KONRAD
ADENAUER
STIFTUNG

From Principle to Paradigm – AI and ethics in concrete application contexts

Antonio Bikić

www.kas.de

At a glance

- For technologies with a high degree of automation, an extended application of the precautionary principle is necessary in order to proactively avoid stress or damage or to reduce them as far as possible.

- The examples listed show various ethical problems. They cannot be resolved only with high-level ethical principles. The principles must be reflected in specific instructions.

- The algorithmic implementation of a fairness principle must be preceded by a discussion on the theory of justice.

- On the long-term view, partnerships between universities and companies are a viable option. Joint projects with research institutes are also possible. In this way, companies can build up the ethical expertise they need to design artificially intelligent systems in a generally compatible manner.

Table of contents

Introduction 3

AI technologies require an advanced application of the precautionary principle 4

Practical example 1: Recommendation systems 4

Practical example 2: Automated driving 6

Practical example 3: Job application processes 8

Ethical principles are insufficient for ethics in the context of artificial intelligence 10

Introduction

The call for the evaluation and regulation of intelligent algorithms from an ethical point of view is getting louder, especially since the recent emergence of self-learning, so-called machine learning algorithms - algorithms that learn based on examples. Contrary to popular belief, this does not necessarily require large amounts of data. So-called one-shot learning approaches try to classify objects using just a few examples (small amounts of data).

To evaluate systems using artificial intelligence (AI) from an ethical perspective, we need to distinguish between two concepts: intelligence and the ability to act.[1] Artificially intelligent systems solve tasks through their ability to act. The criteria for this are either preset or are developed by the system itself. The solutions are then interpreted as intelligent, although machines only have the ability to act without actual intelligence. A machine therefore cannot be described as intelligent in the actual sense of the word. Otherwise it would make sense to equally describe a river flowing around an obstacle as intelligent. Intelligence is based on cognitive abilities, not just on the ability to act.

There are different methods of creating artificially intelligent systems. Machine learning is one of the dominant approaches. At the moment, machine learning is often used to solve classification problems. One successful approach is the so-called sub-symbolic machine learning, which, however, no longer makes the processes within the machines traceable. The machines are becoming black boxes. So-called symbolic approaches have been in use for much longer because they are technically easier to implement. They are used in so-called expert systems (e.g. programs for the support of medical diagnoses). Symbolic approaches are best for solving abstract problems.

Expert systems have been around for decades. Ethical debates, perceived by the general public, only began with the advent of sub-symbolically achieved machine learning. For a long time, actual people were expected to be at critical points in decision chains. Now machines may replace humans in making decisions. In analogy to human action, the machine then “decides”.

AI technologies require an advanced application of the precautionary principle

The search for an ethical justification for the use of software corresponds to the precautionary principle known from environmental and health policy. This is especially true for technologies described as disruptive. If we are to widely use machines, their functioning should reflect the values of European societies, the precautionary principle must come into play. This principle should proactively avoid or reduce as much as possible stress and damage potentially occurring while e.g. using a technology. In practice, the proportionality of risk assumptions is decisive for the weighing of benefits.

Documents such as for instance the ethical guidelines for a trustworthy artificial intelligence[2] published by the High-Level Expert Group on Artificial Intelligence of the European Commission contained very abstract guidelines for an ethical reflection and evaluation of machine functions. However, as soon as we need to analyze concrete situations where ethical expertise is necessary, the different ethical implications become apparent. Among other things, in these cases it must be decided which ethical paradigm should apply to the action.

By an ethical paradigm we understand an ethics system (e.g. an ethic of duty). And it has to be clear that there are different ethics systems, of which principle-based ethics is one system.

The respective companies and institutions must settle on one particular paradigm. An ethical paradigm can also take into consideration ethical principles. However, only structuring these principles will make the ethical reflection process an integral part of product development and evaluation. Some companies do not realize that e.g. the principle of fairness is basically a further developed debate on the theories of justice. And often the diversity of ethical implications only becomes apparent in concrete contexts where artificially intelligent systems are used. Only under such circumstances will, for example, issues of justice become a topic. The principles themselves are abstract and immeasurable items because they have no unit of measurement. There are, however, ways to show the impact of a specific concept of fairness. If they are successful, quota regulations can, for instance, create permeability for certain vulnerable groups.

1

Practical example 1: Recommendation systems

Artificial intelligence algorithms are often integrated into systems without being directly noticeable. For example, recommendation services or recommender systems are common in video streaming services. These programs are often (but not exclusively) based on various forms of machine learning. When recommendation services have collected enough information about a person, they are enabled to recommend products to users. This recommendation is based on previous usage behavior.

Such systems become problematic when, for example, videos of underaged children are uploaded to publicly accessible video portals. This practice relates to so-called family vloggers, i.e. family members who upload videos documenting the daily life of their family. This is often visual material on which children are shown only scantily dressed, e.g. during beach vacations. However, these videos can also be recommended to people with pedophile inclinations, who are looking for related videos.[3] As a result, there are videos with children, which are accessed remarkably often over a short period of time. The children are doing nothing wrong; the videos are uploaded to the network by their legal guardians. Here the parent's naivety conflicts with the targeted nature of certain search queries. One option could now be to stop promoting videos with children through recommendation services. This would have negative effects as well: the family vlogger community consists of millions of users. And not every person who searches for images on such portals has pedophile inclinations. The results of the recommendation services are just proof that the software works very well. It displays exactly what matches the search query. The software is not able to decipher the intentions of the searcher.

If problematic contents of recommendation systems were no longer allowed to be considered, this would not only affect image material with children, but also other areas: e.g. videos with content related to conspiracy theories or extremism. However, since the intentions of the users in the search query are not clear, a general restriction does not prove to be effective.

These examples make it clear how problematic the machine evaluation of content is. Even if machines are by all appearance designed for harmless tasks, they must be able to prevent misuse in a context-sensitive manner. Since this software cannot rule out misuse either, it cannot be fully regulated.

In the case of recommendation services, it is difficult to blame the operators of video portals for the fact that their algorithms work well. There are also good reasons to argue for responsible people putting the videos online. And yet it is problematic when undesirable intentions - such as pedophile tendencies - control a search request. A legally binding analysis of ethical risks on the part of the operator would be a viable option. People who upload video material to the portals must be made aware of the search queries that can be used to find their videos. The user has to be aware of the probability of their videos appearing in problematic search queries. Such search queries should be clearly marked and visible to the person uploading the videos.

Since many parents are not aware of what they are consenting to when uploading files to such portals, they must be told who their - even unintentional - target audience may be.

This knowledge would significantly control their behavior. If they can justify their actions ethically and have media competence, they will act on their own authority. The area that would have to be regulated here can only be regulated to a limited extent, as it is often not clear which content can be problematic. This shows the limits of regulation and technical implementation and/or automation of certain processes. This results in the duty to strengthen the media competence of portal users.

Conceivable supporting options would also be online campaigns or individual advice for users by a chat assistant of the service provider.

2

Practical example 2: Automated driving

From an ethical point of view, highly and fully automated driving is often discussed one-sidedly, as almost exclusively dilemmatic situations occurring in certain ethical paradigms (but not in all) are touched. In reality, a number of other problems require also ethical analysis.

A special topic is the so-called nudging: the intended pushing of human behavior towards a decision that is considered desirable by another person or group of people. A nudge is a non-coerced minimal manipulation of a person‘s behavior.

Various techniques of influencing behavior increase the likelihood that a desired decision will be made. For example, more expensive products are placed on the shelves of a supermarket at eye level, while the cheaper ones are below. Without coercing customers, the more expensive products are bought more often.

It has already been demonstrated with pilots that nudging can save fuel.[4] In the context of highly and fully automated driving, on the other hand, the question arises as to whether drivers may be provided with the option to activate certain functions. Assuming it were statistically proven that an overtaking maneuver carried out by an assistance system is 100 times safer, this would be a good argument to use such a system as often as possible. And let us also imagine a driving situation in which the vehicle detects that the driver‘s concentration is decreasing (e.g. due to microsleep). The driver‘s eyes are monitored by camera systems: a technology that, by the way, comes from German manufacturers.[5]

The significantly increased safety through the use of the assistance system for overtaking maneuvers and the knowledge of their reduced concentration are good reasons to influence the behavior of drivers without applying coercion. How the vehicle triggers such behavioral influence is variable.

Another point in this context is the idea of traffic education: are vehicles allowed to set negative incentives to make certain maneuvers more difficult for a driver? If maneuvers are very risky or not compliant with road traffic regulations, there are some arguments in favor of this idea. In the case of nudging by highly and fully automated vehicles, an ethical assessment is only possible if the precise objectives of fine-tuning the behavior are disclosed. In this way, third-party nudging is avoided and self-nudging is encouraged. However, not all driver behavior should be influenced by the automobile manufacturer or the state.

Should the vehicle‘s algorithms perform some fine tuning, the drivers must also be shown how the vehicle regulates the overall driving behavior. If a vehicle will set negative incentives for certain (e.g. risky) driving maneuvers, these must be designed in such a way that they have a supporting, but never a restrictive, function. The drivers must be able to bypass certain nudges from the vehicle without any special effort. This is the case if, according to their situational assessment, the safety of the journey depends on circumventing the nudges. In this way, safety can be restored by a briefly risky, but necessary driving maneuver. For this reason, the incentives for certain behavior on the part of the vehicle must never be too restrictive.

At this point, economic interests should also be addressed: insurance premiums could be significantly reduced if there is proof that, for instance, a driver continuously and at the right moments uses the assistance systems of their vehicle. The safety of the occupants and all road users would be increased. Not using all assistance systems permanently can also prove to be sensible: drivers have to practice routines in order to steer their vehicle. In this context, however, the amalgamation of economic and private interests can become problematic. It must be clear whether e.g. the vehicle is optimized to achieve the goal of reduced insurance premiums or whether the aim is an extended increase in the safety of occupants and road users. Ideally, these goals will overlap, but the means to achieve them do not necessarily have to be identical. The question here would be whether a choice between these two options should be allowed at all. With such an option, certain driving data would also have to be saved and evaluated. Overall, of course, the principle of accountability applies here as well: The driver must be accountable for following a behavioral influence by a highly automated system, even after the fact. What is meant by this is that behavior must not be influenced in a way that incurs criminal consequences or an unreasonable risk.

The seamless evaluation from ethical standpoints depends on the one hand on the implicit goals of the nudges performed by the vehicle. On the other hand, it must be clear what is to be maximized: product liability or the drivers‘ own responsibility. This relates to the overarching question of the extent to which the vehicles provide responsible maneuvers and behaviors. In this, the principle must apply that only such behavior may be provided as an option, that is accountable by the driver and is also appropriate and in the interests of the driver. In this way, it can be successively ensured that people are introduced to a driving situation that they can cognitively cope with. Only then can they develop the justified feeling that they are making responsible decisions using the machine, since they can see the implications for themselves. Control over the driving process is paramount to the attribution of responsibility. This would also be a way to promote the development of a sense of agency[6] in the context of highly and fully automated systems, i.e. the development of the feeling to be the author[7] of one‘s own actions.

A vehicle that by default maximizes the control to serve manufacturers‘ liability will create a deeply immoral situation: drivers of such a vehicle will develop the feeling that they have no real control over the machine. This undermines taking responsibility. At the same time, in the case of highly and fully automated vehicles, the driver would nonetheless still be required to carry out the driving task, which implies responsibility. It is therefore important to not only document whether it is the assistance systems or the person driving the vehicle who at a certain point in time took over decisive control. We must also record data allowing to deduce whether the recommendations of the assistance systems are responsible. The criteria here must not only be of a legal but also of an ethical nature. It should be possible for the driver to select the risk ethical criteria.

3

Practical example 3: job application processes

It makes sense to use self-learning algorithms, which are particularly well suited for classifications, to optimize job application processes. Without having to use very complex algorithms, CVs that do not meet certain requirements can easily be „sorted out".

In connection with the application process, the concept of fairness is often discussed. This term has been adopted from the justice debate for some time and is now used almost as a synonym for justice.

However, it must be noted that there are numerous other theories of justice that are used today and that do not primarily deal with fairness.[8]

Fairness is regularly associated with the term bias, especially in self-learning systems. This refers to cognitive distortions or prejudices.[9] If it is claimed that certain data sets allowing machine learning are biased, then this does not imply that the data is biased. What is meant is that prejudices can be derived from the data. For example, if a characteristics' value is particularly dominant in a data set, then the self-learning system can judge that this value is desired. If it is only established that a company needs employees, this norm is not a problem. However, when it becomes the norm that employees of any company should be a specific gender, then there is a problem.

Application processes can be optimized for various variables. An attempt can be made to make the application process fair as a whole or to make the result of the process fair. There are also various fairness metrics that can be used to measure the extent to which the results generated by an algorithm are fair in regard to a certain group. These metrics measure the extent to which all users can participate in an appropriate or fair manner in the resources of a system (e.g. jobs). There are also methods that can lower the bias - that is, create fair conditions for assessing a certain group. If these metrics are to be applied outside of an overall concept, the process, even if it eliminates individual prejudices, can ultimately produce unfair or unjust results. The mere presence of these metrics is therefore no guarantee of fairness. Due to the various goals for which optimization is targeted, a non-intended overall result can be achieved.

For example, quota regulations are one form of positive discrimination. This is a form of distributive or distribution justice. The aim here is to ensure that society benefits all people equally. However, one of the problems with these measures, for example in the context of women's quota, is that formulations such as "when equally suited, women will receive preferential treatment" are not uniformly formalized. Formalization allows formulations to be processed by algorithms. But when are two people equally suited? There are numerous criteria that can be used to determine the professional equality of two people. However, these criteria (final grade, place of study, professional experience, etc.) should be standardized and take into account the situation of the respective person. However, when the conditions for establishing the equality of two people are no longer comparable, a situation perceived as unjust is created. It is quite possible that currently, uniform standards used to determine equal suitability of two people for a position are not implemented uniformly even in the same company. This fact could be corrected precisely through the use of algorithmically controlled solutions.

When application processes are automated, it is important to develop an overall perspective of the system. This already starts with the fact that job advertisements, if generated automatically, must be formulated neutrally. Otherwise, the wording itself excludes certain groups that actually belong to the group of applicants.

In addition, general rules must be developed to ensure fairness of distribution (distributive[10]) and at the same time promote fairness of rules (procedural[11]). Depending on how big a company is, different approaches to fairness have to be implemented at various levels, for example to guarantee the advancement opportunities of all social groups in a company. Fairness as a theory of justice can become overall unfair if the fairness metrics are not used coherently. If a very large number of application processes are left to highly automated systems, unfair structures can develop over the years, and the impersonal nature of algorithms makes it difficult to uncover them.

This can be counteracted by establishing standards based on the theory of justice. In certain areas, fair treatment can even be implemented more uniformly by machines than by natural persons.

In addition, for the automation of application processes, binding standards for formalizations such as „with equal suitability" must be worked out. It must be clear whether and to what extent the only relevant criteria should, for example, be the degree and place of study or whether also other things, like for instance the final grade, the professional experience, the entry age and similar elements may play a role. The applicants must feel objectively and equally treated everywhere. Only then can the confidence be built that the artificially intelligent systems effectively implement certain standards of justice.

Systems that process video and audio data represent a separate debate. In this way it is possible to examine a person's body language. It is also possible to use machine learning to estimate based on audio data input whether a person may be suffering from depression.[12]

Using these systems across the board for job interviews is problematic. The current legal situation already covers a lot and regulates what is allowed. A person would very likely have to consent to their video and audio data being processed and possibly also be checked for depression, among others. It would be problematic if the people refuse to give this consent. From this, certain intentions of the applicants can be inferred, which do not have to be true. Therefore, universal use of such systems in application processes would hardly be justifiable. With such systems, a consensus should therefore be found that allows a certain amount of analysis.

It must also be possible to check whether a system used to automate application processes can differentiate between a desired and a factual norm. Thereby, the fact that gender decides whether or not you get a job might actually be the norm. But it is the desired norm that this information should not play a role in the assessment. It should therefore be possible for the machine to autonomously distinguish between these norms. Desired norms could also be called values in this context. There are good arguments for maintaining and enforcing these values. That makes them desirable norms.

Due to their proximity to the automated application system all actors involved in the development, production, operation, evaluation and use of the applications have an expertise hardly achievable by external audits. Therefore, expertise in ethics only becomes visible in the process structures of an automatic application system if the people participating in the system can demonstrate knowledge of ethics. It would make sense to this effect if universities and companies could work together. Concrete problems that can be empirically ascertained could be dealt with by the universities' expertise in ethics. Joint projects with research institutes would also be possible. On the one hand, collaborations to formalize and implement certain quota regulations are conceivable. On the other hand, it would make sense if uniform approaches for the implementation of fairness were used for the entire application process, if it is to run automatically. In this way it would be possible to test all forms of artificially intelligent systems for their overall societal compatibility. This is particularly imperative if these systems are used for a very large number of citizens at neural points in decision-making chains.

Ethical principles are insufficient for ethics in the context of artificial intelligence

Too much focus on ethical principles is not optimal for the current environment in which artificially intelligent algorithms are implemented.[13] This does not mean that a principle-based approach is fundamentally wrong, but that principles need to be formulated in a much more concrete way. The practical examples have shown that in the case of automated application processes, questions of justice theory are also being emphasized. The mere use of fairness metrics does not provide an answer to these questions. Above all, this also means that we need to formulate ethical paradigms, by means of which we can distinguish between ethically correct and incorrect decisions. The necessity of this approach becomes particularly evident when specific application contexts are examined by artificially intelligent algorithms.

Last retrieval of the links mentioned in the sources: March 6th, 2020

[1] Luciano Floridi, J.W. Sanders: „On the Morality of Artificial Agents“, *Minds and Machines* 14, 349–379 (2004). https://doi.org/10.1023/B:MIND.0000035461.63578.9d

[2] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60664

[3] https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html

[4] See http://www.lse.ac.uk/GranthamInstitute/wp-content/uploads/2017/01/ Working-paper-262-Gosnell-et-al.pdf

[5] https://www.nytimes.com/2017/03/16/automobiles/wheels/drowsy-driving-technology.html

[6] See https://www.cell.com/current-biology/pdf/S0960-9822 (12)00191-1.pdf

[7] On the concept of agency over one's own actions: Julian Nida-Rümelin: *Philosophie einer humanen Bildung*, Hamburg 2013.

[8] These theories essentially go back to Rawls, cf. on this John Rawls: *A theory of Justice*, 1971).

[9] *Bias* also occurs independently of self-learning systems. A popular representation is Kah neman's, cf. Daniel Kahneman: *Fast thinking, slow thinking*, Munich 2012.

[10] See https://arxiv.org/pdf/1710.03184.pdf

[11] See http://mlg.eng.cam.ac.uk/adrian/AAAI18-BeyondDistributiveFairness.pdf

[12] See the MIT publication :http://news.mit.edu/2018/neural-network-model-detect-depression-conversations-0830

[13] https://www.nature.com/articles/s42256-019-0114-4

About the author

Antonio Bikić
born 1987, is a doctoral candidate at the Munich Graduate College for Ethics in Practice and is doing his doctorate at the LMU Munich and the ETH Zurich on the feasibility of implementing ethical paradigms. He has a background in philosophy and computational linguistics / computer science and worked for the computing center of the Max Planck Society, at the chairs for practical philosophy / ethics (LMU Munich) and among others for the Association of the Automotive Industry, the Bauhaus Luftfahrt, PwC Munich and for the Fraunhofer Institute for Industrial Engineering and Organization. He regularly gives seminars and lectures on the philosophy of the mind and ethics in the context of artificial intelligence at universities in Germany, Luxembourg and Austria.

Imprint

Contact person:
Tobias Wangermann
Department for Economy and Innovation
Main Department Analysis and Consulting
T +49 30 / 26 996-3380
tobias.wangermann@kas.de

Herausgeberin:
Konrad-Adenauer-Stiftung e. V., 2020 Berlin

Cover photo: © adobeStock/sdecoret; graphiCrash
Design and layout: yellow too Pasiek Horntrich GbR

ISBN 978-3-95721-843-8

www.kas.de