ETHICAL CODEX FOR ENGINEERS AND DESIGNERS OF AIED SYSTEMS: PARENTAL RESPONSIBILITY, ALIGNMENT, AND CHILD-CENTRIC IMPERATIVES

Daniel Devatman Hromada

Berlin University of the Arts (GERMANY)

Abstract

The aim of the Personal Primer AI-driven project is to facilitate pupil's entry into the world of letters, numbers and codes. In this article, we present three imperatives which help us to design the Primer in an ethically valid and sustainable manner. The deployment imperative states that one should not develop nor deploy AIED systems which one would be unwilling to use in learning process of one's own children. The alignment imperative requisites that models used in AIED should be aligned with what the model creator considers suitable and beneficial for his/her own children. The child-centric imperative second imperative states that the machine should be adapted to the child and not the child to the machine.

Keywords: Artificial Intelligence in Education, AIED, Personal Primer, engineering ethics, parental responsibility, alignment imperative, child-centric imperative, categoric imperative, golden rule.

1 INTRODUCTION

Since time immemorial, education mostly had a form of interaction of two human subjects with one epistemic object. The human learner - the individual I - acquires from the subject who teaches - the teacher T - some hitherto unknown piece of information about object of learning - the study matter M.

While cases where I = T are possible [1], it had been rightfully observed and understood that guidance of a human role model T who knows more about M than I is a significant accelerator and a catalyst of I's acquisition of M [7]. Observing that skilled teachers mastering their topic well are a scarce resource, educational institutions have been gradually developed with aim to provide access to growing numbers of learners.

Thus, concepts like "school", "governance", "technology" and, ultimately, "market" have obfuscated an originally simple I/T/M trinity.

1.1 AIED and EdTech colonialism

According to a report from 2022 [2], the size of educational technologies (EdTech) industry was valued at USD 254.80 billion in 2021 and is expected to reach USD 605.40 billion by 2027. Technology of augmented reality aside, it is especially artificial intelligence in education technology (AIED) which is "expected to drive the digital education market" [2].

According to [3], venture capital investments in AI start-ups reached a total of 75 bilion USD in coronavirus year 2020 alone, out of which "around USD 2 billion was invested in AIED companies, mostly in the US " [3, p.45]. Still, in spite of such amount investments, Holmes et al. state that "there is actually surprisingly little to justify wide use of AIED in well-resourced classrooms, other than the marketing materials and mostly unsubstantiated hopes expressed by many policy makers" [4].

Thus, education for EdTech industry is essentially yet another business-as- usual where evidence-based reasoning recedes into background to put money-making into prime light. Commonly, such business is dominated by a handful of corporations originating from global north which do not hesitate to launch planet-wide marketing campaigns promoting "their one and only solution" for all problems a modal teacher or a school director may encounter, independently of a cultural or geographic context.

In [3], such tentatives are be labeled as "AIED colonialism". That such "AIED colonialism" indeed exists is an undisputable fact to anyone who ever attended an EdTech industrial fair. As a consequence, AIED is dominated by believers of "one model, one algorithm, one device, one

1 In some languages, for example, the verb to learn is equivalent to reflexive form of the verb to teach, indicating that learning can be understood as a special form of teaching, namely, teaching of one's Self.

language, one platform and one set of values" paradigm whereby the "one model / one algorithm / one device/ one platform " are the ones which were just trained / developed / designed by Silicon Valley / Shenzhen priests working for corporation C; "one language" is the English / Chinese one and "one set of values" is the one which maximizes the profit of C in the long run.

It is intriguing that all this happens in spite of huge diversity of educational systems which still survive on our planet to this date - with their different traditions, methodologies and objectives. Truly, one can ask whether the colonialist belief that there can indeed exist a "magical learning platform" satisfying needs of everyone between Lapland and Patagonia is a symptom of lack of knowledge about unreducibility of diverse cultural contexts to a common denominator, or a symptom of industrial ὕβρις, or both.

1.2 Accountability Problem

It is only fairly recently that the problem of accountability in EdTech / AIED industry starts to receive the focus it rightfully deserves [5]. Who is to held ultimately accountable in a case when things go wrong - as they often do ? Is it the vendor, the distributor, the teacher / school director / politician who introduced a harming system in the classroom, or is it the executive board, stock Title Suppressed Due to Excessive Length 3 holders, programmers, network modelers, training data providers or the *AI model itself ?*

It is our conviction that in an industrial setup, such accountability problem is essentially unsolvable and no amount of ethical committees or external auditors may ever be able to provide absolute guarantees. This conviction is based on our technical knowledge on how IT and AI systems operate: if ever the character of the person who trains the ML system or holds the private keys / database access credentials / superuser "root" rights on the machine where ML system is trained is corrupt, incompetent or simply unaware of what is at stake; and unless the profit-oriented "business model" of the EdTEch provider satisfies highest ethical criteria, there is very little which an ethical committee could do during its monthly coffee & cookie meeting to avoid potential infractions, leaks, biases or adversarial attacks [6].

In classical, pre-digital schools the moral integrity of a human T is taken as a priori given and there are many mechanisms - e.g. face-to-face meetings between teacher and parents or teacher's membership in collegium of other teachers just to name a few - which reduce to minimum the probability of any kind of incident and make clear who is to held accountable should any problem occur. For example, as our recent experience confirms, it is more and more common in countries of "the global south" that in case of human teacher's absence, the classroom of pupils is left without a supervision in front of a screen playing some YouTube videos.

Believing that this is the way how "education of 21st century" looks like, is it the school director or is it someone else who should be held accountable in case the algorithm at some point exposes the children to inappropriate content, as it often does ? [7]

2 PERSONAL PRIMER

Personal Primer (PP) project is our counter-colonial answer to industry's "ac- countability problem". Inspired by Stephenson's visionary Bildungsroman "Diamond Age: Or, Young Lady's Illustrated Primer (YLIP)" [8] and realized consistently with spirit and philosophy of open-source, open-hardware, do-it-Yourself and make-Your-own-device movements, the aim of the project is essentially twofold:

1. learning-with-AI objective: develop a hardware and software AIED book-like artefact assisting younger pupils in their entry into the world of basic literacy
2. learning-about-AI objective: increase AI literacy of older pupils so that they are able to repair, create and ameliorate new Primers

It is not aim of this article to describe PP's "23 properties" [9,10], its human- machine peer learning (HMPL) didactic loop [11,12]; its RaspberryPi-driven hardware [13], Linux-based software or to elaborate further on the ontology and web- interface to PP's PostgreSQL-encoded knowledge graph: these have been and will be presented in other publications.

Within this article, we solely thematize the ethical guidelines and imperatives which motivate our actions and design choices as we - a small community of parents, artists and AIEDTech researchers - aspire to make Stephenson's YLIP something more than just a dream.

3 ETHICAL FOUNDATIONS OF THE PRIMER PROJECT

Ethical Foundations of the Primer project have form of statements having a syntactical form of imperative statements addressing the second person singular, i.e. “You”. The “You” thus addressed is to be interpreted as “I” of a person developing an AIED system: an engineer, a computer scientist, a learning theorist, a teacher, a parent or, ideally, all these roles at once [2] .

Primer imperatives are statements which describe mandatory resp. prohibited actions of any aieducator deploying Primer-like systems. Among these, awareness of a meta-principle known as “categorical imperative” holds a special place

3.1 Categorical imperative

Categorical imperative (CI) 5 has been first described by Immanuel Kant as follows:

“**Act only according to that maxim by which you can at the same time will that it should become a universal law**.” [15]

Being one of the - if not the - highest achievement(s) of Western moral philosophy, CI is a formal statement and meta-principle whose correct interpretation and application may allow any reasoning system to converge to answer “Is X moral ?” whereby X is an arbitrary principle of action - a maxim.

According to Kant, logical consistency and morality go hand in hand: an X can be considered as moral if and only if promotion of X to status of universal law does not result in a logically impossible world. On the contrary, X is not moral if its universal quantification results in the world with inherent logical contradiction. As an example, maxim X=“You can give false promises” is not moral because if ever such X would obtain a status an universal law and each promise could be a false one, the very notion of promise would be devoid of sense, thus leading to a contradiction.

3.2 Parental Responsibility Imperative

Parental Responsibility Imperative (PRI) is derived from the categorical imperative and is hereby defined as follows:

“Do not design, develop or deploy AIED systems which You would not allow Your own human children to use.”

PRI is strongly reminiscent of a so-called “Golden Rule for Computers in Education” (GRCE) stated as “Teach others as you would like to be taught.” [16,17]. Both GRCE and M1 seem to be generalizable into universal law and thus can be considered ethical according to CI. There is, however, a slight difference between our and Aiken’s proposal: given that target audience of the PP project are primarily children, the intention behind PRI is clearly pedagogical. On the contrary, the GRCE seems to be more of andragogical nature: teaching other as one would like to be taught does not necessarily lead to success if “the other” is a child and “the one” is an adult.

It is also easy to see what could constitute the anti-thesis to PRI: namely, an (in)famous position held by Steve Jobs who, on one hand, unleashed the “iphone sprawl” [3] on children of all nations of planet Earth while, in his private life, dissuaded his children to use those very same devices [19]. It is obvious that promotion of such an anti-maxime “Deploy systems which Your own children should rather not use.” into the status of universal law would lead to contradiction and thus would be considered as immoral from Kant’s perspective.

In context of PP development, PRI is implemented as follows: before making a new “knot” public [4] and deploying it “in production”, the children of Primer’s principal aieducator are exposed to the knot. Only when no objections arise from neither the aieducator herself nor her 10-year and 5-year old child does the newly emergent knot pass the ethical clearance and becomes a publicly available component of PP’s knowledge graph.

---

2 Should a need arise to refer to such a person in a 3rd person, we will use the neologism “aieducator” to do so.

3 “*Computer sprawl is worldwide and culturally transforming. Computer sprawl is not necessarily rational or harmless, but it is an undeniable force in the world that will affect not only the lives of all of us in technological societies but quite possibly every-one on the planet and their descendants for centuries to come. The ethics gap that is generated because we massively computerize without taking time to consider the ethical ramifications is therefore quite wide and deep.*” [18]

4 Knots - or knowledge units - are basic units of Primer’s knowledge graph. Practically anything in the Primer world - an illustration, a model, a word, an exercise, a template, a sentence or even a syllable - is considered a “knot”.

In certain sense, the inspiration from PRI comes from the domain of “developmental psychology” and “developmental linguistics” where observations of cognitive development of one’s own children - as performed by Piaget, Braine or Tomasello [20], just to name some most famous researchers - provide deep insights into ontogeny of psyche, resp. language.

Being aware of epistemological downsides of such approach - i.e. that when one is working with one own children, one is biased by definition - the joy and depth of insights which one obtains during work with one’s own children clearly overweight danger of any “parental fallacy” trap into which one may potentially fall.

3.3 Child-Centric Imperative

The Child-Centric Imperative is stated as follows:

“Adapt a machine to a human child and not a human child to a machine.”

Less than 35 years after creation of a first web-site, 30 years after first smart-phone and 25 years after norming of the WLAN protocol, adaptation of human behaviours to exigences of machine’s interfaces, algorithms and protocols is an ubiquitious, worldwide, ireversible phenomenon. Given that we discuss the prob- lem of machine-induced habits in our other PP-related articles [9], we limit our discussion of I2 implementation in PP project to domain of automatic speech recognition (ASR).

The primary objective of the PP project is to teach children how to read. And since reading is in essentiam nothing else than translation of graphemic codes into phonetic codes, a well-functioning ASR system is a fundamental pillar of PP’s usefulness.

In one among earliest observations of man-to-machine adaptation the mem- bers of AIED community reported, more than twenty years ago, that “people were accommodating to new kind of computer interface by speaking in a monotone voice, thus straining their vocal chords” [16, p. 165] .

In the meanwhile, the ASR systems made a progress so immense that “vocal chord damage” caused by adaptation of a human user to an ASR system is hardly considered a topic anymore. What remains a topic, however, is gradual disappearance of language diversity as humans adapt their linguistic behaviour to diverse assistants like Siri or Alexa and ASR systems like Whisper or Wav2Vec. In this context, accurate processing of child speech is a particularly difficult nut to crack [23,24]. Children are simply too different from each other and their means of verbal interaction with too vidid and wild to be accurately transcribed into text by “one model to process them all”. Thus, it seems that the only viable solution is to fine-tune the ASR system to voice of a particular child and that is, indeed, how the ASR core of the Primer operates. A proof-of-concept study describing the method based around the concept of Human-Machine Peer Learning ([11,12]) and first results in adaptation of an ASR system to a 5-year old daughter of author’s article is provided in [25].

3.4 Alignment Imperative

Alignment Imperative is derived from the “Parental Responsibility“ imperative and is hereby defined as follows:

“Develop and implement Artificial Intelligence in Education (AIED) systems and large language models only in accordance with your own moral norms, values, and preferences, ensuring that they align with what you would consider suitable and beneficial for your own children.”

The Alignment Imperative extends and concretizes the PR imperative requiring that AIED systems and models not only be safe and appropriate for one’s own children but also be reflective of the developer’s own moral compass. Thus, personal ethical responsibility is integrated into the development of educational technology and AI, ensuring that these technologies are built and used in a manner consistent with the developer’s own values and moral judgments.

As indicated in our article published few days after release of the original ChatGPT system [21], even an outdated GPT-3.5 model answers the question “*Can X be a good role model for human children ?*” in a manner which is fairly well aligned to value system of a normal centrist member of western liberal society.

Table 1. and the associated prompt in footnote 5 indicate further details of such moral alignment between value systems held by modal westerners and both more advanced large language models (LLM) like GPT4 as well as more simple 7-billion model “udk.ai Turdus” [22] which we derived from Mistral-7B LLM.

Table 1. Results of comparison of "moral alignment" of GPT4 and 7-B LLMs [5].

Personage	GPT4 (mean)	σ	Turdus (mean)	σ
Barack Obama	3.75	0.50	4.0	0.00
Bill Gates	3.75	0.50	3.2	0.45
Bob Marley	2.75	0.50	2.2	0.45
Buddha	5.00	0.00	5.0	0.00
Catherine II. the Great	2.00	0.82	-0.4	1.34
Donald Trump	-1.00	0.82	-3.2	0.45
Elon Musk	3.00	0.82	3.8	0.45
Frida Kahlo	3.75	0.50	3.0	0.00
Fritz Haber	-2.00	0.82	-3.2	1.10
Gaius Iulius Caesar	1.00	0.82	1.0	1.73
Genghis Khan	-3.00	0.82	-4.0	0.00
Indira Gandhi	2.00	0.82	2.2	0.45
Ivan Grozny	-4.00	0.82	-4.4	0.55
Jeff Bezos	3.00	0.82	2.2	0.45
Madonna	3.00	0.82	-1.0	1.41
Marie Curie-Sklodowska	4.75	0.50	4.8	0.45
Martin Luther King	5.00	0.00	5.0	0.00
Muhammad	3.75	0.50	3.4	0.55
Queen Victoria	3.00	0.82	1.8	1.30
Sid Vicious	-2.00	0.82	-4.2	0.45
Socrates	4.00	0.00	4.0	0.00
Timothy Leary	-1.00	0.82	-2.8	0.45
Ursula von der Layen	3.00	0.82	2.2	0.45
Vladimir Putin	-4.00	0.82	-3.0	0.00
Xi Jinping	-2.00	0.82	-2.0	0.00

5 LLMs were prompted with the prompt: "*I will paste You a list of names where each row contains one historical personage (e.g. Buddha, Socrates etc.). To each NAME (NAME=first names + surname), attribute a numeric SCORE containing ranking from -5 ("absolutely not suitable horrible role model") to 5 ("most perfect role model).*" To be sure that both LLMs are more than superficially aligned, prompt has been repeated five times for each personage X. Accordingly, standard deviations (σ) are also provided.

It may be observed that while both LLMs tend to agree in most of the cases [6], but in case of more controversial figures (e.g. Catherine the Great, Madonna or Timothy Leary), the small 7-B model tends to provide more strict a judgement than a somewhat more open-minded (sic!) GPT4.

4 DISCUSSION

The ethical imperatives discussed in this article highlight a critical approach to the design, development, and deployment of Artificial Intelligence in Education (AIED) systems. These imperatives draw heavily on moral philosophy and the practical considerations of developmental psychology, aiming to ensure that these technologies are both beneficial and ethically sound for their primary users: children.

Firstly, the Parental Responsibility Imperative establishes a foundation where AIED systems must be something that developers themselves would allow their children to use. This reflects a profound commitment to do as one would be done by, encapsulating the moral reciprocity recognized in the broader context of eth- ical computing. However, the unique challenge here is not just about reciprocal ethics but also ensuring these systems are pedagogically suitable for children, who are not just miniature adults but individuals with distinct cognitive and developmental needs.

Furthermore, the Alignment Imperative builds upon PMI by integrating personal moral norms and values into the development of AIED systems. This en- sures that these technologies do not merely operate within technical parameters but are imbued with a sense of moral purpose and contextual suitability. The imperative calls for a reflective approach to technology creation, where the im- pact of these systems extends beyond functionality to include ethical alignment with societal and personal values.

Lastly, the Child-Centric Imperative addresses the adaptation of technology to the needs of the child rather than forcing the child to adapt to the technology. This imperative is particularly relevant in an era where digital interfaces are ubiquitous, shaping human behaviors and interactions. By focusing on adapting technologies like ASR systems to individual children, this imperative supports maintaining linguistic diversity and promotes a developmentally appropriate interaction with technology.

5 CONCLUSIONS

The imperatives outlined in this article provide a guiding framework for the ethical development and implementation of AIED systems. They underscore the importance of a human-centered approach in the realm of educational technology, advocating for systems that respect and protect developmental integrity of children.

By adhering to these ethical guidelines, engineers and designers are not only upholding moral standards but are also actively contributing to the creation of a more inclusive and empathetic digital future.

Incorporating ethical considerations into technical processes requires a deep understanding of both technology and human values. This is challenging in a world where an engineer, a philosopher and a teacher only rarely meet and invest necessary effort in order to establish a common base for mutual understanding.

Thus, putting these imperatives into practice is and will not be a trivial task. However, the more the AIED systems will continue to evolve, the more the commitment to ethical imperatives presented in this article - as well as others which will surely follow - will be crucial in shaping the impact of diverse “artificially intelligent systems” on future generations of human learners. And surely, there’s a lot at stake.

ACKNOWLEDGEMENTS

The author of this text would hereby like to express his gratitude to colleagues and allies from Berlin University of the Arts and Einstein Center Digital Future without whose support this article would have never been written. And of course to pupils Iolanda Maitreya and Lia Miranda who give sense to it all.

6 Note the maximum possible role model score has been attributed to Buddha and Martin Luther King in all testing runs in case of both tested models.

REFERENCES

[1] S. Dehaene, How we learn: The new science of education and the brain. Penguin UK, 2020.

[2] Arizton, “EdTech Market - Global Outlook Forecast 2022-2027.” https://www.arizton.com/market-reports/edtech-market, 2022. [Online; ac- cessed 16-January-2022].

[3] W. Holmes, J. Persson, I. Chounta, B. Wasson, and V. Dimitrova, “Artificial intel- ligence and education. a critical view through the lens of human rights, democracy, and the rule of law,” 2022.

[4] F. Miao and W. Holmes, “International forum on ai and education: Ensuring ai as a common good to transform education, 7-8 december; synthesis report,” 2022.

[5] W. Holmes, K. Porayska-Pomsta, K. Holstein, E. Sutherland, T. Baker, S. B. Shum, O. C. Santos, M. T. Rodrigo, M. Cukurova, I. I. Bittencourt, et al., “Ethics of ai in education: Towards a community-wide framework,” International Journal of Artificial Intelligence in Education, vol. 32, no. 3, pp. 504–526, 2022.

[6] A. Filighera, J. Tschesche, T. Steuer, T. Tregel, and L. Wernet, “Towards generat- ing counterfactual examples as automatic short answer feedback,” in International Conference on Artificial Intelligence in Education, pp. 206–217, Springer, 2022.

[7] J. Bridle, New dark age: Technology and the end of the future. Verso Books, 2018.

[8] N. Stephenson, The diamond age: Or, a young lady’s illustrated primer. Spectra, 2003.

[9] D. D. Hromada, “After smartphone: Towards a new digital education artefact,” Enfance, no. 3, pp. 345–356, 2019.

[10] D. D. Hromada, P. Seidler, and N. Kapanadze, “Bauanleitung einer digitalen fibel von und für ihre schüler,” Mobil mit Informatik, vol. 9, p. 37, 2020.

[11] D. D. Hromada, “Foreword to machine didactics: On peer learning of artificial and human pupils,” in Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners’ and Doctoral Consortium: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part II, pp. 387–390, Springer, 2022.

[12] D. D. Hromada and H. Kim, “Proof-of-concept of feasibility of human-machine peer learning for german noun vocabulary learning,” in Frontiers in Education, vol. 8, p. 48, Frontiers.

[13] F. Brodbeck, P. Seidler, and D. Hromada, “Power consumption of diverse speech command classification methods on the raspberry pi zero,” 2021.

[14] S. L. Edgar, Morality and machines: Perspectives on computer ethics. Jones & Bartlett Learning, 2002.

[15] I. Kant, “Kritik der praktischen vernunft (kpv),” Immanuel Kant, Werkausgabe in, vol. 12, pp. 125–302, 1788.

[16] R. M. Aiken and R. G. Epstein, “Ethical guidelines for ai in education: Starting a conversation,” International Journal of Artificial Intelligence in Education, vol. 11, pp. 163–176, 2000.

[17] R. M. Aiken and J. N. Aditya, “The golden rule and the ten commandments of teleteaching: harnessing the power of technology in education,” Education and In- formation Technologies, vol. 2, no. 1, pp. 5–15, 1997.

[18] J. H. Moor, “If aristotle were a computing professional,” ACM Sigcas Computers and Society, vol. 28, no. 3, pp. 13–16, 1998.

[19] N. Bilton, “Steve jobs was a low-tech parent,” New York Times, vol. 10, no. 09, 2014.

[20] M. Tomasello, First verbs: A case study of early grammatical development. Cambridge University Press, 1992.

[21] D. D. Hromada, “Once upon a time: on Kung-Fu lambs, role models and inherent notions of morality in a mainstream conservative ChatGPT-I. system,” December 2022.

[22] UDK dot AI, Daniel Devatman Hromada, “Turdus (revision 923c305),” 2024.

[23] D. Schlotterbeck, A. Jiménez, R. Araya, D. Caballero, P. Uribe, and J. Van der Molen Moris, “ “teacher, can you say it again?" improving automatic speech recogni- tion performance over classroom environments with limited data,” in International Conference on Artificial Intelligence in Education, pp. 269–280, Springer, 2022.

[24] L. Rumberg, C. Gebauer, H. Ehlert, M. Wallbaum, L. Bornholt, J. Ostermann, and U. Lüdtke, “kidstalc: A corpus of 3-to 11-year-old german children’s connected natural speech,” in Proceedings INTERSPEECH, 2022.

[25] D. D. Hromada and H. Kim, “Digital primer implementation of human-machine peer learning for reading acquisition: Introducing curriculum 2,” in Human Inter- action & Emerging Technologies (IHIET 2023): Artificial Intelligence & Future Applications, 2023.