

Thinking With, Not Through, AI: Philosophical Counseling and Ethical Self-Authorship in Teacher Education

Sivan Sarid-Goldfisher¹

Rotem Waitzman²

Abstract: Generative artificial intelligence is increasingly used to interpret professional situations, formulate arguments, and consider responses to ethical dilemmas. Although these systems can broaden reflection, their persuasive outputs may encourage moral outsourcing, defined as transferring substantive elements of ethical interpretation, justification, or judgment to an external system without adequate examination, endorsement, or responsibility. This conceptual article asks how generative AI can be integrated into philosophical counseling and ethical reflection in teacher education without undermining ethical self-authorship. Drawing on philosophical counseling, moral development, relational autonomy, professional ethics, and AI literacy, it develops the Ethical Self-Authorship Cycle. The model comprises five iterative stages: articulating the dilemma, pluralizing perspectives, interrogating assumptions, testing reasons and consequences, and reclaiming moral authorship. Its contribution is threefold. First, it distributes deliberative functions between the human participant and AI by specifying what the system may contribute and what it cannot establish at each stage. Second, it requires the process to conclude with a human-formulated justification rather than an AI-generated recommendation. Third, it treats ethical self-authorship as a developing capacity requiring pedagogical scaffolding and

¹ Levinsky-Wingate Academic Center, Tel-Aviv, Israel,

² Levinsky-Wingate Academic Center, Tel-Aviv, Israel, Email: rotemw1@gmail.com

ethically and digitally prepared teacher educators. An illustrative case demonstrates the model's application to a dilemma involving student use of AI in assessment. The article argues that AI may generate material for ethical inquiry, but it cannot represent affected people, determine which values should prevail, or assume professional responsibility. Moral authorship and answerability must remain human.

Key-words: ethical self-authorship; moral judgment; generative artificial intelligence; moral outsourcing; philosophical counseling; teacher education;

Introduction

The Emerging Problem of Moral Outsourcing

Generative artificial intelligence is increasingly used in education not only to retrieve information, summarize texts, and support writing, but also to interpret professional situations, formulate arguments, and compare possible courses of action. Teachers and teacher candidates can ask generative systems to identify stakeholders, explain competing values, anticipate consequences, or recommend responses to educational dilemmas. These capabilities mark a shift in the perceived role of AI. A system that was initially approached as an instrument for completing tasks may now appear to function as a dialogical partner in reflection and judgment. This shift is particularly consequential when professional decisions involve moral disagreement. Technical decisions generally concern the selection of effective means for achieving an agreed-upon goal. Ethical dilemmas, by contrast, involve uncertainty or conflict concerning the goal itself, the values that should guide action, or the responsibilities owed to different people. Questions about whether an exception should be made for a student, when confidential information may be disclosed, or how equality should be balanced with individual need cannot be resolved through information retrieval alone. They require contextual interpretation, attention to relationships, and responsibility for the consequences of the decision.

Generative AI may enrich such deliberation by making alternative perspectives and counterarguments more readily available. However, its linguistic fluency can also make its responses appear more authoritative than their epistemic or moral quality warrants. Research indicates that linguistic presentation and perceived AI agency influence users' trust and evaluations of conversational quality (Pan et al., 2025). A response may sound balanced, empathetic, and professionally reasoned even though the system has no lived understanding of the situation and cannot assume responsibility for its consequences. AI is therefore not a moral agent, but it can influence human moral agency by shaping how a problem is framed, which perspectives become visible, and which conclusion appears reasonable.

This article describes the resulting risk as *moral outsourcing*:

Moral outsourcing occurs when a person transfers substantive elements of ethical interpretation, justification, or judgment to an external system without adequately examining, endorsing, or assuming responsibility for them.

Moral outsourcing is not equivalent to consultation. Moral judgment is socially situated and legitimately informed by professional codes, colleagues, experts, philosophical traditions, and affected communities. The critical distinction is between assistance that expands deliberation and assistance that replaces the interpretive and justificatory work through which a person becomes accountable for a decision. A teacher may formally announce an AI-supported decision while being unable to explain its assumptions, defend its reasons, or identify how the system shaped the conclusion.

The risk is especially important in teacher education. Teacher candidates are still developing professional identities, ethical vocabularies, and confidence in judgment. During practicum, they may also operate under time pressure, cognitive load, institutional expectations, and evaluative anxiety. Under such conditions, a fluent AI-generated answer can offer immediate relief from ambiguity by supplying both a structured interpretation and professionally sounding language. Responsible AI use therefore cannot be based on the assumption that candidates already possess the reflective autonomy required to resist premature closure.

This article asks: *How can generative AI be integrated into philosophical counseling and ethical reflection in teacher education without undermining ethical self-authorship and human moral responsibility?*

To address this question, the article develops the *Ethical Self-Authorship Cycle*. Its contribution is threefold. First, it proposes a division of deliberative labor between the human participant and AI. At each stage, the model specifies what AI may contribute, what it cannot establish, and what the human participant must perform independently. Second, it requires an explicit return of moral authorship to the person. The process ends with a human-formulated justification rather than an AI-generated recommendation. Third, it situates this process within teacher education, treating ethical self-authorship as a developing capacity that requires pedagogical scaffolding and ethically and digitally prepared teacher educators. The model's novelty therefore lies not simply in proposing five stages of ethical reasoning. Existing frameworks already provide structured approaches to moral development and professional case analysis. Its distinctive purpose is to preserve human answerability when part of the language, framing, and argumentative content of deliberation is generated by a fluent but non-accountable system.

2. Philosophical and Interdisciplinary Foundations

2.1 Philosophical Counseling as Dialogical Practice

Philosophical counseling is a form of applied philosophical practice in which concepts, assumptions, values, and interpretations of lived experience are examined through dialogue. It emerged from efforts to restore philosophy as a practice relevant to everyday life rather than treating it exclusively as an academic discipline. Although the field includes diverse approaches, many share an emphasis on helping individuals clarify how they understand a problem, what values inform their position, and whether their beliefs and actions form a defensible whole (Lahav, 1996; Raabe, 2001; Schuster, 1999).

The philosophical counselor does not simply supply a solution. The dialogue may examine what an individual means by fairness, responsibility, dignity, autonomy, success, or care; identify contradictions among commitments; or reveal assumptions that make one course of action appear self-evident. Philosophical theories may be introduced, but

they function as resources for examining a lived question rather than as doctrines to be memorized or mechanically applied. The outcome may be a clearer formulation, a revised position, or a more responsible recognition of unresolved uncertainty.

This dialogical orientation distinguishes philosophical counseling from technical consultation. An expert may explain what an institutional policy requires, but that information alone cannot determine whether applying the policy in a particular case is equitable or caring. Philosophical counseling also differs from conventional philosophy instruction because its primary object is the participant's question rather than mastery of a philosophical canon. The distinction from psychotherapy requires greater caution. Philosophical counseling does not diagnose or treat mental disorders, and practitioners must recognize when psychological, medical, legal, or other professional expertise is required. Nevertheless, lived problems do not divide neatly into philosophical and psychological components. Claims that philosophical counseling is always non-clinical, autonomy preserving, or effective cannot be accepted without scrutiny. Amir (2004), for example, criticizes several assumptions through which the field defines itself and calls attention to its possible risks and insufficient empirical foundations. Philosophical counseling is therefore relevant to this article not because it offers an unproblematic method, but because it provides a dialogical ideal whose limits must remain visible.

2.2 Socratic Inquiry and Conceptual Clarification

The Socratic dimension of philosophical counseling is particularly relevant to ethical reflection. Socratic inquiry begins from the possibility that apparent certainty conceals conceptual ambiguity, inconsistency, or an unexamined assumption. Instead of immediately offering an answer, the interlocutor may ask what a central term means, why one obligation should take priority, whether the same principle would apply in another case, or what evidence might justify revising the initial position.

Such questions perform several distinct functions. Conceptual clarification distinguishes competing meanings of terms such as equality and equity. Counterexamples test whether a principle is applied consistently. Alternative formulations reveal that the way a dilemma is described already influences which values and solutions become visible. A teacher candidate who claims that fairness requires identical treatment,

for example, may be asked whether identical conditions remain fair when students face substantially different barriers.

Socratic questioning, however, is not inherently emancipatory. Questions may be leading, adversarial, or structured to produce the questioner's preferred conclusion. In teacher education, this risk is intensified because the teacher educator is also an evaluator. Candidates may attempt to infer the position that will be rewarded rather than articulate genuine uncertainty. Dialogical practice must therefore make its criteria explicit, permit defensible disagreement, and distinguish examination of reasoning from pressure to reach a predetermined answer.

Recognition of uncertainty can be an important educational outcome, but the ability to tolerate ambiguity should not be assumed. Teacher candidates working under evaluative and practical pressure may experience uncertainty as a problem requiring rapid closure. Philosophical inquiry consequently requires scaffolding, including manageable cases, focused questions, peer dialogue, and permission to formulate provisional conclusions.

2.3 Relationship to Moral and Educational Frameworks

The proposed model also draws on three established educational frameworks. Rest's account of moral behavior distinguishes moral sensitivity, moral judgment, moral motivation, and moral implementation (Rest, 1986; Rest et al., 1999). This framework demonstrates that ethical conduct involves more than choosing among abstract principles. A person must first recognize the moral dimensions of a situation and ultimately translate judgment into action. The present article concentrates primarily on interpretation and justification, where generative AI may perform substantive parts of the reasoning for the user.

Baxter Magolda's (2001) theory of self-authorship describes development beyond dependence on external formulas toward an internal capacity to define beliefs, identity, and relationships. Ethical self-authorship extends this idea into moral deliberation. It does not require independence from relationships, traditions, mentors, or professional codes. Instead, it requires individuals to examine how these influences shape their judgment and to remain answerable for the position they adopt. Generative AI adds a distinctive challenge because it can function as an accessible and linguistically persuasive external authority without appearing to occupy a conventional position of power.

Warnick and Silverman's (2011) framework for professional ethics courses in teacher education provides the closest practical precedent. Their case-analysis approach helps candidates identify *prima facie* moral considerations and pursue procedural objectivity without promising absolute solutions. The Ethical Self-Authorship Cycle shares its commitment to structured analysis and respect for moral autonomy. Its additional concern is that AI may now generate the framing, stakeholder map, ethical vocabulary, and justification before the candidate has formed an independent position.

The proposed model does not replace these frameworks. Its distinct contribution is to specify how deliberative functions should be divided when a non-accountable system participates in ethical reflection. Rest clarifies components of moral functioning, Baxter Magolda explains development beyond external authority, and Warnick and Silverman structure professional case analysis. The present model asks what must remain human-authored when AI contributes to these processes.

2.4 Why AI Is Not a Philosophical Counselor

Generative AI can reproduce some observable features of philosophical dialogue. It can ask for clarification, generate counterexamples, compare ethical approaches, and formulate objections. These capabilities may support preparatory or retrospective reflection. Students may also value the accessibility and apparently nonjudgmental character of AI interaction, although human tutors remain better positioned to provide contextual feedback and relational support (Fakour & Imani, 2025).

Yet producing dialogical language is not equivalent to participating in a philosophical relationship. A system may generate a relevant question without understanding why it matters to the person, how it is experienced, or when continued questioning becomes misleading. It does not share the participant's practical world, recognize the lived weight of the dilemma, or become accountable for the direction and consequences of the conversation. Its apparent empathy is a linguistic performance rather than a moral relationship.

AI should therefore be understood as a *bounded dialogical aid*, not as an independent philosophical counselor. It may contribute material for inquiry, but it cannot establish what affected people experience, determine which values should prevail, or assume responsibility for the resulting

decision. Philosophical counseling provides both the methods and the boundary for its use: dialogue should deepen and clarify human judgment, not relieve the participant of the obligation to judge.

3. Moral Outsourcing and the Authority of Fluency

Generative AI can support ethical deliberation without determining what decision should be made. It can organize information, identify possible stakeholders, compare courses of action, and generate objections that a user may not initially have considered. Such assistance may be valuable when relevant considerations are dispersed across professional policies, ethical principles, institutional expectations, and interpersonal relationships. By externalizing possible arguments, AI can make elements of the reasoning process more visible.

Decision support must nevertheless be distinguished from decision substitution. A system supports judgment when it supplies questions, information, or alternatives that the user critically evaluates. It begins to substitute for judgment when its framing or recommendation becomes the conclusion of the deliberative process without adequate examination. The distinction concerns more than who formally announces the decision. It concerns who performs the substantive work of defining the problem, determining what matters, weighing competing claims, and justifying the selected action.

This article defines *moral outsourcing* as follows:

Moral outsourcing occurs when a person transfers substantive elements of ethical interpretation, justification, or judgment to an external system without adequately examining, endorsing, or assuming responsibility for them.

Moral outsourcing should not be confused with consultation. Moral reasoning is socially situated and legitimately informed by colleagues, mentors, professional codes, experts, and affected communities. AI use is not inherently incompatible with ethical self-authorship. The relevant question is whether external assistance expands the user's reasoning or displaces the work required for responsible judgment.

Moral outsourcing may occur at three related levels. *Interpretive outsourcing* occurs when the system effectively determines what kind of moral problem the situation represents. For example, it may frame a case

primarily as a violation of academic integrity rather than as a conflict among integrity, assessment validity, inclusion, and unmet educational need. Because framing determines which values and stakeholders become visible, outsourcing may begin before any recommendation is offered.

Justificatory outsourcing occurs when the user adopts reasons produced by the system without being able to reconstruct or defend them independently. A teacher candidate may reproduce a sophisticated ethical analysis while lacking a clear understanding of why one principle should outweigh another. *Judgment outsourcing* occurs when the system's preferred course of action is accepted as the conclusion. These forms may overlap, but they need not occur together. A user may reject the final recommendation while still relying on the system's framing and argumentative structure.

Generative AI creates a distinctive risk because its influence is exercised through fluent language. Its responses may be coherent, well structured, responsive, and apparently balanced. They can use the vocabulary of fairness, care, autonomy, and professional responsibility in ways that resemble mature ethical reasoning. Research shows that linguistic presentation and perceived AI agency influence trust and assessments of conversational quality (Pan et al., 2025). Perceptions of AI-generated texts are likewise affected by assumptions about authorship and presentation, not only by substantive content (Lermann Henestrosa & Kimmerle, 2024).

It is therefore necessary to distinguish among *linguistic fluency*, *epistemic reliability*, and *moral competence*. Linguistic fluency concerns the clarity, coherence, and readability of a response. Epistemic reliability concerns whether its factual claims and inferences are sufficiently supported. Moral competence would require the capacity to recognize the ethical significance of a situation, understand the people and relationships involved, weigh reasons responsibly, and answer for the consequences. Strength in the first dimension does not establish strength in the other two.

The article uses the term *fluency-authority effect* to describe the proposed risk that linguistic clarity, confidence, and coherence will be interpreted as evidence of understanding, neutrality, expertise, or moral authority. This term is offered as a conceptual proposition requiring empirical examination, not as an already established psychological effect.

It identifies a tendency that may be especially relevant when users face uncertainty and seek cognitive closure.

Fluency can also conceal normative choices. A polished response may present one meaning of fairness or one hierarchy of responsibilities as though it followed naturally from the facts. The transition from description to evaluation becomes difficult to detect because contested assumptions are incorporated into a seamless explanation. Concerns about language models reproducing patterns and biases without making their sources visible further challenge the apparent neutrality of generated texts (Bender et al., 2021).

The problem is consequently not resolved by checking whether an AI-generated recommendation is plausible. A reasonable conclusion may still result from a process in which the user did not formulate the question, evaluate the justification, or assume meaningful ownership of the decision. Preserving moral agency requires making AI's influence visible at each of these levels and ensuring that rhetorical fluency does not become a substitute for ethical judgment.

4. Ethical Self-Authorship as a Developing Capacity

The concept of self-authorship describes a developmental movement from reliance on external formulas toward an internal capacity to define one's beliefs, identity, and relationships (Baxter Magolda, 2001). The present article extends this concept into ethical deliberation and applies it to situations in which generative AI participates in the interpretation and justification of professional decisions.

Ethical self-authorship is defined as follows:

Ethical self-authorship is the developing capacity to recognize, formulate, examine, justify, revise, and answer for a moral position in dialogue with others.

This definition includes several connected abilities. A person must recognize that a situation has an ethical dimension, participate in formulating the dilemma, consider the perspectives of those affected, examine assumptions and competing reasons, and formulate a position that can guide action. The position must remain open to revision when new evidence or stronger arguments emerge. Ethical authorship culminates in

answerability: the person must be prepared to explain the decision to those affected and accept responsibility for its consequences.

The concept should not be understood as moral independence or isolation. Human judgment is always shaped by language, relationships, cultural traditions, professional communities, institutional structures, and unequal distributions of authority. Relational accounts of autonomy show that agency develops through social relations rather than outside them (Mackenzie & Stoljar, 2000). Consultation with mentors, colleagues, professional codes, or AI systems does not necessarily weaken autonomy. Such resources may correct incomplete knowledge and reveal perspectives that the individual could not generate alone.

The relevant question is how these influences are incorporated into judgment. Autonomy is weakened when guidance becomes an unexamined replacement for interpretation or justification. The ethical author is therefore not the person who thinks alone, but the person who can identify how external sources shaped the decision, explain which considerations were accepted or rejected, and remain accountable for the resulting position.

Ethical self-authorship is closely related to critical thinking but cannot be reduced to it. Critical thinking concerns conceptual clarity, the credibility of evidence, the validity of inferences, and the consistency of conclusions. Moral judgment also involves values, relationships, commitments, and responsibility. A person may compare several ethical theories competently without adopting a position on the actual dilemma. Ethical self-authorship requires movement from analysis to a provisional but accountable commitment.

It also extends beyond AI literacy. AI literacy includes understanding the capabilities, limitations, applications, and social implications of artificial intelligence, as well as the ability to evaluate generated outputs critically (Ng et al., 2021). A user may understand that AI is fallible, verify its factual claims, and still adopt its moral framing without examination. Ethical self-authorship therefore asks an additional question: *Can the user explain, contest, endorse, and take responsibility for the moral position developed through interaction with AI?*

The term *developing capacity* is essential. Teacher candidates cannot be assumed to possess mature reflective autonomy when they enter practicum. They may encounter dilemmas while managing

classrooms, attempting to satisfy institutional expectations, and being evaluated by mentors. Under time pressure and cognitive load, ambiguity may be experienced less as an opportunity for philosophical inquiry than as a problem requiring immediate closure. A clear external answer can therefore become particularly attractive.

This creates a circular difficulty. Ethical self-authorship is needed to use AI critically, yet AI-supported reflection is proposed as a means of developing ethical self-authorship. A procedural model cannot resolve this problem merely by instructing candidates to retain responsibility. Participants may complete each formal step while continuing to depend on the system's framing and language. Ethical self-authorship must therefore be scaffold. Initial implementation should use manageable or retrospective cases, educator modeling, structured questions, peer discussion, and explicit comparison of alternative framings. Candidates should document which AI suggestions they accepted or rejected and formulate their final justification without asking the system to write it. Support can be reduced gradually as candidates demonstrate greater capacity to recognize uncertainty, challenge external authority, and defend their reasoning independently.

The model consequently treats ethical self-authorship neither as a prerequisite nor as an automatic outcome of AI use. It is the developmental aim toward which carefully supervised, dialogical, and critically structured reflection should be directed.

5. The Ethical Self-Authorship Cycle

The Ethical Self-Authorship Cycle translates the preceding concepts into a structured process for AI-supported ethical reflection. Its five stages are not intended as an arbitrary checklist. Each corresponds to a distinct task that must occur if deliberation is to result in an accountable moral judgment. A dilemma must first be constructed before perspectives can be expanded; perspectives must be identified before their assumptions can be examined; conceptual examination must precede the normative testing of alternatives; and analysis must ultimately lead to a human-authored position. The sequence can be summarized as:

*Problem construction → Perspective expansion → Conceptual critique
→ Normative evaluation → Accountable commitment*

The stages are analytically distinct but practically iterative. A newly heard perspective may require reformulating the dilemma, while examination of consequences may reveal an assumption that was previously overlooked. The model permits movement between stages, provided that none of the five moral tasks is omitted.

The model's novelty lies in its allocation of deliberative functions. At every stage, it distinguishes among the human task, the contribution AI may legitimately make, and the ethical boundary that the system must not cross. This division is summarized in Table 1.

The Ethical Self-Authorship Cycle: Division of Deliberative Functions

Stage	Human task	Permitted AI contribution	Ethical boundary	Evidence of self-authorship
1. Articulating the dilemma	Distinguish facts, interpretation s, evaluations, and missing information; formulate the moral conflict	Suggest questions, stakeholders, and alternative framings	AI does not determine the definitive moral problem	The participant explains and justifies the selected framing
2. Pluralizing perspectives	Seek and consider the perspectives of those affected	Generate hypothetical viewpoints and identify potentially absent voices	Simulated perspectives do not replace consultation with actual people	Actual testimony, interpretation, and AI-generated speculation are distinguished
3. Interrogating assumptions	Clarify concepts and examine assumptions, inconsistencies, and	Generate Socratic questions, counterexamples, and case variations	AI-generated questions may contain	The participant identifies and evaluates the assumptions

Stage	Human task	Permitted AI contribution	Ethical boundary	Evidence of self-authorship
	generalizations		hidden framing	influencing the judgment
4. Testing reasons and consequences	Compare alternatives using relevant values, duties, evidence, and possible outcomes	Formulate objections, apply ethical perspectives, and suggest possible consequences	Ethical theories and predictions are not decision algorithms	The participant independently weighs the competing reasons and moral costs
5. Reclaiming moral authorship	Formulate, defend, and remain answerable for a revisable position	Help review earlier reasoning, but not generate the final judgment	AI must not provide or compose the final moral answer	The participant explains the decision without relying on the system's authority or wording

5.1 Stage One: Articulating the Dilemma

Ethical deliberation begins by constructing the problem rather than immediately solving it. Professional situations do not arrive with their moral meaning already defined. The way a case is described determines which values, stakeholders, and solutions become visible.

The human participant first distinguishes among facts, interpretations, evaluations, and missing information. Facts concern what is reasonably established. Interpretations concern what those facts may mean. Evaluations concern what is considered fair, harmful, permissible, or required. Missing information identifies what must still be learned before a responsible decision can be made. This separation prevents an intuitive interpretation from being treated as an uncontested fact.

AI may contribute by asking what information is absent, identifying potentially affected stakeholders, and proposing alternative formulations

of the dilemma. For example, it may show that a case framed as rule enforcement could also involve equity, privacy, or assessment validity.

The ethical boundary is that AI cannot determine the definitive moral problem. Framing is a normative act, not a neutral technical operation. The participant must compare the proposed formulations and explain why the selected framing is adequate, what it includes, and what it may still exclude. The outcome is a provisional formulation that remains open to revision.

5.2 Stage Two: Pluralizing Perspectives

The second stage expands the range of perspectives considered. Initial judgments are often shaped by professional role, personal experience, institutional culture, or identification with one stakeholder. Ethical reflection requires attention to people whose interests, rights, or relationships may be affected by the decision.

The human task is to identify relevant stakeholders and, when appropriate, seek their actual perspectives. This does not mean that all positions are equally informed or morally defensible. It means that a position should not be rejected or ignored before its moral claim has been understood. AI may suggest possible stakeholder concerns and identify voices the participant has overlooked. These outputs can function as hypotheses that guide further questions. They may help a teacher candidate consider how a student, parent, classmate, mentor, or school leader could interpret the situation differently.

The boundary is *representational substitution*. A generated account of what a student or community might think cannot establish what that person or community actually experiences. The appearance of several viewpoints may also create a false sense of inclusion. The participant must therefore distinguish actual testimony from personal interpretation and AI-generated speculation. Where consultation is possible, simulation should lead to dialogue rather than replace it.

5.3 Stage Three: Interrogating Assumptions

The third stage examines the concepts and assumptions underlying the available positions. Ethical disagreements often depend on words that appear clear while carrying different meanings. Fairness may refer to identical treatment, equitable opportunity, procedural consistency, or responsiveness to need. Different definitions can lead to different judgments.

The human participant identifies the assumptions connecting the description of the situation to the proposed response. These may concern what students are capable of, what counts as authorship, which responsibilities belong to teachers, or when exceptions are justified. The participant also examines inconsistencies, overgeneralizations, and possible double standards.

AI may support this inquiry by generating Socratic questions, counterexamples, and variations of the case. It might ask whether a stated principle would be applied consistently in another context or what evidence would lead the participant to revise the position.

The boundary is that questions themselves can contain assumptions. Their wording and order may direct the user toward a particular conclusion while preserving the appearance of open inquiry. The participant must therefore examine the framing of the questions as well as the answers they produce. Evidence of self-authorship lies in identifying which assumptions are doing substantive moral work and explaining whether they should be accepted, revised, or rejected.

5.4 Stage Four: Testing Reasons and Consequences

The fourth stage evaluates the available courses of action. The human participant asks which values, duties, rights, relationships, professional standards, and anticipated consequences support each serious alternative. The purpose is not to produce a simple list of advantages and disadvantages, but to determine which reasons are relevant, sufficiently supported, and responsive to those affected.

AI may formulate objections, act as a critical interlocutor, and suggest possible unintended consequences. It may also examine the case through several ethical perspectives. Consequentialist reasoning can focus on benefits and harms; deontological reasoning can consider duties and rights; virtue ethics can ask what forms of professional character the action expresses; and care ethics can attend to relationships, dependency, and vulnerability. These perspectives are not algorithms for generating a correct answer. They illuminate different features of the case and may support conflicting conclusions. Similarly, AI-generated consequences are possibilities rather than reliable predictions. The participant must assess them against evidence, contextual knowledge, professional requirements, and actual stakeholder perspectives.

The stage should not produce false balance. The existence of several arguments does not mean that every option is equally defensible. Some may rely on inaccurate information, violate professional duties, or impose unjustifiable harms. Self-authorship is demonstrated when the participant independently weighs the competing reasons and acknowledges the moral costs and uncertainties attached to the preferred option.

5.5 Stage Five: Reclaiming Moral Authorship

The final stage moves from analysis to accountable commitment. The human participant formulates a position, explains why it is preferable to the alternatives, and acknowledges its limitations. Ethical self-authorship does not require absolute certainty. It requires a position that can guide action, be defended before those affected, and remain open to revision.

The participant identifies the values and obligations supporting the decision, the objections that remain significant, and the evidence that might justify reconsideration. The participant must also explain how AI influenced the process, including which suggestions were accepted and which were rejected.

This stage is governed by the *No-Final-Answer Principle*:

In AI-supported ethical reflection, the system must not provide or compose the final moral answer. The process must conclude with a human-formulated justification.

AI may help review the earlier analysis, but the participant should formulate the final position without asking the system to write it. Adding a statement of personal responsibility to an AI-generated conclusion is insufficient. The participant must be able to reconstruct and defend the reasoning without relying on the system's authority or wording.

The final position should answer five questions:

1. What should be done?
2. Which reasons and values justify this decision?
3. Who may benefit or bear its costs?
4. Which AI-generated claims were accepted or rejected, and why?
5. What new evidence or perspective might justify revision?

The cycle ends when the participant can answer for the decision, not when the system produces its most polished recommendation. AI may assist at every earlier stage, but moral authorship is demonstrated only

when the human participant converts assisted reflection into an independently justified and revisable professional judgment.

6. Application in Teacher Education

The following illustrative case demonstrates how the Ethical Self-Authorship Cycle can be applied in teacher education. It is intended to clarify the model rather than provide empirical evidence of its effectiveness.

A teacher candidate completing her practicum in a secondary school asks students to write an individual essay demonstrating their understanding of a topic discussed in class. The instructions permit generative AI for brainstorming and limited language revision but prohibit its use to produce substantial portions of the submitted text. The instructions do not define “substantial” precisely or explain how permitted assistance should be documented.

While reviewing the assignments, the candidate notices that one student’s essay differs markedly from his previous writing. She does not treat this difference as proof of misconduct. During a private conversation, the student acknowledges using AI beyond the stated limits. He explains that he understands the subject but has considerable difficulty expressing his ideas in writing and that AI allowed him to demonstrate knowledge he could not otherwise communicate. The candidate must decide whether to apply a penalty, require resubmission, consult the mentor teacher, or consider whether the student may need appropriate educational support.

In the first stage, *articulating the dilemma*, the candidate distinguishes facts from interpretations. The student used AI beyond the instructions, but his intention, the extent of AI involvement, the clarity of the policy, and the nature of his writing difficulty remain uncertain. AI may help identify missing questions, including whether the assignment assesses subject knowledge, writing proficiency, or both. The candidate formulates the dilemma as a conflict involving academic integrity, equitable opportunity, assessment validity, care, and the limits of her professional authority.

In the second stage, *pluralizing perspectives*, AI suggests concerns that might be raised by the student, classmates, parents, the mentor teacher, and school leadership. These suggestions are treated as

hypotheses. The candidate does not assume that a generated description represents what any actual stakeholder thinks. She listens to the student and consults the mentor teacher about relevant policy and support procedures. She distinguishes clearly between perspectives that have been expressed and those that remain speculative.

During the third stage, *interrogating assumptions*, the candidate examines what counts as authorship when AI contributes to planning, phrasing, or substantive content. She also questions whether fairness requires identical conditions or meaningful opportunities to demonstrate learning. AI may generate comparisons with spell-checkers, tutoring, or formal accommodations, but the candidate evaluates the limits of these analogies. She rejects both the assumption that every unauthorized use constitutes identical misconduct and the assumption that a learning difficulty automatically justifies unrestricted AI assistance.

In the fourth stage, *testing reasons and consequences*, the candidate compares several responses. An immediate penalty may support consistency but ignore ambiguity in the instructions and a possible unmet need. Accepting the essay unchanged may protect the student from punishment but undermine assessment validity. A supervised resubmission may preserve the requirement for independent work while giving the student another opportunity to demonstrate learning. The candidate considers duties of honesty and confidentiality, likely effects on trust and learning, relevant professional virtues, and the student's particular circumstances.

In the final stage, *reclaiming moral authorship*, the candidate formulates her own position without asking AI to compose it. Subject to the mentor teacher's approval, she proposes that the original essay not be accepted as evidence of independent work. The student will instead complete a supervised resubmission under clearly defined conditions, with any permitted AI assistance documented. The candidate also recommends clarifying the instructions for the whole class and consulting an appropriate support professional about future assessment conditions.

This decision is justified because it preserves academic integrity without treating punishment as the only possible response. It recognizes a potential educational barrier while respecting the candidate's limited authority to determine accommodations. The candidate rejects an AI-generated recommendation for automatic punishment because it reduces

fairness to uniform enforcement. She also rejects unrestricted acceptance because it overlooks authorship and assessment validity.

The case demonstrates the model's practical purpose. AI contributes questions, possible perspectives, and objections, but it does not determine what the student experienced, authorize an accommodation, or compose the final judgment. The teacher candidate remains responsible for consultation, contextual interpretation, and the decision she is prepared to defend.

Discussion, Conditions, and Limitations

The principal theoretical contribution of this article is not the proposal of another general sequence for ethical decision-making. Existing models already describe components of moral functioning, development beyond external authority, and professional case analysis. The distinctive contribution of the Ethical Self-Authorship Cycle is its allocation of deliberative functions between a human participant and a generative system. At each stage, the model specifies what AI may contribute, what it cannot establish, and what the human participant must retain to remain the author of the judgment.

This allocation addresses a problem that earlier frameworks were not designed to confront. Generative AI can now participate in framing a dilemma, representing stakeholders, formulating arguments, and proposing conclusions. The person who formally makes the decision may therefore no longer be the person who performed its substantive interpretive and justificatory work. The model responds by distinguishing interpretive, justificatory, and judgment outsourcing and by requiring evidence of human authorship throughout the process. Its central criterion is not whether AI was used, but whether the participant can explain, contest, and answer for the resulting position.

The model also contributes to philosophical counseling by extending selected dialogical practices into AI-mediated environments without classifying AI as a counselor. Generative systems can produce questions, conceptual distinctions, counterexamples, and alternative formulations. These outputs may support reflection before or after dialogue with a human counselor, peer group, or teacher educator. Nevertheless, producing dialogical language is not equivalent to

participating in a responsible philosophical relationship. AI does not share the participant's practical world, understand the lived significance of the dilemma, or become answerable for the consequences. The model therefore preserves a necessary distinction between generating material for inquiry and functioning as a morally accountable interlocutor.

For teacher education, the cycle provides a structure for examining dilemmas arising during practicum, including assessment, inclusion, privacy, academic integrity, professional boundaries, and institutional policy. It also offers an approach to assessment. Rather than evaluating only the sophistication of a final text, teacher educators can examine whether candidates formulated the dilemma, sought relevant perspectives, exposed assumptions, compared reasons, and justified a revisable position independently. This responds to concerns that AI ethics education often lacks adequate means of assessing whether learners have developed ethical reasoning rather than merely acquired ethical vocabulary (Wiese et al., 2025).

Implementation nevertheless requires safeguards. Cases entered into AI systems must be anonymized, minimized, and processed only through institutionally approved environments. AI-generated stakeholder perspectives must remain hypothetical and must not replace engagement with actual people. Generative systems may reproduce cultural and social assumptions while presenting dominant norms as neutral (Bender et al., 2021; Jenks, 2025). The apparent inclusion of several viewpoints may therefore conceal stereotyping or omission. Cases involving identifiable information, formal legal duties, immediate risks, or matters beyond the candidate's authority should be transferred to appropriate human professionals rather than analyzed independently through AI.

Teacher Educator Readiness

The model assumes not only candidate readiness but also teacher educator readiness. This condition cannot be taken for granted. A teacher educator who lacks relevant competence may focus on prompt formulation, fail to recognize dependence on generated language, or treat AI output as a reliable representation of ethical perspectives. Conversely, an educator may impose a preferred conclusion while describing the process as open reflection.

Implementation requires three interconnected capacities. *Ethical competence* is needed to recognize professional dilemmas, distinguish

factual from normative questions, and identify when consultation or escalation is required. *AI literacy* is necessary for understanding limitations, privacy risks, bias, generated uncertainty, and the difference between linguistic persuasiveness and justified knowledge (Ng et al., 2021). *Dialogical facilitation* is required to support disagreement, question assumptions, and evaluate reasoning without replacing the candidate's judgment. Institutions adopting the model should therefore provide preparation for teacher educators, approved case-handling procedures, and shared assessment criteria. The cycle should initially be used with hypothetical or retrospective cases under supervision. Its introduction as an independent digital exercise would contradict its developmental purpose.

Limitations and Future Research

The model is theoretical and has not been tested empirically. The illustrative case demonstrates possible use but does not establish that the cycle improves moral reasoning or reduces outsourcing. Participants may complete all five stages while continuing to depend on AI framing. They may also learn to perform the language of reflection without engaging in substantive reconsideration. Documentation increases visibility but cannot prove independent authorship. The model's operation will vary across systems, interfaces, prompts, languages, and institutional contexts. Its concept of self-authorship may also reflect cultural assumptions that emphasize individual verbal justification. Further work is required in contexts where responsibility is understood more collectively, relationally, or communally.

Future research should compare structured use of the cycle with unstructured AI consultation, peer discussion, and educator-facilitated dialogue. Outcomes could include recognition of competing values, identification of absent stakeholders, quality of justification, willingness to revise a position, and the ability to defend it without access to the AI transcript. Experimental research should also examine the proposed fluency-authority effect, while longitudinal studies should determine whether scaffolded use leads to greater independence or new forms of procedural dependence. Finally, research must assess teacher educator readiness, since the model cannot support ethical formation if those facilitating it lack the competence to recognize when technological assistance has begun to replace human judgment.

Conclusion

Generative AI is changing how professional situations are interpreted as well as how information is produced. In ethical deliberation, it can help identify missing questions, generate alternative perspectives, expose assumptions, and test possible reasons. These functions can broaden reflection, but they can also shift substantive elements of interpretation and justification away from the human decision-maker.

The risk of moral outsourcing cannot be assessed solely by asking whether an AI-generated recommendation is plausible. A defensible conclusion may still emerge through a process in which the user did not formulate the dilemma, evaluate the reasons, or recognize how the system shaped the judgment. Ethical responsibility therefore depends on more than formal approval of an output.

The Ethical Self-Authorship Cycle addresses this problem through five distinct tasks: constructing the dilemma, expanding perspectives, examining assumptions, evaluating reasons and consequences, and formulating an accountable position. Its contribution lies in assigning AI a limited role within each task while identifying what must remain under human control. The process culminates in the No-Final-Answer Principle, which requires the final moral justification to be formulated by the human participant. In teacher education, this capacity cannot be assumed. Teacher candidates may seek rapid closure while working under uncertainty, workload, and evaluative pressure. Ethical self-authorship must therefore be cultivated through structured cases, educator modeling, peer dialogue, and gradual reduction of support. Implementation also depends on teacher educators who possess ethical competence, AI literacy, and the ability to facilitate disagreement without imposing conclusions.

AI may assist moral inquiry, but it cannot represent affected people, understand the lived significance of a dilemma, or answer for the consequences of a decision. Its educational value lies in helping future teachers examine their judgments more carefully, not in relieving them of the responsibility to judge. The authorship and answerability of professional moral decisions must remain human.

References

- Amir, L. B. (2004). Three questionable assumptions of philosophical counseling. *International Journal of Philosophical Practice*, 2(1), 1–32.
- Baxter Magolda, M. B. (2001). *Making their own way: Narratives for transforming higher education to promote self-development*. Stylus Publishing.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery.
<https://doi.org/10.1145/3442188.3445922>
- Fakour, H., & Imani, M. (2025). Socratic wisdom in the age of AI: A comparative study of ChatGPT and human tutors in enhancing critical thinking skills. *Frontiers in Education*, 10, Article 1528603.
<https://doi.org/10.3389/feduc.2025.1528603>
- Jenks, C. J. (2025). Communicating the cultural other: Trust and bias in generative AI and large language models. *Applied Linguistics Review*, 16(2), 787–795.
<https://doi.org/10.1515/applirev-2024-0196>
- Lahav, R. (1996). What is philosophical in philosophical counselling? *Journal of Applied Philosophy*, 13(3), 259–278.
<https://doi.org/10.1111/j.1468-5930.1996.tb00171.x>
- Lermann Henestrosa, A., & Kimmerle, J. (2024). The effects of assumed AI vs. human authorship on the perception of a GPT-generated text. *Journalism and Media*, 5(3), 1085–1097.
<https://doi.org/10.3390/journalmedia5030069>
- Mackenzie, C., & Stoljar, N. (Eds.). (2000). *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self*. Oxford University Press.
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, Article 100041.
<https://doi.org/10.1016/j.caeai.2021.100041>

- Pan, W., Liu, D., Meng, J., & Liu, H. (2025). Human–AI communication in initial encounters: How AI agency affects trust, liking, and chat quality evaluation. *New Media & Society*, 27(10), 5822–5847. <https://doi.org/10.1177/14614448241259149>
- Raabe, P. B. (2001). *Philosophical counseling: Theory and practice*. Praeger.
- Rest, J. R. (1986). *Moral development: Advances in research and theory*. Praeger.
- Rest, J., Narvaez, D., Bebeau, M. J., & Thoma, S. J. (1999). *Postconventional moral thinking: A neo-Kohlbergian approach*. Lawrence Erlbaum Associates.
- Schuster, S. C. (1999). *Philosophy practice: An alternative to counseling and psychotherapy*. Praeger.
- Warnick, B. R., & Silverman, S. K. (2011). A framework for professional ethics courses in teacher education. *Journal of Teacher Education*, 62(3), 273–285. <https://doi.org/10.1177/00224871110398002>
- Wiese, L. J., Patil, I., Schiff, D. S., & Magana, A. J. (2025). AI ethics education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 8, Article 100405. <https://doi.org/10.1016/j.caeai.2025.100405>