Algorithm Shows Prestigious Job Ads to Men, But Not to Women. Here's Why That Should Worry You', *The Washington Post* (July 6, 2015), online at https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/ (accessed 2019-11-10).

Castelvecchi, Davide. 'Can we open the black box of AI?', *Nature* 538:7623 (2016), pp. 20-23.

Corbett-Davies, Sam, Emma Pierson, Avi Feller and Sharad Goel. 'A Computer Program Used for Bail and Sentencing Decisions was Labeled Biased Against Blacks. It's Actually Not That Clear', *The Washington Post* (October 17, 2016), online at www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas (accessed 2019-11-10).

Crnkovic, Gordana Dodig and Baran Çürüklü. 'Robots: Ethical by Design', *Ethics and Information Technology* 14:1 (2012), pp. 61-71.

Davidson, Donald. 'Handlungen, Gründe und Ursachen', in *Gründe und Zwecke. Texte zur aktuellen Handlungstheorie*, edited by Christoph Horn and Guido Löhrer. Berlin: Suhrkamp, 2010, pp. 46-69.

Dressel, Julia and Hany Farid. 'The Accuracy, Fairness, and Limits of Predicting Recidivism', *Science Advances* 4:1 (2018).

Flores, Anthony W., Kristin Bechtel and Christopher T. Lowenkamp. 'False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks'', *Federal Probation Journal* 80:2 (2016), pp. 38-46.

Floridi, Luciano and Jeff W. Sanders. 'Artificial Evil and the Foundation of Computer Ethics', *Ethics and Information Technology* 3 (2001), pp. 55-66.

Floridi, Luciano and Jeff W. Sanders. 'On the Morality of Artificial Agents', *Minds and Machines* 14 (2004), pp. 349-379.

Floridi, Luciano. 'Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374 (2016), Issue 2083.

Floridi, Luciano. 'Levels of Abstraction and the Turing Test', *Kybernetes* 39 (2010), pp. 423-440.

Fong, Ruth C. and Andrea Vedaldi. 'Interpretable Explanations of Black Boxes by Meaningful Perturbation', *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 3429-3437.

Häußling, Roger. *Techniksoziologie. Eine Einführung*. Opladen, Toronto: Verlag Barbara Budrich, 2019.

Hern, Alex. 'Google's Solution to Accidental Algorithmic Racism: Ban Gorillas', *The Guardian* (January 12, 2018), online at https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people (accessed 2019-11-10).

Horn, Christoph and Guido Löhrer. 'Einleitung: Die Wiederentdeckung teleologischer Handlungserklärungen', in *Gründe und Zwecke. Texte zur aktuellen Handlungstheorie*, edited by Christoph Horn and Guido Löhrer. Berlin: Suhrkamp, 2010, pp. 7-45.

Johnson, Deborah G. and Mario Verdicchio. 'AI, Agency and Responsibility: The VW Fraud Case and Beyond', *AI & SOCIETY* (2018), online at https://doi.org/10.1007/s00146-017-0781-9 (accessed 2019-11-15).

Johnson, Jim. 'Mixing Humans and Nonhumans Together: The Sociology of a Door-Closer', *Social Problems* 35:3 (1988), pp. 298-310.

Kamp, Georg. 'Basishandlungen', in *Handbuch Handlungstheorie. Grundlagen, Kontexte, Perspektiven*, edited by Michael Kühler and Markus Rüther. Stuttgart: J. B. Metzler Verlag, 2016, pp. 69-77.

Kaplan, Andreas and Michael Haenlein. 'Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence', *Business Horizons* 62:1 (2019), pp. 15–25.

Lacave, Carmen and Francisco J. Díez. 'A Review of Explanation Methods for Bayesian Networks', *The Knowledge Engineering Review* 17:2 (2002), pp. 107-127.

Latour, Bruno. 'Social Theory and the Study of Computerized Work Sites', in *Information Technology and Changes in Organizational Work*, edited by W. J. Orlinokowsky and Geoff Walsham. London: Chapman and Hall, 1996, pp. 295-307.

Latour, Bruno. 'Technology is Society Made Durable', in *A Sociology of Monsters? Essays on Power, Technology and Domination*, edited by John Law. London/ New York: Routledge, 1991, pp. 103-131.

Latour, Bruno. *Pandora's Hope. Essays on the Reality of Science Studies*. Cambridge, Mass.: Harvard Univ. Press, 1999.

Latour, Bruno. *Reassembling the Social. An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press, 2007.

Latour, Bruno. *Wir sind nie modern gewesen. Versuch einer symmetrischen Anthropologie*. Berlin: Akad.-Verl., 1995.

[original English version:

Latour, Bruno. *We Have Never Been Modern.* Cambridge, Mass.: Harvard Univ. Press, 1993.]

Mitchell, Tom M. *Machine Learning*. Boston, Mass.: WBC/McGraw-Hill, 1997.

Montavon, Grégoire, Wojciech Samek and Klaus-Robert Müller. 'Methods for Interpreting and Understanding Deep Neural Networks', *Digital Signal Processing* 73 (2018), pp. 1-15.

Quitterer, Josef. 'Basishandlungen und die Naturalisierung von Handlungserklärungen', in *Soziologische Handlungstheorie. Einheit oder Vielfalt*, edited by Andreas Balog and Manfred Gabriel. Opladen: Westdeutscher Verlag, 1998, pp. 105-122.

Rammert, Werner. *Technik – Handeln – Wissen. Zu einer pragmatistischen Technik- und Sozialtheorie*. Wiesbaden: Springer VS, 2016 [2007].

Reuters. 'Amazon Ditched AI Recruiting Tool that Favored Men for Technical Jobs', *The Guardian* (October 11, 2018), online at https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine (accessed 2019-11-10).

Ricken, Friedo. *Allgemeine Ethik*. Stuttgart: W. Kohlhammer, 2013 [1983].

Runggaldier, Edmund. *Was sind Handlungen? Eine philosophische Auseinandersetzung mit dem Naturalismus*. Stuttgart: W. Kohlhammer, 1996.

Russell, Stuart J. and Peter Norvig. *Artificial Intelligence. A Modern Approach*. Boston: Pearson, 2016.

Schlosser, Markus. 'Agency', *The Stanford Encyclopedia of Philosophy* (2015), online at https://plato.stanford.edu/entries/agency/ (accessed 2019-11-15).

Schulz-Schaeffer, Ingo. *Sozialtheorie der Technik*. Frankfurt am Main: Campus-Verlag, 2000.

Sehon, Scott R. 'Abweichende Kausalketten und die Irreduzibilität telologischer Erklärungen', in *Gründe und Zwecke. Texte zur aktuellen Handlungstheorie*, edited by Christoph Horn and Guido Löhrer. Berlin: Suhrkamp, 2010, pp. 85-111.

Verbeek, Peter-Paul. 'Beyond Interaction: A Short Introduction to Mediation Theory', *Interactions* 22 (2015), pp. 26-31.

Verbeek, Peter-Paul. 'Designing the Morality of Things: The Ethics of Behaviour-Guiding Technology', in *Designing in Ethics*, edited by Jeroen van den Hoven, Seumas Miller and Thomas Pogge. New York: Cambridge Univ. Press, 2017, pp. 78-94.

Verbeek, Peter-Paul. 'Materializing Morality. Design Ethics and Technological Mediation', *Science, Technology, & Human Values* 31 (2006), pp. 361-380.

Verbeek, Peter-Paul. 'Some Misunderstandings About the Moral Significance of Technology', in *The Moral Status of Technical Artefacts*, edited by Peter Kroes and Peter-Paul Verbeek. Dordrecht: Springer, 2014, pp. 75-88.

Verbeek, Peter-Paul. *Moralizing Technology. Understanding and Designing the Morality of Things*. Chicago: Univ. of Chicago Press, 2011.

Wieser, Matthias. *Das Netzwerk von Bruno Latour. Die Akteur-Netzwerk-Theorie zwischen Science & Technology Studies und poststrukturalistischer Soziologie*. Bielefeld: transcript, 2012.

Zhu, Jichen, Antonios Liapis, Sebastian Risi, Rafael Bidarra and G. Michael Youngblood. 'Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation', *IEEE Conference on Computational Intelligence and Games* (2018), pp. 1-8.