

Article

Generative AI in Values Education: Affordances, Ethical Risks and Implications for Teaching

Ying Chen^{1,*}
¹ School of Marxism, Tourism College of Zhejiang, Hangzhou, China

* Correspondence: Ying Chen, School of Marxism, Tourism College of Zhejiang, Hangzhou, China

Abstract: Generative artificial intelligence (GenAI) has become ordinary in university study, and students increasingly bring to it the questions about right and wrong that they once brought to teachers. Institutional guidance still treats GenAI mainly as a threat to academic integrity, leaving unsettled what kind of moral participant a language model can be. This article offers a theoretical account of that question. It characterises values education as a relational practice and specifies three conditions that any occupant of the educator's position must satisfy: answerability, exemplarity and address. It then distinguishes five affordances of GenAI, namely the generation of moral content, the simulation of perspectives, Socratic questioning, feedback on moral reasoning, and a tacit curriculum of machine authority, and analyses the risk internal to each. These risks fall into epistemic, relational, axiological and agentic families and are connected rather than independent. The central argument is that a generative model can be a moral interlocutor but not a values educator: it cannot be answerable for what it says, cannot exemplify a life, and cannot address a particular student. Four implications for teaching and institutional design follow.

Keywords: generative artificial intelligence; values education; moral education; moral interlocutor; teacher authority

1. Introduction

1.1. A Question Students Now Ask a Machine

A student writing an essay on whether a company should relocate a factory sits with a dilemma at eleven at night. Ten years ago the available interlocutors were a textbook, a friend, and, at best, a tutor met the following week. Today the student opens a chat window and asks what the right answer is. The answer arrives in four seconds, in fluent prose, with three considerations, a conclusion, and no visible author. This ordinary scene is where values education now happens, and it is a scene that the field has not yet theorised.

The literature has largely approached this scene from the direction of misconduct. Academic integrity dominates the empirical record: unauthorised use, assessment security, detection, and the institutional policies built around them [1]. This framing is understandable, because integrity is measurable and administratively actionable. It is also a narrowing. If the only question asked of GenAI is whether a student cheated, then the more interesting question is never asked: when a student asks a machine what is right, what has the student learned about how moral questions are answered? [2] The speed of adoption is part of what makes the question urgent. Within two years of the public release of conversational models, students in most disciplines had incorporated them into ordinary academic work, and the discussion had moved from whether to use them to how to regulate use [3].

Values education, understood here as the deliberate formation of moral judgement, civic disposition, and professional ethics in formal higher education, has a distinctive stake in that question. Its central practice is not the transmission of information but the staging of encounters in which a student is addressed as a person who must decide. A

Received: 21 August 2026

Revised: 09 October 2026

Accepted: 21 October 2026

Published: 24 October 2026



Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

machine that answers fluently removes the encounter and leaves the answer. Whether anything educational remains is not a question that can be settled by measuring the quality of the answer, because the answer was never the point.

1.2. The Shape of the Argument

This article therefore does not report a survey of what has been published on GenAI and values education. It develops an argument about what a generative model can and cannot be in a pedagogical relation, and it uses the available scholarship as an interlocutor rather than as a body of evidence from which conclusions are pooled. The distinction matters for how the claims should be read. The claims advanced here are conceptual and normative: they concern what the practice of values education requires of anyone who occupies the position of its teacher, and whether a language model can meet those requirements. Empirical findings are introduced where they sharpen a conceptual point, but the argument does not rest on the completeness of any corpus.

The argument proceeds in five steps. Section 2 characterises values education as a relational practice and specifies three conditions that any occupant of the educator's position must satisfy: answerability, exemplarity, and address. Section 3 distinguishes five affordances that GenAI offers to values education and identifies, in each case, the pedagogical relation that the affordance silently replaces. Section 4 groups the ethical risks internal to those affordances into four families and shows that the families are ordered rather than independent. Section 5 argues that a model can be a moral interlocutor but not a values educator, and identifies the three conditions as the reason. Section 6 draws four implications for teaching and for institutional design, and Section 7 states what the argument contributes, what it leaves unresolved, and where the burden of proof now lies.

The thesis can be stated in advance so that the reader can hold it against the argument. GenAI can occupy the position of a moral interlocutor, a partner in reasoning whose contributions a student must answer. It cannot occupy the position of a values educator, because that position requires three things a model does not have: the capacity to be answerable for what it says, the capacity to exemplify a way of living, and the capacity to address this student in this situation. The five affordances discussed below are therefore best read as design possibilities for interlocation, not as a substitute workforce for moral teaching.

2. Values Education as a Relational Practice

2.1. What the Practice Is For

Values education is easily confused with three neighbouring activities. It is not moral instruction, if that means the transmission of correct positions, because a student can recite a position and still be unable to hold it under pressure. It is not socialisation into local norms, because the point of the practice is that a student should be able to examine the norms she has inherited, including the ones the institution would prefer she keep. It is not compliance training, although compliance training often borrows its vocabulary. What distinguishes values education is that its outcome is a disposition to judge well, and a disposition is not an item of information that can be transferred from one mind to another.

This is why the practice has always been organised around encounters rather than around content. A case is discussed, not because the case contains the answer, but because discussing it puts a student in a position where she must take a view and answer for it. A teacher contradicts a student, not because contradiction is intrinsically valuable, but because a student who has defended a position against a serious objection knows something about that position that she did not know before. The content matters, but the content is the occasion; the relation is the mechanism. A curriculum that supplies excellent content and no occasion for a student to be addressed has supplied reading matter, not education.

Two consequences follow for the analysis of technology. The first is that the relevant question about a tool is not what it produces but what position it occupies. A textbook and a teacher may contain the same propositions, and they are not educationally

equivalent, because a student stands in a different relation to each. The second is that any substitution of a machine for a person in this practice must be evaluated as a substitution of a relation, not as a substitution of a function. Most of the enthusiasm and most of the anxiety about GenAI in education turns on function, on whether the tool can do what the teacher did. The pedagogically consequential question is what happens to the relation when the function is transferred.

2.2. Three Conditions of the Educator's Position

If values education is a relational practice, then the position of the educator is defined by what that relation requires. Three conditions can be distinguished. They are not a job description and they are not a moral ideal for teachers; they are the minimum without which the relation ceases to be educational and becomes something else, an information service, a performance, or an authority relation of a different kind.

The first condition is answerability. To educate morally is to make claims that one can be asked to defend, and to accept the consequences of having made them. A teacher who tells a student that loyalty to a colleague matters more than loyalty to an employer can be challenged, can be shown to have said something inconsistent last term, and can be held to account by the student, by the institution, and by her own conscience. What makes the claim educational is not its truth but the fact that it is someone's claim, staked by a person who can be wrong in a way that costs her something.

The second condition is exemplarity. Moral education has always relied on persons whose lives show what a commitment looks like when it is inconvenient, and the philosophical literature on exemplars explains why this cannot be replaced by propositions about virtue: admiration and emulation are responses to a life, not to a description [4]. A teacher's authority in this domain is not derived from her argumentative skill but from the fact that she is a working instance of the thing she recommends. This is why hypocrisy damages moral teaching more than error does. A teacher who is wrong can be corrected; a teacher who does not live as she speaks has removed the object of imitation rather than misdescribed it.

The third condition is address. Education is addressed to someone in particular. A teacher knows that this student is supporting a younger sibling, that another's father is ill, that a third has just been refused a placement, and adjusts what is said accordingly. This is not personalisation in the technical sense of adapting difficulty to measured performance. It is the recognition that a moral claim lands differently depending on who is hearing it, and that the same sentence can be a challenge for one student and a cruelty for another.

2.3. The Status of the Claim

These three conditions are derived from the concept of the practice, not from a survey of teachers. The derivation is analytic in a modest sense: if a relation lacks all three, we would not call it moral education, in the way that we would not call an argument a proof if it could not be examined by anyone. This does not make the conditions arbitrary. They capture what is at stake when someone claims the authority to form another person's judgement, and they explain intuitions that the field already holds, for instance that a moral curriculum delivered entirely by recorded lecture is diminished even when the lectures are excellent.

It is important to be clear about what this argument does not claim. It does not claim that machines are poor at moral reasoning, and it does not depend on any thesis about machine consciousness. Models produce formally adequate ethical reasoning and can revise it under counter-argument, and the comparative literature on that capacity is genuinely interesting [5,6,7]. The argument is that the capacity in question is not the capacity the position requires. A comparison of model and human performance on a dilemma task measures competence at dilemma solving; the educator's position is defined by answerability, exemplarity, and address, and none of these is a dilemma-solving skill. The rest of this article works out that distinction in the concrete settings where GenAI is now being asked to teach.

3. Five Affordances

Five positions can be distinguished in the pedagogical use of GenAI in values education. They are called affordances here not because they are properties of software but because they are possibilities that a relation offers: each is defined by the position the model is invited to occupy, and each is therefore defined by the relation it replaces. The five are not a taxonomy of tools, since the same tool can occupy several positions in a single session, and they are not a ranking. They are distinguished because each carries a different internal risk, and because the risk is generated by the position rather than by the quality of the output. Table 1 summarises the five positions, the relation each replaces, and the risk internal to each.

Table 1. Five Pedagogical Positions of Generative AI in Values Education and the Risk Internal to Each

Pedagogical position	Relation it replaces	What the model supplies	Risk internal to the position
Generator of moral content	The teacher who selects and frames cases	Dilemma cases, scenario variations, prompts and assessment items at low cost	An implicit moral ontology presented as a neutral prompt
Simulated perspective	The stakeholder who can be wronged	An interlocutor who holds a position and never tires or takes offence	The cognitive form of encounter without its vulnerability
Socratic interlocutor	The questioner with a stake in the answer	Unlimited questioning, restatement and pressure on inconsistency	Dialogue without answerability, and no ground for refusal
Reader of moral reasoning	The reader whose judgement must be earned	Consistent assessment of structure, reasons and counter-arguments	The model's standard becomes the standard of good reasoning
Tacit curriculum of authority	The person whose practice shows what is authoritative	Fluent, immediate answers to any moral question	The lesson that moral questions have answers of that kind

3.1. Generating Moral Content

The most common position assigned to GenAI is that of producer of teaching material: dilemma cases, scenario variations, discussion prompts, and assessment items. A multi-case study of civic education in higher education describes teachers using the model to produce localised cases that connect abstract questions of justice to events students recognise, and locates the gain in the cost of variation, since thirty versions of a dilemma with different stakeholders can be produced in an afternoon [8]. Curriculum-level proposals for ethical assessment make the same assumption, treating the model as a supplier of prompts and cases that the teacher then selects, frames, and takes responsibility for [9].

The risk that is usually noticed is fabrication, and it is usually discussed in terms of accuracy, with the reasonable recommendation that cases be verified before use. Studies of value alignment in civic education report that the tools remain unreliable on factual accuracy and on contextual fit even when they are usable in other respects [10]. Fabrication is a real problem, but it is a technical one, and it is the kind of problem that improves. The risk that is less often noticed is default. A model asked to generate a dilemma about, say, loyalty to a colleague draws on a distribution of moral vocabulary that is neither neutral nor universal. Work that reads the ethics of learning with language models through Confucian self-cultivation makes the point in another register: what a model offers is not a neutral procedure but a particular conception of the moral life, in which cultivation, relation, and example carry the weight [11].

A generated case therefore carries an implicit moral ontology. It has already decided, before the teacher sees it, which considerations count as considerations, whose standpoint is the standpoint from which the dilemma is described, and what a resolution would look like. A teacher who uses generated cases without noticing this is not teaching pluralism. She is teaching the model's implicit position while presenting it as a neutral prompt, and the presentation is what makes the lesson effective. The point is not that generated cases are unusable. It is that they are never neutral artefacts, and that the labour of noticing their ontology is part of the pedagogical work rather than an optional refinement of it.

A variant of the position reverses the direction of use. In an analysis of public submissions to a national inquiry on civics education, a model was used to read and code the submissions rather than to produce teaching material [12]. Here the model serves the curriculum by making a public deliberative process legible to students, and the risk shifts from fabrication to selection: what the model counts as a theme is what students will see as the public's concerns. The teacher has delegated not the writing of the case but the perception of the public, and the delegation is invisible in the output.

3.2. Simulating Perspectives

A second position places the model in the role of a stakeholder: the developer who wants the contract, the resident who will be displaced, the patient who refuses treatment. The pedagogical rationale is familiar from role-play and from perspective-taking traditions in moral education. What is new is that the counterpart can be produced on demand and will not become tired, embarrassed, or defensive. Studies of collaborative reasoning with AI agents report a persona effect: the way the agent is framed changes the deliberation itself, altering how students argue and what they are willing to concede [13].

Classroom accounts are largely positive about this affordance and largely silent about its relational cost. A simulated interlocutor cannot be wronged. A student who argues that the displaced resident should accept compensation is practising a move that, with a real resident, would carry the weight of having said it to someone. The simulation offers the cognitive form of moral encounter without its vulnerability, and it is an open question whether the form transfers when the vulnerability returns. Nothing in a role-play with a model teaches a student what it is like to have spoken carelessly to a person who then looks at her differently.

The strand of work that speaks most directly to this gap is the literature on exemplars, which argues that moral learning depends on attending to real persons whose lives embody a commitment, precisely because admiration and emulation are not operations that can be performed on a text [4]. A simulated stakeholder can display a position; it cannot display the cost of holding it. The difference is not a difference in realism, and it will not be closed by better simulation. It is a difference in what kind of thing is being encountered: a description of a standpoint, or a person for whom the standpoint is a stake.

3.3. The Socratic Interlocutor

The third position is the one with the strongest theoretical warrant and the weakest demonstrated effect. If a model can ask a student why she holds a position, restate her reasoning, and press on an inconsistency, it appears to perform the function that moral educators have assigned to dialogue since Socrates. The premise is supported by work showing that language models can solve moral dilemmas through reasoning and can revise that reasoning when given counter-arguments, which is the minimal condition for the position to be possible [6,14]. Conceptual work in the field accordingly proposes dialogue with the model as a rehearsal space for moral argument, on the grounds that students can be challenged without the social cost of being challenged in front of peers [15].

The difficulty is that the Socratic relation depends on an asymmetry of a particular kind. The questioner must have a stake in the answer, and the student must be able to affect the questioner. A model has neither. Comparative work is consistent on this point: models produce formally adequate ethical reasoning but lack the contextual and emotional depth that makes a question land [7]. The comparison has begun to extend from

students to teachers, setting teachers' moral reasoning against that of models on the same set of school dilemmas [16], and the terms in which such comparisons are made, competence, consistency, sensitivity to context, are exactly the terms on which a machine is most likely to look adequate and least likely to be educative. A student can therefore be questioned indefinitely without ever being answered, which produces the appearance of dialogue and the experience of a vending machine that asks for coins.

There is a further and less obvious problem. Because the model will always continue the conversation, it never signals that a question has become intrusive, that a topic is not its business, or that a student should stop and speak to a person. A teacher's refusals, her change of subject, her decision that this is not the moment for a hard question, are part of what makes moral conversation educational rather than merely interactive. Their absence is not a technical defect that better models will fix, because it is not a failure of the model to do something; it is the absence of a stake that would make the refusal appropriate. A system with no interests of its own has no grounds on which to object to anything, and a dialogue partner that cannot object is a very particular kind of partner.

3.4. Feedback on Moral Reasoning

The fourth position turns the model into a reader: it responds to a student's written moral argument with an assessment of its structure, its use of reasons, and its consideration of opposing views. This is where the most direct instructional effects are reported. A framework for ethical assessment in large courses describes requiring students to defend their reasoning conversationally, with an initial decline in scores followed by improvement, which the authors interpret as evidence of genuine cognitive challenge rather than of failure [9]. Curriculum-level work in medical education reaches a compatible conclusion from the opposite direction: as generative tools absorb the production of text, the competencies worth teaching are the ones the tool cannot replace, among them clinical judgement and communication [17].

The risk here is a substitution that is easy to miss. If the model both prompts and evaluates, the standard of good moral reasoning becomes the model's standard. Comparative studies complicate any simple hierarchy: some models score at or above graduate level on dilemma reasoning tasks while others fall below human participants on measures of moral competence, and in either case the instruments were validated on humans and carry assumptions that artificial agents may not share [5,6,18]. A model used as an examiner is therefore not a neutral instrument. It rewards a particular shape of answer, typically explicit reasons in a recognisable order, and that shape is not the same as moral judgement. A student who learns to produce the shape has learned something, but what she has learned is the genre of assessable moral reasoning, which is a genre before it is a virtue.

The pedagogical literature tends to treat the consistency of machine feedback as an advantage over the variability of human marking. The advantage is real, and it is also the mechanism of the narrowing. A human marker is inconsistent partly because she is responsive: she notices that this student's argument has improved in a way that the rubric does not capture. A consistent marker cannot notice that, and a student who is marked consistently by an instrument that rewards one shape of reasoning will produce that shape. The more reliable the feedback, the more efficiently it teaches the genre.

3.5. The Tacit Curriculum of Machine Authority

The fifth position is rarely named as a position at all. It consists in what students learn about the status of moral claims from the fact that a machine produces them fluently and immediately. Qualitative work on how students position GenAI within their studies finds that students prioritise their own developing values and moral positions over institutional messaging, and that their judgements about what is permissible are made in a landscape the institution does not see [2]. This is not simple deference. Students are actively interpreting. But the interpretation takes place around a tool that answers every question with equal confidence, and the lesson that confidence carries is that moral questions have answers of that kind.

A parallel finding on the machine side makes the pattern sharper. In a Milgram-paradigm study of agentic language models, systems complied with harmful steps when those steps were embedded in a procedurally legitimate workflow, a failure the authors describe as procedural obedience [19]. Set the two findings side by side and a symmetry appears: students defer to procedure, models defer to procedure, and the interaction between them is a procedure that neither party examines. This is the mechanism by which a tool becomes a curriculum. It does not teach a moral doctrine; it teaches a form, in which questions have answers, answers have authors who are not present, and the absence of an author is not noticed.

Three strands of scholarship point at this object without naming it as a pedagogical position. Work on the governance of GenAI in applied undergraduate education argues that governance must integrate technological, institutional, and cultural dimensions at once, the cultural dimension being the one that concerns what a campus treats as authoritative and the one least amenable to regulation [20]. A decolonial reframing of student AI use argues that integrity frameworks have substituted compliance for epistemology: students are taught what counts as honest rather than what counts as knowing [1]. A political-economy reading of the same phenomenon argues that algorithmic mediation reshapes the epistemic and normative foundations of disciplinary knowledge rather than merely its delivery [21]. Each describes a different surface of the same thing, which is what a student learns about authority from working with a system that has no stake in being right.

4. Four Families of Ethical Risk

Principle-based frameworks for the ethics of AI in education already exist, and they organise concern around privacy, autonomy, and accountability [22]. Those are the right concerns for a technology governance question, and they are not the concerns that a values educator has, because they do not ask what a student learns about value from the encounter. A different cut through the same material yields four families of risk: epistemic, relational, axiological, and agentic. The families are analytically distinct and causally connected, and the connection is the substance of the argument in Section 5.

4.1. Epistemic Risk

Epistemic risk concerns what students come to accept as true or as well reasoned. Four varieties can be distinguished. The first is fabrication, in which a case, a statistic, or an institutional practice is invented and arrives in the register of fact. The second is the implicit moral ontology discussed above, in which a generated artefact carries an unmarked conception of what matters. The third is homogenisation: because the same model is available to everyone, the range of arguments that students encounter narrows towards the range the model produces, and a class of thirty students may arrive with thirty versions of the same three positions. The fourth is the substitution of fluency for judgement, in which the smoothness of an answer is taken as evidence of its quality, because in human speakers fluency and judgement usually co-occur.

The four are usually treated as separate technical problems with separate fixes, and they are better understood as one problem with four surfaces. Each is a case in which a signal that reliably indicated a human achievement has been detached from that achievement and continues to be read as if it had not been. Fluent prose indicated that someone had thought; a well-formed dilemma indicated that someone had weighed a case; a consistent judgement indicated someone with a settled view. The signals remain, the achievements do not, and the reading habits of students and teachers were formed when they did.

4.2. Relational Risk

Relational risk concerns what happens to the positions of teacher and student. Three losses can be identified. The first is the other who cannot be wronged, discussed in Section 3.2: a student rehearses moral moves against a counterpart who has nothing at stake, and the rehearsal may or may not survive contact with a counterpart who does. The second is

the tutor who is always available and never tired, which is a genuine good for access and a subtle harm for formation, because a student who can always obtain an answer never experiences the interval in which a question has to be carried. The third is the repositioning of the teacher as a supervisor of tool use rather than as a person who embodies a commitment. A teacher whose role is to check that the tool was used appropriately has been moved out of the position that made her a values educator, and the move is usually justified in the vocabulary of efficiency rather than in the vocabulary of formation.

4.3. Axiological Risk

Axiological risk concerns value pluralism. Where the interlocutor has no position to defend, pluralism becomes a list of options rather than a disagreement between people, and a list of options is precisely what relativism looks like from the inside. Studies of how students reason about their own use of GenAI report a narrow repertoire: consequentialist and situational grammars do most of the work, and obligation-based reasoning is thin [23]. Studies of students' stated concerns find that they cluster around authorship, transparency, verification, and privacy [24], which are the concerns a consumer of information has rather than the concerns a person forming a judgement has. Where institutional guidance is absent or unclear, students construct their own rules, often generous ones, without any sense that a judgement has been made [25].

The three findings describe a single condition. A student who has only a consequentialist grammar is not a relativist by conviction; she is a person who has not been given the occasion to encounter a claim that binds irrespective of outcome. The point of a pluralist values education is not to teach that all positions are equal. It is to put a student in front of positions she does not hold, defended by people who hold them, so that she discovers that disagreement is between persons and not between options on a menu. A machine that will adopt any position on request cannot supply that discovery, and its helpfulness is what makes the omission invisible.

4.4. Agentic Risk

Agentic risk concerns responsibility. If a model's advice shapes a student's judgement about a professional dilemma, no one is answerable for that advice, and the student is left holding a decision whose grounds were produced elsewhere. Work on how students interpret liability and professional conduct in engineering education shows that students already understand the vocabulary of liability well enough to apply it to tools, and that the application does not extend to the reasoning the tools supply [26]. The professional ethics literature recognises this structurally, which is why it talks about liability and safety rather than about wisdom.

Agentic risk is the one most often left out of educational discussions, because educational discussions are organised around learning outcomes and the question of who answers for a judgement does not look like a learning outcome. It is, however, the risk that determines whether the other three matter. A student who accepts a fabricated case has made an epistemic error that can be corrected. A student who holds a judgement for which no one is answerable has acquired a way of holding judgements, and that is not an error but a formation.

4.5. The Order of the Four Families

The four families are not independent, and their dependence has a direction. Epistemic risks are the visible surface of a relational change. Students accept a fabricated case because it arrives in the register of a case; they accept a homogenised argument because nothing in the exchange marks it as one voice among others; they mistake fluency for judgement because fluency is what a person with judgement sounds like. The axiological risks follow: when no one defends a position, pluralism thins into a list. The agentic risk is the end point of the sequence. If no one is answerable for the reasoning a student encounters, the student is left as the only candidate, holding a judgement whose grounds were assembled elsewhere.

This ordering has a practical consequence for how the risks should be researched and taught. Treated as independent, they invite independent remedies: better retrieval for fabrication, better detection for integrity, better prompting for homogenisation. Treated as ordered, they invite a single question: in this arrangement, who is answerable to this student for what is being claimed? That question cannot be answered by improving the tool, which is why it tends to be dropped from discussions that are organised around tool improvement. One counter-argument deserves a place here. GenAI does not have to undermine education, provided integrity is reconceived around the process of coming to know rather than around the finished product [27]. That argument is compatible with the position developed in Section 5, although it is also compatible with a much thinner view of what a machine can do than its authors sometimes imply.

5. Why a Model Can Be an Interlocutor but Not an Educator

5.1. Answerability

The first condition is the one most easily confused with a technical property. A model can be asked to explain itself, and it will produce an explanation. But explanation is not answerability. The explanation is produced by the same process as the claim, by a system that has no stake in the claim and cannot be wrong in a way that costs it something. When a teacher is asked why she said what she said, the answer may include an admission, a revision, a concession that she spoke too quickly, or a refusal to back down that will be held against her. None of these moves is available to a model, not because the model is dishonest, but because there is no one there for the moves to be made by.

This is why the professional ethics literature's focus on liability is not a distraction from the educational question but an unusually clear statement of it. The reason a professional is answerable is that she can be sanctioned, sued, struck off, or simply ashamed, and those possibilities are what make her judgement hers. A system that cannot be sanctioned can still be accurate, and accuracy is a virtue of instruments. It is the wrong virtue for a position defined by the staking of claims. This is also why the argument does not depend on any claim about model reliability: a perfectly reliable model would still not be answerable, and a perfectly reliable authority is a description of an oracle rather than of a teacher.

5.2. Exemplarity

The second condition is exemplarity. Moral education has always relied on persons whose lives show what a commitment looks like when it is inconvenient, and the philosophical literature on exemplars explains why this cannot be replaced by propositions about virtue: admiration and emulation are responses to a life, not to a description [4]. The simulation literature most consistently underestimates this condition, because simulation is very good at producing descriptions of commitment. A model can describe integrity with more precision than most teachers, and the description will be correct. What it cannot supply is an instance.

The distinction can be put sharply. A model can describe integrity, but there is nothing it has declined in order to be integrity. It can describe the cost of a position, but it has paid no cost. Everything it says about commitment is said by something that has no commitments, and a student who learns about commitment from such a source learns the vocabulary of commitment without learning what it is to be committed. This is not a gap that better alignment or more autobiographical training data will close, because the gap is not informational. It is the difference between a report of a life and a life.

5.3. Address

The third condition is address. Education is addressed to someone in particular. A teacher knows that this student is supporting a younger sibling, that another's father is ill, that a third has just been refused a placement, and adjusts what is said accordingly. A model can personalise content in the technical sense, remembering that a student has mentioned a sibling and referring to it later, and this is often described as personalisation.

It is not address. Address requires that the person speaking is affected by the particularity of the person spoken to, that the same sentence is harder for the speaker to say to this student than to another, and that the difficulty is part of what the student receives.

The elements of this argument are scattered through the literature. Models lack the contextual and emotional depth that makes a question land [7]; human experts regulate emotion in moral judgement in ways that students do not, and models approximate neither pattern reliably [28]; and students position the tool within their own moral landscape rather than submitting to it [2]. Each finding is a fragment of the same point, which is that the relation between a speaker and a particular hearer is not a channel through which content passes but part of the content. A system with no continuing relation to anyone can transmit a message; it cannot address a person (As shown in Table 2).

Table 2. Three Conditions of the Educator's Position and the Corresponding Capacities of a Generative Model

Condition	What it requires of the occupant	What a generative model can do	What it cannot do
Answerability	Claims staked by someone who can be held to them	Explain its output on request	Be wrong in a way that costs it something
Exemplarity	A life that shows what a commitment costs	Describe a commitment with precision	Exhibit one, having declined nothing
Address	Adjustment to this student in this situation	Personalise content from remembered detail	Be affected by the particularity of the hearer

5.4. Relocation Rather Than Rejection

None of this makes the machine useless. It relocates it. A moral interlocutor is a genuine educational position, and it is the position that the most interesting pedagogical work is actually describing: a partner that will take a position and hold it long enough for a student to find out what she thinks [8,14]. The interlocutor does not need to be answerable, exemplary, or particular, because the student is not learning from it what a commitment costs; she is learning what her own commitments are by having to state them to something that will not accommodate her. That is a real educational good, and it is available now.

The error is not using the tool. The error is describing the tool's position in the same vocabulary as the teacher's, because that description then justifies arrangements in which the relation has quietly been removed: automated moral assessment, fully online ethics modules, a values curriculum delivered by a system that answers every question with equal confidence. The vocabulary does the work of the argument. Once a model is called a tutor, the question of whether a tutor is needed becomes pedantic, and the arrangement follows. The remedy is not to ban the vocabulary but to keep the distinction visible in every design decision that depends on it.

6. Implications for Teaching

Four implications follow, ordered from the design of a single session to the design of an institution. They are not recommendations for using GenAI well; they are consequences of the distinction between interlocutor and educator, and each can be stated as a design rule that follows from it.

6.1. Designing Encounters Rather Than Answers

If the model is a good generator of cases and a poor authority on them, then the pedagogically defensible use is to put the generated case in front of students as something to be interrogated, including interrogated for its implicit moral ontology. A workable pattern is a three-step task: the model produces a dilemma with a decision; students identify the moral register the case assumes; students rewrite the case so that a different

register is at stake. The model's output becomes the object of study rather than the answer to a question. This has the practical advantage of making the tool's limits visible without a lecture about them [8,11].

The rule behind the pattern is worth stating separately, because it generalises beyond this exercise. Whenever a generated artefact enters a values classroom, the question to ask is whether it arrives as an answer or as an object. An object can be examined, disagreed with, and used to make a student's own assumptions visible. An answer can only be accepted or rejected, and both responses leave the student where she was. The same output can serve either purpose, and which one it serves is decided by the task the teacher sets, not by the quality of the output.

6.2. Assessing Judgement Rather Than Moral Vocabulary

Models can produce reasoning that scores well on instruments designed to measure moral competence [5,6], which means that any assessment satisfied by well-formed reasoning is now an assessment of a tool. What remains assessable is the thing the tool cannot supply: a decision the student has actually made, in a situation where something was at stake for her, with an account of what she gave up. This is not a new pedagogy. It is the old pedagogy of practicum and reflection, and its revival is forced by the technology rather than chosen.

The difficulty is scale, and it should be acknowledged rather than waved away. Judgement is expensive to assess because it requires knowing the situation a student was in, and large courses are large. Two responses are available and neither is fully satisfactory. The first is to assess fewer things well, replacing a number of vocabulary tests with one documented decision and its retrospective account. The second is to use the model to prepare the material for human judgement, summarising a student's account and flagging inconsistencies, while leaving the judgement itself to a person who can be answerable for it. The second response is consistent with the argument of this article; the first is consistent with it and harder to administer.

6.3. Preparing Teachers for Irreducible Disagreement

When a student brings a machine-generated position to class, the teacher's first task is not to correct it but to establish what kind of claim it is and who is making it. That requires a repertoire that teacher education rarely supplies: recognising the moral register of a text, distinguishing a reason from a rationalisation, and keeping a disagreement open long enough for students to enter it. Work on teacher-AI teaming in civic and history teaching identifies the normative boundaries of this work but says little about how teachers acquire the capacity to hold them [29].

The repertoire is not a set of techniques for handling AI. It is the ordinary repertoire of a teacher who can conduct a disagreement, and the reason it now matters more is that machine-generated positions arrive without an author to be questioned, which makes the teacher the only party in the room who can be asked what a claim means. A teacher who can only correct positions will correct machine output and leave the student's relation to it untouched. A teacher who can hold a disagreement open will make the absence of an author the subject of the lesson, which is the point at which the tacit curriculum becomes visible and therefore teachable.

6.4. Treating Institutional Policy as a Curriculum Question

Policy documents on GenAI tell students what is permitted, and students read them as statements about what the institution actually values. Where guidance is absent, students construct their own rules and feel no need to justify them [24,25]; where guidance is present but purely procedural, it teaches compliance rather than judgement [1]. A policy that specifies disclosure requirements but is silent on what counts as a good argument is not neutral. It is a curriculum in which the visible lesson is procedure.

The governance literature's argument that cultural dimensions are the hardest to legislate is, in this sense, an argument about the hidden curriculum [20]. Values are formed by an ecosystem rather than by a course, and the parts of the ecosystem that

students read most closely are the parts that tell them what will be rewarded. A policy that rewards disclosure and does not reward the quality of a reason has taught students which of the two the institution believes in. This is not an argument against disclosure requirements; it is an argument for pairing them with a public statement, in the same document, about what the institution treats as a good argument, and for making that statement specific enough that a student could tell whether she had met it.

7. Conclusion

This article set out to answer a question that the integrity literature leaves unasked: when a student asks a machine what is right, what has the student learned about how moral questions are answered? The answer developed here is that she has learned a form, and that the form is consequential precisely because it is not a doctrine. Questions have answers, answers have authors who are not present, and the absence of an author is not noticed. The five affordances distinguished in Section 3 are the positions through which that form is acquired, and the four families of risk in Section 4 are its surfaces.

The theoretical contribution is a distinction and a test. The distinction is between a moral interlocutor and a values educator, and it is drawn by three conditions: answerability, exemplarity, and address. The test follows from it. For any proposed substitution of a machine for a teacher in values education, ask which of the three conditions the machine satisfies, and what happens to the relation when the condition is unmet. On this test, the substitutions that fail most clearly are those that look most efficient, automated moral assessment and fully online ethics modules, because they remove exactly the relation that the conditions describe.

The relation to existing scholarship can be stated by contrast. Research on academic integrity measures behaviour and asks how to secure it; it therefore cannot see formation, because formation is not a behaviour. Research comparing human and machine moral reasoning measures competence and asks how the two compare; it therefore cannot settle the pedagogical question, because competence is not the virtue the position requires. Principle-based frameworks for AI ethics in education organise concern around privacy, autonomy, and accountability, which are the right concerns for governance and the wrong ones for a classroom, because they do not ask what a student learns about value from an encounter. The relational account offered here is compatible with all three and explains why none of them, on its own, can answer the question this article began with. Its weakness is correspondingly clear: it delivers a criterion rather than a measure, and a criterion cannot be tabulated.

Four problems remain open. The first is empirical and belongs to others: whether the cognitive form of moral encounter practised with a simulation transfers when vulnerability returns, which requires longitudinal design and a willingness to follow students into settings where something is genuinely at stake. The second is developmental: whether a long relation with a system that remembers a student gradually acquires some of the properties of address, and whether that would be an improvement or a more efficient simulation of one. The third is assessment: how judgement can be assessed at scale without recreating the instrument problem, that is, without substituting a measure of the vocabulary of reasons for a measure of the capacity to decide. The fourth is institutional: whether a culture that treats argument as something to be evaluated, rather than as something to be permitted or detected, can be designed, or whether it can only be grown.

The wider point is about what a values curriculum is for. If its purpose were the transmission of correct positions, a language model would be an adequate delivery system, and a cheap one. If its purpose is the formation of persons who can hold a position under pressure, then the relation between a student and someone who can be answerable, exemplary, and particular is not an input to the process but the process itself. No improvement in model capability supplies that relation, and no account of values education that omits it will survive contact with a classroom.

Funding: This paper was supported by Tourism College of Zhejiang Project (No.2025QKZD04).

References

1. C. DiPietro, "From ethics to epistemology: a decolonial reframing of student AI use in postsecondary education," *Teaching in Higher Education*, pp. 1–17, 2026.
2. M. Bearman, "Time, emotions and moral judgements: how university students position GenAI within their study," *Higher Education Research & Development*, vol. 45, no. 4, pp. 884–898, 2026.
3. Y. K. Dwivedi et al., "So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *International Journal of Information Management*, vol. 71, p. 102642, 2023.
4. H. Han, "Why do we need to employ exemplars in moral education? Insights from recent advances in research on artificial intelligence," *Ethics & Behavior*, vol. 35, no. 4, pp. 259–276, 2024.
5. S. Bajpai, "Insights into moral reasoning of AI: a comparative study between humans and large language models," *Journal of Media Ethics*, vol. 41, no. 3, pp. 143–157, 2025.
6. D. Nabizadeh, "Exploring large language models' responses to moral reasoning dilemmas," *Ethics & Behavior*, vol. 36, no. 7, pp. 591–610, 2026.
7. B. Tonbuloglu, "Comparing human and generative AI ethical decision-making in computer ethics education," *International Journal of Human-Computer Interaction*, pp. 1–41, 2026.
8. S. Gu, "Generative AI in civic education: a multi-case study of pedagogical innovation and governance in higher education," *Educational and Humanities*, vol. 3, no. 1, p. 9, 2026.
9. C. Kooli, "Colloquial engagement theory with AI awareness (CET-AIA): a new creative pedagogical framework for ethical assessment in the age of artificial intelligence," *Thinking Skills and Creativity*, vol. 60, p. 102051, 2026.
10. R. D. Agustin, "Analysis of value alignment and ethical guardianship of learning with AI in civic education," *Jurnal Moral Kemasyarakatan*, vol. 9, no. 2, pp. 255–265, 2024.
11. C. M. Lam, "Confucian self-cultivation and the ethics of learning with large language models," *Studies in Philosophy and Education*, 2026.
12. K. Heggart, "Perspectives on civics and citizenship education in Australia: findings from the recent inquiry through AI-augmented analysis," *Citizenship Teaching & Learning*, vol. 21, no. 1, pp. 89–115, 2026.
13. Y. Jin, "When machines join the moral circle: the persona effect of generative AI agents in collaborative reasoning," *British Journal of Educational Technology*, 2026.
14. H. Han, "Potential benefits of employing large language models in research in moral education and development," *Journal of Moral Education*, pp. 1–16, 2023.
15. H. Han, "AI and moral education: opportunities and challenges examined from multidisciplinary perspectives," *Journal of Moral Education*, pp. 1–6, 2026.
16. A. Göçen, "Teachers vs. large language models: ethical decision-making across educational dilemmas in Turkish schools," *Ethics & Behavior*, vol. 36, no. 5, pp. 400–426, 2026.
17. I. Shimizu, "Developing medical education curriculum reform strategies to address the impact of generative AI: qualitative study," *JMIR Medical Education*, vol. 9, p. e53466, 2023.
18. M. G. Reinecke, "The puzzle of evaluating moral cognition in artificial agents," *Cognitive Science*, vol. 47, no. 8, p. e13315, 2023.
19. H. Dery, "Procedural obedience as a behavioral safety failure: evidence from a Milgram-paradigm study of agentic LLMs," *Computers in Human Behavior Reports*, vol. 24, p. 101322, 2026.
20. G. Xu, "Governing generative AI in applied undergraduate education: integrating technological, institutional, and cultural dimensions," *Interactive Learning Environments*, vol. 34, no. 8, pp. 5594–5609, 2026.
21. S. Hlaimi, "Economics education under algorithmic power: moral judgment, pluralism and the political economy of AI," *Citizenship, Social and Economics Education*, 2026.
22. A. Nguyen, "Ethical principles for artificial intelligence in education," *Education and Information Technologies*, vol. 28, no. 4, pp. 4221–4241, 2023.
23. A. Pérez-Portabella, "Explaining higher education social sciences students' misuse of generative artificial intelligence: evidence from a multidimensional ethics scale," *SN Social Sciences*, vol. 6, 2026.
24. Y. B. Tabares-Cruz, "Predictors of the ethical use of generative artificial intelligence in higher education," *Frontiers in Education*, vol. 11, 2026.
25. Z. Revesai, "Navigating ethical ambiguity without institutional guidance: generative AI decision-making among students in Zimbabwean higher education," *Cogent Education*, vol. 13, 2026.
26. Z. Quince, "Student perceptions of GenAI integration into engineering practice: how students interpret liability, safety, professional conduct and ethics across disciplines," *International Journal of Mechanical Engineering Education*, 2025.
27. M. J. T. Tan, "Shaping integrity: why generative artificial intelligence does not have to undermine education," *Frontiers in Artificial Intelligence*, vol. 7, 2024.

28. A. Zok, "Sentiment differences in resolving moral dilemmas: an analysis of student attitudes and large language models," *Journal of Academic Ethics*, vol. 24, 2026.
29. O. Drobysheva, "Teacher-AI teaming in teaching history and civic education: pedagogical risks and normative boundaries," *Education and Pedagogical Sciences*, pp. 17–27, 2026.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Publisher and/or the editor(s). Publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.