Ethics and Technology of the Mimetic AI Professor

*James Brusseau (jbrusseau@pace.edu) Pace University, New York, USA & University of Trento, Trento, Italy

Sara Falcone Pace University, New York, USA

Anirudh Kolanupaka Pace University, New York, USA

Thilak Narasimhamurthy Pace University, New York, USA

Nidhi Pandya Pace University, New York, USA

Abstract

A mimetic AI professor reproduces a specific human professor's pedagogical method. Within an academic domain, the vocabulary, conceptual framing, and interaction patterns defining the AI mirror those of the human professor it emulates. In this model, knowledge is represented as accurate replication of a subject-matter expert, as opposed to generic subject-matter expertise.

This paper provides an ethical and technical foundation for mimetic AI pedagogy by addressing three questions. First, it defines the concept of the mimetic AI professor and outlines why it may provide learning enhancement. Next, it outlines potential instructional roles the AI could play. Then, core human dilemmas that mimetic systems provoke are analyzed, and design responses are proposed. Finally, the paper consolidates as an ethics framework for mimetic AI professors containing these categories of ethical analysis: Autonomy, Dignity, Privacy, Fairness, Equity, Social wellbeing, Explainability, Performance. The framework intends to be useful for professors, engineers, and universities designing their own mimetic AI tools in the education domain.

Keywords: Mimetic AI, Mimetic AI professor, ethics, ethics framework, education

1. Introduction to the mimetic AI professor

There are AIs that will play generic but skillful chess against human opponents, and there are also AIs that will play specifically in the style of Magnus Carlsen or Bobby Fischer [McIlroy-Young et al., 2022]. This second approach is mimetic in the sense that it replicates a specific human's unique style of capability more than level of capability. It copies players' strengths and also their idiosyncrasies [Bukingolts 2025]. For this second AI, what *counts* as knowledge is accurately imitating a subject-matter expert, as opposed to gaining subject-matter expertise.

The distinction between expertise in the subject-master and in the subject-matter is significant in university classrooms where professors from diverse intellectual backgrounds employ distinct vocabularies and methods to explore similar ideas [Zinn 2004]. For example, the perennial philosophical question asking whether events exist objectively and outside of human perception (If a tree falls in a forest and no one hears it, does it make a sound?) can be addressed from a continental European orientation that contrasts idealism and perspectivism, and then works within a historical context. Or, the problem can be addressed by a professor grounded in more analytic thought who prefers to contrast realism and anti-realism, and then depends on more recent scholarship to address the dilemma [Braver 2007].

For students, these differences are significant. Those who have taken courses in one tradition may be

baffled by the lessons of professors framing their thought inside the other [Prado 2003]. More, there is not a single divide. Many professors develop their own personal keywords and paradigmatic examples to lend a personalized coherence to their work [Lowenstein 2015].

These coherences create rich intellectual human communities, but they are dissonance for language models oriented toward generic subject mastery because their knowledge-production functions by blending philosophy training data into homogenized language. The *way* conventional AI language models produce understandings is by flattening applicable data into a single set of mathematical embeddings that sheer away human particularities. There are manifest efficiency advantages to this amalgamizing, but the result may nevertheless verge on babel in the real world where unique students learn from specific teachers. [Xu et al., 2024].

Translating the conclusion into AI for pedagogical support in philosophy – and perhaps not only philosophy – students may do better to access an AI agent trained to accurately represent the particular expertise of their known professors, instead of a model homogenizing the expertise of all professors even if that second source is more knowledgeable in a cumulative or aggregative sense.

There are further advantages to the mimetic AI professor, though they are not unique to the kind of AI that is built to replicate a specific person. First, the tool is available on-demand for every student, and simultaneously to many students. Further, research shows that some students find AI agentic interlocutors to be less intimidating than humans, and consequently they are more likely to seek help with misunderstandings. [Kweon et al., 2025].

To prioritize these human-centric solutions to pedagogical challenges, the mimetic AI professor emerges from consistent design thinking. There is an empathy step that observes and understands the student struggle with learning support availability, and with the homogenized language of generic models. The project also employs test and refine cycles by proposing iterative responses to mitigate ethical risks like autonomy loss and grading unfairness. The innovation of cascading fidelity to the human professor will be pertinent here, and also representative of a design approach that ensures AI is not merely functional technology but also aligned with perennial university values.

Finally, there are drawbacks to AI in the university, but the potential benefits require a systematic ethical and technical analysis of the digitally replicated professor. The purpose of this paper is to introduce that analysis by describing emerging ethical dilemmas surrounding the technology, and by proposing design responses. To begin, the forms and roles of this particular AI are outlined, and then the ethical implications and design responses are elaborated. Finally, the result will be an ethics framework capable of managing the emergent technical reality. To authors' knowledge, this is the first work to comprehensively address the ethics of the mimetic AI professor.

2. What forms could the mimetic AI professor take?

Within the restraints of current technology, a high-fidelity digital twin professor could operate in three modes: textual, verbal, and visual interface.

Minimally, the tool interacts textually. The challenges of replicating the human professor cluster around word choices, the type of examples used, historical framing, analogies and counter-examples. So, fidelity at the level of text is about choosing and arranging words to produce a familiar instructional presence.

Technically, this mimetic professor will operate along a course-bounded retrieval-augmented generation pipeline. Initially, course materials must be gathered to train the machine, including recordings of lectures, slides used in the sessions, and written materials including textbooks or the

professor's own writings. This multimodal information is converted by automated processes into text strings, and then those strings may be chunked into citation-ready passages. They are labelled with their subject matter, their origin, and other relevant meta-data, and then indexed in a database such as ChromaDB. At runtime, a query is submitted, and the system retrieves candidate chunks, filters low-quality or off-topic text, selects supporting evidence sentences, and generates a response. Concretely, a two-stage LoRA fine-tuned model could be employed, with the first stage trained on about 3,305 conversational pairs and the second stage trained on 580 citation-grounded pairs. Then, the responses could be formulated to include paragraph-level citations via vector similarity to retrieved chunks. Students who query the system, consequently, would not only receive a response but also be directed back to the specific course materials supporting the response. In accord with the metadata originally attached to the textual groupings, the source may be a timestamp in the recording of a lecture, or a passage in a textbook. Finally, in cases where retrieval confidence is insufficient to link with a query, the course-bound system would abstain and a more general answer may be provided, along with appropriate notification.

It must also be added that rapidly advancing technology quickly renders past tools and practices outmoded. But, even while the capabilities of specific software may be eclipsed, the larger design strategy of producing responses that are course-bound and traceable to course materials remains in effect.

One level above textual interaction there is verbalization. It requires textual fidelity and then also aurally echoing the professor's distinctive vocal characteristics, including tone, cadence, habitual emphasis, rhetorical timing, and similar. Then, all these characteristics need to be translated into high-quality synthetic speech [Muskaan 2025].

One way to implement verbalization is by wrapping the grounded text pipeline with speech interfaces. Student questions are captured via speech-to-text, routed through the same course-bounded retrieval and generation stages, and the resulting answer is delivered through text-to-speech technology. Current off-the-shelf tools include those offered by 11Labs, and Tavus. Because the conversion of text to speech is so widely applied across industries, there is no apparent reason to develop a proprietary approach for a mimetic AI professor. Simply using industry state of the art tools should serve the project well. Of course, the grounding rules remain unchanged: the voice response is constrained by retrieved course evidence, and the system explicitly signals when course support is insufficient rather than improvising or hallucinating responses to queries.

Finally, in authors' experience, there is a danger in the voice stage. Because the sound of a particular human's voice can now be reproduced with high fidelity, it can be easy to overlook content insufficiencies. For that reason, a mimetic AI professor's technical performance should be validated on the textual level before adding multimodal wrappers.

One level above verbal interaction there is visual embodiment of the mimetic AI, which could be understood today as a screen-based presence, compatible with the experience offered by a human professor in a synchronous, online course. To reach this level, the textual and verbal imitation is ideally supplemented by a photorealistic (or, potentially stylized) avatar. This pixelated agent would exhibit synchronized non-verbal cues—such as micro-expressions, hand gestures, and eye contact—creating an immersive face-to-face instructional environment [Longoni et al 2026].

Here too state of the art industry technology may be sufficient to meet educational design choices. As layered over the grounded dialogue system, the visual user interface would provide pedagogically relevant usability and transparency features such as captions/transcript, turn-taking cues, and source viewing for cited materials. To distinguish course-grounded statements from lower-confidence continuations, visual cues could be engineered. For example, the face could simply freeze when the words being generated cannot be confidently attributed to the professor's own course materials.

Regardless of the replication mode, the underlying and driving textual AI remains unchanged: course-bounded retrieval, constrained generation, and auditable citations.

3. What roles could the mimetic AI professor play?

Three roles for the mimetic AI professor have been established: resource, lecturer, evaluator [Woolf 2010].

As a resource, the tool could respond to student questions on-demand. This availability holds value according to research showing that questions are prime learning opportunities, but also ones that collapse into frustration when no authority is readily available to respond. Further, the AI may positively substitute the human even when both are available in circumstances where students feel intimidated or judged by the human teacher. When this awkwardness exists in face-to face interaction, students may find an AI agent to be less psychologically challenging, and so be open to support from a mimetic AI professor even when the opportunity to interact with a human is refused [Kweon et al., 2025].

As a lecturer, a sufficiently capable AI could directly replace a human professor in specific instances and in limited ways. Of course, *entirely* replacing the human is non-sensical as there needs to be an original for the tool to emulate. That said, a high-level mimetic AI could be called upon to present a subject to students, it could also provide reviews for students who find specific material challenging, or present supplemental lectures entering subjects at greater depth.

In technical terms, this lecturing role extends the same grounded language model from single answers to structured teaching sequences. To make that jump, the system could be designed to generate short mini-lectures following a repeatable pattern - i) outline; ii) explanation; iii) example; iv) brief check-in with audience - while retrieving supporting course evidence for each segment to prevent drift. This would enable supplemental explanations, review sessions, and deeper dives while remaining consistent with the professor's vocabulary and framing.

The third mimetic AI role is evaluation. Currently, and as humanly managed, grading is tedious for professors, and the actual responding to an exam provides few pedagogical benefits for students. It has also become clear that AI is subverting the entire process of assessment by making cheating easy in many cases [Gray et al 2025]. These problems can be addressed by mimetic AI, first by changing what is assessed, and then by changing how assessment functions.

3.1 What does mimetic AI assess?

When a mimetic AI grades a student, what is measured is *not* how correct or incorrect an answer is generically. That model for achievement functions better for the impersonal AI agent professor built from the aggregation of all available training data on the broad subject domain. Student work would then be classified by referring to the homogenized information.

Mimetic AI is different because quality is measured not as correctly reflecting the aggregated data, but by gauging fidelity to a subject-matter expert. So, just as the lecturing mimetic agent tries to reproduce what a specific professor would say in a defined context, so too the grading agent aims to measure student performance within the personalized constraints. The right answer is not the one cohering with homogenized data, but the one most accurately evoking the subject-matter expert's own thought and words. *What a student tries to get right is the professor, not the material*.

3.2 How does assessment function?

Besides the potential change in what is graded, the mimetic AI professor could also shift how grading typically occurs: from written to oral examinations. While this distinction is among the most venerable subjects of academic pedagogical discussion [Huxham et al., 2012], the practical reality is that few human professors today have the luxury of time to evaluate students individually [Eachempati et al 2025]. This is just the kind of problem that could be resolved by the force multiplier that is AI, however. So, the technology breathes new significance into the old discussion about exam formats.

First, how would oral examination by a mimetic AI professor function? In accord with a standard human oral exam [Theobold 2021], the tool could pose a question, the student would respond, and the tool then evaluates the response. If sufficient, the next question is addressed. If the response is inadequate, the mimetic AI could potentially replicate the human practice of providing cues and nudges to aid the student's understanding and provoke a second, improved response. Then, the process repeats until reaching a satisfying answer. Finally, the resulting score would derive from the scaffolding required by the student to reach proficiency. For the AI, that may translate into the number of rounds needed to reach the satisfactory answer, with fewer rounds corresponding to a higher grade.

More technically, the mimetic scoring combines semantic alignment to the professor's concepts with checks for required course terms/distinctions and simple reasoning-structure cues.

If mimetic AI technology met the practical challenges of implementing this dynamic assessment, the positive disruption would be significant in at least three ways. First, the experience of assessment would shift from interruption to prolongation of learning. Conventional written exams allow only limited, asynchronous feedback and no guarantee that students will reach the correct answers. Oral exams, by contrast, function by providing immediate feedback, and only end when the correct answer has been reached [Eachempati 2025]. In a sense, an oral exam means everyone gets 100%, even while a meaningful hierarchization is achieved. Learning, consequently, is primary, and hierarchization is secondary. For written exams, hierarchization is primary and learning secondary.

The second positive disruption attributable to the move from written to oral exams is a shift in the professor-student relationship from antagonistic to cooperative [Joughin 1998; Dicks et al., 2001]. Instead of the professor understood as a critic of the students' work and a detractor from it in the sense of points deducted for each imperfection in a written answer, the relation converts into a collaboration when the student and professor – whether human or mimetic AI – struggle to reach the final result together.

Finally, traditional written exams, at least those written on computers instead of paper, have become increasingly vulnerable to cheating. While this problem can be solved with conventional oral exams, or with pencil and paper testing, those strategies resurrect the time and tedium problems, while the mimetic AI professor dissolves them.

Overall, mimetic AI may positively disrupt teaching and grading practices as depicted in Table 1.

Table 1

Mimetic AI as positively disrupting teaching and grading

	Human professor	**Mimetic AI professor**
Availability for student questions and discussion	Limited	On-demand
Intimidation factor in student interaction	May be perceived as judgmental, intimidating	May be perceived as judgement free, risk-free, accessible

Range of instruction	Limited to needs of overall class	Personalized to each student's challenges, aptitudes, and interests
Logic of hierarchization in assessment	How much is wrong?	How much scaffolding to get the response right?
The relation between assessment and learning	Interruption (Hierarchization primary, learning secondary)	Contribution (Learning primary, hierarchization secondary)
Assessor's relation to student	Antagonistic	Collaborative
Assessment vulnerability to cheating	Increasing	Low

4. What are the ethics of the mimetic AI professor?

With forms and roles of the mimetic AI professor established, the range of ethical dilemmas surrounding the technology can be projected.

To provide an initial frame, there exists theoretical work done in broader AI fields, starting with the 2019 publication of the *Ethics Guidelines for Trustworthy AI* [European Commission, 2019]. That work, along with related advances including the German Data Ethics Commission recommendations [Data Ethics Commission 2019], organized standards for the human effects of recent information systems as tri-partite [Brusseau 2020]. First, there are a set of moral principles surrounding the individual, including human autonomy, dignity and privacy. There follows a cluster of principles applying to the wider society, including fairness, equity, and social wellbeing. Finally, there are principles joined to machines and their operation, including explainability, safety, and performance [Brusseau and Turchet 2026].

This broad ethics framework has served well in AI healthcare where technical advances intersect with moral complications in coronary testing [Zicari et al 2021b], dermatology [Zicari et al 2021b], and radiology [Allahabadi et al 2022].

More recently, focused work has demonstrated how the baseline ethical theories can be pruned and supplemented to account for narrow ethical contexts. The internet of musical things [Brusseau and Turchet 2024], for example, as well as the musical metaverse emphasize traditional AI ethics concerns including privacy, while also introducing an ethical inclination toward creativity as a value and principle.

With respect to mimetic AI in university classrooms, the standard set of ethical values can be slightly edited to not include safety, as that imperative better suits technology entailing potential physical harm. Driverless car engineers, for instance, face potentially lethal ethical challenges, as do healthcare AI tools when statistical methods determine whether chest x-ray anomalies or skin blemishes merit flagging as cancer risks [Zicari et al 2021b]. The classroom, by contrast, presents no severe biological risks.

Finally, some introductory work has been done in the area of ethics and mimetic AI pedagogy. In one paper, historical figures are digitally cloned to provide history lessons to students - Caesar discusses his military campaigns – and the authors underline ethical concerns about student dignity and the AI agent as a source of human truth [Oger et al., 2025]. In another paper, authors summarize the equitable benefit of AI on-screen instructors insofar as they reduce lecturing costs and so contribute to greater educational access, notably in remote areas [Pang et al., 2025]. Still, authors of this paper are unaware of any framework developed to comprehensively account for the ethics of the mimetic AI professor.

The following proposed framework responds to that absence, and is based on the following established catalog of AI ethics values: Autonomy, Dignity, Privacy, Fairness, Equity, Social wellbeing, Explainability, Performance. Each principle will be briefly defined, and then its perspective on the mimetic AI professor described. Finally, design-based responses to the ethical context will be proposed.

Autonomy

Autonomy is self-determination, both physically and mentally, as described by John Locke [Schouls 2018]. It implies the ability to set one's own purposes and interact intentionally with mimetic AI tools.

For the human professor employing a mimetic AI in the classroom, there exists a familiar autonomy dilemma. First, the decision to train and use the tool is an autonomy amplification: it broadens the professor's potential in terms of availability and capacity. Once that decision is made, however, there can also follow an autonomy narrowing, especially if the human professor becomes dependent on the stand-in, or if students begin taking the word of the AI as more authoritative than the human.

On the students' side, a similar duality emerges. Autonomy increases through on-demand support by converting frustrating uncertainties into immediate learning opportunities. And, it is also true that some students may fear being judged by a human professor, but find that querying an agent occurs without awkwardness. These autonomy increases are paired, however, with at least the following potential dependence. If answering questions becomes as automatic as asking them, then there is no reason for students to learn to answer for themselves.

In healthcare AI, an analogous form of the paradox between increased autonomy and dependency has been observed. Studies have shown that radiologists can use AI to accelerate and improve their diagnoses [Allahabadi 2022], but the tools may gradually dull their skills even while patients begin treating the digital replica as the higher authority, thus undermining the human doctor's actual agency [*Lee et al.* 2023].

Design response

How can the mimetic AI professor be designed to minimize autonomy risks? One core element is *cascading fidelity*. Building an AI to mimic a person also means designing it to accurately estimate when it is *not* replicating that person. When confidence in the replication accuracy is low, the tool can be designed to adjust. For example, if the AI is interacting only textually, this may mean that one colored font demonstrates confident mimetic accuracy, while another color shows that the machine has reverted to a broader, generic large language model process. If the interaction is verbal, the tone of voice could represent the change by switching from the professor's synthesized speech, to an anonymous tone, or the tool could simply announce that it is reverting from mimetic to generic modes.

Design for cascading fidelity could also be accomplished by tagging the textual output of the mimetic tool with the output's source. That is, outputs retrieved from the training data tied directly to the professor (recorded lectures, writings) are marked with those specific origins, while more general information drawn from the foundational language model is tagged as generic.

Regardless, the ethical concept is that the tool is helpful and also intentionally less helpful by presenting cascading levels of fidelity with the human professor. This keeps the humans at the center of learning by making them active participants in the interactions, and also in the judgement of the quality of those interactions.

The autonomy ethics opportunities, risks, and design responses are captured in Table 2 below.

Table 2

Autonomy opportunities, risks, and design responses

	Benefits	Risks	Design to mitigate ethical risks
Autonomy	Professor's force is multiplied Student on-demand learning, and judgement-free learning	Dependence on the machine reduces professors' capability and student independent learning	*Cascading fidelity* to require active instead of passive interaction.

Dignity

Dignity as described by Kant [1785/2012] is composed of two elements. On one side, human dignity represents protection against being treated as an instrument for others' purposes. It is protection against being used, commodified, rendered anonymous and replaceable, like a tool. On the other side, dignity is the right to accountability. It implies the ability to take responsibility for one's own actions, as opposed to the condescension of being treated as incapable, like a baby, or like those suffering mental decline.

In both senses, the mimetic AI professor is challenged by the ethics of dignity. First, digital twin technology commodifies human professors and students by reducing their interactions to numerical optimization measures [Tarnoff 2023]. There is the accuracy measured between the human professor and the digital twin in terms of teaching and of grading. Then, there comes the optimization of AI-student learning as compared – quantitatively – to human-student learning. Almost definitionally, the use of a mimetic AI professor violates the dignity prohibition of human reduction to anonymous, commodified, informational instruments in larger processes.

Stronger, for the AI to function, that dignity violation must have *already* occurred: quantification of qualitative human experience is the technical foundation of the project, it is how data is gathered and how the AI is built. So, it is true that mimetic AI technology can serve human uses, but it is equally true that the process reduces humans to tools of the technology

In the second sense of dignity, the aspect of accountability, there too it is clear that mimetic AI *performs* education, but without responsibility, even without understanding what responsibility for teaching would mean. In this sense, the work of the mimetic AI is a kind of counterfeit language of education [Pasquale 2020], and one that also has the character of diminishing the value of human-to-human learning. Stronger, just as counterfeit money is simultaneously illegitimate while also devaluing the legitimate currency, so too a mimetic professor undercuts education twice. It provides illusory pedagogy in the sense that it teaches without accountability for what is taught or how the learning is used and, second, it sows doubts about the value of human teaching. This is especially true if uncertainty is allowed to fester about whether a professor's words are generated by a human or by artificial intelligence.

Both these objections are serious, and difficult to directly rebut. Probably, the strongest ethical arguments against the mimetic AI professor stem from human dignity. The argument can be made, however, that there is a dignity loss for students in the AI classroom, but insofar as the pedagogical outcomes are superior to what they would have been without the AI, there is also a dignity gain in the sense that students have greater capabilities and potential to take accountability for their own decisions and lives in the future.

Design responses

Because some dignity loss is a condition of the possibility of building a mimetic AI professor, this kind of ethical objection cannot be completely overcome, only limited.

Two design steps can be taken. First, the tool can professionalized in the following sense. It performs narrowly, only within the domain of its expertise. For an agent tasked with a philosophy class, that means, questions will be entertained about ethical values, but if a student asks about the economic value of a used car, the mimetic agent will cease participating. This is not only because of performance concerns (the AI has not been trained to price cars) but also to ensure that the reification of human experience in terms of pedagogical data occurs across the narrowest possible slice of experience. The mimetic AI professor may limit students' dignity, but that limitation is itself limited.

The second step is transparency. As opposed to the sycophantic language characterizing interactions with many of today's mainstream language modes [Turner and Eisikovits 2026], the mimetic professor can sternly remind human interlocutors that its words are generated from statistical probabilities and information databases. In Europe, a version of this strategy for dignity protection has been instantiated by legally requiring AI tools to announce their nonhuman nature [Gstrein 2022].

While the imperative to transparency is morally powerful, it must also be stated that there exists a strong practical argument against it. Users may benefit from smooth, conversational interactions and find that the tool's awkward announcements of its own limitations becomes an impediment to learning. The preservation of human dignity, in other words, may twist the machine against itself.

The dignity ethics opportunities, risks, and design responses are captured in Table 3 below.

Table 3

Dignity opportunities, risks, and design responses

	Benefits	Risks	Design to mitigate ethical risks
Dignity	Increased learning	Commodification of the human when pedagogy is reduced to numerical scores Learning seems real but is counterfeit, an appearance without genuine human accountability	Narrow applicability Transparency about the professor's human/AI status

Privacy

Privacy is control over access to one's personally identifying information [Westin 1970]. As opposed to the rudimentary idea of secrecy, privacy is the ability to *decide* what aspects of life are, and are not secret. In this sense, privacy worries on the professors' side dimmish to the extent that they select the material for inclusion in their digital twin. Assuming knowledgeable consent, the requirement for privacy is technically fulfilled regardless of how much may be selected for divulging. Stronger, if the tool is well made, the potential for professors to curate and then project those aspects of themselves that they select for broad diffusion is a vigorous privacy benefit.

On the student side, questions arise along two vectors. First, persistence. Should the mimetic tool maintain information regarding past interactions with students? Of course, the information could be pedagogically useful. Future understandings build well on those of the past, and many students would benefit from continuity of discussion across a semester, or longer. Then, along the other vector, there is training. While it is obvious that students; own words will improve the AI performance over the course of an exchange, should one student's exchanges be employed to improve the tool's functioning for others? This is, of course, an established debate surrounding AI platforms, and the various arguments have been researched extensively [Mittelstadt et al. 2016].

Design responses

Design responses to the privacy imperative in personalized AI are well developed under the heading of privacy-by-design [Morley et al. 2020]. Most tenets of the approach apply well to the mimetic AI professor. On one side, steps can be taken to ensure knowledgeable consent about data use. Leaving users to decide for themselves how much personal information may be taken and stored is the fundamental move of privacy protection.

The accompanying design step is data minimization, and the engineering challenge is to do the most with the least. Assuming protection of personally identifying information as the paramount concern, minimization means, first, that no more data is collected than that required to perform the tool's function. Next, there is the accompanying imperative to limit data preservation: once personal information has been employed to the extent useful, it should be deleted. Of course, the definition of "useful" will be open to debate.

The privacy ethics opportunities, risks, and design responses are captured in Table 4 below.

Table 4

Privacy opportunities, risks, and design responses

	Benefits	Risks	Design to mitigate ethical risks
Privacy	Professors empowered to choose which aspects of themselves are projected Persistent personal information allows more effective educational scaffolding for students and training for future use	Personal information may escape control of professors/students	Privacy-by-design (Informed consent and data minimization)

Fairness

Fairness as formulated by Aristotle is the imperative to treat equals equally and unequals proportionately unequally within the decision domain [Aristotle 1934]. In the context of the mimetic AI professor, questions about equal and unequal treatment rise most immediately around grading where fairness implies that students with dissimilar abilities and dedication should receive proportionally unequal grades.

The fairness benefits of AI grading begin with objectivity. No human could match the machine's cold, computational regularity. AI assessors never grade generously after lunch, or become increasingly irritable as they progress through the thick exam stack.

It is also true, however, that probabilistic engines are prone to the occasional, wildly erroneous output [Manning and Schütze. 1999]. In a general sense, of course, even this unfairness is fair in the sense that episodes of automated mistreatment (and overly benevolent treatment) should balance over time. So, automated grading, even when prone to breakdown, may be fair in the same way as a coinflip: while one person wins and another loses, the initial odds were 50/50, and more coinflips will inevitably lead to balance. Still, that statistical assurance means little to the student whose paper has received a manifestly inaccurate evaluation.

Finally, there is the well-developed field of fairness research in bias stemming from imbalanced training data [Barocas et al 2019]. While most of that work has centered on algorithmic decisions about hiring, finance, and similar data-intensive decisions, there is no reason to believe that mimetic AI will sidestep the fairness vulnerabilities common to other AI models.

Design responses

Two established design responses fairness ethics dilemmas are explainability, and redress-by-design. Explainability is transparency about how the machine grades, so that errors can be traced back to their source and corrections made. The tradeoff, however, may be lower performance [Rudin 2019].

Redress-by-design could be more directly helpful to specific students. It is the incorporation of methods for students to appeal grades they believe to be unfair. Technologically, this appeal may entail strategies like running the same data numerous times with small perturbations to ensure the assigned grade is not a statistical outlier [Green 2018]. More generally, an appeals process could be formalized within the assessment mechanism as an appeal for grades to be re-assigned – and therefore confirmed or disconfirmed – by a human professor.

The fairness ethics opportunities, risks, and design responses are captured in Table 5 below.

Table 5

Fairness opportunities, risks, and design responses

	Benefits	Risks	Design to mitigate ethical risks
Fairness	Consistency of the machine	As black box, could produce any grade for any reason	Explainability Redress-by-design

Equity

Equity as based on John Rawls's philosophy is the principle of distributing the greatest advantage to the least advantaged [Rawls 1971]. In ordinary language, this is the principle that no one be left behind. Within the context of AI and education, disproportionate distribution in favor of the neediest suits the mimetic AI professor in several ways. First, the tool may be less financially expensive – and therefore more accessible – than a human expert. Further, as the AI always has infinite time and patience, increased inclusiveness in education may be understood as expanding opportunities for those with uncommon or disadvantaged learning abilities, styles, and interests.

The AI may also improve educational outcomes for students who may feel judged or intimidated by human professors, including those who may feel the insecurity for broader social, cultural, and historical reasons that may include being English language learners, or suffering racial and gender stereotyping [Feng et al 2025].

At least theoretically and pedagogically, the AI professor could operate underneath the Pareto curve, meaning more resources can be delivered to one person without subtracting them from others, and to the extent that premise is correct, no immediate equity risks arise, and no significant design implementations are required. In other words, it is difficult to imagine ethical objections to AI in terms of resource distribution so long as everyone is benefiting.

Social Wellbeing

Social Wellbeing aims to distribute the greatest good or happiness to the greatest number of people, in accord with the utilitarian philosophy of Jeremy Bentham and John Stuart Mill [Bentham and Stuart Mill. 2004]. The principle converts ethics into a happiness calculation (with the term "happiness" understood broadly) as summed across society's members, with higher total happiness implying ethical higher value. In the area of technological support for education, greater overall happiness generated by mimetic AI would convert into a higher ethics score.

As social wellbeing is a consequentialist philosophy, tensions will exist with rigid deontological stances, including Kant's categorical imperative and its prohibition of dishonesty. As Kant's detractors note, lying to others can find justification when it serves the happiness of the deceived (parents and Santa Claus, doctors and placebos).

For the mimetic AI professor, the tension between honesty and beneficent disinformation informs the question about whether the tool should inform users of its nonhuman status. Plausibly, some students achieve better learning outcomes when believing their interlocutor is human. For example, if a mimetic professor is tasked with writing brief responses to student participation on an online discussion board, students may respond more reflectively to those comments when believing they came from a human [Nass and Moon 2000]. If so, the positive consequence justifies opacity about the comments' origin. More generally, if educational outcomes improve when students believe they are working with human professors, then the illusion of humanity may be intentionally perpetuated and ethically justified by social wellbeing theory. The mimetic AI should claim to be human.

Of course, the same logic works the other way. While ambiguity about whether an AI is machine or human may lead students to broadly accept machines as human, it may also lead to the widespread rejection of humans under the suspicion that they may be machines.

Design responses

As originally engineered, the mimetic AI professor aims to digitally recreate a unique human's identifying particularities, which means that misleading users is an inevitable effect of technical success. Creating a good liar is a design goal. The ethical risk is that pedagogical social wellbeing may not be maximized because of failure to achieve the optimal level of transparency, whatever that level may be.

So, the design task is not to decide *whether* the tool should misrepresent itself as human, the task is to build a machine capable of implementing both truth and illusion about its own status.

There are levels of illusion. Students may not be informed whether an interlocutor is human or machine, but remain free to ask and then receive true information. Or, students may not be told, and not be provided with information if they ask. Or, the machine may answer that it is, in fact, human. (Hypothetically, the human could also claim to be a machine, in the right circumstances.) In any case, within the ethics of social wellbeing, the larger decision will be made in terms of which option generates the most overall happiness for a society when everything is taken into account, including educational outcomes, economic realities, and other ethical values.

The social wellbeing ethics opportunities, risks, and design responses are captured in Table 6 below.

Table 6

Social wellbeing opportunities, risks, and design responses

	Benefits	Risks	Design to mitigate ethical risks
Social wellbeing	Increased happiness resulting from pedagogical success	Limiting happiness outcomes by restricting AI ability, notably the ability to impersonate the human professor	Enable the tool to transparently reveal, and fully conceal its mechanical status

Explainability

Explainable AI can provide human-understandable reasons and justifications for its specific outcomes,

decisions, and internal logic [Adadi & Berrada 2018] When the mimetic AI professor lectures or assesses student work, what must be explained is how the database of professorial lectures and writings is converted into outputted words or grades.

Explainability discussions lead to accountability. When, for example, an AI assessment veers away from accurately reflecting a student's work, who is to blame? The human professor who provided the initial training data? The algorithms that processed the data? The software engineers who created the computing stack? The company that hired the software engineers? It can be hard to discern where responsibility lies when the earlier step has not been taken to understand exactly why the error occurred.

So, to ensure accountability for AI decisions in the classroom – or anywhere – there exists an ethical demand for explainability. It is also true, however, that understanding probabilistic systems is inherently difficult, and the requirement to make them transparent can lead to diminished performance [Rudin 2019].

Finally, explainability also holds legal relevance since systems that steer learning or evaluate learning outcomes can fall into EU AI Act high-risk education uses (Annex III), with major obligations and full roll-out foreseen by 2027 [Manganello et al 2025].

Design responses

Several design responses exist to the twin ethical risks of insufficient explainability and inadequate accountability. First, the tool can be engineered to clearly label the source of its claims within the training data. While these citations will not fully describe the technical process, they may confirm or disconfirm the accuracy of the response with respect to the ground truth, which is the professor's own recorded lectures and writings.

The second potential design response to explainability and accountability risks is interpretability [Sengupta et al 2025]. Even if it is impossible to understand the inner workings of an AI tool, what may be discernable is the kinds of inputs that would lead to preferable outputs. In assessment, this is especially important since students may be less interested in *why* one of their answers was found to be wrong than in what needs to be changed to make it right. This corrective knowledge is what is meant by interpretability, and it can exist independently of transparency into the machine's deeper workings.

Finally, the most direct design response to the opaque functioning of AI models is a redress-by-design element, one that builds mechanisms into the AI to facilitate human review of its outputs [Kaminski 2019]. These design mechanisms include efficient and direct ways for students to appeal grades to the human professor.

The explainability ethics opportunities, risks, and design responses are captured in Table 7 below.

Table 7

Explainability opportunities, risks, and design responses

	Benefits	**Risks**	**Design to mitigate ethical risks**
Explainability	Explainability helps to provide accountability – credit and blame – for AI outputs.	Forcing explainability can reduce model performance.	Output citation Interpretability Redress-by-design

Performance

Performance is the art of the technology itself, it is the value held *because* it functions in accordance with its purpose, and regardless of what the purpose is [Brusseau 2026)]. Of course, the purpose of mainstream humanities education rarely draws objections, but in a purely theoretical sense, a mimetic AI professor trained to teach the skills of art forgery or financial fraud would qualify for the same performance value as an AI trained for philosophy, assuming it performed equally well. As a more hyperbolic example, there is Nick Bostram's infamous paperclip machine, the technological tool programmed to make paperclips that executed so flawlessly as to convert the entire earth into a desolate paperclip manufacturing plant. While that outcome is undesirable, the attainment of the purpose – the performance – remains exemplary.

One performance risk facing the mimetic AI professor echoes the paperclip disaster, it is that the machine fails because it performs *too* well, it is not less than perfect but more than perfect. Concretely, if the mimetic AI truly replicates the human, then it should copy all the shortcomings as well, the implicit biases, the short-temperedness on Mondays, the decreasing care to carefully grading student exams as the afternoon wears on. In other words, the tool should intentionally forsake some of its own positive pedagogical qualities in the name of mimetic fidelity.

Design responses

The central design response to the performance risk of mimetic AI being *too* mimetic is *curated fidelity*. The optimum aspects of the unique professor are those embodied by the digital twin. The challenge is to isolate the positive qualities without falling into the trap of creating a generic professor, one designed to counter all possible objections through bland neutrality, in the same way that unremarkable cafeteria food is made acceptable to everyone by being objectionable to no one.

One strategy focuses on quality and selection of training data. Recordings produced by the professor for asynchronous online class lectures, for example, are well planned, with complete definitions and helpful examples, and this kind of material could be placed at the center of the training process. Recordings of live classes could also be used, potentially with some editing of those exchanges with students that do not go as planned. (Every professor has had the experience of choosing an example extemporaneously and poorly, and then talking themselves into a knot in front of an increasingly confused room of students.) No doubt different professors are characterized by their own kind of in-class mistakes and stumbling recoveries, but it is nevertheless possible to cut out those episodes without neutering the professor's uniqueness, without subtracting from the vocabulary, approach, and intellectual framing that characterize the professor in the classroom.

The performance ethics opportunities, risks, and design responses are captured in Table 8 below.

Table 8

Performance opportunities, risks, and design responses

	Benefits	Performance Risks	Design to mitigate ethical risks
Performance	Mimetic AI serves pedagogically	Uncontrolled replication of the human professor diminishes AI benefits	Curated fidelity (Curation of training data)

An ethics framework for mimetic AI professors

This paper's core purpose is to develop an ethics framework for analyzing mimetic AI professors. The

framework is meant to help professors, developers, students, and stakeholders design for human dilemmas as they build their own digital twins.

Future work may further elaborate the rules for measuring the ethical performance of mimetic AI technology, as well as develop efficient schema for reporting those measurements, perhaps replicating the color-coding results that have been applied to the ethics of musical performance embedded in the internet [Turchet 2024], and in the metaverse [Brusseau and Turchet 2026].

Table 9

Ethics opportunities, risks, and design responses for the mimetic AI professor

	Benefits	Risks	Design to mitigate ethical risks
Autonomy	Professor's force is multiplied Student on-demand learning, and judgement-free learning	Dependence on the machine reduces professors' capability and student independent learning	*Cascading fidelity* to require active instead of passive interaction.
Dignity	Increased learning	Commodification of the human when pedagogy is reduced to numerical scores Learning seems real but is counterfeit, an appearance without genuine human accountability	Narrow applicability Transparency about the professor's human/AI status
Privacy	Professors empowered to choose which aspects of themselves are projected Persistent personal information allows more effective educational scaffolding for students and training for future use	Personal information may escape control of professors/students	Privacy-by-design (Informed consent and data minimization)
Fairness	Consistency of the machine	As black box, could produce any grade for any reason	Explainability Redress-by-design
Social wellbeing	Increased happiness resulting from pedagogical success	Limiting happiness outcomes by restricting AI ability, notably the ability to impersonate the human professor	Enable the tool to transparently reveal, and fully conceal its mechanical status
Explainability	Explainability helps to provide accountability – credit and blame – for AI outputs.	Forcing explainability can reduce model performance.	Output citation Interpretability Redress-by-design
Performance	Mimetic AI serves pedagogically	Uncontrolled replication of the human professor diminishes AI benefits	Curated fidelity (Curation of training data)

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Allahabadi, H., Amann, J., Balot, I., Beretta, A., Binkley, C., Bozenhard, J., Bruneault, F., et al. (2022). Assessing trustworthy AI in times of COVID-19: Deep learning for predicting a multiregional score conveying the degree of lung compromise in COVID-19 patients. *IEEE Transactions on Technology and Society, 3*(4), 272–289.

Aristotle. (1934). *Nicomachean ethics* (H. Rackham, Trans.). Harvard University Press.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Bentham, J., & Mill, J. S. (2004). *Utilitarianism and other essays*. Penguin.

Braver, L. (2007). *A thing of this world: A history of continental anti-realism*. Northwestern University Press.

Brusseau, J. (2020). What a philosopher learned at an AI ethics evaluation. *AI Ethics, 1*(1). https://doi.org/10.47289/AIEJ20201214

Brusseau, J. (2026). Acceleration AI ethics and the Telus GenAI conversational agent. *Law Ethics Technol.* 2026(2):0006, https://doi.org/10.55092/let20260006.

Brusseau, J., & Turchet, L. (2024). An Ethics Framework for the Internet of Musical Things, 2024, *IEEE Transactions on Technology and Society*. DOI: 10.1109/TTS.2024.3398423.

Brusseau, J., & Turchet, L. (2026). *An ethics framework for the musical metaverse* (Forthcoming).

Bukingolts, N. (2025). Mimetic AI systems: Understanding and regulating the use of generative models for impersonation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(1), 469–485. https://doi.org/10.1609/aies.v8i1.36564

Data Ethics Commission. (2019). *Opinion of the data ethics commission: Executive summary*. German Federal Government.

Dicks, B., Williams, M., & Coffey, A. (2001). Finding a new way: The use of oral examinations in undergraduate geography. *Journal of Geography in Higher Education, 25*(2), 221–231.

Eachempati, P., Komattil, R., & Arakala, A. (2025). Should oral examination be reimagined in the era of AI? *Advances in Physiology Education, 49*(1), 208–209.

European Commission. (2019). *Ethics guidelines for trustworthy AI*. Publications Office of the European Union.

Feng, Y., Hang, Y., Wu, W., Song, X., Xiao, X., Dong, F., & Qiao, Z. (2025). Effectiveness of AI-driven conversational agents in improving mental health among young people: Systematic review and meta-analysis. *Journal of Medical Internet Research, 27*, e69639.

Gray, L., Edsall, S., & Parapadakis, D. (2025). AI-based digital cheating at university and the case for new ethical pedagogies. *Journal of Academic Ethics, 23*(4), 2069–2086.

Green, B. (2018). *The smart enough city: Putting technology in its place to reclaim our urban future*. MIT Press.

Gstrein, O. J. (2022). European AI regulation: Brussels effect versus human dignity? *Zeitschrift für Europarechtliche Studien*.

Huxham, M., Campbell, F., & Westwood, J. (2012). Oral versus written assessments: A test of student performance and attitudes. *Assessment & Evaluation in Higher Education*.

Joughin, G. (1998). Dimensions of oral assessment. *Assessment & Evaluation in Higher Education, 23*(4), 367–378.

Kaminski, M. E. (2019). The right to contest AI. *Columbia Law Review, 121*(7).

Kant, I. (2012). *Groundwork of the metaphysics of morals* (M. Gregor & J. Timmermann, Eds. & Trans.). Cambridge University Press. (Original work published 1785)

Kweon, S., Nam, S., Lim, H., Hong, H., & Choi, E. (2025). A large-scale real-world evaluation of an LLM-based virtual teaching assistant. In *Proceedings of the Association for Computational Linguistics* (pp. 850–864).

Longoni, F., Rocca, S. L., Passalacqua, G., & Riva, G. (2026). MAPS: A multidimensional alignment of presence in systems. *Cyberpsychology, Behavior, and Social Networking*.

Lowenstein, M. (2015). General education, advising, and integrative learning. *The Journal of General Education, 64*(2), 117–130. https://doi.org/10.1353/jge.2015.0010

Manganello, F., Nico, A., Ragusa, M., & Boccuzzi, G. (2025). Testing the applicability of a governance checklist for high-risk AI-based learning outcome assessment in Italian universities under the EU AI act annex III. *Frontiers in Artificial Intelligence, 8*, 1718613. https://doi.org/10.3389/frai.2025.1718613

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

McIlroy-Young, R., Kleinberg, J., Sen, S., Barocas, S., & Anderson, A. (2022). Mimetic models: Ethical implications of AI that acts like you. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 479–490). https://doi.org/10.1145/3514094.3534177

Mittelstadt, B. D., et al. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). The ethics of AI in health care: A mapping review. *Social Science & Medicine, 260*, 113172.

Muskaan, M. (2025). Synthetic voice and the philosophy of agency, authenticity, and ethics in AI-mediated speech. *AI and Ethics, 5*(6), 5889–5908.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103.

Oger, A., Gorisse, G., Fleury, S., & Christmann, O. (2025). MemorIA: An architecture for creating interactive AI historical agents in educational contexts. *Computer Animation and Virtual Worlds, 36*(3), e70032.

Pang, C. C., Zhao, Y., Yin, Z., Sun, J., Mogavi, R. H., & Hui, P. (2025). Artificial human lecturers: Initial findings from Asia's first AI lecturers. In *International Conference on Human-Computer Interaction* (pp. 105–124). Springer.

Pasquale, F. (2020). *The new laws of robotics: Defending human expertise in the age of AI*. Harvard University Press.

Prado, C. G. (Ed.). (2003). *A house divided: Comparing analytic and continental philosophy*. Humanity Books.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Schouls, P. (2018). *Reasoned freedom: John Locke and enlightenment*. Cornell University Press.

Sengupta, A., Seth, P., & Sankarapu, V. K. (2025). Interpretability as alignment: Making internal

understanding a design principle. *arXiv*. https://doi.org/10.48550/arxiv.2509.08592

Tarnoff, B. (2023). The moral panic of AI. *Logic(s) Magazine*.

Theobald, A. S. (2021). Oral exams: A more meaningful assessment of students' understanding. *Journal of Statistics and Data Science Education, 29*(2), 156–159. https://doi.org/10.1080/26939169.2021.1914527

Turchet, L. (2024). Entangled internet of musical things and people: A more-than-human design framework. *IEEE Transactions on Technology and Society, 5*(4), 355–367. https://doi.org/10.1109/TTS.2024.3443540

Turner, C., & Eisikovits, N. (2026). Programmed to please: The moral and epistemic harms of AI sycophancy. *SSRN*.

Westin, A. F. (1970). *Privacy and freedom*. Atheneum.

Woolf, B. P. (2010). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.

Xu, Y., et al. (2024). AI, ethics, and cognitive bias: An LLM-based synthetic simulation for education and research. *MDPI*.

Zicari, R. V., et al. (2021a). On assessing trustworthy AI in healthcare. *Frontiers in Human Dynamics, 3*. https://doi.org/10.3389/fhumd.2021.673104

Zicari, R. V., et al. (2021b). Co-design of a trustworthy AI system in healthcare. *Frontiers in Human Dynamics, 3*. https://doi.org/10.3389/fhumd.2021.688152

Zinn, L. M. (2004). Philosophy of adult education inventory (PAEI).

End