Ethical procedures for responsible experimental evaluation of AI-based education interventions

Dekker, Izaak; Bredeweg, Bert; te Winkel, Wilco; van de Poel, Ibo

DOI
10.1007/s43681-024-00621-4

Publication date
2025

Document Version
Final published version

Published in
AI and Ethics

Citation (APA)
Dekker, I., Bredeweg, B., te Winkel, W., & van de Poel, I. (2025). Ethical procedures for responsible experimental evaluation of AI-based education interventions. *AI and Ethics*, *5*(3), 2977-2986. https://doi.org/10.1007/s43681-024-00621-4

AI and Ethics (2025) 5:2977–2986
https://doi.org/10.1007/s43681-024-00621-4

ORIGINAL RESEARCH

Ethical procedures for responsible experimental evaluation of AI-based education interventions

Izaak Dekker[1] · Bert Bredeweg[1,2] · Wilco te Winkel[3] · Ibo van de Poel[4]

Received: 11 August 2024 / Accepted: 10 November 2024 / Published online: 30 November 2024

Abstract
Many have suggested that AI-based interventions could enhance learning by personalization, improving teacher effectiveness, or by optimizing educational processes. However, they could also have unintended or unexpected side-effects, such as undermining learning by enabling procrastination, or reducing social interaction by individualizing learning processes. Responsible scientific experiments are required to map both the potential benefits and the side-effects. Current procedures used to screen experiments by research ethics committees do not take the specific risks and dilemmas that AI poses into account. Previous studies identified sixteen conditions that can be used to judge whether trials with experimental technology are responsible. These conditions, however, were not yet translated into practical procedures, nor do they distinguish between the different types of AI applications and risk categories. This paper explores how those conditions could be further specified into procedures that could help facilitate and organize responsible experiments with AI, while differentiating for the different types of AI applications based on their level of automation. The four procedures that we propose are (1) A process of gradual testing (2) Risk- and side-effect detection (3) Explainability and severity, and (4) Democratic oversight. These procedures can be used by researchers and ethics committees to enable responsible experiment with AI interventions in educational settings. Implementation and compliance will require collaboration between researchers, industry, policy makers, and educational institutions.

Keywords Artificial intelligence · AIED · Responsible artificial intelligence · Ethics · Institutional review board · Experiment

1 Introduction

Countries throughout the world invest in the development of Artificial Intelligence (AI) applications, expecting that this may grant gains in efficiency, control, or benefits to human flourishing [31]. Within the field of education, AI-based applications are claimed to be able to (1) enhance learning by personalization, (2) improve teacher effectiveness, or (3) optimize educational processes [18, 19, 27]. Developments in generative AI further expand expectations and promises, e.g., by promises of providing helpful roles for students or teachers [28, 29]. The technological possibilities are developing rapidly, scientific publications on the topic proliferate, and large investments are made to develop AI-based EdTech because of the expected benefits that the investors believe they can provide [10]. However, rigorous evidence on the effectiveness of AI-based EdTech is scarce [7, 18, 19], and recent studies indicate that AI can, in practice, just as well undermine learning as enhance it [1].

"I exposed our main defect, that is, the incapacity of our imagination to grasp the enormity of what we can produce and set in motion" Gunther Anders, 1980, p.19.

✉ Izaak Dekker
i.dekker@hva.nl

Bert Bredeweg
b.bredeweg@hva.nl

Wilco te Winkel
wilco.tewinkel@eur.nl

Ibo van de Poel
I.R.vandePoel@tudelft.nl

[1] Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

[2] University of Amsterdam, Amsterdam, The Netherlands

[3] Erasmus University Rotterdam, Rotterdam, The Netherlands

[4] Delft University of Technology, Delft, The Netherlands

The uncertain effects of AI-based EdTech in combination with the large investments and high expectations warrant careful and thorough experimental research that evaluates both direct and side-effects. Yet, according to a white paper from IEEE on ethically aligned AI design "It is unclear how research on the interface of humans and AI [.] will impact research ethical review boards. Norms, institutional controls, and risk metrics appropriate to the technology are not well established in the relevant literature and research governance infrastructure" [21, p. 128].

The current framework and codes that research ethics committees (also known as institutional review boards, for readability here called 'ethics committees') apply are based on the assumption that we can predict or anticipate social consequences beforehand. Van de Poel [34] argues that this approach does not sufficiently deal with the uncertainties and unknowns that are inherent to social changes introduced by technological development. He proposes an approach that conceives the introduction of new technology as a social experiment and offers an ethical framework for the acceptability of such experiments based on the existing bioethical principles for experiments with human subjects: non-maleficence, beneficence, respect for autonomy, and justice [5]. The sixteen conditions that Van de Poel derived through this method were formulated for experimental technology in general and need to be further specified for specific technologies and domains.

In order to specify Van de Poel's conditions for AI, the conditions should address its specific characteristics. AI is, by definition, characterized as technology that automates (aspects of) (human) intelligence [14]. It can perform goal-directed tasks in interaction with its environment, in a more or less automated way. This makes the effects of its introduction inherently less predictable. The degree of automation that AI-based interventions have can vary enormously, and has fundamental consequences for its role, effects, and risks. Molenaar [27] applies the six levels of automation model from the vehicle industry to education to distinguish these degrees. Each level of automation implies a different relationship between the teacher, technology, and the learner. AI-based EdTech can offer assistance with supportive information while the teacher has full control (level 2) or it can control all tasks automatically (level 6), for example. These different levels have substantial ethical importance because levels of supervision and control determine responsibility. For example, if AI EdTech assists the teacher with assessments, the teacher is still responsible and accountable. The levels of automation are, additionally, of great pedagogical importance because they affect the relationship between the teacher and the learners. For example, replacing group work with individualized assignments can crowd out social interaction and relations between students and teachers. Treating all AI EdTech the same will either create unnecessary red tape and stifle the much needed experiments, or underestimate the ethical and pedagogical consequences. Differentiating based on levels of automation allows ethics committees to process lower risk experiments quickly and detect and assess (or escalate) higher risk cases thoroughly.

Education is a field in which rigorous experiments are relatively complicated and uncommon [8, 11, 13], for example, because many effects are heterogeneous and context dependent, and because it is controversial to randomize children into 'placebo education'. The field of AI and education similarly lacks rigorous experimental studies with relevant control conditions [7, 18, 19]. If we want to encourage responsible experimentation with AI-based EdTech as a research field, then we need to develop usable and clear procedures.

In this conceptual article we apply the lens of the six levels of automation from Molenaar [27] to specify the ethical framework of Van de Poel [34] for AI educational interventions. We translate each condition into a procedure that can be used by researchers and ethics committees to assess these conditions when designing and evaluating experimental studies with AI-based EdTech. Within each procedure we differentiate for the different levels of automation in order to prevent unnecessary regulation.

2 Differentiating between AI educational interventions: 6 levels of automation

AI is often defined as the development of computer programs that can solve problems and achieve goals in the world that typically require humans [26]. This encompasses both the ambition of creating general artificial intelligence and already existing specialized forms of autonomously operating technology. The degrees of autonomy vary and make all the difference. The applications, therefore, also range from applications that aid humans with a small degree of automation to nearly fully automated support for learning processes. Molenaar [27] defined these different levels of automation in the application of AI in education using the six levels of automation model from the automotive industry (Fig. 1).

In the case of automation of vehicles, the different levels are often perceived as a hierarchy with the 6th level as a potential goal. In the domain of education, it is important to note that while some imperfect versions of level 6 already exist, it might never be the final goal. Some video games and educational games can be seen as fully automated learning environments. They do not (yet), however, manage to obtain the learning outcomes that standardized curricula strive for and subsequently do not (yet) fulfill formal roles within

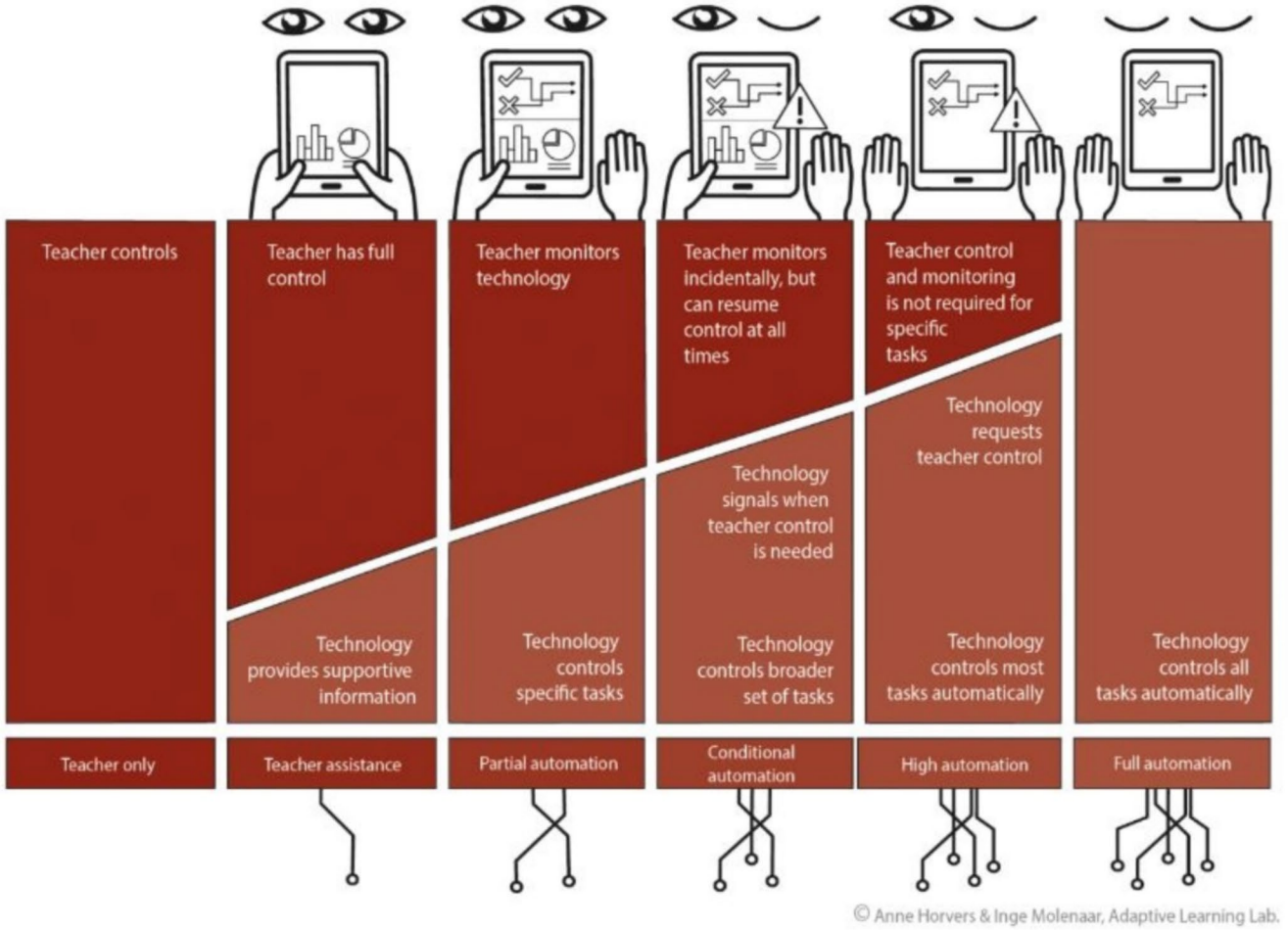

Fig. 1 Six levels of automation model of personalized learning

education. Already since the development of the AI in Education field, many have argued that AI interventions should not strive to replace what teachers do, but rather supplement or assist them as a form of 'hybrid intelligence' [2, 9].

The degree of automation matters from the perspective of learning, but it also matters from an ethical perspective. Higher levels of automation can both be more untransparent and, thereby, unpredictable [28], and can have a different intended and unintended impact on the autonomy and relationships of the teacher and learner. An application that automatically provides feedback and a grade for written essays based on previous feedback or programmed feedback from a teacher might improve judgment consistency at the cost of less professional autonomy. An application that automatically provides feedback and grades based on probabilistic technology and a big and evolving dataset of internet users might further enhance perceived quality, but will be less transparent and understandable. Furthermore, it could be used not only for feedback but for first or even final drafts of essays, which might lead the professional to reconsider how to grade essays and or use them for assessment [e.g., 15].

3 Application of framework to AI educational interventions

When evaluating the uses of different types of automated applications in education, we should strive to take the mentioned ethical risks into account. Currently this is done by trying to predict or anticipate social consequences and to use these as a basis for moral and regulatory appraisal. Such an approach can, however, not deal with the uncertainties and unknowns that are inherent in social changes induced by technological development [18, 34]. Van de Poel, therefore, proposes conceiving the introduction of new technology as a social experiment. He introduces a framework for this purpose that translates the moral principles for experiments with human subjects based on the Nuremberg Code, Helsinki Declaration, and Common Rule into 16 conditions

that researchers and ethics committees can use to assess and deal with unexpected risks (Table 1).

These sixteen conditions are relevant for experiments with AI in education, however, they do not yet provide practical guidance on how they might be met. Furthermore, they do not yet differentiate between different types of risk categories. The risks and unpredictability of AI-based EdTech is contingent on the degree of autonomy that the technology has [18]. Higher levels of automation, unpredictability, and intransparency should call for higher levels of risk management and scrutiny [22, 23]. Below we provide a first exploration of how they could be translated into concrete actions or procedures by describing how each condition could procedurally be checked by ethics committees. For each condition we differentiate potential procedures with risk categories based on the six levels of automation model. The sixteen different conditions are organized in four different procedures that can be used to apply and assess them: (1) A process of gradual testing (2) Risk and side-effect detection (3) Explainability and severity, and (4) Democratic oversight. In Sect. 3 we focus on ethics committees and researchers as primary users of the procedures, and in Sect. 4 we will expand this to other stakeholders and the broader domain, which are essential for eventual successful implementation.

Table 1 Ethical Framework for Experimental Technology [29]

	Moral principle
1 Absence of other reasonable means for gaining knowledge about risks and benefits	Non-maleficence
2 Monitoring of data and risks while addressing privacy concerns	Non-maleficence; Beneficence
3 Possibility and willingness to adapt or stop the experiment	Non-maleficence
4 Containment of risks as far as reasonably possible	Non-maleficence; Beneficence
5 Consciously scaling up to avoid large-scale harm and to improve learning	Non-maleficence
6 Flexible set-up of the experiment and avoidance of lock-in of the technology	Non-maleficence
7 Avoid experiments that undermine resilience	Non-maleficence
8 Reasonable to expect social benefits from the experiment	Beneficence
9 Clear distribution of responsibilities for setting up, carrying out, monitoring, evaluating, adapting, and stopping of the experiment	Beneficence; Procedural justice
10 Experimental subjects are informed	Respect for autonomy
11 The experiment is approved by democratically legitimized bodies	Respect for autonomy; Procedural justice
12 Experimental subjects can influence the setting up, carrying out, monitoring, evaluating, adapting, or stopping of the experiment	Respect for autonomy
13 Experimental subjects can withdraw from the experiment	Respect for autonomy
14 Vulnerable experimental subjects are either not subject to the experiment or are additionally protected or particularly profit from the experimental technology (or a combination)	Distributive justice
15 A fair distribution of potential hazards and benefits	Distributive justice
16 Reversibility of harm or, if impossible, compensation of harm	Distributive justice

3.1 A process of gradual testing

The first seven conditions are based on the non-maleficence principle. These conditions call for regulations in the form of processes. Conditions 1 (absence of other reasonable means for gaining knowledge about risks and benefits), 4 (containment of risks as far as reasonably possible) and 5 (consciously scaling up) could be met by first using explorations of impact, then small-scale prototype tests in controlled environments, and finally field experiments that are replicated with representative samples. Risks and benefits can be explored with research designs such as the Responsible Research and Innovation method in which stakeholders are consulted in order to envision the desired impact of the technology (condition 1). Eventually, however, causal information is required to reliably gain knowledge about risks and benefits. Scientific experiments are the least unreliable way of inferring causality [13]. The combination of small controlled lab experiments that test the theory and efficacy, with large field experiments which test external validity and ecological validity is the best bet for reliable information about risks and benefits. Conditions 3, 4, 5, 9, and 15, additionally call for a step-wise approach because this allows for reliably assessing risks, while limiting the risks to which the participants are exposed, and because this makes it easier to stop before a large sample is affected.

The ethical considerations during the experiments with human subjects (condition 10, 13, 14, and 16) are largely similar to the principles and questions that ethics committees currently apply to ensure that informed consent is acquired, vulnerable subjects are additionally protected, and that compensation and withdrawal options are readily available.

The step-wise scaling up sequence should ideally be used for all interventions that use some form of automation (level 2–6). However, from level 4 on up there are moments were the teacher delegates oversight and responsibility to the AI based EdTech. This is a more principled distinction which calls for more thorough scaling up of risks and more strenuous effect tests. Therefore we recommend the procedure for level 1–3 and advise to make it a requirement for AI based EdTech with level 4–6 of automation.

The stakeholders who would have to enact this procedure are researchers and ethics committees. Past developments, however, show that these changes can be initialized and

catalyzed once grant funders, journals, and policy makers stimulate them by setting them as conditions for funding or publication [16]. The required collaboration of stakeholders will be further discussed in Sect. 4.

3.2 Risk and side-effects detection

Condition 2 (monitoring of data and risks while addressing privacy concerns) combines monitoring risks during the experiment and aspects of good data-stewardship. Monitoring risks during the experiment requires researchers to design their study in a way that (i) problems/incidents can be detected and (ii) that side-effects can be explored based on expected trade-offs or more exploratory qualitative methods and long-term impact measurement [13, 36]. Providing options for the users to report problems/incidents is mostly similar to current requirements for experiments. The level of automation, however, does require additional measures to ensure that this condition is met. From the second level on, the teachers and the learners can interact directly with AI. If this occurs without the presence of the researcher, there need to be visible ways or manners to report issues within the application interface. From the fourth level of automation on this needs to be actively asked by the researchers conditional on additional consent because of limited oversight from the teacher. The argument here is that researchers in those cases take over the responsibility of oversight where that of the teacher diminishes.

Researchers could extend the incident reporting infrastructure beyond the current AI incident database [30] and learn from procedures and infrastructure for incident reporting in the more experienced fields of aviation or biosafety. Exploring side-effects in experimental studies should be a requirement for condition 2 because it allows researchers to formally detect whether the experiment should be stopped or whether there are unexpected risks or benefits.

Data stewardship can partly be taken into account with the current procedures for data management plans and data privacy impact assessments (in the case of new applications). A concern that is specifically voiced in the context of AI-based applications that collect large amounts of data, is data ownership and the potential consequences for agency and dependency that this entails. This dependency is a specific type of dependency that could undermine resilience (condition 7). It could be addressed by the piece-meal approach, by taking side-effects such as dependency into account, and possibly with conscious democratic decisions depending on the risk category (condition 11, further specified below).

Conditions 3 (possibility and willingness to adapt or stop the experiment), and 6 (flexible set-up of the experiment and avoidance of lock-in of the technology) are connected to the oversight required by condition 2 and require the technology to be limited in availability (or facilitation thereof) during the experimental process. Exploring side-effects is required in order to detect lock-in because this is only rarely the direct subject of the investigation [e.g., potential 'crutch' effect: 4]. The possibility to stop the experiment and the requirement of a flexible setup are also related to condition 9 (clear distribution of responsibilities) and additionally require these aspects of the process of the experiment to be organized and described beforehand. This could be done by creating practical guidelines for researchers about designing experiments with AI that could be integrated with questions that are asked in a pre-registration. Given that this is already partly checked by ethics committees, this could be a relatively minor addition to the current process. The risk of lock-in of technology partly overlaps with the principle of resilience (condition 7) because a lock-in could create dependency and, thereby, undermine resilience. It requires risk monitoring, side-effect detection. If these are detected in earlier stages, it could mean that the stage II or stage III experiments would require the ethics committee to escalate the decision to a public committee (condition 11).

3.3 Explainability and severity

The conditions related to beneficence such as condition 8 (Reasonable to expect social benefits from the experiment) require the experimenters to have a rationale for expecting and explaining social benefits. Merely expecting benefits based on previous results is not sufficient because this runs a high risk of confusing correlation with causation. As long as the effectiveness of AI Edtech is not explainable, it is not reliable. Explainability in AI educational research can be improved by theory building: iterations of theorizing and data gathering for verification or falsification. This requirement should be included in the pre-registration of the experiment in order to increase the severity of the experiment [24], the explainability of the underlying AI technology [12] and the scientifically theorized mechanism [13, 17]. Doing so would also take parts of condition 3, 6, and 9 into account if the process, responsibilities and stopping rules are clearly described. This is a process that should be applied to AI interventions of all levels of automation from 2 until 6 because each level of automation can be based on untransparent probabilistic technology. Theorizing the mechanisms and testing the mechanism can help to safely transfer its effectiveness (and risk category) to larger and other contexts.

3.4 Democratic oversight

Condition 11 (the experiment is approved by democratically legitimized bodies) is dependent on the type of 'social

experiment' that is conducted. In Jordans [22, 23] analyses of the requirements of ethics committees for judging the risks of AI, she advises mirroring a combination of the current infrastructure that is used for assessing research with biosafety risks. In the United States, Institutional Biosafety Committees can escalate proposals that fall within a high risk category to a national body that is organized by the National Institutes of Health and that conducts public hearings. Regardless of whether there should be a separate ethics committee that specifically addresses research with AI, or whether it should fall under the responsibility of the existing ethics committees, having a specialized public overarching committee that can handle cases that are escalated because of higher risks, would be a way to take this condition into account. Even when the risks are relatively low, condition 12 demands that there is a form of internal democracy, by allowing experimental subjects to influence the setting-up or stopping of the experiment. This could be organized by setting up an advisory committee for the experiment that includes at least one member of each type of experimental subject (e.g., student, teacher, both). For automation level 2, having an internal advisory committee can suffice because the responsibility and oversight still lies with a specific and accountable teacher. From level 3 on, the teacher no longer exerts full control. Ethics committees should then be able to escalate these cases to a national committee based on the risk categories that Jordan [23] proposed or a risk-indexation that is currently enacted within that national jurisdiction (e.g., the EU AI-act for European countries).

Summarizing the above application of the different conditions, we identify four overarching procedures that can practically take the 16 different conditions Van de Poel [34] suggested into account, while differentiating for the six levels of automation of AI-based interventions [27]. These four procedures are respectively: (1) A process of gradual testing (2) Risk- and side-effect detection (3) Explainability and severity, and (4) Democratic oversight (Table 2).

4 Implementation of the procedures

In the previous section we focused our attention on the research process and on researchers and ethics committees as primary audience. The procedures we proposed are directed at responsible research practices, and as such should be used by researchers who initiate studies with AI in education, ethics committees who judge research proposals, and research institutions who manage their research procedures. The first three procedures we proposed can be implemented with the current ethical research infrastructure, the fourth would require the organization of a new (most likely national) public ethics committee to which institutional research ethics committees can cases with less oversight and a higher risk category.

In order to support implementation and for these procedures to have serious impact on the field, however, collaborations between the different stakeholders within AI in education are required. AI in education is an interdisciplinary field which requires collaborations between educators, students, AI developers, ethicists, and policy makers. This section will, therefore, describe the role of the procedures in the broader domain and outline three directions outside of academia that could initiate and or catalyze the impact of the procedures.

Table 2 Application of conditions for different levels of automation

Procedures and Conditions (#)	Level of automation			
	2	3	4	5 6
Process of gradual testing 1, 3, 4, 5, 7, 9, 10, 13, 14, 15, 16	Recommended: Prototype trials (Stage I) + Field Experiment (Stage II) + Field experiment with representative sample (Stage III)		Required: Prototype trials (Stage I) + Field Experiment (Stage II) + Field experiment with representative sample (Stage III)	
Risk and side-effects detection 2, 6, 7, 15	Data management plan, DPIA + monitoring side-effects	Learners and teachers can easily report problem/incidents	Active problem/incident detection procedure with additional informed consent	
Explainability and severity 3, 6, 8, 9, 10	Theory and design-driven pre-registered experimentation with a clear description of the process and responsibilities			
Democratic oversight 11, 12	Ethics committee + Institutional student and professional advisory council	Either the ethics committee or the advisory council can escalate issues to a public committee which could include a voice for organized professional and student bodies, or democratic representatives		Escalation to public committee is required

4.1 Regulatory bodies and government

Regulatory bodies and governments played an essential role in the creation and implementation of ethics committees. The first independent ethics committee for research was organized in 1953 by the National Institutes of Health [16]. In 1964, the world medical association published the declaration of Helsinki which provides a set of ethical principles for experiments with human subjects [16]. In 1991 the government of the United States enacted the Common Rule as a baseline ethics standard to which all government funded research should be held [16]. Other countries, funding organizations, scientific organizations, and journals followed suit by requiring approval from an independent ethics committee. This brief history shows how associations, grant providers, and a national government can effectively initiate ethical oversight through funding requirements. The medical domain has had an additional impetus through jurisdiction enforced by the U.S. Food and Drug Administration (FDA). Any new drug needs to be approved by such a national agency before being allowed access to its market. The combination has ensured that both experiments with humans conducted at the university and in companies (or collaborations) comply with ethics committee requirements. AI based EdTech does not yet have to comply with requirements from specific national agencies such as the FDA before being allowed access to a market. However, data and privacy are legally protected through laws such as General Data Protection Regulation (GDPR) in the European Union. These requirements are reviewed by data management officers and the ethics committees. Regulations of AI such as the recent AI act of the European Union could lead to a similar market barrier for AI-based EdTech. This could eventually mean that organizations that publish AI technology should undergo a similar process as pharmaceutical companies who bring medical products to market. The first procedure proposed in this article, is inspired by this process and adjusts it to the educational context by replacing (in vitro) lab tests that are common in pharmaceutical trials with design-based research or educational research design to create prototypes and use cases that are then tested in small controlled settings. Regulatory bodies can, therefore, play an essential role both in initiating and catalyzing these procedures throughout the community. If, in contrast to the medical field, the introduction of new technology is interpreted as uncontrollable ("what can be made must be made" [3], then it could still be seen as the responsibility of (inter)governmental bodies to organize or require these trials post-hoc in order to formulate fitting policy measures for regulation [5].

4.2 Industry standards

When it comes to industry standards there are two processes to consider: research and development processes and quality management. Although this varies based on the domain, in many cases a substantial part of research and innovation is organized by companies. Some companies have exclusive access to unique large datasets and some run several large-scale field experiments on a daily basis [25]. Outcomes of this research are not often publicly available or published because of corporate interests. When it comes to ethical treatment of human subjects, however, these research practices should be held to the same standards. This could require regulation such as mentioned in the previous paragraph, but it can also be done in collaboration with companies or, in some cases, even be initiated by companies. Collaboration with companies can improve the feasibility of and compliance with the procedure. When regulation is too stringent it can stifle innovation or drive companies to more accommodating countries, while too little regulation can lead to monopolies and undermine consumer interests or rights. Currently, China, the EU and the US approach this balancing act in remarkably different ways [20]. The second process is quality management: decades of development in quality management show how it can be in the interest of companies to continuously improve and optimize working processes. There are similarities between the piecemeal step-by-step approach that is proposed as procedure 1 in this article and corporate practices such as developing beta-versions and improving them in small controlled environments before eventually releasing them for public use. When it comes to developing and using AI systems, several relevant existing and new standards are available: general quality management standard ISO 9000 (international set of standards for meeting regulatory requirement, customer satisfaction and continual improvement) and more specific standards for AI and ethics such as ISO/IEC 42,001 (standard for maintaining and continually improving of AI management systems) and IEEE 7000 (process for addressing ethical concerns during system design). These standards can prove to be a great industry-wide tool for improving responsible use of AI. However, they do not provide requirements for responsible research practices, do not address specific educational concerns, and they are voluntary, which makes them complimentary and not sufficient for the purposes addressed in this paper.

4.3 Educational institutions

As a customer, or an intermediary to learners and teachers, educational institutions should demand reliable scientific evidence about the benefits and side effects of AI-based

EdTech before purchasing or endorsing it. Not doing so could infringe on universal rights of children such as the right to non-discrimination, for example by exposing them to applications based on biased data (e.g., proctoring software with racial bias) [35]. It could also undermine resilience by making them dependent on the software of large tech companies. These are the risks that should stimulate educational institutions to have a high standard. Yet, simultaneously, they have a stake in stimulating responsible innovation and improvement of educational practices through technology. They have both a stick and a carrot with which they can stimulate responsible experiments with AI in education. It is important to note here that there are large differences between the different domains in education, based on the specific populations that they serve. Parents will be a more important stakeholder in pre-K domain than in post-secondary education, for example. The different types of educational institutions are also organized differently: this often means that individual institutions have to organize and collaborate in associations in order to advance their interests. There are, however, interests that they have in common, by being the wardens of human learners. The proposed procedures in this article are based on universal rights for human subjects. This allows them to function as the ethical baseline that could function for all educational domains, with potential additional amendments for special or vulnerable populations.

5 Discussion

The conditions that Van de Poel [34] formulated for responsible experiments with experimental technology are highly relevant now that AI developments have taken flight with the recent breakthroughs in generative AI. Those conditions were, however, not yet translated into practical procedures that could be applied to design and judge experiments with AI-based EdTech. According to the AI in education literature, both rigorous experiments and procedures for organizing them are wanting but essential for further development of the field [7, 18, 19, 21]. This article contributed to the literature by deducing practical procedures from the conditions that Van de Poel [34] proposed, and by differentiating the required procedures based on the levels of automation model of Molenaar [27] in order to reduce unnecessary bureaucratic requirements.

An issue with AI-based technology that these procedures do not fully 'solve', is the matter of 'dual-use'. For example, AI-based technology that is designed to develop antigens, can also be used to create chemical weapons [32]. Similarly, within the domain of AI-based EdTech, for example, large language models can be used by students to receive feedback, but they can also be used to commit fraud [1]. The proposed procedures in this article provide a manner in which a specific user-case can be studied. Unexpected other user-cases might be overlooked by both the researcher and the ethics committee. However, we do believe that the proposed procedure for risk and side-effect detection offers a responsible way to address this and explore what unintended dual-uses might exist. The process of gradual testing additionally ensures that this is done in a relatively controlled and responsible way. As mentioned above, current regulations (or lack thereof) might mean that this only occurs after the introduction of technology. We argue that, even so, this would be worthwhile as a clinical approval and requirement for public support or uptake by state funded organizations.

The procedures proposed in this article offer guidelines for designing and assessing responsible research into AI-based EdTech. They are prior and complementary to regulation of the use of AI technology such as a DPIA and governmental AI regulation. As such, the proposed procedures do not focus on the impact or scope of the technology. Rather, they provide the infrastructure that generates the factual basis for decision making and regulation. This proposal should, therefore, be distinguished from and seen as complementary to national or intergovernmental regulations such as the EU AI-act.

In this article we used the levels of automation model to differentiate between AI-based EdTech. This is a model that has mostly been used in the context of personalized learning, which is the most prevalent but not the only type of AI-based EdTech [18]. One could question whether it translates or applies to the other types: teacher assistance and optimization of the educational organization. Indeed, the levels of automation model is too crude to do justice to all the dimensions in which the applications that can be found within these types differ from each other. However, when it comes to ethical and legal responsibility, the model offers the most important distinctions that also apply outside of personalized learning. As soon as AI-based teaching assistants take over responsibilities of teachers with less or no teacher oversight, a morally fundamental line is crossed. The same holds for instances in which dashboards or scheduling software lose human oversight.

One could argue that an alternative or additional dimension to automation could be 'explainability': the degree to which AI applications allow human users to comprehend and trust the output or results [12]. More untransparent applications are less comprehensible and predictable, which makes it harder to prevent or detect errors, thereby increasing risks of unwarranted conclusions or solutions. Explainability could be improved by, for example, providing the sources that were used to derive answers. Within the scope of this article, however, we are interested in the scientific

process of designing evaluations of the effectiveness of AI-based EdTech. From this meta-perspective we argue that mechanisms should be defined and tested to help explain the effects of the intervention on the subjects. We propose addressing this with the (third) procedure of 'explainability and severity'.

While this article provides a complete theoretical overview of required moral procedures, it does not yet test these empirically. Future studies can build on this theoretical contribution by further specifying which stakeholders should be involved in which part of the process and by consulting them about the feasibility and desirability of their involvement. User friendliness of the procedures could be tested by applying the procedures to different cases in order to find out whether these processes could be further streamlined. Even though the proposed procedures attempt to reduce any unnecessary requirements, they could still deter researchers as 'yet another requirement'. In addition to these procedures, a clear and practical primer for researchers on designing experiments with experimental technology could help researchers to navigate through the legal and ethical requirements.

Finally, the procedures in this article are supposed to be used by researchers and ethics committee members, but their implementation would require collaboration with and between policy makers, AI developers, educators, and educational institutions.

6 Conclusion

In this paper we derived four procedures that can be used to design and assess responsible experiments with AI-based EdTech. The first procedure is a process of gradual testing. This entails that AI-based EdTech are recommended to undergo three different stages before they are to be approved for public and widespread use by public educational institutions. From the fourth level of automation on, when there is no longer guaranteed oversight from a teacher, this should be required. The second procedure is risk and side-effects detection. This means that researchers should design their study in a way that (i) problems/incidents can be detected and (ii) that side-effects can be explored based on expected trade-offs or more exploratory qualitative methods and longer term impact measurement. Based on the level of automation additional requirements are needed for problem reporting. The third procedure helps to increase the explainability of AI-based EdTech and the severity of their evaluation. This requires, on the one hand, a clear description of the process, responsibilities and stopping rules. On the other hand, this requires substantive design principles and a prediction about the mechanism through which the intervention is expected to work. Both can be integrated in a pre-registration. Finally, the fourth procedure is targeted at democratic oversight. The relevant stakeholders should be involved if experiments with AI-based EdTech can have public and lasting repercussions. For ('level 2') interventions which only assist a teacher, current research ethics committees and local advisory councils which include members of relevant stakeholders can suffice. For interventions of the third level of automation and further, either ethics committees or local councils should be able to escalate decision making to a public council that includes both experts and representatives of the relevant stakeholders. With automation level 6, when there is no longer any teacher oversight, escalation to the public council is required.

Educational institutions and educational researchers are looking for ways to responsibly experiment with the rapidly evolving applications of AI in education in order to provide the much needed facts for evidence-informed decision making. By translating the 16 ethical conditions for responsible experimentation with experimental technology into four practical procedures, this article provides a specific proposal for guidelines that can be used to design and judge these experiments.

Acknowledgements The authors would like to thank Bhoomika Agarwal, Anders Bouwer, Wayne Holmes, Inge Molenaar, and Marthe Stevens for valuable feedback on concept versions of the manuscript.

Author contribution Izaak Dekker: Conceptualization; Writing - original draft; Writing - reviewing & editing. Bert Bredeweg: Writing - reviewing & editing. Wilco te Winkel: Writing - reviewing & editing. Ibo van de Poel: Writing - reviewing & editing.

Funding This work was supported by the Nationaal Regieorgaan Onderwijsonderzoek of the Dutch Ministry of Education, Culture and Science [Grant Number 40.5.23960.028].

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Abbas, M., Jam, F.A., Khan, T.I.: Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. Int. J. Educ. Technol. High. Educ. **21**(1), 10 (2024). https://doi.org/10.1186/s41239-024-00444-7
2. Akata, Z., Balliet, D., De Rijke, M., et al.: A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer. **53**(8), 18–28 (2020). https://doi.org/10.1109/MC.2020.2996587
3. Anders, G.: The obsolescence of man. Vol. 2. On the destruction of life in the epoch of the third industrial revolution. Chapter 2:

The obsolescence of appearance. (1980). https://files.libcom.org/files/ObsolescenceofManVol%20IIGunther%20Anders.pdf
4. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, O., Mariman, R.: Generative Ai can harm learning. SSRN. 4895486 (2024). https://doi.org/10.2139/ssrn.4895486
5. Beauchamp, T.L., Childress, J.F.: Principles of Biomedical Ethics. Oxford University Press (2001)
6. Bockting, C.L., van Dis, E.A., van Rooij, R., Zuidema, W., Bollen, J.: Living guidelines for generative AI—why scientists must oversee its use. Nature. **622**(7984), 693–696 (2023). https://doi.org/10.1038/d41586-023-03266-1
7. Bond, M., Khosravi, H., De Laat, M., et al.: A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. Int. J. Educ. Technol. High. Educ. **21**(1), 4 (2024). https://doi.org/10.1186/s41239-023-00436-z
8. Brady, A.C., Griffin, M.M., Lewis, A.R., Fong, C.J., Robinson, D.H.: How scientific is educational psychology research? The increasing trend of squeezing causality and recommendations from non-intervention studies. Educ. Psychol. Rev. **35**(1), 37 (2023). https://doi.org/10.1007/s10648-023-09759-9
9. Bredeweg, B., Kragten, M.: Requirements and challenges for hybrid intelligence: a case-study in education. Front. Artif. Intell. **5**, 891630 (2022). https://doi.org/10.3389/frai.2022.891630
10. Chaudhry, M.A., Kazim, E.: Artificial intelligence in education (AIEd): a high-level academic and industry note 2021. AI Ethics. **2**(1), 157–165 (2022). https://doi.org/10.1007/s43681-021-00074-z
11. Cook, T.D.: Randomized experiments in educational policy research: a critical examination of the reasons the educational evaluation community has offered for not doing them. Educ. Eval Policy Anal. **24**(3), 175–199 (2002). https://doi.org/10.3102/01623737024003175
12. Cortese, J.F.N.B., Cozman, F.G., Lucca-Silveira, M.P., Bechara, A.F.: Should explainability be a fifth ethical principle in AI ethics? AI Ethics. **3**(1), 123–134 (2023). https://doi.org/10.1007/s43681-022-00152-w
13. Dekker, I., Meeter, M.: Evidence-based education: Objections and future directions. Front. Educ. 7941410 (2022). https://doi.org/10.3389/feduc.2022.941410
14. Ezenkwu, C.P., Starkey, A.: Machine autonomy: Definition, approaches, challenges and research gaps. In: Arai, K., Bhatia, R., Kapoor, S. (eds.) Intelligent Computing. Comp Co. 2019. Advances in Intelligent Systems and Computing, p. 997. Springer (2019). https://doi.org/10.1007/978-3-030-22871-2_24
15. Fleckenstein, J., Meyer, J., Jansen, T., Keller, S.D., Köller, O., Möller, J.: Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. Comput. Educ. Artif. Intell. **6**, 100209 (2024). https://doi.org/10.1016/j.caeai.2024.100209
16. Grady, C.: Institutional review boards: purpose and challenges. Chest. **148**(5), 1148–1155 (2015). https://doi.org/10.1378/chest.15-0706
17. Greene, J.A.: What can educational psychology learn from, and contribute to, theory development scholarship? Educ. Psychol. Rev. **34**(4), 3011–3035 (2022). https://doi.org/10.1007/s10648-022-09682-5
18. Holmes, W.: The unintended consequences of artificial intelligence and education. (2023)
19. Holmes, W., Tuomi, I.: State of the art and practice in AI in education. Eur. J. Educ. **57**(4), 542–570 (2022). https://doi.org/10.1111/ejed.12533
20. Hutson, M.: Conflicting visions for regulation: China, the EU, and the US have different approaches to reining in artificial intelligence. Nature. **620**, 260–263 (2023). https://doi.org/10.1038/d41586-023-02491-y
21. IEEE. White paper - Ethically aligned design-A vision for prioritizing human well-being with autonomous and intelligent systems: (2019). https://ieeexplore.ieee.org/servlet/opac?punumber=9398611
22. Jordan, S.R.: Designing an Artificial Intelligence Research Review Committee. In: Future of Privacy Forum (2019)
23. Jordan, S.R.: Designing artificial intelligence review boards: Creating risk metrics for review of AI. In: 2019 IEEE International Symposium on Technology and Society (ISTAS). IEEE;:1–7. (2019). https://doi.org/10.1109/ISTAS48451.2019.8937942
24. Lakens, D.: When and how to deviate from a preregistration. Collabra Psychol. **10**(1), 117094 (2024). https://doi.org/10.1525/collabra.117094
25. List, J.A.: The voltage effect: how to make good ideas great and great ideas scale. Currency (2022)
26. McCarthy, J.: From here to human-level AI. Artif. Intell. **171**(18), 1174–1182 (2007). https://doi.org/10.1016/j.artint.2007.10.009
27. Molenaar, I.: Towards hybrid human-AI learning technologies. Eur. J. Educ. **57**(4), 632–645 (2022). https://doi.org/10.1111/ejed.12527
28. Mollick, E.R., Mollick, L.: Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts. SSRN. (2023). https://doi.org/10.2139/ssrn.4391243
29. Mollick, E., Mollick, L., Assigning, A.I.: Seven approaches for students, with prompts. arXiv. (2023). https://doi.org/10.48550/arXiv.2306.10052
30. Nasim, S.F., Ali, M.R., Kulsoom, U.: Artificial intelligence incidents & ethics: a narrative review. Int. J. Technol. Innov. Manag. **2**(2), 52–64 (2022). https://doi.org/10.54489/ijtim.v2i2.80
31. Stahl, B.C., Andreou, A., Brey, P., et al.: Artificial intelligence for human flourishing–beyond principles for machine learning. J. Bus. Res. **124**, 374–388 (2021). https://doi.org/10.1016/j.jbusres.2020.11.030
32. Urbina, F., Lentzos, F., Invernizzi, C., Ekins, S.: Dual use of artificial-intelligence-powered drug discovery. Nat. Mach. Intell. **4**(3), 189–191 (2022). https://doi.org/10.1038/s42256-022-00465-9
33. Umbrello, S., Van de Poel, I.: Mapping value sensitive design onto AI for social good principles. AI Ethics. **1**(3), 283–296 (2021). https://doi.org/10.1007/s43681-021-00038-3
34. Van de Poel, I.: An ethical framework for evaluating experimental technology. Sci. Eng. Ethics. **22**(3), 667–686 (2016). https://doi.org/10.1007/s11948-015-9724-3
35. Yoder-Himes, D.R., Asif, A., Kinney, K., Brandt, T.J., Cecil, R.E., Himes, P.R., Cashon, C., Hopp, R.M.P., Ross, E.: Racial, skin tone, and sex disparities in automated proctoring software. Front. Educ. **7**, 881449 (2022). https://doi.org/10.3389/feduc.2022.881449
36. Zhao, Y.: What works may hurt: side effects in education. J. Educ. Chang. **18**(1), 1–19 (2017). https://doi.org/10.1007/s10833-016-9294-4