A Framework for Socio-Developmental Ethics in Educational AI

Ilkka Tuomi
Meaning Processing
ilkka.tuomi@meaningprocessing.com

Abstract

In recent years there have been many attempts to create ethical frameworks for AI. Theoretical concepts, such as privacy, fairness, transparency, explainability, responsibility, risk, and trustworthiness have been used as key elements in these frameworks. The use of these concepts is often justified by their wide use in similar frameworks and guidelines but does not seem to result from any coherent shared theoretical foundation. Educational and developmental theories and research have so far had little impact on ethical debates but become important when AI is used in education and learning (AIEd). A socio-developmental view on ethics naturally emerges in the educational context, and the paper shows that it has important implications also beyond the education sector. This paper describes an ethical framework structured in three thematic domains: agency, social fairness, and justified choice, that links AI with theories of education and human development, opening new ways to understand ethics of AI and the social and technical challenges and opportunities in AIEd.

Keywords: artificial intelligence, ethics, education, capability development, learning

1. Introduction

In the last five years there has been an avalanche of reports and reviews of ethical AI guidelines. An influential early initiative was the Asilomar Conference on Beneficial AI, held in 2017, but over hundred national, international and commerce-driven ethics reports and a large body of research literature has been produced since then (EC, 2021; Floridi & Cowls, 2021; Hagendorff, 2019; Jobin et al., 2019). Among the most globally visible outcomes of these activities have been the AI ethics guidelines developed by the EU High-Level Expert Group (AI HLEG, 2019), the OECD AI Principles (OECD, 2019), and the UNESCO Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021).

An often acknowledged aim of these guidelines and recommendations has been to define ethical frameworks based on expert consensus. In practice, this has meant that reports have been based on a careful analysis of existing ethical guidelines and frameworks. For example, Singapore's model AI framework builds on 12 ethical principles commonly found in existing AI guidelines. These are accountability, accuracy, auditability, explainability, fairness, human centricity and well-being, human rights alignment, inclusivity, progressiveness, responsibility, accountability and transparency, robustness and security, and sustainability (IMDA & PDPC, 2020). Floridi and Cowls (2021), in search for unifying AI ethics principles, find five common principles, beneficence, non-maleficence, autonomy, justice, and explicability. The EU guidelines, in turn, were to an important extent based on the Charter of Fundamental Rights of the EU that defines relations between European institutions and its citizens, adding explainability as a novel element relevant for AI. AI ethics researchers have also tried to formulate principles relevant for education based on meta-level studies about ethics guidelines (e.g. Adams et al., 2021).

A methodological challenge in these reports, guidelines and frameworks is that their rather eclectic combinations of concepts are essentially based on syntactic similarity. It is far from clear that experts would agree on a shared interpretation of the used terms, also because different ethical traditions understand these terms in different ways. Concepts such as fairness, justice, beneficence, autonomy, and privacy have not been discussed in any theoretical sophistication in many of these guidelines or frameworks, perhaps partly because they are expected to be "policy-relevant" and aimed for practical use in system design. As a result of this theory-avoidance, the existence of large bodies of empirical and theoretical research on concepts such as privacy, fairness, risk, transparency, and trustworthiness is rarely noted.

In educational applications of AI, this lack of coherent theoretical foundation becomes a problem. What is autonomy or privacy for a child in compulsory education? What means equality or equity in differentiated education based on educational performance, innate aptitude, or disability? What is explainability and who can ask for explanation? As soon as general principles are operationalized, it becomes clear that there are many context-dependent

URI: https://hdl.handle.net/10125/103386
978-0-9981331-6-4

interpretations and, as Mittelstadt (2019) has suggested, a coherent interpretation may only be possible if the community already shares the same system of values. In effect, the multidisciplinary initiatives that have involved mainly ethicists and computer scientists, have fallen back to principles derived from bioethics and universal human rights declarations and their legal formulations. Ethics of education or theories of human development and learning have played a marginal role in these discussions (Aiken & Epstein, 2000; Holmes et al., 2021).

A different point of view emerges when these key ethical concepts are put in a developmental and cultural context. Privacy has important social and cultural functions (Roessler & Mokrosinska, 2015; Sax, 2018). Autonomy is deeply embedded in social contexts that make individuals competent agents in specific socio-technical settings, and therefore closely related to human capabilities (Nussbaum, 2000; Sen, 1993). The limits of explainability have been known for millennia, and different cultures have different ways in explaining causes and effects and in attributing agency and responsibility. For example, Pyrrhonian skepticism was to an important extent based on the observation that the chain of explanations is infinite (Barnes, 1990), and Evans-Pritchard's (1976) ground-breaking work on social anthropology was largely focused on different forms of causal explanation. In general, ethical theories assume competent adults who make rational ethical choices, neglecting developmental questions central to pedagogy and developmental psychology.

When AI is put in an educational context, many of these concepts appear in a new light. A key starting point becomes the question of the development of human agency. To the extent that education aims at the development and growth of human agency, the capabilities that underpin agency are key concerns for ethics. In contrast to ethical traditions where ethics is about social contracts between fully competent adults, a capability-based approach highlights factors that enable developing individuals to become agents capable for ethical action in their social and technological environments.

2. What is AIEd for?

Educational philosophy and theory of education are vast areas of study, and the objectives of education are highly contested within and across cultures. It is, however, important to note that without a clear characterization of the purpose of education, the search for "ethics of AI" in education is futile. Important ethical concepts, such as beneficence, require shared understanding about what counts as "benefit," "improvement," or "progress." In the context of AIEd, such terms can only be interpreted if we have a clear understanding about the aims of education.

Many existing AIEd systems have been influenced by the instructional strategy of mastery learning. As described by Bloom in the 1960s (cf. Guskey, 2012), mastery learning builds on the idea that students need to achieve a sufficient level of mastery in prerequisite knowledge to be able to move ahead and learn subsequent information. Different students require different amounts of time to get to the level of mastery. Bloom (1984) further argued that the mastery learning approach combined with individual tutoring results up to two standard deviations increase in learning gains, measured as the relative improvement of attainable increase in test results. Later reviews have reduced the learning gains towards one standard deviation and claimed that this is similar to gains attainable using intelligent tutoring systems (VanLehn, 2011).

Mastery learning has been an important starting point for many AIEd systems, and it at least implicitly underpins the common belief that there is great potential to use AI for personalized instruction (Blikstein et al., 2022). AIEd is often contrasted with teacher-led lecturing and K-12 classrooms where information flows from the teacher and the textbook to the head of students. Adaptive learning environments based on AIEd have been viewed as a move beyond the "one-size-fits-all" approach to learning (Baker & Smith, 2019). In general, many influential intelligent tutoring systems and adaptive learning environments are based on the assumption that computers can be used to sequence and pace learning materials according to the learning needs of individual students (Holmes et al., 2019).

The instructional strategy of mastery learning has been particularly influential in mathematics and science education, where conceptual structures are built on hierarchical abstraction. To know how to solve an equation, one needs to know addition, multiplication, and fractions. Many leading AIEd systems, therefore, have built extensive domain-specific knowledge models, where knowledge components form complex hierarchical structures, with associated learning materials and knowledge mastery tests. The criterion of learning success in this approach has most often been test score, typically related to curriculum requirements (Holmes & Tuomi, 2022).

Among educational researchers, such an objective is usually viewed as a too limited one. Biesta (2010), for example, suggests three different domains of purpose for education. One is qualification. Here education is concerned with the transmission and acquisition of knowledge, skills, dispositions and understandings that qualify people to do certain things. Another important purpose is socialization. Through education, people become part of existing traditions, cultures, ways of doing and ways of being. This is sometimes called the "hidden curriculum." The third function and domain of purpose of education Biesta calls subjectification. This has to do with how education contributes to how we can exist as human subjects.

There is also an important difference between education and learning. According to Biesta:

> "Education, to put it differently, is not designed so that children and young people might learn – people can learn anywhere and do not really need education for it – but so that they might learn particular things, for particular reasons, and supported by particular (educational) relationships."

This means that education is characterized by describing the purposes of education (curriculum) and relationships that organize the process (pedagogy). Education, according to Biesta, is not only practice characterized by the presence of purposes: it is practice constituted for its purposes.

Due to historical factors such as globalization and the increasing economic importance of innovation and knowledge creation, inter-generational transfer of accumulated knowledge has declined in importance, and competence-based views have gained visibility in defining curricula. It is now often claimed that knowledge is becoming obsolete at accelerating speeds, and children need 21st century skills and competences (Tuomi, 2022). The assumption has been that, instead of domain specific content, education must develop more general capabilities, such as problem solving, critical thinking, and communication and collaboration skills. At the same time, life-long learning has become a key element for educational policy, with a similar focus on domain-independent objectives for learning. For example, the influential Delors report (UNESCO, 1996) defined four pillars for life-long learning: learning to know, learning to do, learning to be, and learning to live together.

Table 1 re-frames the four UNESCO pillars as the development of epistemic capacity, development of personal agency, development of inter-personal capabilities, and as social progress. Epistemic capacity is related to UNESCO's "learning to know," whereas the development of personal agency and interpersonal capabilities are related to "learning to do" and "learning to be." Biesta's "qualification" dimension captures elements from epistemic capacity and generic competences related to the development of personal agency.

It should be noted that personal agency depends also on the capability to mobilize social resources. The separation of competences related to personal agency and interpersonal capabilities, therefore, mainly reflects the common assumption that the subject of learning and education is an individual person. This conventional unit of analysis, of course, is far from obvious in social and progressive learning theories, in organizational learning and knowledge creation research, and for research on distributed and augmented cognition.

Table 1 groups some common research constructs under each of its pillars. Many of these have also been studied in AIEd research (du Boulay, 2019), although the traditional focus has clearly been on the development of epistemic capacity and, more specifically, in knowledge development (Holmes et al., 2019; Holmes and Tuomi, 2022; Luckin, 2018).

Table 1. Developmental pillars for life-long learning

Domain dependent	Domain independent		
Development of epistemic capacity	Development of personal agency	Development of interpersonal capabilities	Social progress
• skill • knowledge • experience	• meta-cognition • self-efficacy • curiosity • motivation • interest • emotional control • executive function • self-reflection • grit	• communication • collaboration • "emotional intelligence"	• cohesion • participation • social and economic renewal • inter-cultural interaction

The above discussion highlights the point that the traditional focus on epistemic capacity in AIEd research is too narrow for a coherent ethical framework. Education is not only teaching for test or accumulation of knowledge, and informal and non-formal learning are becoming increasingly important for individual and social development. It is in this broader context where coherent conceptual foundations for ethics of AI in education can be found.

3. Elements of the ethical framework

Given the above discussion on the broad aims of education, one can ask what the elements of an improved ethical framework would be. I suggest three thematic starting points around which such a framework can be built: agency, social fairness, and justified choice. Ethics spans a vast area of human thought, and any attempt to structure this field is bound to generate rich and deep discussions. The aim here is to do exactly that and facilitate deeper discussion than before.

The proposal is to use these three top-level concepts to organize the various challenges of ethical use of AI and data in education. The aim is to organize ethics so that we can discuss relevant principles that need to be applied in the educational domain. A perhaps surprising outcome is that ethics of AI in education is not a special case of a more general ethics of AI. On the contrary, ethics of AI needs to be based on developmental concerns, making "general" ethics of AI a special case of ethics of AI in education.

Ethics is a very complex field, and we need to make it as simple as possible, but not simpler. Much of this complexity could be avoided by building on principles articulated in the various human rights declarations and conventions. This is the strategy followed in many of the existing human-centric frameworks. In the educational sector, such a starting point would, however, have problems. For historical reasons, these declarations mainly focus on the relations between governments and citizens and do not cover all the ethical challenges that are important. They have their roots in the experiences of the Second World War, and many of their statements can be understood as a post-war review of what should never happen again. As their essence is in making clear the limits of the use of state power, they can easily be translated into law, and, in fact, the EU Charter (Charter of Fundamental Rights of the European Union) and the European Convention of Human Rights are legally binding for the participating states. They express shared values, but in their proper domains of validity, ethical problems easily become reduced to questions about following existing laws. An example of this is the European General Data Protection Regulation (GDPR), which now in practical and operational terms defines what privacy and identity mean in Europe and beyond.

In the educational domain, laws remain important and particularly so in compulsory education that is tightly regulated and, in many countries, a part of the public sector. The relations between teachers, students, and other stakeholders, however, are complex human relations where law provides only generic constraints. This is why guidance is needed that goes beyond reiterating existing rules and regulations. In the ethics of education, both human development and contested social objectives play central roles. Whereas good public governance often aims at predictability, universality, and neutrality, education aims at realizing and expanding idiosyncratic human potential. It also transfers culture and its values across generations and facilitates social, cultural, and economic renewal.

As the development of human agency is a fundamental objective of education, the impact of AI on learner and teacher agency, therefore, provides a useful starting point to organize ethical debates.

3.1. Agency

Agency refers here to an individual's capability for action and characterizes what a person is free to do and achieve in pursuit of goals and values the person regards as important (Sen, 1993, 2009). In the context of ethics, agency, therefore, subsumes concepts such as autonomy and self-determination, fundamental freedoms, and responsible action. A person can act in a responsible way only if the person can choose among ways to act and is able to understand their consequences. Education has an important role in generating alternatives for action, enabling informed action and, more broadly, in developing capacity for ethical behavior. Increasing freedom for achieving one's life goals comes with increasing possibility for responsibility.

The concept of agency also highlights a fundamental problem in many ethical traditions. In consequentialist and deontological ethics, the subjects of ethical choice are typically assumed to be rational and competent adults, with various levels of moral education and character. In such settings, social contracts can be agreed based on shared universal principles and knowledge (e.g., Rawls, 1999). In a more developmental view, such a history- and context-independent image of equal competent partners in an ethical endeavor begs the question how, exactly, they developed the necessary competences for rational choice and ethical judgment.

Ethical frameworks often understand agency in terms of autonomy. In a literal sense autonomy is the ability to live under your own laws. The concept of autonomy is therefore closely related to freedom, liberty, free will, and oppression (cf. Kane et al., 2021). In philosophical and political literature, the concept of positive freedom is used to denote the capacity to act upon one's free will, whereas negative freedom refers to freedom from external coercion.

Freedom, thus, refers to an imagined set of alternative actions and the capability to imagine such alternatives. From a capability perspective, education has an important role in expanding the capacity to imagine possible futures and alternatives. Autonomy can be interpreted as the possibility to realize one of these imagined alternatives, based on the chooser's evaluation of the alternatives. An autonomous agent, therefore, makes a choice that reflects her preferences, including values and values about values. Because of this, autonomous acts reflect and express value choices, and the chooser can be held morally accountable for the act.

Autonomy is a concept that is often used in ethical discussions on AI and data use without linking it to pedagogy and individual development. Piaget (1932), for example, argued that at the age of seven or eight, the moral thinking of a child moves from heteronomy to autonomy. Heteronomy can be characterized as the "morality of obedience," where a child uses authoritative rules as the basis for her moral assessments. In this stage, no particular reasoning for moral judgments is needed. Acts are considered moral if they follow the rule, and, for example, the intentions of the actor do not influence moral judgments.

According to Piaget, this stage is followed by the development of autonomous moral reasoning. Piaget emphasized that autonomy develops in relation with social cooperation. On a more philosophical level, the interpersonal nature of agency underpins dialogical ethics. For example, Bakhtin built his theory of ethical action on the idea that humans are fundamentally open, unknowable and continuously changing. As a result, Bakhtin located responsibility and "answerability" at the center of ethics (Bakhtin, 1993).

In child psychology, the development of identity, self-image, self-efficacy, and growth mindset are closely linked with the development of autonomy. Agency, in turn, has been a widely used concept in childhood studies over the last two decades, sometimes understood from an individualistic perspective (Sutterlüty & Tisdall, 2019). This had lead to claims that that children are social actors who need similar freedoms and rights as the other members of society. In ethics, identity, in turn, is often associated with human dignity. Much of the contemporary debate on human dignity has centered on limits on coercion and self-determination at the individual level (e.g., body, health, human trafficking, assisted suicide), but also on identity as membership in self-determined social groups (e.g., religion, sexual identity).

An important aspect of agency can also be understood as the development of affective and cognitive self-regulation. Piaget and Inhelder (1979, p. 159) summarize this noting:

> "It is impossible to interpret the development of affective life and of motivations without stressing the all-important role of self-regulations, whose importance, moreover, all the schools have emphasized, albeit under various names."

3.2. Social fairness

Another key element in the proposed framework is social fairness. Strictly speaking, in the proposed framework fairness is always social; we note this explicitly in an attempt to reduce confusion.

Fairness has been a central topic in AI research over the last years, to a large extent because of many cases of unacceptable bias that have been extensively discussed in the media. Many existing ethical frameworks for AI, in fact, can be viewed as attempts to alleviate public worries about the impact of AI on the future of jobs and the potential unfairness of algorithmic decision-making. Although the concept of fairness is widely used, it is not always clear how it should be interpreted. In a limited computational sense, there have been many proposals to establish design principles, processes, and systems that would improve fairness in AI systems, but it has also been noted that computer scientists use at least 21 different formal definitions of algorithmic fairness (Mitchell et al., 2021; Kizilcec et al., 2023), and distinguish several allocative and representational harms (Wang, 2022).

In contrast to much of extant research on algorithmic fairness, the proposed framework takes a more socio-cultural view on fairness. Fairness requires that we understand the society through the duality of diversity and equality. Fairness subsumes questions concerning diversity, inclusion, representation, and the right to classify and profile people. It also relates to the complementary questions about equality and equity, including non-discrimination, allocation of opportunity, and the treatment of socially distinguished minorities and groups that are deemed to be vulnerable. Fairness, therefore, is deeply linked to the structures of power and social categories in use (Bowker & Star, 1999). Equality somewhat paradoxically assumes that we locate people in social groups. Different cultures use different ways to do this (Douglas & Wildawsky, 1982; Lakoff, 1987). The fact that categories such as race, gender, religion, ethnic origin, and, for example, sexual orientation, are listed in human rights declarations and conventions, simply highlight the problematic historical use of these groups.

Fairness, therefore, is fundamentally about how we as individuals and through social institutions treat people as representatives of social groups. The groups are formed based on available information, and fairness therefore reflects existing knowledge structures, informational classifications, and observable data. Data-driven machine learning systems generate new fine-grained social categories in ways not possible before, and history does not always inform us whether the use of these categories is appropriate or not.

In everyday speech, we often say that an individual has been treated unfairly, without any regard to social groups or categories. Such "unfair" treatment reflects our broken expectations about acceptable behavior, and as they are about argument and justification, they are discussed in the next section. A purely criminal act is rarely described as "unfair," but acts of God and human acts that are covered only by normative cultural rules are often considered to be unfair. Here we simply maintain that fairness is a social concept, and the types of unfairness described above should more properly be understood as a lack of justification for choices and acts.

3.3. Justified choice

What counts as evidence, and whose evidence counts, is a foundational question for ethics. Although it has rarely been noted, ethics rests on an epistemic foundation. Ethical choices are about distinctions we make, and these are based on observation, data, and knowledge. An important function of education is to develop our capability to make distinctions that are relevant to us and to our socio-cultural environment. In important ways, the development of human capabilities is about the increasing sophistication of making distinctions relevant for specific domains of knowing.

All ethical choices require justification. Much of our daily behavior is automatic, and we may struggle in finding an explanation for it. Ethics, in contrast, requires addressing the question why. Ethics, therefore, requires Aristotle's "final causes," explicitly rejected in science since Newton.[1] Such final causes are also central to information systems design theories that aim at changing the world, based on judgements on what is good or bad for a given design situation (Hanseth & Lyytinen, 2008).

As Habermas (1993) argued in discourse ethics, justification does not have to be based on universally agreed value statements. More important is that the argument is coherent and there is a reasonable attempt to negotiate disagreements. Justified choice, therefore, can be understood as choice that has a coherent justification. This, however, requires shared understanding about what counts as a coherent argument.

Democracy is one way to make common choices in a society where people have many incompatible value systems. As long as the participants can agree on a shared process, they can disagree on how to value existing evidence and alternatives. From the point of view of social fairness, a minimal requirement is that different voices are heard. Modern concepts of democracy, therefore, emphasize participation. In technology design, participatory design practices, including co-creation and user-centric design, have also this ethical dimension. In addition to bringing in various sources of knowledge, they also bring in different stakeholder perspectives and value systems. With somewhat lesser user involvement, also value sensitive design aims for this (e.g., Umbrello & van de Poel, 2021).

Ethics, therefore, becomes a process, instead of a set of content-oriented principles and associated checklists. A challenge in such participatory practices, however, is that as the various value-systems involved may not be compatible, the outcomes may reflect only surface-level agreement on terms, something that this paper argued was the case in many existing AI ethics frameworks. The three domains of ethics: agency, social fairness, and justified choice, are proposed to structure discussions on ethics of AI and AIEd in a way that admits the lack of consensus among the participants, at the same time allowing them to clarify agreements and disagreements. In information systems research, this naturally leads to questions about technologies that can mediate negotiation, collaboration, and argumentation. As such, this is nothing new and, as Star (1992) noted long time ago, collaboration is possible also without consensus.

[1] The formal structure of Newtonian physics only embeds effective, formal, and material causes, and describes "how" things happen, not "why" they happen. This is what Newton means with "hypotheses non fingo" in the Principia. Computer algorithms have the same formal structure, which makes explainable AI a challenging problem (cf. Rosen, 1985).

4. Operationalizing the framework

The framework presented above is an abstract one, and a begs the question how it can be made actionable. The answer depends on our definition of the actor. This, in turn, depends on the underpinning models of innovation and action.

The proposed framework is informed by science and technology studies (e.g., Bijker et al., 1987), research on the social bases of innovation and technology use (e.g., Brown & Duguid, 1991), and, in particular, the view that technology-enabled social practices are articulated in multi-stakeholder processes of learning and knowledge creation (Engeström, 1999; Tuomi, 2002). This means that there are many actors involved in technology articulation and adoption. The operationalization of the framework, therefore, requires addressing several interlinked stakeholder groups. In educational settings these stakeholders include technology developers and providers, educational researchers, educators, administrators, students, and, for example, policymakers, parents, and the general public.

Social and historical studies on technology and innovation have noted that unintended consequences are common, and the future uses of technology are difficult to predict. Consequentialist ethics, in particular, assumes that it is possible to assess future benefits and the consequences of acts. In contrast, the proposed framework builds on the view that innovation makes the future inherently unpredictable (Tuomi, 2012). Instead of universal ethical maxims and checklists derived from these, the proposed framework, therefore, adopts a capability-based approach (Sen, 1993, 2000). In this view, instead of universal ethical principles, we can have adequate capability to act in an ethical way in specific and often idiosyncratic contexts where decisions and actions are needed. The fundamental unknowability of others is a key starting point in dialogical ethics (e.g., Bakhtin, 1993), and here we simply extend it to the outcomes of human action and—potentially—also non-human agentic action (Newman et al., 2019).

The required capabilities can be actualized using rubrics that describe competences that the various stakeholders need for ethical action. Some examples of such rubrics are outlined in the forthcoming European Commission guidelines for the ethical use of AI and data in education and training (to be published in October 2022).

For ethical design, also the ethical rationale needs to be documented in a form that allows the stakeholders to negotiate their potentially incompatible values. Whereas existing frameworks include checklists for ex-ante assessment and governance models for ex-post assessment of ethical impact, in the proposed framework competence development prepares the participants for ethical use and system development, and ongoing "ex-post" observation enables the participants to develop their understanding of ethical challenges that need to be addressed. Technology development and deployment, therefore, are understood as an ongoing learning process. Conceptually, such a learning process underpins also Dewey's model of experiential learning (cf. Miettinen, 2000).

5. Alternative kernel theories in ethical frameworks

The table below summarizes the proposed framework (SDF) and contrasts it with three well-known AI ethics frameworks. The first of these is the "Five Principles" framework (5P), originally developed in the AI4People initiative and articulated in Floridi and Cowls (2021). This framework consolidates six well-known AI ethics frameworks in an attempt to reduce confusion and find common themes in the selected documents. The second framework is the one developed by the European High-Level Expert Group on AI (AI HLEG). The AI HLEG guidelines (AI HLEG, 2019) derive ethical principles from human rights, and translates these into seven generic ethical requirements, which, in turn, are operationalized as the Assessment List of Trustworthy AI (AI HLEG, 2020). In contrast to these two AI ethics frameworks, the third one was developed specifically for education in a two-year project organized as The Institute for Ethical AI in Education, located at the Buckingham University in the U.K. (IEAIE, 2021). The IEAIE framework does not include an explicit set of ethical principles; instead, it distinguishes nine objectives that ethical AIED should address.

Table 2. Four alternative AI ethics frameworks

	5P (2021)	AI HLEG (2019)	IEAIE (2021)	SDF
Ethical domains / objectives	Beneficence Non-maleficence Autonomy Justice Explicability	Respect for human autonomy Prevention of harm Fairness Explicability	Achieving educational goals Broader forms of assessment Organizational improvement Equity Learner autonomy Balanced data privacy Transparency and accountability Informed participation Ethical design (participation, diversity)	Agency Social fairness Justified choice
Main target groups	Policy developers, best practice developers	System developers	Educational organizations	Researchers, policy developers, stakeholders
Kernel theory	Bioethics	Human rights	Educational practice	Capability-based model of development
Operationalization	Recommendations	Checklist questions	Checklist questions	Competence rubrics, process orchestration

6. Conclusion

A basic assumption that underpins the present paper is that AI will have a profound impact on learning, education, and social processes of knowledge creation. The use of AI in education will have important effects on the development and expression of human agency. Based on the concept of agency, many challenges discussed in AI ethics, such as autonomy, freedoms, privacy, transparency, and responsibility, can usefully be linked to theories of learning and development. Similarly, recognizing the social nature of fairness and the informational bases that underpin it, allows us to link fairness with information system design.

Today we are far from the pre-industrial world where, according to Durkheim (1933), communities were organized around shared values. The often-observed difficulty in operationalizing ethical guidelines reflects the fact that a superficial consensus breaks down as soon as universal principles are domesticated in the various stakeholder communities. Forms of justification, evidence, argument, and knowledge, therefore, are important elements in an ethical framework. Above we consolidated these under the theme of justified choice.

In the many existing attempts to develop ethics of AI, many ethical traditions have been explicitly and implicitly used, including deontological duty ethics, consequentialism and utilitarianism, virtue ethics, ethics of care, and dialogical ethics. Previous work on biomedical ethics and human rights have played a central role. In contrast, the present paper builds mainly on the capability-based approach as this directly links with human and social development. There exist large bodies of literature on philosophical, political and economic research on this approach, and the relevance the capability-based approach for education and AIEd has also been noted before (Poquet & de Laat, 2021; Tuomi, 2015).

Education has many important social functions, and the use of AI in education may address these all. The ethics of AIEd must explicitly address these different functions and aims of education. A developmental view on human-AI interaction is necessary in educational settings and also in the broader context of life-long learning. Existing frameworks miss this developmental view.

The present paper is an attempt to refocus and re-frame ethical discussion on AI and AIEd. This has consequences for research and practice. At present, the European Commission is about to publish guidelines for educators on the ethical use of AI and data in education and training, partly influenced by the ideas presented above. UNESCO is also working on ethical guidelines aimed at educators and technology developers, focusing on the future of learning.

In this paper, we only in very broad terms map some key elements of this emerging landscape. Further development and elaboration is, of course, needed. A socio-developmental approach to ethics, however, helps us see where AIEd is today and where we should be going in the future.

5. References

Adams, C., Pente, P., Lemermeyer, G., & Rockwell, G. (2021). Artificial intelligence ethics guidelines for K-12 education: A review of the global landscape. In I. Roll et al. (Eds.), *Artificial Intelligence in Education* (pp. 24–28). Springer International Publishing. https://doi.org/10.1007/978-3-030-78270-2_4

AI HLEG. (2019). *Ethics Guidelines for Trustworthy AI*. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

AI HLEG. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI)*. European Commission. https://doi.org/10.2759/002360

Aiken, R. M., & Epstein, R. G. (2000). Ethical guidelines for AI in education: Starting a conversation. *International Journal of Artificial Intelligence in Education*, 11(2), 163–176.

Baker, T., & Smith, L. (2019). *Educ-AI-tion Rebooted? Exploring the future of artificial intelligence in schools and colleges*. NESTA.

Bakhtin, M. (1993). *Toward a Philosophy of the Act*. University of Texas Press.

Barnes, J. (1990). *The Toils of Scepticism*. Cambridge University Press.

Biesta, G. (2010). *Good Education in an Age of Measurement: Ethics, Politics, Democracy*. Routledge.

Bijker, W. E., Hughes, T. P., & Pinch, T. J. (1987). *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. The MIT Press.

Blikstein, P., Zheng, Y., & Zhou, K. Z. (2022). Ceci n'est pas une école: The discourses of artificial intelligence in education through the lens of semiotic analytics. *European Journal of Education*, 57(4), *in press*

Bloom, B. S. (1984). The 2 Sigma Problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16.

Bowker, G., & Star, S. L. (1999). *Sorting Things Out: Classification and its Consequences*. The MIT Press.

Brown, J. S., & Duguid, P. (1991). Organizational learning and communities of practice: Toward a unified view of working, learning, and innovation. *Organization Science*, 2(1), 40–57.

Douglas, M., & Wildawsky, A. (1982). *Risk and Culture*. University of California Press.

du Boulay, B. (2019). Escape from the Skinner Box: The case for contemporary intelligent learning environments. *British Journal of Educational Technology*, *50*(6), 2902–2919. https://doi.org/10.1111/bjet.12860

Durkheim, E. (1933). *Division of Labor in Society*. The Free Press.

EC. (2021). *Proposal for a Regulation of the European Parliament and of the Council: Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts* (COM(2021) 206 final). European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

Engeström, Y. (1999). Innovative learning in work teams: Analyzing cycles of knowledge creation in practice. In Y. Engeström, R. Miettinen, & R.-L. Punamäki (Eds.), *Perspectives in Activity Theory* (pp. 377–404). Cambridge University Press.

Evans-Pritchard, E. E. (1976). *Witchcraft, Oracles, and Magic among the Azande* (Abridged with an introduction by Eva Gilles). Claredon Press.

Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. In L. Floridi (Ed.), *Ethics, Governance, and Policies in Artificial Intelligence* (pp. 5–17). Springer International Publishing. https://doi.org/10.1007/978-3-030-81907-1_2

Guskey, T. R. (2012). Mastery Learning. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 2097–2100). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_1553

Habermas, J. (1993). *Moral Consciousness and Communicative Action*. The MIT Press.

Hagendorff, T. (2019). The ethics of AI ethics—An evaluation of guidelines. *ArXiv:1903.03425 [Cs, Stat]*. http://arxiv.org/abs/1903.03425

Hanseth, O., & Lyytinen, K. (2008). Theorizing about the design of information infrastructures: Design kernel theories and principles. *All Sprouts Content*, 4(12). https://aisel.aisnet.org/sprouts_all/68

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching & Learning*. The Center for Curriculum Redesign.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2021). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00239-1

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, 57(4), *in press*.

IEAIE. (2021). *The ethical framework for AI in education*. The institute for ethical AI in education. https://www.buckingham.ac.uk/wp-content/uploads/2021/03/The-Institute-for-Ethical-AI-in-Education-The-Ethical-Framework-for-AI-in-Education.pdf

IMDA & PDPC. (2020). *Model Artificial Intelligence Governance Framework: Second Edition*. Infocomm Media Development Authority and Personal Data Protection Commission. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kane, G. C., Young, A. G., Majchrzak, A., & Ransbotham, S. (2021). Avoiding an oppressive future of machine learning: A design theory for emancipatory assistants.

MIS Quarterly, 45(1b), 371–396. https://doi.org/10.25300/MISQ/2021/1578

Kizilcec, R. F., & Lee, H. (2023). Algorithmic fairness in education. In W. Holmes & K. Porayska-Pomsta (Eds.), *Ethics in Artificial Intelligence in Education*. Routledge. *in press*

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press.

Luckin, R. (2018). *Machine Learning and Human Intelligence: The Future of Education for the 21st Century*. UCL Institute of Education Press.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Applications*, *2021*(8), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Miettinen, R. (2000). The concept of experiential learning and John Dewey's theory of reflective thought and action. *International Journal of Lifelong Education*, 9(1), 54–72.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Newman, S., Birhane, A., Zajko, M., Osoba, O. A., Prunkl, C., Lima, G., Bowen, J., Sutton, R., & Adams, C. (2019). AI & Agency. UCLA: The Program on Understanding Law, Science, and Evidence (PULSE). https://escholarship.org/uc/item/8q15786s

Nussbaum, M. C. (2000). *Women and Human Development: The Capabilities Approach*. Cambridge University Press.

OECD. (2019). *Recommendation of the Council on Artificial Intelligence*. OECD. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Piaget, J. (1932). *The Moral Judgement of the Child*. Kegan, Paul, Trench, Trubner & Co.

Piaget, J., & Inhelder, B. (1979). *Memory and Intelligence*. Routledge & Kegan Paul.

Poquet, O., & de Laat, M. (2021). Developing capabilities: Lifelong learning in the age of AI. *British Journal of Educational Technology*, *52*(4), 1695–1708. https://doi.org/10.1111/bjet.13123

Rawls, J. (1999). *The Law of Peoples: With "The Idea of Public Reason Revisited."* Harvard University Press.

Roessler, B., & Mokrosinska, D. (2015). *Social Dimensions of Privacy: Interdisciplinary Perspectives*. Cambridge University Press.

Rosen, R. (1985). *Anticipatory Systems: Philosophical, Mathematical and Methodological Foundations*. Pergamon Press.

Sax, M. (2018). Privacy from an Ethical Perspective. In B. Van der Sloot & A. De Groot (Eds.), *The Handbook of Privacy Studies: An Interdisciplinary Introduction* (pp. 143–173). Amsterdam University Press.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Sen, A. (1993). Capability and well-being. In M. C. Nussbaum & A. Sen (Eds.), *The Quality of Life* (pp. 30–53). Clarendon Press.

Sen, A. (2000). *Development as Freedom*. Anchor Books.

Sen, A. (2009). *The Idea of Justice*. The Belknap Press of Harvard University Press.

Star, S. L. (1992). Cooperation without consensus in scientific problem solving: Dynamics of closure in open systems. In S. Easterbrook (Ed.), *CSCW: Cooperation or Conflict?* (pp. 93–106). Springer-Verlag.

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. EEAMO '21*: Equity and access in algorithms, mechanisms, and optimization*, https://doi.org/10.1145/3465416.3483305

Sutterlüty, F., & Tisdall, E. K. M. (2019). Agency, autonomy and self-determination: Questioning key concepts of childhood studies. *Global Studies of Childhood*, *9*(3), 183–187. https://doi.org/10.1177/2043610619860992

Tuomi, I. (2002). *Networks of Innovation: Change and Meaning in the Age of the Internet*. Oxford University Press.

Tuomi, I. (2012). Foresight in an unpredictable world. *Technology Analysis & Strategic Management*, 24(8), 735–751. https://doi.org/10.1080/09537325.2012.715476

Tuomi, I. (2015). Epistemic literacy or a clash of clans? A capability-based view on the future of learning and education. *European Journal of Education*, *50*(1), 21–24. https://doi.org/10.1111/ejed.12101

Tuomi, I. (2018). *The impact of artificial intelligence on learning, teaching, and education: Policies for the future.* Publications Office of the European Union. https://doi.org/10.2760/12297

Tuomi, I. (2022). Artificial intelligence, 21st century competences, and socio-emotional learning in education: More than high-risk? *European Journal of Education*, 57(4). *in press*

Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, *1*(3), 283–296. https://doi.org/10.1007/s43681-021-00038-3

UNESCO. (1996). *Learning: The Treasure Within*. UNESCO.

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence* (41 C/73). UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000379920

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Wang, A., Barocas, S., Laird, K., & Wallach, H. (2022). Measuring representational harms in image captioning. 2022 *ACM Conference on Fairness, Accountability, and Transparency*, 324–335. https://doi.org/10.1145/3531146.3533099