# AI Ethics - Critical Reflections on Embedding Ethical Frameworks in AI Technology

Henrikki Salo-Pöntinen[1]

[1] University of Jyväskylä, Jyväskylä, Finland
lncs@springer.com

**Abstract.** Embedding ethical frameworks in artificial intelligence (AI) technologies has been a popular topic for academic research for the past decade. [1-7] The approaches of the studies differ in how AI technology, ethics, role of technical artefacts and socio-technical aspects of AI are perceived. In addition, most studies define insufficiently what the connection between the process of embedding ethical frameworks to AI technology and the larger framework of AI ethics is. These deficiencies have caused that the concept of AI ethics and the construct of embedding ethical parameters into AI are used in an ambiguous, rather than in a complementary manner.
One reason for the ambiguity within this field of research is due to a lack of a comprehensive conceptual framework for AI ethics in general. I intend to fill this void by grounding AI ethics as a subfield of philosophy of technology and applied ethics and presenting its main issues of study by examining recognized spheres of activities through the method of levels of abstraction [8]. I put forward an initial hierarchical conceptual framework for AI ethics as an outcome. After this, I discuss the connection between the process of embedding ethical frameworks in AI and the larger AI ethics framework, leading to presenting basic requirements for the sphere of activity hereafter known as embedded ethics.

**Keywords:** AI Ethics, Embedded Ethics, Applied Ethics, Human-Technology Interaction.

## 1 Introduction

The need to design AI technology that embodies ethical frameworks has risen due to the ever-increasing role of AI in today's societies. It has been recognized that non-technical governance methods (e.g., legislation or guidelines) are in many cases insufficient in assuring that AI technologies function in a morally desirable manner. [2-4] Therefore, the sustainable development and deployment of AI systems require installing parameters into the AI systems themselves that guide their decision-making processes from a moral perspective.

However, it is not clear how this should be done. Some advocates propose to develop so called moral guards that are placed in the technical artifacts of AI to assess their outputs and authorize only those outputs that are morally acceptable. [2, 5-7] This leads to a dilemma of the kind of moral code the moral guards should follow and how would

it be possible to develop the moral guards in an efficient way, since they would have to be able to authorize outputs that might not be possible to predict beforehand. [2, 3] In addition, it is seen as a decentralization of the moral responsibility of humans and by doing so obscuring the discourse of AI ethics. [9]

Luciano Floridi and Mariarosaria Taddeo consider it to be misleading to treat ethical dilemmas as rising from the functioning of the technical artifacts, but to emerge from the myriad ways of processing and using data. And therefore, they should also be solved by analyzing forms of data used in AI and its further processing and by imposing codes of conducts, standards, and professional ethics to guide the discipline of data science so that its achievements will be morally acceptable. [10] In his proposition for systematizing the process of embedded ethics, Ibo van de Poel suggests that in addition to studying AI as technical artifacts, the sociotechnical aspects (as organizational institutions) of AI should be noted in the act of embedding ethical frameworks in the technology. Otherwise, AI is understood in a too narrow manner leading to imposing insufficient actions to successfully embed ethical frameworks in AI. [3]

All forementioned views perceive the central information processing of AI as a technical process[1] and hence do not account for aspects related to human-technology interaction, which then again are perceived as pivotal in systems engineering. [4, 11] In addition, the view of Floridi and Taddeo[2] leads to a somewhat technocratic view of AI development and deployment, which is not durable [12, 13] considering AI's role as a central technology in the ongoing societal change towards intelligent societies [14, 15].

This scatteredness of views is partly result from looking at the issue from the perspective of single disciplines, partly due to the non-foundational approach of some of the views and mostly due to the yet unstructured multifaceted nature of AI ethics. We need to discover a comprehensive basis for the process of embedding ethical frameworks in AI technology, for it to be possible to produce complementary theories for it.

I agree with van de Poel in that there is a common ground to be found and I concentrate in two critical aspects to further the discourse in this paper. Firstly, I will clarify the general framework for AI ethics to facilitate more rigorous discourse. Secondly, I discuss the role of embedding ethical parameters to AI technology as part of the larger framework of AI ethics and evaluate what possibilities and boundaries its placing imposes for the ethical parameters that are to be instantiated in AI.

In the next section of this chapter, I provide a definition for AI and shortly discuss its further dimensions and connections with autonomous systems. The produced observations serve as an important part in the further chapters. In the second chapter I examine the multifaceted nature of AI ethics as to be revealed through spheres of activities recognized in earlier research related to the ethics of technology and AI ethic, but which has not yet been examined as forming a comprehensive whole. I use the method of levels of abstraction to further the examination and to illustrate the form of the AI ethics

---

1 Even though Floridi and Taddeo focus on the information sphere, they examine it as phenomena emerging from the combination of data, algorithms and hardware and software applications. [10]

2 Floridi and Taddeo most likely understand this risk, since they propose that practices related to responsible research and innovation should be considered when examining important practices related to data ethics.

framework. In the third chapter I discuss the central sphere of activity for this paper, embedded ethics, as part of the larger framework of AI ethics and propose some basic requirements for a comprehensive theory of embedded ethics. In the conclusions I will discuss focal findings of this paper and suggest further research that would advance the discourse of embedded ethics and provide important information for the needs of AI development and deployment.

### 1.1 AI and Autonomous Systems

Providing a definition for AI is a fundamental, but notorious task within the field of AI ethics. For this paper, I use a definition that combines common features from definitions most used in AI research. Samoili et al. provide a comprehensive analysis of commonly used features to describe AI in their paper *Defining Artificial Intelligence* (2020). When comparing Samoili et al.'s analysis with Tony Gillespie's analysis of autonomous systems which he provides in his book *Systems Engineering for Ethical Autonomous Systems* (2019), I noticed the challenge of clarifying how AI and autonomous systems differ. Both refer to technological systems capable of autonomous goal attaining[3] through the means of observing their environment and adaption to changes in it. Therefore, it is not a surprise that much of research does not pursue for their distinction, but place AI as a heading which also accounts autonomous systems within it [17] or defines autonomous systems formally and AI informally, as is done in IEEE's global initiative on ethics of autonomous and intelligent systems *Ethically Aligned Design* (2019).

However, if we are to talk about AI ethics, we need to have a definition that is not too vague. When comparing the analysis of Samoili et al. and Gillespie, the most distinguishable difference between autonomous systems and AI seems to be in how the capability to process information is produced. For AI, the capability to process large masses of data is gained by algorithms that are formed through differing techniques (e.g., symbolic systems and machine learning). [16, 18] The information processing of autonomous systems is then again produced by control systems (networks of nodes that react to fixed inputs) and feedback loops, which can be built to provide the system a capability to carry out very sophisticated functions. [4]

From this observation I construct the following definition. In this paper artificial intelligence (AI) refers to technology that can achieve given goals through data collection (i.e., perception of its environment) and interpretation (reinforced by algorithm techniques) which enables it to perform tasks e.g. in the form of adapting its behavior to changes in the environment.

For further purposes it is important to acknowledge that AI and autonomous systems are parallel concepts and therefore research observations may benefit them both. In addition, AI is an umbrella concept which means that depending on the context of examination, further distinctions might be called for. Further distinctions may refer to the information processing techniques used (e.g., Symbolic AI, guided machine learning, re-enforced machine learning, unguided machine learning and deep learning), [18-19] levels of reflected intelligence (weak AI – strong AI/super artificial intelligence), or the

[3] The goals are always defined by human operators.

scope of tasks AI can manage autonomously (narrow AI - artificial general intelligence) [18, 20].

Due to the lack of explicit reference on the forementioned further definitions of examined AI, researchers may state their offset for a study as examining current or near future developments of AI but end up argumenting based on futuristic expectations.[4] Therefore, many discussions are done implicitly about AI that refers to AGI or superintelligence, even though it might be that they are never reached. [9, 18-19, 21] This kind of misleading argumentation produces alarmism[5], false expectations and obscures academic, as well as societal discourses of AI ethics.

For the sake of coherence, I consider it to be important to acknowledge the different information processing techniques and their differing influence on AI ethics. However, this paper considers the technical artifact of AI to be capable of reflecting only weak intelligence and narrow autonomy [4, 18-19], which means that AI should not be perceived as to have a will of its own, but to always act on orders given by humans, its functional role in a process is and should be very precisely defined and limited [4] and AI should not be perceived as to hold moral agency [3, 9, 21].

## 2 Mapping the Conceptual Framework for AI ethics

Applied ethics issues often rise from actors within or closely linked to the field in question noticing ethical dilemmas rising from status quo course of actions. This is one reason why studying manifestation of the dilemmas may take room from a more systematic analyzation of the whole sphere of activity related to the field – which is more likely to provide understanding about the roots that lead to the noticed problems. [22-23] This is also evidential in the field of AI ethics, where most of the literature is concerned about describing key issues of accountability, explainability, fairness, privacy, safety and security. [16] While an important phase in acknowledging the need to take ethical aspects into consideration, perceiving the role of ethics as only detecting and describing evident issues recalls the analogy about looking for ways to get rid of smoke without detecting fire.

Another misleading way to understand applied ethics is to see it merely as applying some existing foundational moral theory (e.g., deontology, virtue ethics, utilitarism, ubuntu, emotivism or eastern ethical traditions) to describe and solve observed dilemmas. While important in developing high dimensional viewpoints to issues [2], it is important to understand that foundational moral theories have strong underlying ontological and epistemological presumptions and are structured as universal theories, which may prevent from providing holistic conceptualizations that have context sensitive practical value for the issues at hand.

---

[4] This is the case for *Moral Machines* (2008), which I will elaborate in the third chapter of this paper.

[5] The research community has a responsibility to explicitly elaborate what their research considers. Not stating what the current development phase of technology is and not informing when one's paper considers theoretically possible, but unlikely scenarios, researchers legitimize pseudo problems, such as the closeness of singularity.

Understanding applied ethics broadly as outlining and systematizing issues and solutions to practical problems through the approaches and concepts known in the field of moral philosophy gives us a larger variety of tools to frame and analyze issues with case sensitivity and more flexibility to satisfy the need for practical solutions. It can be simplified, that the central study subject of moral philosophy is the connection of forms of human activities[6] and moral agency. Moral agency then again brings us to the concepts of good and bad (evaluative aspects of ethics) and right and wrong (normative aspects of ethics), which provide us with the possibility of framing standards upon which we measure the moral nature of our actions. [24-27]

However, understanding the role of applied ethics in a broad manner is not sufficient as itself to produce a useful framework for AI ethics. It leaves too many questions open about how ethical argumentation can have a meaningful impact on development of AI. This void tends to lead to argumentation aiming to validate universal values or codes of principles that should guide actions within the studied field. [3, 22] Thus, also the field of AI ethics is abundant with lists of principles. [28] This type of argumentation leaves a similar gap between high-level principles and practical needs of the real world as the two first mentioned views of applied ethics. [29]

I argue that we need to recognize focal spheres of activites related to AI ethics to avoid obscuring the connection between high level principles and practice. To do so, the spheres should account for different levels of abstractions and so distinguish study subjects that form a spectrum where one can move from high abstraction towards practical solutions. In this section I aim to provide such a framework by introducing spheres of activities as a key to understand focal observables for AI ethics and analyzing how the different spheres are positioned on different levels of abstractions.

### 2.1 Sphere of Activity

With the concept of sphere of activity, I refer to entities that reflect operational wholes and as such can be understood to be a focal subject of interest for applied ethics. However, distinct fields of applied ethics require analyzing what are the central spheres of activities for each of them separately. The most typical way of describing central activities related to AI ethics is design (including redesign) and development, implementation, deployment, and disposure. [2] This reflects the general idea of a product's lifecycle which helps to plan and impose for example design and management requirements on each phase.

I argue that to understand the multilayered role of ethics in AI, we need to recognize spheres of activities that support the needs of ethical contemplation and its operationalization towards practicality. One reason for this is that the lifecycle division recognizes only very practical phases, which inherently do not account for more abstract[7]

---

[6] Human activities always take place in social contexts, which are shaped by varying cultural, historical and political backgrounds [2, 24].

[7] The design phase is about bringing abstract ideas to exist in the real world, but to understand it only as a phase of a product's lifecycle is not enough for applied ethics.

phases necessary for AI ethics. To do this, we need to take a couple steps back from current discourses and look at the philosophical basis of AI ethics.

**AI Ethics.** AI ethics is a subsection of ethics of technology – which in turn is a subfield of philosophy of technology. Therefore, to provide backdrop for understanding what the basis of AI ethics is, we need to understand what technology is and how this general understanding reflects on AI. In this paper, technology is understood as a combination of technical artefact(s)[8] and human action(s) to fulfill defined objectives. Furthermore, technology development – and AI development with it – is a practice for changing the world to what one sees as ought to be. [31-33]

This leads us to the obvious but often implicit first sphere of activity that will be called (AI) deployment ethics. This sphere comprises from explicitly forming the abstract notions that are to guide what we want to accomplish through the deployment of AI technology. It is obvious for it is ever present in the development of technology, but often driven by tacit impressions, because technology is easily perceived as to develop as separated from other social development. [13-14]

In addition to the core definition of technology, it is important to acknowledge that technology is inevitably rooted in wider[9] sociotechnical contexts. In its narrow sense, the concept of sociotechnical refers to the institutions and organizations in which the technology is utilized in [3, 11] and its broader meaning refers also to the communities and societies where the technology's utilization takes place. [12-13, 32, 34]

For ethics of technology, it is pivotal to understand the concept of sociotechnical through its broader meaning. Otherwise, social impacts of technology are not taken into consideration and their steering becomes an ambiguous and reactive process. [14, 28] Additionally, it would lead to untenable power accumulation of influencing what kind of life should be pursued through technology development to the hands of few people[10]. [12, 32, 34]

AI is distinct from many other technologies [12] in the sense that there is a wide consensus about the importance of considering the societal role and possible larger impacts of AI development and deployment. A reason for this may be that AI is seen as one of the central driving forces in the transition from information societies to intelligent societies, which will have large disruptive effects. [14-15]

Understanding the importance of sociotechnical aspects of technology and AI development requires us to add impact assessment and the process of enabling impactful societal discourse as central parts of deployment ethics. [12, 26] This means that the

---

8 Technical artifacts consist of the artifact (tangible or intangible) and its use-plan. [30-31] This definition shows how the artefacts are always a means to instantiate human intentions.

9 By wider, I mean that the concept of sociotechnical stretches to refer to relations outside of mere user(s) - technical artefact(s) relation.

10 Even in the situation that the few people have good intentions, the strong narrative that technology development is a morally neutral activity [13, 14], and the fact that a few people cannot in any shape perceive the needs and desires of large populations water down the possibility of accepting that kind of power to a small group of people. [13, 23, 35] Therefore, enabling meaningful societal discourse is one of the corner stones of (AI) deployment ethics. [12, 32]

nature of deployment ethics requires balancing with the pursuance for good and prevention/reduction of harm.

Now we have the basis for the first sphere of activity of AI ethics. However, as is implied in the term AI ethics, we need to distinguish the (AI) deployment ethics from technology ethics in general. Therefore, the next sphere of activity should be one that contemplates the ethical dimensions emerging from the nature of AI itself. As was noted in the introduction section of this article, AI and the varying forms of autonomous systems are distinguished by the data processing and refining capabilities invested in AI.

The field of data ethics [10] is committed to analyzing moral dimensions that are inherent in or emerge from the processes of deploying different types of data, their refinement, and the used information processing techniques (e.g., Symbolic systems and the various forms of machine learning) and contemplates their societal impacts. Therefore, data ethics is the second sphere of activity recognizable for AI ethics.

The reciprocal relation between deployment ethics and data ethics produces a cycle where high dimensional questions related to possibilities and risks of AI deployment is imposed on the sphere of data ethics and respectively viewpoints of data ethics provide flesh to the basis of deployment ethics, distinguishing its viewpoints from other fields of ethics of technology.

An example of how the reciprocal nature of deployment ethics and data ethics manifests is how the large volume of principles produced under AI ethics can be defined to originate from the five principles of beneficence, non-maleficence, justice, human autonomy and explicability, when they are understood through the lenses of data ethics. [28] For example, explicability should be understood as a principle that enables the attainment of the other principles, since many information processing techniques used in AI are inherently opaque. [4, 28] Additionally, explicability in AI includes two semantically distinct requirements – intelligibility and accountability. Intelligibility refers to being able to understand[11] and predict the behavior of used AI. Accountability then again refers to establishing mechanisms by which developers and deployers of AI technology can be held accountable for the technology's functioning even though the complexity of placing responsibility rises as the ecosystems responsible for the functioning of AI become more complicated. [2, 28] They both can be understood as components of justified trustworthiness of AI. [36]

As we have distinguished two central spheres of activity for AI ethics – which provide the basis for ethical enquiry of AI and distinguishes AI ethics from other fields of ethics of technology – it is time to pursue further examination through the method of levels of abstraction for the reason of its illustrative power.

[11] It is important to notice that actors with different roles require different level of intelligibility. [2, 28] For example, operators of an AI system require understandable information explaining states relevant for the used basic functions, whereas engineers taking part in the AI systems redesign require explanations that reach the underlying phases of information processing behind the functions.

### 2.2 Levels of abstraction in AI ethics

Levels of abstraction is a well-known method in the field of computational sciences for modelling complex entities into more intelligible ones through distinguishing levels of knowledge representation of the observed subject matter. Luciano Floridi has clarified and refined the method in his paper *The Method of Levels of Abstraction* (2008) in which he demonstrates how the method can be used to elaborate empirical as well as purely conceptual studies. According to Floridi, levels of abstraction should be understood as a method which elaborates the epistemological form of observed entities, since using it to distinguish ontological, or methodological levels of abstractions is on a more unstable basis.

The method consists of three main phases. Recognizing sets of observables that form distinct levels of abstractions (hereafter LoA's) in the studied subject, analyzing the relations between observables within a LoA to gain knowledge about the behavior of a single LoA, and finally analyzing relations between the recognized LoA's which gives understanding about the gradient of abstraction. The gradient of abstraction is a holistic illustration of how the LoA's are connected throughout the different abstraction levels starting from approximate LoA behavior description towards more comprehensive and detailed descriptions, "until the final LoA accounts for the desired behaviors" [8; 314].

Some additional definition of the concepts of observables and LoA behavior is required in addition to explaining what is meant by relations between LoA's. Floridi defines observables as "typed variables[12] together with a statement of what feature of the system under consideration they represent" [8; 306. Footnote added]. For example, when forming basis of deployment ethics, one must consider variables such as *values* and *guiding principles* in addition to *risks* and *opportunities* that are to guide the formation of *goals* and *codes of conduct* for technology development. These variables form the observables of ethical argumentation.

LoA behavior reflects how the system of observables in a certain LoA is to behave in order to work properly. It is needed, since otherwise LoA's would consist merely of sets of observables that can take any value within their given set of values.[13] [8] In other words, by taking LoA behavior into account we avoid producing models with ambiguous functioning.

Floridi illustrates as an example of LoA behavior that in theory traffic lights (LoA) in a single crossroad could all hold the typed variable of showing green light (color as an observable) at the same moment of time, but that would not be a functional traffic

---

[12] A typed variable is a variable that can hold only certain explicitly stated data. Data in this case can refer to either symbols of empirical perception or symbols related to purely conceptual theories. [8]

[13] There are two types of information that are possible to depict by the levels of abstraction method: analogous and discrete. Analogous information refers to information used in natural sciences to depict the basis of natural phenomena. In analogous information the observables can take infinite number of values and their behavior is described with differential equations. The other type of information is discrete, meaning that the observables have a finite number of values they can take. [8] This research considers discrete information.

light system. To avoid such dysfunctionality in a model, the LoA behavior should be described as exactly as possible. The description is provided in the form of a predicate, which reflects the connections and possible values that the observables may take in each LoA. [8] For a functional traffic light example, the describing predicate could be safety securing.

As I stated earlier, the perceived LoA's of a whole are to be connected to each other and their connections are illustrated as a gradient of abstraction (hereafter GoA). Floridi describes two types of connections that can be perceived between LoA's. Disjoint GoA describes a connection where the constituent observables differ between LoA's and nested GoA describes a connection where the constituent observables are common on each LoA. [8] To continue illustration through the traffic light example, a nested GoA for traffic light could consist of a first LoA in which the observable is color and the second LoA on a more concrete abstraction level where the observable would be wavelengths of color. A disjoint GoA for traffic light could consist of a LoA with the observable of color and another LoA where the observable would be the orientation of the lights. In the latter example, the LoA's can be perceived to be on the same level of abstraction from knowledge representation point of view but are complementary as they consider features of the same system.

These two types of relations between LoA's can be combined in a tree like form, in which the GoA contains hierarchical surjective[14] information about several perspectives of the observed issue. [8] The GoA of AI ethics as presented in Figure 1. on the next page depicts a GoA with combined LoA relations.

[14] Surjective means that an abstract observation can be traced back to at least one concrete counterpart. Its strict meaning would allow only a single concrete counterpart per abstract observation but as Floridi points out, abstract information in the field of humanities is often traced back as a connection between several concrete counterparts.

**AI Ethics Conceptual Mapping.** For conceptual mapping of AI ethics, the method of levels of abstraction provides structure and the perceived spheres of activities provide content for distinct LoA's. In figure 1. I have placed central spheres of activities in the GoA of AI ethics to illustrate the conceptual mapping of AI ethics.

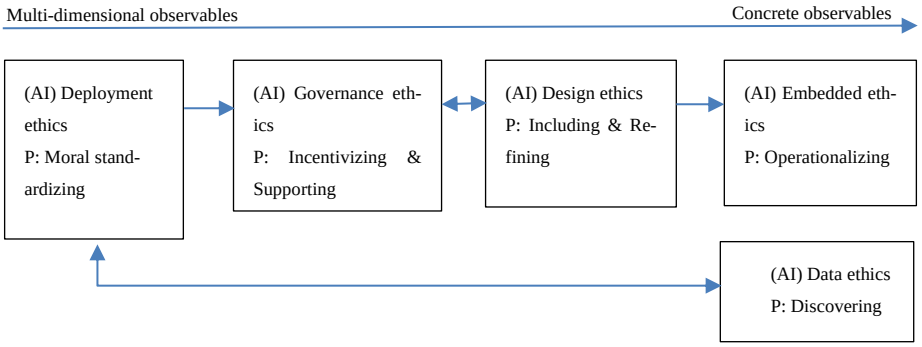


**Fig. 1.** Gradient of abstraction for AI ethics. The arrow on top illustrates the hierarchical relation of the GoA to go from LoA's consisting of multidimensional observables towards LoA's consisting of more concrete observables. Each level of abstraction (LoA) has a predicate (P) to describe its system behavior.

Figure 1. shows that I have added governance ethics, design ethics and embedded ethics as the remaining central spheres of activity for AI ethics. The reason for this is that they depict the line of activities required for integrating AI ethics into action. They are not arbitrarily chosen but relate to existing research on the topic.

As mentioned earlier, AI is always used and developed in social contexts (communities and organizations). Therefore, governance of the social is an inseparable part of how the high-level goals and concepts of the deployment ethics LoA disseminate into the technology. [3, 14, 31] The governance ethics LoA is currently the least developed sphere of activity related to AI ethics. Its importance is widely noticed for the purpose of incentivizing and supporting the integration of ethical frameworks to AI [2, 36], but so far studies have centered in describing issues, whilst providing little concrete recommendations. [37-38]

The sphere of design ethics is then again often regarded as the central sphere of activity for AI ethics since it is the stage where abstractions and real use cases are to be combined. [2, 3, 32] As for the sphere of activity illustrating the most concrete level of knowledge representation (together with data ethics), embedded ethics depicts the activity in which the operationalizations of high dimensional concepts and objectives described in deployment ethics are instantiated in the real world. [2-3]

**Behaviors of the spheres of activity.** As it is out of the scope of this – or any single article – to exhaustively uncover the sections of AI ethics described in figure 1., I shall

pursue to depict their central functions and connections as parts of the gradient of abstraction of AI ethics.

As mentioned earlier, there are three main phases in the method of levels of abstraction. So far, I have illustrated spheres of activities as providing insight about the sets of observables for each LoA, without explicitly defining the observables for each sphere of activity. Moreover, I have slightly, yet not comprehensively provided explanations about the behaviors of the LoA's. And finally, I have illustrated the central features and connections of the AI ethics GoA (figure 1.) but have not described all the relations the GoA depicts (i.e., explicating the two symbols of bidirectional relations).

I will start further explanation with the LoA highest in the AI ethics gradient of abstraction hierarchy, deployment ethics. I regard the paper *AI4People—An Ethical Framework for a Good AI Society* (2008) as a good resemblance of what can be perceived as well-developed deployment ethics. The paper is coauthored by members of the scientific committee of AI4people initiative[15] and parallelly to the whole initiatives purpose, the authors state their aim as producing an ethical framework for a good AI society. To do so, they analyze perceived opportunities[16] and risks that can emerge from pursuing human flourishment and promotion of human dignity by using AI. They suggest using the five principles of beneficence, non-maleficence, human autonomy, justice and explicability to successfully balance between the opportunities and risks while pursuing the forementioned goals of deploying AI.

From the example of AI4people, it would seem obvious to understand the high-level goals (e.g., human flourishment), opportunities, risks and moral principles as observables for deployment ethics. However, the real observables are the ethical arguments formed by these variables. For the notion of ethical argumentation explains the role and meaning of the forementioned variables in the system of deployment ethics.

The AI4people working groups paper provides in addition a good example of the behavior of deployment ethics. In its last section, the working group announces 20 action points based on their ethical argumentation that are to guide governance and use of AI. [28] This can be perceived as the behavior of providing moral standards for the use of AI. Moral standardizing in this sense corresponds to the act of explicating guidance for what we want to achieve through AI development and describes the behavior of deployment ethics.

Next, I will further explain the LoA of (AI) data ethics as consisting of the observables of moral problems of data and moral problems of information processing. As forementioned, this section of AI ethics provides understanding of the ethical problems emerging from AI's nature as a technology based on data usage and its refining through information processing techniques.

In their paper *What is Data Ethics* (2016) Luciano Floridi and Mariarosaria Taddeo introduce data ethics as to represent a new focal level of abstraction in the continuum of information ethics. They describe data ethics as a paradigm shift which directs the

---

[15] AI4people is an Atomium European Institution for Science, Media, and Democracy (EISMD) initiative which pursues to produce frameworks for a good AI society. For more information see https://www.eismd.eu/ai4people/.

[16] Opportunities can turn to missed opportunities if AI is underused for the sake of misleading argumentation.

disciplines focus to the invariant sphere of information formation, and as such, it recognizes moral dimensions that were earlier left unnoticed. This refers to acknowledging "even data that never translate directly into information but can be used to support actions or generate behaviors" [10; 1].

The moral problems related to data may emerge from the generation, recording, curation, processing, dissemination, sharing and use of data which makes them important variables for data ethics. In addition, Floridi and Taddeo consider algorithms (e.g., those used in symbolic systems and different machine learning techniques) and practices related to information processing (such as programming and hacking) as central variables for understanding the moral problems related to information processing. [10] The multifaceted observables of moral problems of data and information processing are used to discover moral dimensions that emerge from the basic nature of AI. Therefore, discovering is the behavior of (AI) data ethics.

To understand the central functions of (AI) governance ethics we need to distinguish ethical aspects of governance from judicial aspects related to AI as they are often implicitly dealt as analogous. Both, governance and judicial actions are normative forces that guide the development and deployment of AI. However, governance refers to establishing and implementing infrastructures and practices that support goal attaining of a given organization or government and judicial aspects of AI refer to the established institutionalized regulations that relate to AI, and their further processing. [14, 39]

The regulative institutions are part of the larger frame of governance of a nation and regulations reciprocally set boundaries for governing actions in general. Therefore, it is understandable why they are sometimes misleadingly used in an analogous manner. Underlining their distinction is not to say that governance and judicial aspects would not be interrelated but to underpin that the LoA of governance ethics refers to all governance actions and not only to the judicial aspects of AI.[17]

Simply put, the LoA of governance ethics consists of the observables of infrastructures and processes that correspond to the further organizational dissemination of the moral standardization constructed in the LoA of deployment ethics. Management and success measurement, organizational structures, codes of conduct, standards, regulations, and auditing are examples of the typed variables forming the corresponding infrastructures and processes. [2, 4, 14, 32] The role of governance is not just to support but also to incentivize the forementioned dissemination and therefore the behavior of the governance ethics LoA is incentivizing and supporting. [3, 14, 28]

Before clarifying the LoA of design ethics, I want to point out to the dissemination processes of the concepts and high-level goals adopted in the LoA of deployment ethics towards the lower abstraction levels starting from the LoA of governance ethics. The reciprocal relation between the LoA's of governance ethics and design ethics is important for illustrating differing disseminations. The reciprocal relation can be understood through an example of how certain values disseminate into technology through the social structure of design processes [13, 35] while others may be considered to result

[17] Some aspects of the governance ethics LoA may be perceived to best serve its meaning if they were regulated nationally or by an intergovernmental covenant [40], but that is a whole other discourse and out of the scope of this article.

from an intentional implementation process [3] within the design phase. Consider for example the notion of non-discrimination[18], which has been acknowledged to be almost impossible to tackle without ensuring that the design teams developing AI are constructed as intercultural and gender balanced. [35, 41] This applies even though there are design tools to include multifaceted perspectives through user-inclusion to the design phase.

As governance actions are the ones which structure the social sphere of the organization, it is also the sphere of activity in which the values that disseminate through the social structure of design are to concretize. [13] Therefore, acknowledging the varied nature of how high-level goals disseminate to AI technology through the structure of the design phase as well as an intentional act of designers [3] requires us to understand the influencing relation of the LoA's of governance and design ethics to be inherently reciprocal. To put more simply, one must understand what kind of requirements value dissemination in the design phase imposes to governance actions to be able to construct impactful governance ethics frameworks. In figure 1. this is depicted with the bidirectional arrow.

The design ethics LoA illustrates the sphere of activity, in which the concrete use cases and abstract models and concepts for AI are integrated. [2, 4, 30] In this article, design is understood broadly as being inherently a science of problem solving. [30-32] Therefore, problem identification and definition is the starting point for a design process. In addition, the problem definition guides further design actions and thus must incorporate the ethical dimensions of deployment ethics for them to be realized in the developed technology. [30, 32]

From the AI ethics point of view, the task for problem definition is to include the recognized ethical dimensions into the design process and refine them to suite the use case context. Thus, the design ethics LoA's behavior is including and refining. Design thinking and approaches must be suitable for this to be possible in such a way that the role of ethics is understood as part of the whole [32] instead of a compulsory, yet ineffective, checklist ticking [42] task within the design phase. Consequently, design thinking and approaches are the observables of the design ethics LoA and problem definitions are their typed variables. Problem definitions can be perceived as cognitive models that explicate how the moral abstractions and use cases can be – and ought to be – integrated. [31-32]

**The GoA of AI ethics.** As figure 1. illustrates, the gradient of abstraction of AI ethics is a system describing hierarchical normative relations. It describes what is needed for the prescriptive information of the high dimensional LoA's to realize in the most concrete LoA. Therefore, it differs from GoA's describing phenomenal systemic wholes in how modifications in lower abstraction levels do not influence LoA's on a higher abstraction level but are caused by them. However, there are certain exceptions which have been described above and are depicted in figure 1. with bidirectional arrows.

[18] It would be more accurate to talk about prejudice discrimination, since discrimination in its broad meaning refers to distinguishing groups of information from a mass of data. Therefore, discrimination in its broad meaning is a non-separational function of AI.

The GoA depicts spheres of activities that are always relevant when AI ethics is considered. Therefore, it can be used to uncover necessary discourses within any context treating AI, be it a single team, organization or discipline, discourse of a single technology or a discourse about sector specific needs. Floridi describes in his paper considering the method of levels of abstraction that "specifying LoA's means clarifying from the outset the range of questions that a) can be meaningfully asked and b) are answerable in principle" [8; 315]. Due to the normative nature of ethics, I would add a third effect of the AI ethics clarification to be that it elaborates c) questions that ought to be asked.

In the next chapter, I will elaborate the LoA of embedded ethics and discuss how its relation to the other LoA's are to be taken into consideration when forming its conceptual understanding.

## 3 Embedded Ethics

As resembling the most concrete level of abstraction of AI ethics, embedded ethics should depict the concrete instantiations of the high-level goals and concepts of the higher abstractions. Therefore, the central questions for embedded ethics are what the ethically relevant parameters of AI technology are, and how can they be instantiated to AI. Existing research gives multiple distinct answers to these questions. I sum them as computational, norm-sensitive, information driven and systemic approaches.

The computational approaches emphasize the role of forming technical solutions to observed issues or high-level goal instantiations. I examine the concept of ethical governors as providing an example of a computational approach. The idea of ethical governors is laid down within a research field at present known as machine ethics. Its advocates bear concern for ethical implications of the increasing amount of highly complex automated systems in contemporary societies and especially in everyday social contexts. The concerns culminate to the question of how we can assure that the AI systems function in a morally desirable manner. [5-7]

The introduction of ethical governors is based on understanding the problem of machine functioning to be first and foremost computational, which is why it is perceived that the means of providing an answer should also be computational. [7] The idea of ethical governors is that there should be formed technical subsystems that would assess the outputs of the AI system and only authorize functions that are morally acceptable according to moral codes that are encoded in them. This way, for example opaqueness of the systems functioning would not matter.

There are three ways in how the ethical governors could be built. Top-down refers to encoding specified moral codes as a symbolic system. Bottom-up refers to using machine learning techniques to provide the artifacts the ability to develop moral codes by observing their environment. These both are seen to have restrictions, as the top-down method requires the designers to code acceptable functions to all possible situations and the bottom-up method may lead to the system learning contradictory or unwanted action patterns. A hybrid approach is suggested to redeem the concept of ethical governors, as it would combine the adaptive features of machine learning with the predictive nature of symbolic systems as prescribing restrictions to what is learnt. [2, 5-7]

The idea of ethical governors is tempting as it gives a promise of a carefree possibility to produce highly autonomous AI systems. [7] However, it appears as a superficial when considered through the AI ethics framework of this paper. It narrows technology as to focus on technical artefacts and the role of ethics as over-gluing action instructions to otherwise ready technology. This way it misses the larger ethical implications of technology development and its effects on societies. As the theories do not form proper understanding of how the aspects of deployment ethics relate to the encoded morality, they seem to concentrate on the how part of embedded ethics on the expense of what.

The approach to embedded ethics I refer to as norm-sensitive is from the IEEE document *Value Aligned Design* (2019). Its writers of the section concerning about embedding values into AI takes a step back from the technical aspects of AI and start the discourse with asking what are the parameters that ought to be embedded into AI. Its writers end up proposing social norms of communities as the instantiations of meaningful values. [2] This view focuses on the social interaction between humans and technology to form basis for the process of embedded ethics. Therefore, it differs from the computational approaches.

The writers of Value Aligned Design propose to use the same technical methods as is considered in ethical governors for the process of how to embed the recognized social norms. However, they add the requirement for redesign possibilities for actors of the community in which the AI systems are used. [2] This way the norm-sensitive approach supplements the computational approach by providing an answer to both questions, what and how, required in embedded ethics.

Nevertheless, when embedded ethics is examined as a continuum of the AI ethics framework, the embedding of social norms reflects merely a descriptive nature of AI ethics even though the other sections of the Value Aligned Design -document invoke normative measures to guide the development and deployment of AI systems. Therefore, it can be said that the deployment ethics LoA of Value Aligned Design and its LoA of embedded ethics are controversial, and that the embedded ethics of Value Aligned Design is too narrow. If the normative aspects of AI ethics are left to a higher abstraction level, there is a gap between values/codes and practical reality. [29]

As another observation, both forementioned approaches of embedded ethics, computational and norm-sensitive, assume that AI systems are highly autonomous and have tangible appearances. [2, 7] Therefore, they leave out many AI systems such as decision aid systems that may not have complex tangible appearances, but will have major implications for future work, healthcare, exercise of justice and application processing within any sector just to name a few examples. [43-44]

The information approach refers to Floridi & Taddeo's introduction of data ethics as a response to the oversimplification of the multifaceted moral nature of data and information processing of theories that simply focus on the functioning level of AI. [10] Data ethics provides knowledge of concrete variables for how the nature of AI as an information processing system may cause the emergence of moral implications. It for example focuses on the differences of computational language and human intentionality as mediators of morally relevant information. [3, 10] This way it also provides good insight about how certain high-level goals such as fostering of privacy should be examined for it to be best instantiated in AI as computational means. [10]

However, when observing AI ethics holistically, we need to acknowledge AI as forming from the interaction between humans and technical artefacts. And even though Floridi and Taddeo argue that the focus should be shifted to the lowest abstraction level of information formation, instead of the functioning of the technical artefacts, [10] they study information formation mostly through computational means and neglect human information processing as a component of technology.

The last approach towards embedded ethics that I examine is the one which I call the systemic approach. The theories of embedded ethics that I depict as part of this approach do not emphasize some aspect of AI as a pivotal area of embedded ethics but consider the process of successfully embedding ethical frameworks in AI as a sum of multiple factors. Ibo van de Poel represents this kind of approach in his article *Embedding Values in Artificial Intelligence (AI) Systems* (2020) where he sees the act of embedding values to AI to consist of perceiving organizational institutions as guiding the use of AI, users as intentional actors and the purpose of the use of the technical artefact and technical norms guiding the artefacts functioning. According to van de Poel, the combined effect of these parameters should be in place for the wanted value to be instantiated in AI.

Even though van de Poel brings many important aspects to the discourse, such as mental states of people and the need to count the immediate sociotechnical surrounding accordingly, he perceives human aspects of AI too narrowly. For example, he argues that the insufficiency of the operator in using the technical artefact in intended ways is in principle the user's fault. [3] The truth is that usability problems are often manifestations of not taking human factors into account in the design of the artifact. [4, 32, 45] Van de Poel perceives intentionality as the only important mental state of users, which causes him to neglect other aspects of human information processing, such as situational awareness or cognitive workload, which are vital in understanding human-technology interaction of AI systems. [4, 43]

Van de Poel also builds his theory on grounds that all AI systems evolve during their use. [3] This is true only for certain machine learning techniques such as reinforced and unsupervised machine learning. [18-19] The way an AI system adapts to its environment should be well limited in its design phase. [4]

As a last example of embedded ethics, I examine approaches provided by systems engineering research, which also depict a systemic approach. To do so, I combine point of views from Tony Gillespie as the author of *Systems Engineering for Ethical Autonomous Systems* (2019) and Eric Hollnagel and David D. Woods as the authors of *Joint Cognitive Systems* (2005). The first significant perspective for embedded ethics stems from the fact that AI technology is always used to fill a functional task(s) in a given process. [4, 11] When embedded ethics is understood as a continuum of the design ethics LoA, process analyzation can be perceived as a lower abstraction level continuum for problem definitions. Accordingly, the process analyzation has a central role in combining abstractions with the use case.

For joint cognitive systems, the design process starts from accurately defining the process that is to be fulfilled. It is then examined as a continuous flow of events (continuous control process) required to follow through the process. This way actions and lines of actions related to the process are perceived as being part of the same flow of

events, preventing from examining them as separate parts of the process. This type of understanding focuses on how the joint system of humans and technical artefacts can stay in control of a complex process[19]. It requires examining the human-technology interaction while not emphasizing the role of one or the other. [11] Tony Gillespie emphasizes that in addition to considering staying in control of the process, the functions included in the flow of events should be considered through a decide to delegate method, which seeks to ask if there are ethical implications of delegating functions for AI. As an example, Gillespie argues that on ethical grounds, an autonomous weapon system's control process can be automated at any stage, excluding the decide to act function, which is perceived as the point of making a possible human harming decision. [4]

The systems engineering point of view provides us process definition as the observable for embedded ethics as it makes it possible to examine ethical aspects of the intended AI technology in the holistic manner that the AI ethics framework invokes. The typed variables for the process definition are the required flow of events for goal attaining and the decide to delegate examination. The flow of events can be further divided as functions carried out by the joint cognitive system of human (plural) actions and technical artefacts. In addition, the decide to delegate process can be divided to case sensitive ethical implications that connects the moral standardization of the deployment ethics phase and the use case context.

The exact nature of the ethical implications related to the decide to delegate examination must be contextually defined. For ethics has to do with the metalevels of action, meaning that there are no ethical actions per se, but the morality of an action is based on the context, agents, their roles, relations, and intentions. [24] For example, stabbing a burglar with a kitchen knife as self-defense versus stabbing someone in pursuance of robbery. Same action, different moral implications. For this reason, we should avoid about talking of general ethical actions or implications within embedded ethics but produce understanding about how to analyze the ethical aspects of the process and its context as in relation to the (AI) deployment ethics.

## 4 Conclusions

The purpose of this paper was to put forward an initial conceptual framework for AI ethics that considers distinct levels of abstractions and discuss the connection between the process of embedding ethical frameworks to AI systems and the larger framework of AI ethics. Both objectives were fulfilled and the compiled conceptual framework consisting of (AI) deployment ethics, (AI) data ethics, (AI) governing ethics, (AI) design ethics and (AI) embedded ethics proved as an insightful tool for providing rigor to discussions about AI ethics and forming practical steps for ethical design of AI. The observations made in this article highlight the need for such discussions to form understanding of the landscape related to AI ethics and avoiding its oversimplifications as well as the oversimplifications of the process related to embedded ethics.

[19] Compare to the concept of meaningful human control as a HCI grand challenge. [46]

As the examination of existing approaches of embedded ethics pointed out, the deployment ethics stage and embedded ethics stage of AI ethics are often in contradiction even if they are considered in the same document. The conceptual framework can be used as a pragmatic tool to help form narratives about how to embed ethical frameworks to AI systems. It can also be used to study parts of AI ethics without rendering the whole as considering only questions related to that aspect, which is unfortunately common. The framework also makes it possible for researchers, policymakers, and organizational actors – who inevitably look at issues emphasizing the angle related to their position – to contemplate how their aspect is interrelated to the whole.

When perceived as part of the AI ethics framework, the process of embedded ethics must consider the basic features of technology and more precisely basic features of automated systems design as it being a parallel technology. This is often lost when embedded ethics is perceived as an extension of data ethics, or as the act of solving a single evident issue related to AI deployment. Other obscuring perceptions include that embedded ethics is often approached through a presumption of high-level autonomy for the AI system as well as it is presumed that the ethical or moral aspects of the systems decision making could be examined separately from the other system behavior.

As the behavior of the embedded ethics LoA is operationalizing the abstractions provided on the higher levels of AI ethics, its observable(s) should connect a systemic understanding of AI technology and use case sensitive ethical knowledge. As the problem definition provided in the design ethics phase should count for refining the goal of the automatable process to reflect the moral standardization, it is left for the embedded ethics phase to examine the functions within the process. Therefore, process analyzation and the decide to delegate examination are observables of embedded ethics.

This article concentrated in creating the frame for discourse about embedded ethics, and therefore further studies should pursue to refine details within it. It is not a simple task, for the complexity of the field necessitates interdisciplinary approaches, which combine engineering-, natural-, behavioral- and social sciences and humanities. It is not justifiable nor feasible to thrust this whole on engineering disciplines as is implicitly or explicitly often suggested in AI research and national AI strategies.

## References

1. European Commission.: Ethics guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
2. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. First Edition, IEEE (2019).
3. van de Poel, I.: Embedding Values in Artificial Intelligence (AI) Systems. Minds & Machines 30, 385–409 (2020). https://doi.org/10.1007/s11023-020-09537-4.
4. Gillespie, T.: Systems Engineering for Ethical Autonomous Systems. SciTech Publishing, London (2019). ISBN-13: 978-1-78561-372-2.
5. Arkin R.: EMBEDDED ETHICS - “Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture.” Proceedings of the 2008 3rd ACM/IEEE

International Conference on Human-Robot Interaction (HRI), Amsterdam, Netherlands, March 12 -15, 2008, IEEE, pp. 121–128, (2008).
6. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, New York (2008).
7. Anderson, M., Anderson, SL. (eds.): Machine Ethics, Cambridge University Press, New York (2011).
8. Floridi, L.: The Method of Levels of Abstraction. Minds & Machines 18, 303–329 (2008). https://doi.org/10.1007/s11023-008-9113-7.
9. Johnson, D. G., Miller, K. W.: Un-making artificial moral agents. Ethics and Information Technology, *10* (2), 123–133, (2008). https://doi.org/10.1007/s10676-008-9174-6.
10. Floridi L., Taddeo M.: What is data ethics? Phil. Trans. R. Soc. A.3742016036020160360 (2016). http://doi.org/10.1098/rsta.2016.0360.
11. Hollnagel, E., Woods, D.D.: Joint Cognitive Systems – Foundations of Cognitive Systems Engineering. CRC Press, Taylor & Francis Group, London (2005). ISBN-13: 978-0-367-86420-0.
12. Schomberg, R.V. (Ed.): Towards Responsible Research and Innovation in the Information and Communication Technologies and Security Technologies Fields. Luxembourg: Publication Office of the European Union. (2011). http://ec.europa.eu/research/science-society/document_library/pdf_06/mep-rapport-2011_en.pdf.
13. Jasanoff, S.: Future Imperfect: Science, Technology and the Imaginations of Modernity. In: Jasanoff S., Kim, S. (eds.): Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power. The University of Chicago Press, London (2015).
14. Floridi, L. Soft Ethics and the Governance of the Digital. Philos. Technol. 31, 1–8 (2018). https://doi.org/10.1007/s13347-018-0303-9.
15. Cabinet Office of Japan: Society 5.0. https://www8.cao.go.jp/cstp/english/society5_0/index.html, last accessed 2021/3/12.
16. Samoili, S., Lopez C., M., Gomez G. E., De Prato, G., Martinez-Plumed, F. and Delipetrev, B.: AI WATCH. Defining Artificial Intelligence. EUR 30117 EN, Publications Office of the European Union, Luxembourg, (2020). ISBN 978-92-76-17045-7, doi:10.2760/382730.
17. Minsky, M. L.: Computation: Finite and Infinite Machines. Prentice-Hall, Englewood Cliffs, NJ (1967).
18. Norvig, P., Russel S.: Artificial Intelligence – A Modern Approach, Third edition. P., Pearson, Boston (2010).
19. Pietikäinen M., Silven, O.: Tekoälyn haasteet: koneoppimisesta ja konenäöstä tunnetekoälyyn. Oulun Yliopisto, Oulu (2019). ISBN: 978-952-62-2482-4.
20. Bostrom, N., Yudkowsky E.: The ethics of artificial intelligence. In: Keith Frankish & William Ramsey (eds.), The Cambridge Handbook of Artificial Intelligence (pp. 316-334). Cambridge University Press, Cambridge (2014). doi:10.1017/CBO9781139046855.020.
21. Kostopoulos, L.: Decoupling Human Characteristics from Algorithmic Capabilities. The IEEE Standards Association (2014).
22. Beauchamp, T., Childress J.: Principles of Biomedical Ethics. 7th ed. Oxford University Press, New York (2013).
23. Hansson, S. O.: Theories and Methods for the Ethics of Technology. In: The Ethics of Technology. Sven Ove Hansson (ed.). Rowman & Littlefield; London (2017). ISBN:978-1-7834-8658-8.
24. Hallamaa, J.: Yhdessä Toimimisen Etiikka. Gaudeamus, Helsinki (2017).
25. Westermarck E. The origin and development of the moral ideas (Vol. 2). Macmillan, London (1908).

26. Habermas, J.: The theory of communicative action: Vol. 1, Reason and the rationalization of society. Heinemann, London (1984).
27. Velasquez, M., Andre, C., Shanks, T., Meyer, M.: What is Ethics? Markkula center for Applied ethics (2019). https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/what-is-ethics/, last accessed 2021/3/12.
28. Floridi, L., Cowls, J., Beltrametti, M. et al.: AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds & Machines 28, 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5.
29. Hallamaa, J., Snell, K.: Ethics in AI research – what and how? Finnish Center for Artificial Intelligence (2020). https://fcai.fi/eab-blog/2020/9/4/ethics-in-ai-research-what-and-how, last accessed 2021/3/12.
30. Gregor, S., Jones, D.: The Anatomy of a Design Theory. Journal of the Association for Information Systems. Vol. 8, Is. 5, Article 2, pp.312-335, May 2007.
31. Simon, H. A.: *The sciences of the artificial*. M. I. T, Cambridge (MA) (1970).
32. Saariluoma, P., Cañas, J., Leikas, J.: Designing for life. MacMillan, London (2016).
33. Franssen, M., Gert-Jan L., van de Poel, I.: Philosophy of Technology. In: The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/fall2018/entries/technology/ (2018).
34. Saariluoma, P., Oulasvirta, A.: User Psychology: Re-assessing the Boundaries of a Discipline. Scientific Research, vol.1, No.5, 317-328 (2010).
35. Homepage of Black in AI: https://blackinai.github.io/#/, last accessed 2021/3/12.
36. European Commission: White Paper on Artificial Intelligence: a European approach to excellence and trust. Brussels (2020).
37. ETAIROS -project homepage: https://etairos.fi/en/front-page/, last accessed 2021/3/12.
38. AIGA -project homepage: https://des.utu.fi/projects/aiga/, last accessed 2021/3/12.
39. Canca, C.: AI & Global Governance: Human Rights and AI Ethics – Why Ethics Cannot be Replaced by the UDHR. United Nations University, Center for Policy Research (2019).
40. Isaac Ben-Israel, Cerdio, J., Emal, A., Friedman, L., Ienca, M., Manteleroe, A., Matania, E., Muller, C., Shiroyama, H., Vayena, E.: Towards Regulation of AI Systems. Council of Europe (2020).
41. Buolamwini J., Gebru, T.: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research 81:1–15 (2018).
42. Vakkuri, V., Kemell, K., Abrahamsson, P., Franch, X., Männistö, T., Martínez-Fernández, S. et al.: Implementing Ethics in AI: Initial Results of an Industrial Multiple Case Study. Technology, F. o. I. 2019, Springer (2019).
43. Ruff, H., Narayanan, S., Draper, M.: Human Interaction with Levels of Automation and Decision-Aid Fidelity in the Supervisory Control of Multiple Simulated Unmanned Air Vehicles. Presence: Teleoperators and Virtual Environments 2002 11:4, 335-351 (2002).
44. Jordan H., Berk, R.: Machine Learning Forecasts of Risk to Inform Sentencing Decisions. Federal Sentencing Reporter (2015). 27. 222-228. 10.1525/fsr.2015.27.4.222.
45. Norman, D.: The design of everyday things: Revised and expanded edition. Basic Books (AZ) (2013). ISBN: 9780262525671.
46. Chairs C. S., Salvendy, G., et al.: Seven HCI Grand Challenges, International Journal of Human–Computer Interaction, 35:14, 1229-1269, (2019). DOI: 10.1080/10447318.2019.1619259.