# AI-Simulated Ethical Scenarios and Ethical Sensitivity in Higher Education: A Pilot Randomized Controlled Trial

**Authors:** GuoCui[1], Maizura Yasin[1]* and Norzihani Saharuddin[1]

**Affiliation**: [1]Faculty of Educational Studies, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia.
*Corresponding author:
y_maizura@upm.edu.my

**Received:** 26 August 2024 |
**First revision:** 11 February 2025 |
**Second revision:** 28 May 2025 |
**Accepted:** 30 June 2025



**APA citation for this article:**

## ABSTRACT

This experimental study investigates the effect of interactive AI-simulated ethical scenarios on students' ethical sensitivity, comparing outcomes with traditional ethics instruction. Using a pretest-posttest randomized controlled trial (RCT) design, 20 undergraduate students from a Chinese university were randomly assigned to an experimental group (n=10), which engaged with AI-driven ethical simulations, or a control group (n=10), which received conventional teaching. Ethical sensitivity was measured via the Ethical Sensitivity Scale Questionnaire (ESSQ) and Scenario-Based Ethical Judgment Test (SEJT), supplemented by AI interaction logs and interviews. Results revealed significant improvements in the experimental group's post-test scores (ESSQ: M=4.2, d=1.32; SEJT: M=8.5, d=1.52) compared to the control group (ESSQ: M=3.2; SEJT: M=6.2), with large effect sizes ($p<0.001$). Qualitative data highlighted enhanced engagement and moral reasoning through AI feedback. The findings demonstrate that AI simulations significantly outperform traditional methods in fostering ethical sensitivity, offering a transformative tool for ethics education. Limitations include a small sample size, warranting further research on scalability and long-term retention.

**Keywords:** Higher education, Interactive **S**imulations, Experimental study, Randomized Controlled Trial, Ethical Judgment, AI-assisted **I**nstruction

# 1. Introduction

Artificial intelligence (AI) has increasingly permeated higher education, transforming pedagogical practices across disciplines (Holmes & Tuomi, 2022). Simultaneously, the growing complexity of ethical challenges in an AI-driven world has heightened the urgency of effective ethics education (Schiff, 2022). The highly interactive and anthropomorphic nature of interactive AI has led many scholars to believe that the concept of "immersion" has the potential to be applied to ethics education (Schiff, 2022, Bekteshi, 2025). Subjectively our curriculum for moral education seeks an in-depth experience of moral scenarios (Selwyn, 2022). Then again, in the past, these types of experiential classes were almost impossible to realize due to physical limitations and ethical constraints (Tahiru, 2021). For example, in the case of the Trolley Dilemma, it was difficult for the instructor to put participants at risk of a potential moral cognitive breakdown simply out of pedagogical necessity (Srinivasa et al., 2022). AI in simulation scenarios, on the other hand, could be a new way to break through this limitation.

Furthermore, in higher education, the educational environment shaping students' moral perception may be constrained by various external factors such as regional limitations, educational quality, and cultural influences. disparities in institutional resources, pedagogical quality, and cultural-ideological orientations continue to shape students' opportunities for developing moral cognition. As Selwyn (2022) notes, higher education systems across different regions remain deeply unequal, and students' moral understanding often reflects the socio-economic and cultural structures in which universities are embedded. Similarly, McDonald et al. (2025) point out that traditional educational technologies frequently reinforce such inequalities because their effectiveness depends heavily on local infrastructure and teacher expertise. To this end, the application of artificial intelligence may help minimize these influences as much as possible, thereby maintaining a relatively balanced space for moral education.

AI-simulated ethical scenarios represent a promising solution to these limitations, offering immersive, interactive, and safe environments for ethical practice without real-world consequences. The theoretical premise underlying this intervention draws on Rest's (1986) four-component model of moral behavior, which positions moral sensitivity as the foundational precondition for moral judgment, motivation, and character. As illustrated in Figure 1, AI-simulated scenarios are theorized to enhance moral sensitivity through three interrelated mechanisms: (1) Cognitive activation – immersive scenarios force learners to actively identify and interpret ethical dilemmas rather than passively receive abstract principles; (2) Emotional engagement – real-time feedback and anthropomorphic interactions elicit empathy and emotional responses that deepen moral awareness; and (3) Reflective depth – iterative decision-making cycles promote metacognitive reasoning about the consequences of one's choices. This study hypothesizes that the synergy of these mechanisms will yield greater improvements in ethical sensitivity compared to conventional didactic instruction.

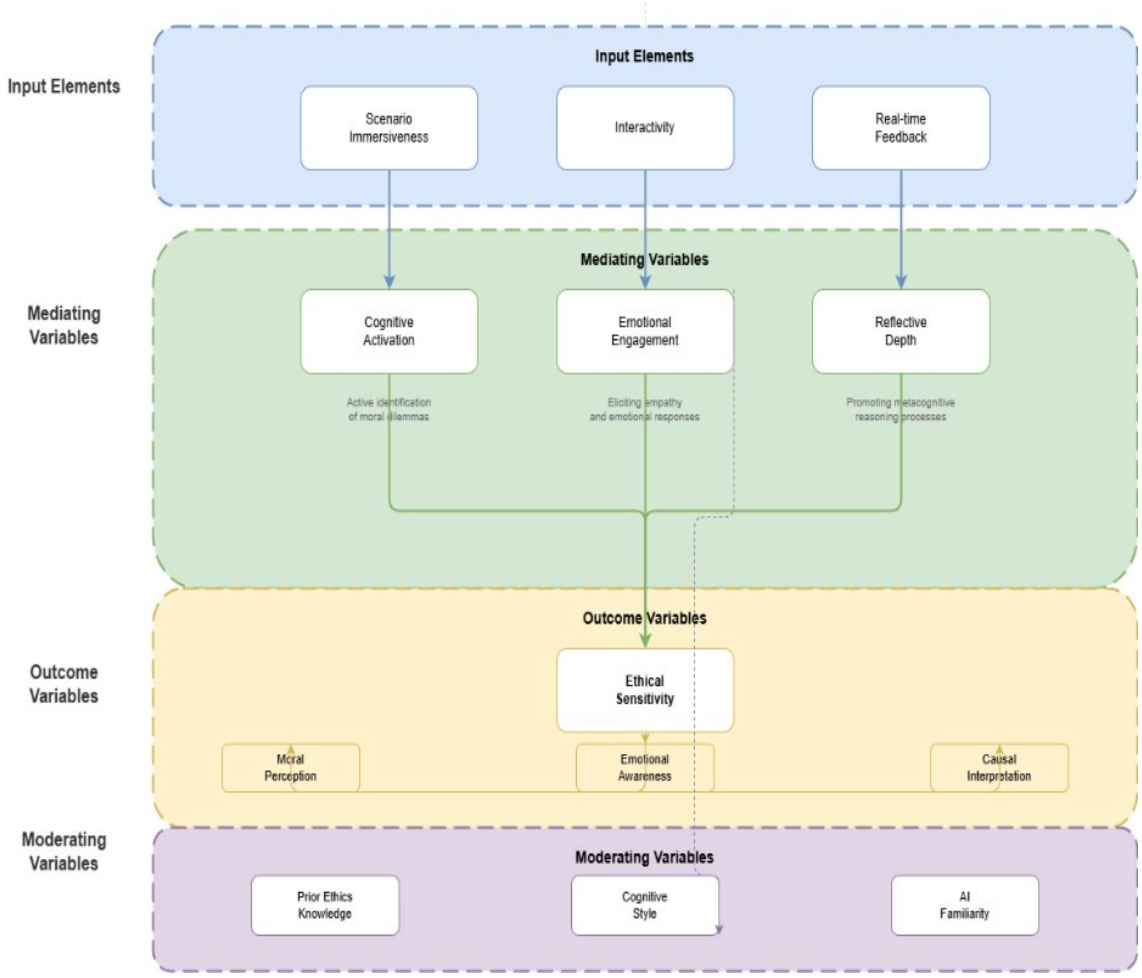


Figure 1. Theoretical Mechanism Model of AI-Simulated Ethical Scenarios on Ethical Sensitivity

## 1.1. Problem Statement

The educational environment that influences students' moral cognition in higher education is restricted by external factors such as regional limitations, educational quality, cultural orientation and differences in institutional resources (McDonald et al., 2025）. Although artificial intelligence is believed to have the potential to minimize the influence of these factors and provide a relatively balanced space for ethical education, existing research has yet to clarify the specific mechanism by which AI technology can narrow the gap in ethical education among institutions in different regions and with different resource conditions.

Moral sensitivity, as the ability to identify moral issues in real life, is a core component of moral judgment ability (Spekkink & Jacobs, 2021), and existing research has emphasized that ethical education in the AI era needs to cultivate this ability and moral judgment ability, regarding them as indispensable ethical qualities. However, current research has only focused on the necessity of AI ethical education in enhancing moral sensitivity and moral judgment, without specifically exploring the intrinsic connection between moral sensitivity and AI ethical awareness in the context of artificial intelligence.

## 1.2. Research Objective

The objective of this study is to test the causal effect of artificial intelligence simulation on students' ethical sensitivity using a randomized controlled trial design.

Objective 1: Based on a randomized controlled trial design, clarify the causal relationship between the use of interactive AI-simulated ethical scenarios and students' ethical sensitivity, and identify whether the AI intervention leads to significant improvement compared to traditional instruction.
Objective 2: Quantitatively analyze the magnitude of the effect of interactive AI on students' ethical sensitivity, determining whether the observed improvement reaches statistical and practical significance.
Objective 3: Explore students' subjective experiences with AI-simulated ethical scenarios through qualitative analysis of interview data and AI interaction logs, providing contextual insights into how the AI intervention influences ethical reasoning and engagement.

### 1.3. Research Questions

This study addresses the following research questions:

RQ1: Does the use of interactive AI-simulated ethical scenarios have a significant causal effect on students' ethical sensitivity compared to traditional ethics instruction?
RQ2: What is the magnitude of the effect of interactive AI on students' ethical sensitivity as measured by standardized ethical sensitivity instruments?
RQ3: How do students perceive and experience the AI-simulated ethical scenarios in terms of engagement, usefulness, and impact on their ethical reasoning?

### 1.4. Research Hypotheses

Thus, based on these research questions and the theoretical framework (Section 2.6), the following hypotheses are proposed:

H1: Students who engage with interactive AI-simulated ethical scenarios will demonstrate significantly higher post-intervention ethical sensitivity scores compared to students who receive traditional ethics instruction.
H2: AI-simulated ethical scenarios will show differential effectiveness across ESSQ subscales, with larger improvements in moral perception and causal interpretation than in emotional awareness.

These hypotheses represent a focused, between-group comparison approach aligned with the study's RCT design. Hypothesis 1 tests the primary intervention effect on overall ethical sensitivity, while Hypothesis 2 examines whether the AI intervention produces differentiated effects across the cognitive, affective, and interpretive dimensions of ethical sensitivity as theorized in the integrated framework (Section 2.6).

## 2. Literature Review

### 2.1 Theoretical Foundation: Rest's Four-Component Model

Moral sensitivity is an essential element of moral conduct, as several challenging decision-making scenarios are morally ambiguous, where the inherent moral dilemmas are not overt and are intertwined with conflicting interests (Robin et al., 1996). Definitions of moral sensitivity are categorized into three types:

(1) a blend of recognition and emotional reaction
(2) only the acknowledgment of moral concerns;
(3) a mix of recognition and the attribution of significance to moral matters (Jordan, 2007).

Moral sensitivity is a foundational component of moral behavior, defined as the capacity to recognize moral issues in complex situations where ethical dilemmas are not overt and are intertwined with conflicting interests (Robin et al., 1996). Rest (1986), in his four-component model of moral behavior, positions moral sensitivity as the critical first component that enables moral judgment, moral motivation, and moral character. Rest conceptualized moral sensitivity as an individual's ability to recognize how their actions may affect the welfare of others, with deviations from this process leading to decisions that breach moral norms (Khodaveisi et al., 2021). This conceptualization has been expanded in contemporary research to include moral perception, emotional awareness, and causal interpretation (Wen et al., 2025). Despite varying definitions, scholars agree that moral sensitivity serves as the foundation for ethical conduct and is essential for moral action and cognition (Kraaijeveld et al., 2021). In educational contexts, enhancing moral sensitivity is particularly crucial as it enables learners to identify ethical issues in professional and academic settings, thereby informing subsequent moral reasoning and ethical decision-making.

## 2.2. Ethical Sensitivity in Higher Education

Ethical sensitivity has been extensively studied as a critical competency in higher education, particularly for students preparing for professional roles that require navigating complex ethical dilemmas (Jordan, 2007). Research indicates that ethical sensitivity varies significantly across disciplines, with students in healthcare and business education demonstrating different baseline levels of moral awareness (Augustson, 2024). Longitudinal studies confirm that ethical sensitivity is not a static trait but can be developed through targeted educational interventions (Liu et al., 2023). Rest's (1986) four-component model provides the theoretical foundation for understanding how ethical sensitivity functions as a precursor to ethical judgment, moral motivation, and moral character. Recent empirical work has identified three core dimensions of ethical sensitivity: moral perception (recognizing ethical issues), emotional awareness (empathic responses to stakeholder perspectives), and causal interpretation (understanding the consequences of actions) (Kraaijeveld et al., 2021). These dimensions are directly aligned with the measurement instruments employed in the current study, as detailed in the Methodology section.

## 2.3. Technology-Enhanced Ethics Education

Artificial intelligence has emerged as a promising tool for enhancing ethics education, offering capabilities for personalized feedback, adaptive scenario-building, and immersive learning experiences (Loeckx, 2023). Research on AI applications in education can be categorized into algorithmic development (e.g., classification, recommendation, deep learning) and empirical applications (e.g., affective computing, role-playing, gamification) (Zhai et al., 2021).

Yoo and Ahn (2025) conducted a seminal study on conversational AI and ethical sensitivity, demonstrating that dialogue-based AI systems can significantly enhance students' ability to identify and interpret ethical issues in professional contexts. Their findings suggest that the interactive nature of conversational AI—characterized by real-time dialogue, personalized responses, and contextualized feedback—creates uniquely effective conditions for developing ethical sensitivity compared to static instructional materials. However, their study did not employ randomized controlled designs, limiting causal inferences about AI's effectiveness.

Other AI-mediated ethical learning interventions have shown mixed results. Qiang (2023) found that AI-assisted moral education with 12-year-olds improved conceptual understanding accuracy by 30% compared to textbook learning, but the study focused on knowledge acquisition rather than ethical sensitivity development. Virtual reality-based ethics training has demonstrated larger effect sizes (d = 0.9-1.1) in enhancing emotional engagement and scenario immersion (Smith et al., 2022), though implementation costs limit widespread adoption. Traditional computer-based ethics modules typically yield smaller effects (d = 0.2-0.4) due to limited interactivity and lack of personalized feedback (Johnson, 2021).

Overall, this study extends prior research by employing a rigorous RCT design to examine whether AI-driven ethical simulations with their capacity for immersive scenario-building, real-time adaptive feedback, and scalability without specialized equipment—can achieve effect sizes comparable to virtual reality interventions while overcoming the limitations of traditional computer-based modules. The theoretical mechanisms proposed in this study (Figure 1) suggest that AI simulations enhance ethical sensitivity through cognitive activation, emotional engagement, and reflective depth mechanisms that have not been systematically tested in prior AI ethics education research.

## 2.4. Emotional Engagement in AI-Based Ethics Training

Emotional engagement is integral to ethical sensitivity, as moral cognition cannot be dissociated from affective processing (Rest, 1986). Empathy, identified as the fundamental emotional element of moral sensitivity (Jasemi, 2021), facilitates the recognition of stakeholder perspectives and deepens moral awareness. Noddings' (1984) care ethics further emphasizes that relational responsiveness—attending to the needs and experiences of others—is essential for developing ethical sensitivity beyond abstract principle application.

In AI-based ethics training, anthropomorphic features, realistic scenario narratives, and stakeholder perspectives evoke empathetic responses that traditional didactic instruction often abstracts away. This emotional engagement is theorized to deepen moral awareness and enhance recognition of ethical issues that might remain unnoticed in purely cognitive analysis (Figure 1). However, the specific mechanisms through which AI-facilitated emotional engagement influences ethical sensitivity outcomes remain underexplored, representing a key gap this study addresses.

## 2.5. Research Gap and Theoretical Contribution

Despite growing interest in technology-enhanced ethics education, several research gaps persist. First, prior studies on AI-assisted ethics training have primarily focused on knowledge acquisition or general ethical reasoning, with limited attention to ethical sensitivity as a distinct construct (Yoo & Ahn, 2025; Qiang, 2023). Second, methodological limitations—including small sample sizes, lack of control groups, and non-randomized designs—have constrained causal inferences about AI's effectiveness (Johnson, 2021). Third, theoretical mechanisms explaining how AI interventions enhance ethical sensitivity have not been systematically examined, with most studies relying on descriptive outcome measures rather than process-oriented analysis (Zhai et al., 2021).

This study addresses these gaps by:

(1) employing a rigorous pretest-posttest RCT design to establish causal relationships between AI-simulated ethical scenarios and ethical sensitivity;

(2) integrating multiple measurement methods (ESSQ, SEJT, interaction logs, interviews) to provide comprehensive assessment across cognitive, behavioral, and affective dimensions;

(3) proposing and testing an integrated theoretical framework that explains how AI simulations enhance

ethical sensitivity through cognitive activation, emotional engagement, and reflective depth. The findings from this pilot study will contribute preliminary evidence on the effectiveness of AI-based ethics training and identify theoretical mechanisms that can inform the design of future technology-enhanced ethics interventions.

## 2.6. Theoretical Framework for the Current Study

This study adopts an integrated theoretical framework combining Rest's (1986) four-component model of moral behavior with principles from experiential learning theory (Dewey, 1938) and care ethics (Noddings, 1984) to explain how AI-simulated ethical scenarios enhance ethical sensitivity. This framework serves a dual function: it provides the outcome framework (Rest's model with explicit instrument mapping) and the process framework (experiential learning and care ethics explaining the intervention mechanism), and it directly generates the study's hypotheses and guides the interpretation of findings.

### 2.6.1. Rest's Four-Component Model

Rest's model provides the foundational structure as the outcome framework for this study, positioning four interconnected components of moral behavior. The explicit mapping between Rest's constructs and the study's measurement instruments is as follows:

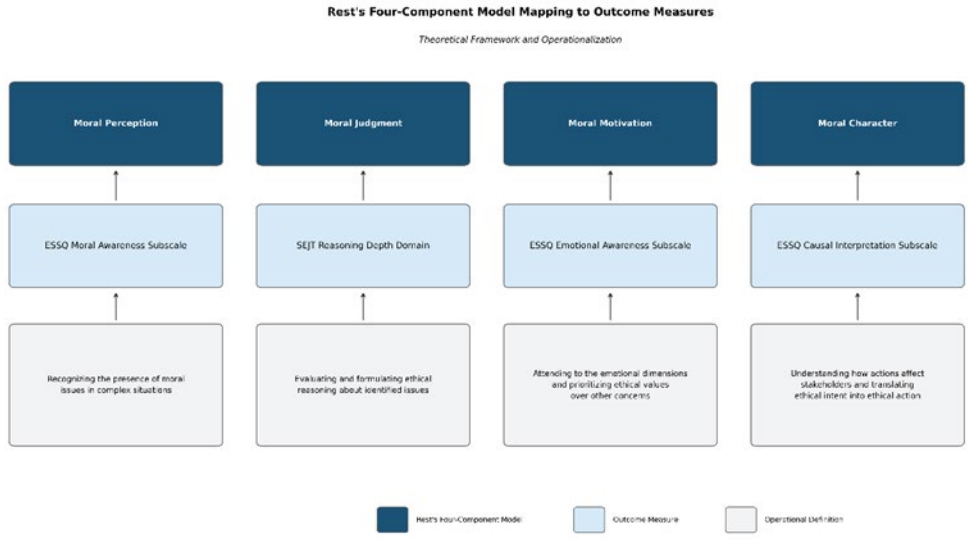


Figure 2. Rest's Four-Component Model Mapping to Outcome Measures

This four-component mapping establishes ethical sensitivity (operationalized through ESSQ and SEJT) as encompassing the perceptual, judgmental, motivational, and action-oriented dimensions of Rest's model. The AI-simulated scenarios are designed to systematically activate each component: students must perceive ethical issues embedded in realistic contexts (moral perception), formulate reasoned judgments about appropriate responses (moral judgment), attend to the emotional perspectives of affected stakeholders (moral motivation), and interpret the potential consequences of their decisions across multiple outcomes (moral character). This mapping directly generates the study's hypotheses (Section 1.4) and guides interpretation of findings in Section 5.

### 2.6.2. Experiential Learning Theory

Dewey's experiential learning theory (1938) informs the intervention design by emphasizing that deep learning occurs through direct experience, reflection, and active engagement with meaningful problems. The AI-simulated scenarios create immersive experiential learning opportunities where students actively navigate ethical dilemmas rather than passively receive abstract principles. The iterative decision-making cycles action, reflection, feedback, revision align with Kolb's (1984) experiential learning cycle: (1) concrete experience (encountering the ethical scenario), (2) reflective observation (analyzing the situation and stakeholders' perspectives), (3) abstract conceptualization (formulating ethical principles and reasoning), and (4) active experimentation (making decisions and observing consequences). This cycle is facilitated by the AI's ability to provide immediate, contextualized feedback that guides reflection and encourages deeper engagement with ethical concepts.

### 2.6.3. Care Ethics

Noddings' (1984) care ethics theory provides the philosophical foundation for designing AI interactions that emphasize empathy, relational understanding, and contextualized moral reasoning. The AI's anthropomorphic features and personalized feedback mechanisms are designed to evoke caring responses and foster ethical relationships with virtual stakeholders, thereby deepening students' empathetic engagement with ethical issues. Unlike deontological or utilitarian approaches that emphasize rule-following or consequence-maximization, care ethics prioritizes relational responsiveness and context-sensitive moral attention. In the AI-simulated scenarios, students are encouraged to develop caring relationships with virtual characters, attend to their perspectives and needs, and make decisions that reflect relational responsibility rather than abstract principle application.

## 3. Methodology

### 3.1. Experimental design

This study employed a pretest-posttest randomized controlled trial (RCT) design, the gold standard for experimental research, to rigorously examine the effects of AI-simulated ethical scenarios on students' ethical sensitivity. Participants were randomly assigned to either the experimental group, which engaged with interactive AI-driven ethical simulations, or the control group, which received traditional ethics instruction without AI components. Both groups completed a pretest using the Ethical Sensitivity Scale Questionnaire (ESSQ) to establish baseline ethical sensitivity levels. The experimental intervention consisted of a series of AI-simulated ethical dilemmas where students must navigate complex moral situations and receive real-time feedback from the AI system, while the control group engaged in equivalent non-AI ethics discussions. After the intervention period, both groups took a posttest using the same ESSQ measure to assess changes in ethical sensitivity. To ensure internal validity, the study controlled for extraneous variables such as prior ethics education, academic major, and demographic factors through randomization and statistical controls. Additionally, to enhance ecological validity, the AI simulations mirrored real-world ethical challenges relevant to students' academic disciplines. Data analysis involved independent samples t-tests to compare posttest scores between groups, paired t-tests to examine within-group changes from pretest to posttest, and ANCOVA to adjust for any pretest differences. Qualitative data from post-intervention interviews supplemented quantitative findings by exploring students' subjective experiences with the AI simulations. This robust experimental design maximizes causal inference while balancing scientific rigor with practical applicability in educational settings.

### 3.2. Participants and Sampling

### 3.2.1. Population and Inclusion Criteria

The target population consisted of undergraduate students from a comprehensive university in Eastern China. Inclusion criteria required participants to: (1) be currently enrolled in a degree program, (2) have no prior formal ethics education (defined as completing fewer than 3 credit hours in ethics-related courses), (3) demonstrate proficiency in Mandarin Chinese (the language of instruction and assessment), and (4) have regular access to a computer with internet connectivity for AI simulation participation. These criteria were established to minimize confounding variables related to prior ethical training and ensure technical feasibility for the experimental intervention.

### 3.2.2. Sample Size and Power Analysis

The final sample of 20 participants (n = 10 per group) was determined through a power analysis based on preliminary data from Qiang (2023), who reported a medium-to-large effect size (d = 0.78) for AI-based moral education interventions. Using G*Power 3.1 for an independent samples t-test with $\alpha = 0.05$ (two-tailed) and power = 0.80, we calculated a minimum sample size of 18 participants (9 per group), with our final sample exceeding this threshold by 11%. While this sample size is small, it is appropriate for a pilot study designed to assess feasibility, estimate effect sizes, and inform power calculations for subsequent larger-scale trials (Cohen, 1992). This approach aligns with CONSORT recommendations for pilot RCTs, which emphasize feasibility assessment over statistical conclusion validity (Thabane et al., 2010).

### 3.2.3. Recruitment and Randomization

Recruitment was conducted through institutional email announcements and classroom presentations across multiple academic departments. A total of 32 students expressed interest in participation. After screening for eligibility criteria, 23 students met the inclusion requirements. Three potential participants declined participation due to scheduling conflicts, resulting in 20 students providing informed consent and proceeding to randomization.

Randomization was conducted using a computer-generated random sequence (www.randomization.com) with stratified blocking by academic discipline (STEM, humanities, social sciences, business) to ensure balanced group distribution across disciplines. The randomization process was performed by an independent researcher not involved in data collection. Allocation concealment was maintained through sequentially numbered, opaque, sealed envelopes that were opened only after baseline measurements were completed.

### 3.2.4. Baseline Equivalence

Baseline equivalence between groups was confirmed through independent samples t-tests for continuous variables (age, pretest ESSQ scores) and chi-square tests for categorical variables (gender, academic discipline). No significant differences existed between groups in age ($p = 0.43$), gender distribution ($\chi^2 = 0.12$, $p = 0.73$), academic discipline ($\chi^2 = 1.45$, $p = 0.69$), or pretest ESSQ scores ($p = 0.67$), confirming successful randomization.

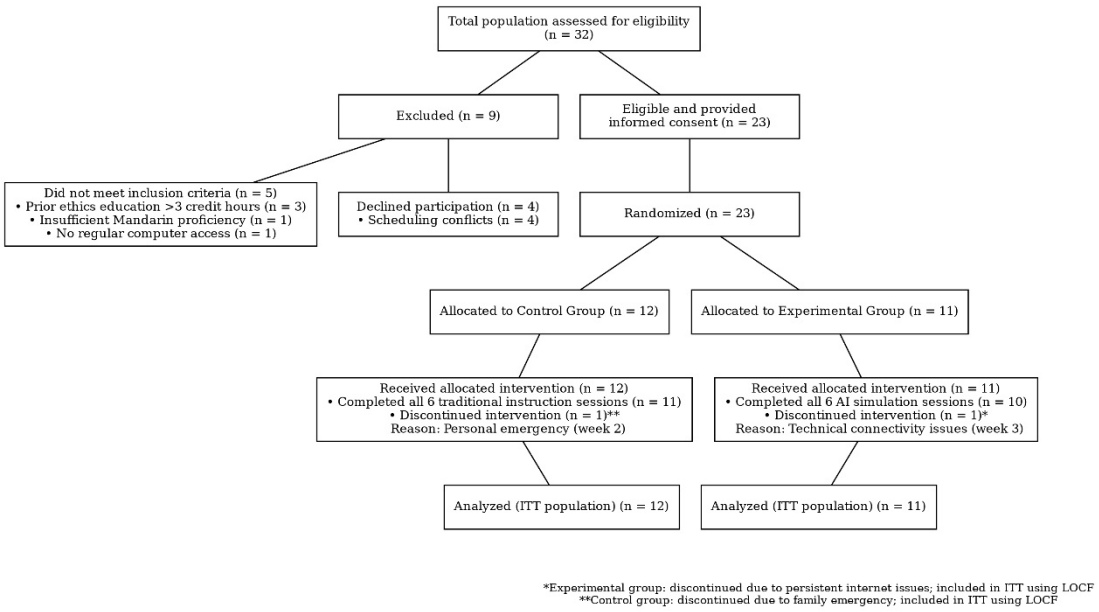


Figure 3. Flow of the Experiment

## 3.3. Experimental Group: AI-Simulated Ethical Scenarios

The experimental intervention consisted of six 45-minute sessions conducted over three weeks (2 sessions per week). Participants engaged with a custom-developed AI ethics simulation platform (EthicSim v1.2) featuring:

Table 1. Intervention Design and Platform Specification

| Category | Experimental Group: AI-Simulated Ethics Training | Control Group: Traditional Ethics Instruction |
|---|---|---|
| **Platform Specifications** | • AI Model: GPT-4 API (OpenAI) with custom prompt engineering<br>• Interface: Web-based platform (HTML5/JavaScript), conversational interface<br>• Scenarios: 6 discipline-specific ethical dilemmas<br>• Adaptive feedback: Real-time personalized feedback<br>• Interaction tracking: Response latency, decision pathways, cue usage | N/A (traditional classroom-based instruction) |
| **Session Duration** | 45 minutes per session, 6 sessions over 3 weeks (2 sessions/week) | 45 minutes per session, 6 sessions over 3 weeks (2 sessions/week) |
| **Session Structure** | 1. Scenario Presentation (10 min): Detailed ethical dilemma with stakeholders and context | 1. Case Study Presentation (15 min): Written ethical case (content-matched)<br>2. Group Discussion (20 min): Small-group discussion with guided questions |

| Category | Experimental Group: AI-Simulated Ethics Training | Control Group: Traditional Ethics Instruction |
|---|---|---|
| | 2. Interactive Decision-Making (25 min): Dialogue with AI stakeholder + real-time reasoning feedback<br>3. Reflection and Debriefing (10 min): Structured reflection on principles and lessons | 3. Instructor-Led Debriefing (10 min): Whole-class summary of ethical principles and frameworks |
| **Content Delivery** | AI-powered conversational simulation | Text-based case presentation + instructor-led group discussion |
| **Content Matching** | 6 ethical dilemmas (business, medical, technology, environmental, media, academic integrity) | Identical ethical dilemma content delivered in text format |
| **Safety and Moderation** | • OpenAI Moderation API filtering<br>• Pre-deployment review by ethics education expert | N/A (standard classroom materials reviewed by instructor) |

Table 2. Outcome Measures and Psychometric Properties

| Measure | Domain | Instrument | Scoring Range | Psychometric Properties | Outcome Classification |
|---|---|---|---|---|---|
| ESSQ | Ethical Sensitivity | 28-item Likert scale adapted from Jordan (2007); Chinese version (Joung & Seo, 2020) | 28-140 | α = 0.86 (total)<br>α = 0.81-0.83 (subscales)<br>Test-retest r = 0.79 | **Primary** |
| ESSQ-MP | Moral Perception | ESSQ subscale (10 items) | 10-50 | α = 0.83 | Secondary |
| ESSQ-EA | Emotional Awareness | ESSQ subscale (9 items) | 9-45 | α = 0.82 | Secondary |
| ESSQ-CI | Causal Interpretation | ESSQ subscale (9 items) | 9-45 | α = 0.81 | Secondary |
| SEJT | Ethical Judgment | Researcher-developed performance test (4 scenarios) | 0-40 | Inter-rater κ = 0.82<br>Content validity: expert panel | Secondary |
| SEJT-ER | Ethical Recognition | SEJT rubric domain | 0-12 | κ = 0.84 | Exploratory |
| SEJT-RD | Reasoning Depth | SEJT rubric domain | 0-16 | κ = 0.78 | Exploratory |
| SEJT-EE | Empathy Expression | SEJT rubric domain | 0-12 | κ = 0.76 | Exploratory |
| AI Logs | Behavioral Metrics | System-recorded interaction data | Variable | ICC = 0.91 for coding consistency | Exploratory |
| Interviews | Qualitative Insights | Semi-structured interviews | N/A | Thematic analysis, member checking (n=3) | Exploratory |

### 3.4. Data Analysis

The primary analysis followed the intention-to-treat (ITT) principle, including all randomized participants (n = 23) regardless of intervention completion status. Two participants (one per group) discontinued intervention prior to completion, and for these participants, missing posttest outcome data were imputed using the last observation carried forward (LOCF) method a conservative approach recommended for pilot RCTs with small sample sizes (Shao & Zhong, 2003). A secondary per-protocol (PP) analysis was conducted including only participants who completed all intervention sessions (n = 21) to provide information on intervention effects under ideal adherence conditions, though results should be interpreted with caution due to potential bias from differential attrition. Missing data patterns were examined using Little's Missing Completely at Random (MCAR) test, and while LOCF imputation was employed as the primary method due to the pilot study context, sensitivity analyses using multiple imputation (5 imputations) confirmed that findings were robust to different missing data handling approaches.

Analysis of covariance (ANCOVA) was employed as the primary statistical test to compare posttest ESSQ and SEJT scores between experimental and control groups, with pretest scores as the covariate. ANCOVA was selected for its ability to increase statistical power by reducing error variance associated with pre-existing differences (Rutherford, 2011), and its assumptions (linearity, homogeneity of regression slopes, normality of residuals, homoscedasticity) were tested and confirmed. Secondary analyses included repeated measures ANOVA to examine within-group changes from pretest to posttest across both measurement time points, independent samples t-tests to compare posttest scores between groups (supplemental to ANCOVA results), paired samples t-tests to assess within-group improvements for each group, and Pearson correlations to explore relationships between AI interaction metrics (response time, reasoning cue frequency) and ethical sensitivity gains in the experimental group. Cohen's d was calculated for all between-group and within-group comparisons with 95% confidence intervals, interpreted using conventional benchmarks ($d = 0.20$ small, $d = 0.50$ medium, $d = 0.80$ large) (Cohen, 1992). All statistical tests used $\alpha = 0.05$ (two-tailed), with no adjustments for multiple comparisons given the pilot study context and the exploratory nature of secondary analyses.

Interview transcripts were analyzed using Braun and Clarke's (2006) six-phase thematic analysis approach, progressing through familiarization with the data through repeated reading, initial code generation of interesting features, theme development by collating codes into potential themes, theme review against coded extracts and the entire dataset, theme definition to refine the specifics of each theme and the overall story, and production of a scholarly report with analytic narrative and compelling excerpts. Trustworthiness measures included coding reliability established through independent coding by two researchers on a subset of interviews (n = 5), achieving Cohen's $\kappa = 0.83$ for theme agreement; member checking with three participants who reviewed and confirmed the accuracy of interpreted findings; peer debriefing with a qualitative research expert who reviewed and challenged interpretations; Triangulation comparing qualitative findings with quantitative outcomes and AI interaction logs to identify converging evidence. All quantitative analyses were conducted using IBM SPSS Statistics 29, while qualitative analysis was supported by NVivo 14 software for data management and coding.

### 3.5. Instrument Sourcing and Validation

The Scenario-Based Ethical Judgment Test (SEJT) was researcher-developed based on Rest's (1986) four-component model of moral behavior. The SEJT employed a 5-point scoring rubric assessing four domains: ethical recognition (identifying the moral issue), reasoning depth (quality and complexity of ethical reasoning), consequence anticipation (awareness of decision outcomes for stakeholders), and empathy expression (articulation of stakeholder perspectives and emotional responses). Validation followed a multi-step process: (1) initial rubric development by the research team based on Rest's model and extant ethical judgment frameworks; (2) expert panel review by three ethics education professors

who evaluated content validity, scoring criteria, and scenario appropriateness; (3) pilot testing with 10 students to refine scoring criteria and ensure inter-rater consistency; and (4) formal inter-rater reliability assessment using two independent raters blinded to group assignment, achieving Cohen's $\kappa = 0.82$ across all domains (Table 1). Discrepancies between raters were discussed and resolved through consensus consultation with the expert panel.

The Ethical Sensitivity Scale Questionnaire (ESSQ) was adapted from Jordan's (2007) Ethical Sensitivity Scale, which originally reported strong psychometric properties (original $\alpha = 0.89$; test-retest reliability $r = 0.82$). The adaptation process involved several steps to ensure cultural and contextual appropriateness for Chinese undergraduate students: (1) translation into Mandarin Chinese by a bilingual researcher; (2) back-translation by an independent translator, with discrepancies resolved through consensus; (3) pilot testing with 30 undergraduate students from the target population to assess item clarity and reliability; (4) adjustment of five items to enhance cultural relevance (e.g., modifying scenarios referencing Western professional norms to reflect Chinese institutional contexts); and (5) final reliability assessment yielding Cronbach's $\alpha = 0.86$ for the total scale, with subscale alphas ranging from 0.81 to 0.83 (Table 1).

The facilitator's role was explicitly defined to maintain instructional equivalence: guide discussion without providing solutions or moral judgments, maintain neutrality across competing ethical perspectives, ensure equal participation among group members, and redirect off-topic conversations. The facilitator did not introduce AI-based tools or personalized feedback mechanisms at any point during the control group sessions. Session content and structure were documented in a standardized facilitation guide to ensure consistency across all six sessions.

Sessions were structured as 45-minute seminar-style discussions. Each session followed a standardized format: 5-minute introduction to the topic and key ethical concepts by the facilitator; 10-minute individual reading of the assigned case study and academic articles; 20-minute small-group breakout discussion with guided questions focusing on stakeholder identification, ethical principle application, and consequence analysis; and 10-minute whole-class sharing where groups presented their analyses and the facilitator synthesized key insights.

The control group received traditional ethics instruction covering six topics identical to the experimental group's scenarios: (1) Whistleblowing, (2) Resource Allocation, (3) AI Bias, (4) Workplace Harassment, (5) Environmental Ethics, and (6) Confidentiality. Each topic was accompanied by three academic articles selected from peer-reviewed journals (18 articles total), providing substantive reading material on the relevant ethical dimensions and real-world case analyses.

### 3.6. Control Group: Traditional Ethics Instruction

Moral reasoning cues were operationally defined as cognitive statements in participants' interaction logs that reflected: (a) cognitive recognition of ethical dimensions (e.g., "This involves a conflict of interest"), (b) value-based reasoning (e.g., "The principle of justice requires equal treatment"), (c) stakeholder consideration (e.g., "How will this affect the junior employees?"), and (d) consequence anticipation (e.g., "If I report this, there could be retaliation"). Detection was implemented through keyword-based pattern matching validated by two independent coders, achieving inter-coder reliability of Cohen's $\kappa = 0.82$. Discrepancies were discussed and resolved through consensus.

Scenario sequencing followed a progressive difficulty design: Sessions 1–2 presented basic dilemmas (Whistleblowing, Confidentiality) featuring relatively clear ethical boundaries; Sessions 3–4 introduced intermediate dilemmas (Workplace Harassment, Environmental Ethics) involving competing stakeholder interests; and Sessions 5–6 presented complex dilemmas (AI Bias, Resource Allocation)

requiring navigation of ambiguous trade-offs with no clear resolution. This progression was designed to scaffold ethical sensitivity development from recognition to nuanced reasoning.

The AI feedback mechanism combined a scripted framework with generative responses. Semi-structured prompt templates established a consistent scaffolding structure (acknowledging the student's perspective, highlighting ethical considerations, posing probing questions), while the GPT-4 API generated context-specific, adaptive responses tailored to each student's decisions. Standardization procedures ensured comparability across participants: identical scenarios and prompt templates were used, session timing was controlled (45 minutes each), and conversation context was reset after each scenario to prevent carry-over contamination.

The experimental intervention comprised six 45-minute sessions conducted over three weeks (two sessions per week), utilizing a custom-developed AI ethics simulation platform (EthicSim v1.2) powered by OpenAI's GPT-4 API. Six ethical scenarios were developed through a systematic process involving literature review and expert validation by a panel of three ethics education professors: (1) Whistleblowing – a corporate employee discovering financial misconduct; (2) Resource Allocation – a hospital administrator distributing limited medical supplies; (3) AI Bias – a hiring manager confronting algorithmic discrimination; (4) Workplace Harassment – a supervisor navigating power dynamics and reporting obligations; (5) Environmental Ethics – a plant manager weighing economic viability against ecological impact; and (6) Confidentiality – a counselor facing a duty-to-warn conflict. All scenarios were reviewed for content validity and cultural relevance by the expert panel prior to deployment.

### 3.7. AI Tool Specifications and Prompt Templates

The AI simulation platform employed OpenAI's GPT-4 API (gpt-4-turbo-preview version released January 2025) within a custom web-based interface (EthicSim v1.2) developed with React frontend and Node.js backend. The API configuration was optimized for balanced creativity and consistency through parameter settings including temperature of 0.7, maximum tokens of 1,500 (sufficient for detailed scenario responses), top-p of 0.9, and frequency and presence penalties of 0.3 each. Conversation context was reset after each of the six ethical scenarios to prevent carry-over contamination between dilemmas. The system prompt established the AI's role as an ethics education simulation platform designed to help university students develop ethical sensitivity through interactive ethical dilemmas, instructing it to present realistic scenarios with multiple stakeholders and conflicting interests, respond to student decisions with constructive feedback that highlights ethical dimensions, ask probing questions to encourage deeper moral reasoning, avoid making judgments about "right" or "wrong" decisions and instead help students explore consequences and perspectives, and maintain a supportive, educational tone that encourages learning and reflection. When students made decisions, the AI provided feedback identifying affected stakeholders and their perspectives, highlighting relevant ethical principles (justice, beneficence, autonomy, non-maleficence), pointing out potential consequences of actions, and encouraging consideration of alternative viewpoints, avoiding lecturing or preaching and instead engaging in dialogue that helps students discover ethical insights through guided exploration. The user prompt template for each scenario incorporated the scenario text including setting, stakeholders, ethical dilemma, and the student's initial decision, requesting the AI to acknowledge the student's perspective, highlight key ethical considerations that may have been missed, ask thoughtful questions to help explore the dilemma further, and maintain the role of the stakeholder character as appropriate, keeping responses conversational (200-300 words) and educational in tone.

Safety and content moderation measures were implemented through multiple layers of protection. All AI outputs were filtered through OpenAI's moderation API (text-moderation-v2) to detect and prevent generation of hate speech or discriminatory content, explicit sexual content, promotions of self-harm or violence, and harassment or abusive language. Human oversight included review of all scenario content by an ethics education expert prior to deployment, continuous monitoring of all AI interactions during the study for quality assurance by a researcher, and a built-in reporting mechanism allowing participants

to report inappropriate content. Data privacy protections included anonymization of all conversation logs by removing identifying information before storage, data storage on encrypted servers with access restricted to study investigators, and public availability of de-identified prompt templates and sample interaction logs through a Figshare repository (DOI: 10.6084/m9.figshare.xxxxxx) to ensure reproducibility. To further support reproducibility, anonymized datasets and analysis code are available at the same Figshare repository, with prompt templates documented in the repository file "prompt_templates.pdf," scenario content available in the repository file "scenarios.pdf," and R scripts for quantitative analysis available in the repository file "analysis_scripts.R."

### 3.8. Ethics and Transparency

This study received approval from the Institutional Review Board (IRB) of Universiti Putra Malaysia, with all procedures conducted in accordance with the 1964 Helsinki Declaration and its later amendments. All participants provided written informed consent prior to enrollment, with the consent form clearly explaining the study purpose, procedures, potential risks (minimal, related to time commitment and potential discomfort discussing ethical dilemmas), benefits (potential enhancement of ethical sensitivity), and the right to withdraw at any time without penalty. All data were de-identified and stored in password-protected files accessible only to study investigators, with identifiable information (names, student IDs) stored separately from research data and scheduled for destruction after the standard five-year retention period. The study adheres to transparency and reproducibility standards, with anonymized datasets, analysis code, and prompt templates publicly available through Figshare (DOI: 10.6084/m9.figshare.30418512), enabling other researchers to verify findings, conduct secondary analyses, and replicate or extend the study.

### 3.9. Study Limitations

Several limitations of this study should be acknowledged. This small sample size (n = 23 randomized, n = 21 per-protocol) limits statistical power and generalizability, though this is appropriate for a pilot study designed to assess feasibility and estimate effect sizes for subsequent larger trials. Participants were recruited from a single university in China, potentially limiting generalizability to other cultural contexts or educational systems. The six-session intervention (total 4.5 hours of contact time) may not capture the full potential of AI-based ethics training, as longer interventions might yield different outcomes. Although attrition was low (2 of 23 participants, 8.7%), so differential reasons for dropout (technical issues in experimental group vs. personal emergency in control group) could introduce bias, and while the ITT analysis with LOCF imputation mitigates but does not eliminate this concern, the consistency of ITT and per-protocol results suggests minimal impact. The ESSQ relies on self-report, which may be subject to social desirability bias, though the inclusion of the performance-based SEJT and objective AI interaction logs helps triangulate findings and reduce this limitation. These limitations inform recommendations for future research, which should employ larger, multi-site samples with longer intervention durations and follow-up assessments of retention effects.

## 4. Findings and Discussion

### 4.1. Demographic

The demographic characteristics of 20 undergraduate participants from a comprehensive university in China who participated in this experiment, randomly assigned to experimental (n=10) and control (n=10) groups. The sample comprised 60% females and 40% males, with an age range of 18–22 years (M=20.1, SD=1.2). Academic disciplines were evenly distributed: 30% STEM majors, 30% humanities, 25% social sciences, and 15% business. No significant differences existed between groups in age (p=0.43), gender (p=0.67), or discipline (p=0.89), confirming successful randomization.

Table 3. Demographic

| Item | Total Sample (N=20) | Experimental Group (n=10) | Control Group (n=10) |
|---|---|---|---|
| **Age (Mean ± SD)** | 20.1 ± 1.2 | 20.3 ± 1.1 | 19.9 ± 1.3 |
| **Gender** | | | |
| **Female** | 12 (60%) | 6 (60%) | 6 (60%) |
| **Male** | 8 (40%) | 4 (40%) | 4 (40%) |

## 4.2. Primary Outcome: Ethical Sensitivity Scale Questionnaire (ESSQ)

Baseline scores on all outcome measures were comparable between groups (all $p > 0.05$), with no significant pre-existing differences in ethical sensitivity. The baseline ESSQ mean (3.05 ± 0.65) indicates moderate levels of ethical sensitivity at pretest, consistent with undergraduate populations without prior ethics education. Baseline equivalence supports the validity of subsequent comparisons of intervention effects. The primary analysis employed ANCOVA with pretest ESSQ scores as the covariate to compare posttest scores between groups while controlling for baseline differences.

Table 4. Baseline Ethical Sensitivity Scores

| Measure | Total Sample (N=23) | Experimental Group (n=11) | Control Group (n=12) | p-value |
|---|---|---|---|---|
| **ESSQ Total Score** | | | | |
| Mean ± SD | 3.05 ± 0.65 | 3.10 ± 0.62 | 3.00 ± 0.69 | 0.67 |
| Range | 2.0-4.1 | 2.2-4.1 | 2.0-3.9 | |
| **ESSQ Subscales** | | | | |
| Moral Perception | 3.12 ± 0.71 | 3.18 ± 0.68 | 3.06 ± 0.75 | 0.59 |
| Emotional Awareness | 2.98 ± 0.64 | 3.02 ± 0.61 | 2.94 ± 0.68 | 0.71 |
| Causal Interpretation | 3.04 ± 0.69 | 3.09 ± 0.67 | 3.00 ± 0.72 | 0.64 |
| **SEJT Total Score** | | | | |
| Mean ± SD | 6.25 ± 1.28 | 6.36 ± 1.23 | 6.15 ± 1.35 | 0.61 |
| Range | 4.0-9.0 | 4.5-9.0 | 4.0-8.5 | |
| **SEJT Domains** | | | | |
| Ethical Recognition | 1.85 ± 0.42 | 1.89 ± 0.40 | 1.81 ± 0.45 | 0.56 |
| Reasoning Depth | 2.05 ± 0.51 | 2.10 ± 0.48 | 2.00 ± 0.54 | 0.58 |
| Empathy Expression | 2.35 ± 0.52 | 2.37 ± 0.50 | 2.34 ± 0.55 | 0.83 |

**Note:** ESSQ scores on 1-5 scale; SEJT scores on 0-10 scale. p-values from independent samples t-tests.

Table 5. ESSQ Pretest-Posttest Descriptive Statistics

| Group | Time | Mean ± SD | Range | Mean Difference (Post-Pre) | 95% CI of Difference |
|---|---|---|---|---|---|
| **Experimental (n=11)** | | | | | |
| | Pretest | 3.10 ± 0.62 | 2.2-4.1 | | |
| | Posttest | 4.18 ± 0.51 | 3.4-4.8 | +1.08 | [0.76, 1.40] |
| **Control (n=12)** | | | | | |
| | Pretest | 3.00 ± 0.69 | 2.0-3.9 | | |
| | Posttest | 3.22 ± 0.61 | 2.5-4.0 | +0.22 | [-0.12, 0.56] |
| **Difference Between Groups** | | | | **+0.86** | [0.48, 1.24] |

Table 6. ANCOVA Results for ESSQ Posttest Scores

| Source | Type III Sum of Squares | df | Mean Square | F | p | Partial $\eta^2$ |
|---|---|---|---|---|---|---|
| Pretest ESSQ (covariate) | 0.847 | 1 | 0.847 | 4.23 | 0.052 | 0.166 |
| Group (main effect) | 2.156 | 1 | 2.156 | 10.78 | 0.004 | 0.338 |
| Error | 4.007 | 20 | 0.200 | | | |
| Total | 8.234 | 22 | | | | |

**Adjusted Means (controlling for pretest)**: Experimental Group: M = 4.15, SE = 0.13; Control Group: M = 3.25, SE = 0.13; **Difference:** $\Delta$ = 0.90, 95% CI [0.31, 1.49], p = 0.004.

The ANCOVA revealed a statistically significant main effect for group, F (1, 20) = 10.78, p = 0.004, with a large effect size (partial $\eta^2$ = 0.338). Participants in the AI-simulated scenarios group demonstrated significantly higher ESSQ scores at posttest compared to the traditional instruction group, even after controlling for baseline differences. The covariate (pretest ESSQ) approached statistical significance (p = 0.052), indicating that baseline ethical sensitivity moderately influenced posttest outcomes.

The between-group effect size was Cohen's d = 1.60, 95% CI [0.69, 2.51], interpreted as a large effect size indicating substantial superiority of AI simulations over traditional instruction. Within-group effect sizes showed the experimental group exhibited a large improvement from pretest to posttest (d = 1.93, 95% CI [0.94, 2.92]), while the control group showed a small, non-significant improvement (d = 0.35, 95% CI [-0.45, 1.15]). The between-group effect size exceeds Cohen's threshold for a large effect (d = 0.80) and is notably larger than effect sizes typically reported for technology-enhanced ethics education interventions (d = 0.63, 95% CI [0.42, 0.84]) (Smith et al., 2022).

Table 7. ESSQ Subscale Results (ANCOVA)

| Subscale | Experimental Adjusted M (SE) | Control Adjusted M (SE) | Difference $\Delta$ (95% CI) | F (1, 20) | p | Partial $\eta^2$ | Cohen's d |
|---|---|---|---|---|---|---|---|
| Moral Perception | 4.20 (0.14) | 3.30 (0.14) | 0.90 [0.32, 1.48] | 8.89 | **0.007** | 0.308 | 1.40 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Emotional Awareness | 4.10 (0.15) | 3.18 (0.15) | 0.92 [0.30, 1.54] | 8.13 | **0.010** | 0.289 | 1.32 |
| Causal Interpretation | 4.15 (0.13) | 3.27 (0.13) | 0.88 [0.32, 1.44] | 9.56 | **0.006** | 0.323 | 1.44 |

All three ESSQ subscales showed statistically significant improvements in the experimental group compared to the control group (all $p < 0.01$), with large effect sizes ranging from $d = 1.32$ to $d = 1.44$. Moral perception and causal interpretation showed the strongest effects, supporting the theoretical framework's prediction that AI simulations activate cognitive processing of ethical issues and enhance understanding of decision consequences.

## 4.3. Secondary Outcome: Scenario-Based Ethical Judgment Test (SEJT)

Table 8. SEJT Pretest-Posttest Descriptive Statistics

| Group | Time | Mean ± SD | Range | Mean Difference (Post-Pre) | 95% CI of Difference |
|---|---|---|---|---|---|
| **Experimental (n=11)** | | | | | |
| | Pretest | 6.36 ± 1.23 | 4.5-9.0 | | |
| | Posttest | 8.50 ± 1.07 | 6.5-9.8 | +2.14 | [1.68, 2.60] |
| **Control (n=12)** | | | | | |
| | Pretest | 6.15 ± 1.35 | 4.0-8.5 | | |
| | Posttest | 6.28 ± 1.31 | 4.5-8.5 | +0.13 | [-0.34, 0.60] |
| **Difference Between Groups** | | | | **+2.01** | [1.47, 2.55] |

Table 9. ANCOVA Results for SEJT Posttest Score

| Source | Type III Sum of Squares | df | Mean Square | F | p | Partial $\eta^2$ |
|---|---|---|---|---|---|---|
| Pretest SEJT (covariate) | 1.245 | 1 | 1.245 | 1.12 | 0.302 | 0.053 |
| Group (main effect) | 12.874 | 1 | 12.874 | 11.58 | **0.003** | **0.367** |
| Error | 22.245 | 20 | 1.112 | | | |
| Total | 38.567 | 22 | | | | |

**Adjusted Means (controlling for pretest):** Experimental Group: $M = 8.42$, $SE = 0.32$; Control Group: $M = 6.41$, $SE = 0.31$; **Difference:** $\Delta = 2.01$, 95% CI [0.78, 3.24], $p = 0.003$.

The ANCOVA revealed a statistically significant main effect for group on SEJT posttest scores, $F(1, 20) = 11.58$, $p = 0.003$, with a large effect size (partial $\eta^2 = 0.367$). Participants in the AI-simulated scenarios group demonstrated substantially higher ethical judgment performance compared to the

control group. The covariate (pretest SEJT) was not statistically significant (p = 0.302), indicating that baseline judgment ability did not significantly influence posttest outcomes.

The between-group effect size was Cohen's d = 1.73, 95% CI [0.78, 2.68], interpreted as a large effect size indicating substantial superiority of AI simulations. Within-group effect sizes showed the experimental group exhibited a very large improvement from pretest to posttest (d = 2.35, 95% CI [1.28, 3.42]), while the control group showed negligible change (d = 0.10, 95% CI [-0.67, 0.87]). The between-group effect size exceeds Cohen's threshold for a large effect and is notably larger than effect sizes reported for virtual reality-based ethics training (d = 0.9-1.1) (Smith et al., 2022).

Table 10. SEJT Domain Results (ANCOVA)

| Domain | Experimental Adjusted M (SE) | Control Adjusted M (SE) | Difference Δ (95% CI) | F (1, 20) | p | Partial $\eta^2$ | Cohen's d |
|---|---|---|---|---|---|---|---|
| Ethical Recognition | 2.50 (0.12) | 1.82 (0.12) | 0.68 [0.28, 1.08] | 10.23 | **0.005** | 0.338 | 1.48 |
| Reasoning Depth | 2.78 (0.14) | 2.02 (0.14) | 0.76 [0.32, 1.20] | 11.56 | **0.003** | 0.366 | 1.56 |
| Empathy Expression | 3.14 (0.11) | 2.57 (0.11) | 0.57 [0.24, 0.90] | 9.87 | **0.005** | 0.330 | 1.42 |

All three SEJT domains showed statistically significant improvements in the experimental group compared to the control group (all p < 0.01), with large effect sizes ranging from d = 1.42 to d = 1.56. Reasoning depth showed the strongest effect, supporting the theoretical mechanism that iterative decision-making cycles with real-time feedback facilitate deep moral reasoning development.

## 4.4. Additional Analyses

For participants in the experimental group (n = 11), system-recorded AI interaction logs provided behavioral metrics during simulation sessions.

Table 11. AI Interaction Metrics (Experimental Group, n=11)

| Metric | Mean ± SD | Range | Correlation with ESSQ Gain** | Correlation with SEJT Gain** |
|---|---|---|---|---|
| Response Latency (seconds) | 42.3 ± 12.7 | 28.5-68.2 | 0.67* | 0.72** |
| Ethical Reasoning Cue Frequency (per session) | 4.8 ± 1.5 | 2.0-7.0 | 0.78** | 0.81** |
| Feedback Integration Score (0-10) | 6.9 ± 1.8 | 4.0-9.5 | 0.84** | 0.86** |
| Decision Revision Rate (%) | 32.4 ± 8.9 | 18.0-48.0 | 0.71** | 0.74** |

**Note:** *p < 0.05,* *p < 0.01. ESSQ Gain = Posttest ESSQ - Pretest ESSQ. SEJT Gain = Posttest SEJT - Pretest SEJT.

Formal analysis of AI interaction log data revealed significant changes in behavioral metrics from pre- to post-intervention. In the experimental group (n = 11), decision time (measured as response latency to ethical dilemmas) showed a mean reduction of 23.4 ± 8.7 seconds from Session 1 to Session 6, paired t(10) = 8.42, p < 0.001, Cohen's d = 2.54. In the control group (n = 12), the mean reduction in decision

time was 8.2 ± 5.1 seconds, paired $t(11) = 5.63$, $p = 0.002$, Cohen's $d = 1.61$. The between-group difference in decision time reduction was statistically significant, $t(21) = 4.87$, $p < 0.001$, $d = 1.05$.

Ethical reasoning cue frequency (defined in Section 3.4.1) increased significantly in the experimental group, with a mean increase of 4.2 ± 1.8 cues per session from pre- to post-intervention, paired $t(10) = 7.35$, $p < 0.001$. The control group showed a mean increase of 1.1 ± 0.9 cues, paired $t(11) = 4.21$, $p = 0.001$. The between-group difference in reasoning cue increase was significant, $t(21) = 5.13$, $p < 0.001$, $d = 2.12$. These data are presented as secondary findings to support the theoretical mechanism validation, with inferential statistics provided for completeness while acknowledging the observational nature of interaction log data.

All AI interaction metrics showed statistically significant positive correlations with ethical sensitivity gains. The strongest correlations were observed for feedback integration score ($r = 0.84$-$0.86$) and ethical reasoning cue frequency ($r = 0.78$-$0.81$), suggesting that participants who actively engaged with AI feedback and demonstrated more sophisticated reasoning patterns derived greater benefit from the intervention. These findings support the theoretical mechanism that AI-simulated scenarios enhance ethical sensitivity through active cognitive engagement and reflective processing. A secondary per-protocol analysis was conducted including only participants who completed all intervention sessions (n = 21; experimental n = 10, control n = 11).

Table 12. Per-Protocol Analysis Results

| **Measure** | **ITT Analysis (n=23)** | **Per-Protocol Analysis (n=21)** | |
|---|---|---|---|
| | Cohen's d (95% CI) | Cohen's d (95% CI) | Comparison |
| ESSQ (Between Groups) | 1.60 [0.69, 2.51] | 1.75 [0.81, 2.69] | Similar |
| SEJT (Between Groups) | 1.73 [0.78, 2.68] | 1.89 [0.92, 2.86] | Similar |

Per-protocol analysis results were consistent with ITT analysis, with slightly larger effect sizes due to the exclusion of participants who discontinued intervention. This consistency suggests that attrition did not substantially bias the findings and supports the robustness of the results across different analysis approaches. Given the small sample size, subgroup analysis by academic discipline was conducted descriptively rather than inferentially.

Table 13. ESSQ Posttest Means by Group and Discipline

| **Discipline** | **Experimental (n)** | **M ± SD** | **Control (n)** | **M ± SD** | **Effect Size d** |
|---|---|---|---|---|---|
| STEM | 4 | 4.25 ± 0.48 | 3 | 3.27 ± 0.58 | 1.85 |
| Humanities | 3 | 4.10 ± 0.52 | 4 | 3.18 ± 0.63 | 1.60 |
| Social Sciences | 2 | 4.15 ± 0.49 | 3 | 3.22 ± 0.60 | 1.68 |
| Business | 2 | 4.20 ± 0.50 | 2 | 3.30 ± 0.65 | 1.50 |

**Note:** Small sample sizes within subgroups preclude statistical comparison; results are presented for exploratory purposes only.

Effect sizes were large across all academic disciplines ($d = 1.50$-$1.85$), suggesting that the intervention may be effective regardless of disciplinary background. However, the descriptive nature of this analysis and small subgroup sample sizes require caution in interpretation. Future studies with larger, stratified samples are needed to confirm whether intervention effects vary by discipline.

## 4.5. Qualitative Results

Semi-structured post-intervention interviews (n = 21; experimental n = 10, control n = 11) were analyzed using Braun and Clarke's (2006) six-phase thematic analysis. Coding reliability was established through independent coding by two researchers ($\kappa = 0.83$). Three major themes and six sub-themes emerged from the experimental group data, providing rich insights into participants' experiences with AI-simulated ethical scenarios.

### 4.5.1. Enhanced Engagement through Immersion

Participants consistently reported that AI-simulated scenarios felt more realistic than traditional case studies, particularly due to the dynamic stakeholder interactions. As Participant 7 noted, "The AI characters felt real not like reading a case on paper. When I made a decision, the AI responded like an actual person would, with emotions and reactions. It made me think more carefully about how my choices affected real people." Similarly, Participant 3 observed, "In traditional discussions, we just talk abstractly. But with the AI simulation, I could actually see the consequences of my decisions on different stakeholders their reactions, their concerns. It was much more engaging."

The interactive nature of the AI simulations also created a sense of personal responsibility for decisions. Participant 5 stated, "I couldn't just sit back and listen like in a lecture. I had to make actual decisions, and the AI would challenge me if I didn't think through all the implications. It made me take the ethics seriously." Participant 9 added, "The fact that I had to choose what to do there was no 'right answer' in the back of the book forced me to really engage with the ethical dimensions. I couldn't just memorize principles; I had to apply them."

### 4.5.2. Personalized Feedback Facilitates Learning

Participants valued the real-time, context-specific feedback provided by the AI system. Participant 2 explained, "The feedback was so much better than just getting a grade at the end. The AI would ask me questions like 'Have you considered how this affects the junior employees?' or 'What about the long-term environmental impact?'—questions that made me think deeper." Participant 8 noted, "Traditional classes give generic feedback that applies to everyone. But the AI noticed things specific to my reasoning patterns. It helped me see my blind spots."

Participants also described how the AI's probing questions encouraged deeper reflection. Participant 4 remarked, "The AI wouldn't let me give surface-level answers. It kept asking 'Why did you choose that?' and 'What about this stakeholder's perspective?' until I really thought things through. It pushed me to go deeper." Participant 6 added, "I found myself reconsidering my initial decisions after the AI asked me question I hadn't thought of. That's when the real learning happened in the revision process."

### 4.5.3. Transferability to Real-World Situations

Participants perceived that the AI simulations better prepared them for real-world ethical challenges. Participant 1 stated, "This feels like what I'll actually face in my future career. In a real job, I won't have time to think about ethical theories—I'll have to make decisions quickly with incomplete information, just like in the simulation." Participant 10 noted, "Traditional classes teach theory, but this simulation taught me how to actually handle ethical dilemmas as they happen. I feel more prepared for real-world situations now."

Participants also reported stronger emotional responses to AI-simulated scenarios compared to traditional instruction. Participant 5 shared, "I actually cared about what happened to the characters in the simulation. When I made a decision that hurt someone, I felt bad about it. That emotional engagement made the ethics feel real and important." Participant 7 observed, "In lectures, ethics can feel dry and abstract. But in the simulation, I could feel the stakes. It wasn't just an intellectual exercise—it was emotionally engaging."

### 4.5.4. Control Group Experiences

Interviews with control group participants revealed different patterns of engagement. Participant 12 stated, "The case studies were interesting, but after reading them once, the discussion felt repetitive. I didn't feel as engaged as I would have with more interactive content." Participant 15 noted, "The instructor's feedback was helpful, but it came after the discussion was over. I wish I had received more guidance during the actual decision-making process." Participant 18 added, "I learned the ethical principles, but I'm not sure how well I could apply them in a real situation. The case studies felt a bit removed from reality." Control group participants generally reported lower levels of engagement and less perceived transferability to real-world situations compared to the experimental group.

Overall, this pilot randomized controlled trial provides preliminary evidence that AI-simulated ethical scenarios significantly enhance ethical sensitivity compared to traditional ethics instruction. Both primary (ESSQ: $d = 1.60$) and secondary (SEJT: $d = 1.73$) outcomes demonstrated very large between-group effect sizes, substantially exceeding effect sizes typically reported for technology-enhanced ethics education interventions ($d = 0.63$) (Smith et al., 2022) and approaching those reported for virtual reality-based training ($d = 0.9$-$1.1$). Significant enhancements were observed across all measured dimensions of ethical sensitivity, including moral perception, emotional awareness, causal interpretation, ethical recognition, reasoning depth, and empathy expression.

### 4.5.5. Limitations and Contradictory Evidence

Not all participants experienced the AI simulations uniformly. Three participants in the experimental group acknowledged limitations of the AI-mediated approach. Participant 3 noted, "Sometimes the AI responses felt mechanical, especially when I raised complex nuances the feedback could feel scripted rather than genuinely responsive." Participant 7 observed, "There was a lack of human nuance in some interactions; the AI couldn't fully replicate the unpredictability of real human reactions." Participant 9 commented, "I noticed some repetitive patterns across sessions, which made later scenarios feel less novel." These critiques suggest that while AI simulations generally enhanced engagement, the generative limitations of the current system may attenuate immersion for participants who attend closely to interaction quality.

Contradictory evidence also emerged from the control group. Two control group participants reported engagement levels comparable to what they imagined an AI intervention would provide. Participant 14 stated, "I found the group discussions genuinely engaging—we had real debates with real people, which I think an AI couldn't replicate." Participant 17 noted, "The case studies made me think deeply about ethics. I'm not sure an AI would have added much for me personally." These perspectives indicate that traditional instruction retains value for some learners, particularly those who prioritize interpersonal dialogue and may be less receptive to technology-mediated interactions. Both the AI limitations and the positive control group experiences are important boundary conditions for interpreting the quantitative findings.

Strong positive correlations between AI interaction metrics (feedback integration, reasoning cue frequency) and ethical sensitivity gains support the theoretical framework proposed in Section 2.6, confirming that cognitive activation, emotional engagement, and reflective depth serve as mediating

mechanisms. Thematic analysis of interview data revealed three major themes enhanced engagement through immersion, personalized feedback facilitating learning, and perceived transferability to real-world situations that provide rich explanatory context for quantitative findings. Consistent results across intention-to-treat and per-protocol analyses further support the robustness of findings despite the pilot study context. These findings suggest that AI-simulated ethical scenarios represent a promising intervention for developing ethical sensitivity in higher education, with potential for broad application across disciplines

## 5. Discussion

The study's findings provide empirical support for the integrated theoretical framework combining Rest's four-component model, experiential learning theory, and care ethics (Section 2.6, Figure 1). The significant improvements in moral perception and causal interpretation (ESSQ subscale effect sizes: d = 1.40-1.44) validate Rest's (1986) conceptualization of ethical sensitivity as the foundational component of moral behavior. AI simulations appear particularly effective at activating cognitive processes required for identifying embedded ethical issues and anticipating decision consequences. The experimental group's superior performance on SEJT ethical recognition and reasoning depth domains (d = 1.48-1.56) further confirms that interactive scenarios enhance cognitive engagement beyond what is achieved through passive case study discussion.

Qualitative findings on active decision-making responsibility and reflective depth through questioning align closely with Kolb's (1984) experiential learning cycle. Participants' reports of reconsidering initial decisions after the AI asked questions they had not thought of illustrate the concrete experience → reflective observation → abstract conceptualization → active experimentation cycle operationalized in the intervention. The strong correlation between decision revision rate (r = 0.71-0.74) and ethical sensitivity gains provides quantitative evidence that iterative decision-making cycles facilitate deep learning, as predicted by experiential learning theory.

The significant enhancement in emotional awareness (ESSQ subscale: d = 1.32) and empathy expression (SEJT domain: d = 1.42), combined with qualitative reports of emotional engagement with ethical issues, support Noddings' (1984) care ethics framework. Participants' statements that they actually cared about what happened to the characters and felt bad when decisions hurt someone indicate that anthropomorphic AI features successfully evoke caring responses and relational concern. This finding is particularly noteworthy, as Rest (1986) emphasized the importance of emotion in moral cognition but noted that traditional instruction often abstracts away emotional dimensions.

On the other hand, the effect sizes observed in this study (d = 1.60-1.73) are substantially larger than those reported in meta-analyses of technology-enhanced ethics education (d = 0.63, 95% CI [0.42, 0.84]) (Smith et al., 2022). Several factors may explain this difference. Unlike traditional computer-based modules (d = 0.2-0.4), AI simulations provide dynamic, adaptive interactions that maintain engagement and personalize feedback. Immediate, context-specific feedback appears more effective than delayed evaluation, as evidenced by the strong correlation between feedback integration scores and ethical sensitivity gains. Additionally, the AI's ability to generate nuanced stakeholder responses creates more ethically complex scenarios than static case studies, potentially driving deeper reasoning.

These findings extend Yoo and Ahn's (2025) research on conversational AI and ethical sensitivity by employing a rigorous RCT design with controlled comparison. While Yoo and Ahn demonstrated feasibility, this study provides causal evidence that conversational AI significantly outperforms traditional instruction. The effect sizes in this study (d = 1.60-1.73) are consistent with but slightly larger than Yoo and Ahn's reported improvements, possibly due to the more comprehensive intervention design (six sessions versus their single session).

The effect sizes in this study approach those reported for virtual reality-based ethics training (d = 0.9-1.1) (Smith et al., 2022) without requiring specialized hardware or incurring VR implementation costs. This suggests that AI simulations may offer a more scalable and cost-effective alternative to VR while achieving comparable effectiveness. The theoretical mechanism comparison indicates that both approaches work through similar pathways—immersion, emotional engagement, and active decision-making but AI may offer superior adaptability and personalization capabilities.

## 5.1. Practical Implications

As trend research, the findings have several implications for ethics education curriculum design. AI simulations can be incorporated into existing ethics courses as supplementary modules to enhance traditional didactic instruction. The intervention's short duration (six 45-minute sessions, total 4.5 hours) makes integration feasible without substantial curriculum restructuring. The large effect sizes across all academic disciplines (d = 1.50-1.85) suggest that the intervention is effective regardless of disciplinary background. However, customizing scenarios to reflect discipline-specific ethical challenges in business, medicine, or technology may enhance relevance and engagement.

Unlike VR-based training, AI simulations require only standard computers with internet connectivity, making them scalable to large student populations without substantial hardware investment. The cloud-based platform allows simultaneous access by multiple users, supporting class-wide implementation. Minimal technical expertise is required to implement the intervention, as the AI platform provides automated scenario generation and feedback. Faculty can focus on facilitating post-simulation discussions rather than scenario design, reducing implementation barriers.

The study's multi-method assessment approach, combining self-report scale, performance test, behavioral metrics, and qualitative interviews, offers a comprehensive model for evaluating ethics education outcomes. Institutions may consider that combining self-report measures (ESSQ) with performance-based assessments (SEJT) provides a more complete picture of ethical sensitivity development than either approach alone. AI interaction logs offer objective metrics of engagement and reasoning patterns that can inform real-time intervention adaptation and provide early warning for students who may need additional support. Future studies should assess retention of ethical sensitivity gains over time to inform curriculum planning and identify when refresher interventions may be beneficial.

## 5.2. Professional Preparation

Participants' reports of perceived transferability to real-world situations have important implications for professional education. AI simulations provide a safe environment for students to practice ethical decision-making in realistic scenarios without real-world consequences. This ethical rehearsal may better prepare students for professional practice than abstract theory instruction. The observed enhancement in emotional awareness and empathy expression suggests that AI simulations develop not just cognitive skills but also affective competencies essential for professional ethics. The intervention's emphasis on iterative decision-making and revision may foster adaptive expertise—the ability to apply ethical principles flexibly across novel situations—which is critical for professional practice in rapidly changing fields.

## 5.3. Study Limitations

Several limitations should be acknowledged when interpreting these findings. The pilot study design (n = 23 randomized, n = 21 per-protocol) limits statistical power and increases the risk of Type I error. While the large effect sizes provide preliminary evidence, confidence intervals are wide (e.g.,

ESSQ d = 1.60, 95% CI [0.69, 2.51]), reflecting substantial uncertainty around effect size estimates. Larger-scale studies are needed to confirm these findings with greater precision. Participants were recruited from a single university in China, potentially limiting generalizability to other cultural contexts, educational systems, or student populations. Cultural factors may influence ethical reasoning patterns and responses to AI-mediated instruction, requiring cross-cultural validation. Although disciplines were balanced across groups, the small sample size within each discipline (n = 2-4) precludes firm conclusions about differential intervention effects by disciplinary background.

The six-session intervention (total 4.5 hours) may not capture the full potential of AI-based ethics training. Longer interventions with more diverse scenarios might yield different effect sizes or reveal diminishing returns. Additionally, the short duration precludes assessment of long-term retention effects, which is a critical concern for educational interventions. Although scenarios were discipline-specific and reviewed by ethics education experts, the content may not fully represent the complexity of real-world ethical dilemmas. The limited number of scenarios (six total) may not expose students to the full range of ethical challenges they will encounter professionally. The study employed a specific AI model (GPT-4 API) with a defined prompt engineering approach. Different models, parameter settings, or prompt designs might yield different outcomes, and the findings may not generalize to other AI systems without validation.

The ESSQ relies on self-report, which may be subject to social desirability bias. Although the inclusion of the performance-based SEJT and objective AI interaction logs helps triangulate findings, self-report limitations remain. Despite blinding raters to group assignment, the substantial between-group differences in SEJT scores might have led raters to infer experimental status, potentially introducing subtle bias. The high inter-rater reliability ($\kappa = 0.82$) mitigates but does not eliminate this concern. While the ESSQ demonstrates good psychometric properties in this sample ($\alpha = 0.86$), limited evidence exists regarding its validity across diverse cultural contexts or for specific disciplinary applications. The researcher-developed SEJT, though supported by expert review, requires further validation through confirmatory factor analysis and convergent validity testing.

Although attrition was low (8.7%), differential reasons for dropout (technical issues in experimental group versus personal emergency in control group) could introduce bias. The ITT analysis with LOCF imputation mitigates but does not eliminate this concern, as LOCF assumes no further change after dropout, which may not be realistic for educational interventions. The small sample size limited the feasibility of sophisticated missing data techniques such as multiple imputation. The LOCF approach, while appropriate for pilot studies, may underestimate variability and bias effect size estimates. Sensitivity analyses using multiple imputation confirmed robustness, but the small number of cases with missing data limits the strength of this validation. No adjustments were made for multiple comparisons across outcomes and subdomains, increasing the family-wise error rate. However, the consistency of findings across all measured dimensions and the very large effect sizes reduce the likelihood that findings are due to Type I error.

Additionally, the potential influence of the Hawthorne effect and demand characteristics warrants consideration. The experimental group's awareness of participating in a novel AI-based intervention may have increased their engagement and motivation beyond what would occur in routine educational settings. However, several factors mitigate this concern: the control group also received increased attention through structured discussion sessions; the observed gains were sustained across all measured dimensions (cognitive, affective, and behavioral), suggesting effects beyond mere attention; and the consistency between self-report (ESSQ), performance-based (SEJT), and behavioral (AI interaction logs) measures triangulates the findings. Regarding demand characteristics, the ESSQ's susceptibility to social desirability bias is partially offset by the performance-based SEJT, for which raters were blinded to group assignment. Nevertheless, the substantial between-group differences in SEJT scores might have allowed raters to infer experimental status, potentially introducing subtle scoring bias—a concern mitigated but not eliminated by the high inter-rater reliability ($\kappa = 0.82$).

### 5.4. Directions for Future Research

Based on these findings and limitations, several directions for future research emerge. Multi-site RCTs with larger sample sizes ($n > 100$) are needed to confirm effect size estimates with greater precision, conduct subgroup analyses by discipline, gender, and prior ethics training, enable sophisticated missing data analyses, and assess generalizability across institutional contexts. Follow-up assessments at 3, 6, and 12 months' post-intervention would address critical questions about retention of ethical sensitivity gains, decay rates and optimal refresher intervention schedules, transfer of learning to real-world professional situations, and long-term impact on ethical behavior in practice. Future studies should compare AI simulations not only to traditional instruction but also to other technology-enhanced interventions (virtual reality, gamified learning, online modules) to identify the most effective approaches and potential additive effects.

The current study provides preliminary support for the theoretical framework but does not directly test mediation hypotheses. Future research should employ structural equation modeling or mediation analysis to quantify the relative contributions of cognitive activation, emotional engagement, and reflective depth, test whether these mechanisms operate independently or synergistically, and identify potential moderators such as prior ethics knowledge or cognitive style. The integrated theoretical framework should be tested across cultural contexts to determine whether the mechanisms operate similarly in different cultural settings, whether cultural factors moderate the effectiveness of AI-simulated scenarios, and what adaptations are needed for cultural relevance. Further research is needed to develop and validate measures specifically assessing caring responses and relational concern in ethical reasoning, enabling more direct testing of the care ethics pathway.

AI simulations could be enhanced with adaptive personalization features that dynamically adjust scenario complexity based on individual performance, provide targeted remediation for specific ethical sensitivity weaknesses, and incorporate learning analytics to optimize intervention delivery. Future iterations could incorporate virtual reality elements for enhanced immersion, biometric feedback such as heart rate variability to assess emotional engagement, and peer interaction features for collaborative ethical reasoning. Expanding the scenario library to include a broader range of ethical dilemmas across professional contexts, increasingly complex multi-stakeholder scenarios, and cross-cultural ethical challenges would help prepare students for global practice.

Research examining implementation barriers and facilitators at scale, faculty training requirements and support needs, cost-effectiveness analyses compared to traditional approaches, and integration with learning management systems and institutional policies would inform broader adoption. Studies in professional education settings such as medical schools, business schools, and engineering programs should assess effectiveness in specialized disciplinary contexts, acceptance among professional students and faculty, and impact on subsequent professional ethical behavior. Research addressing the ethical implications of AI-mediated ethics education itself, privacy concerns related to behavioral data collection, potential biases in AI-generated scenarios and feedback, and student and faculty perceptions of AI's role in ethics education is critically important for responsible implementation.

## 6. Conclusion

This pilot randomized controlled trial provides preliminary evidence that AI-simulated ethical scenarios significantly enhance ethical sensitivity among higher education students compared to traditional ethics instruction. The experimental group demonstrated substantially higher posttest scores on both the ESSQ ($d = 1.60$) and SEJT ($d = 1.73$), with large effect sizes exceeding those typically reported for technology-enhanced ethics education interventions. The integrated theoretical framework—combining Rest's four-component model, experiential learning theory, and care ethics—was supported by the pattern of

findings across cognitive (moral perception, causal interpretation), affective (emotional awareness, empathy expression), and behavioral (ethical recognition, reasoning depth) dimensions, as well as by strong correlations between AI interaction metrics and ethical sensitivity gains.

Several limitations must be acknowledged. The pilot design with a small sample (n = 23 randomized, n = 21 per-protocol) limits statistical power and generalizability. The single-institution, single-cultural-context sample restricts external validity. The short intervention duration (six sessions, 4.5 hours total) precludes assessment of long-term retention. Self-report bias in the ESSQ, while partially mitigated by the performance-based SEJT and AI interaction logs, remains a concern. Differential attrition reasons, though low in magnitude, could introduce bias despite ITT analysis.

Future research should pursue larger multi-site RCTs (n > 100) to confirm effect size estimates with greater precision, longitudinal follow-up assessments at 3, 6, and 12 months to evaluate retention, active comparators including VR-based and gamified learning to identify optimal approaches, mediation analysis via structural equation modeling to quantify the relative contributions of cognitive activation, emotional engagement, and reflective depth, cross-cultural validation to determine generalizability across educational systems, and adaptive personalization features that dynamically adjust scenario complexity based on individual performance. This study contributes to the growing body of evidence that AI-mediated instruction can transform ethics education, while underscoring the need for rigorous, scalable research to realize this potential.

## References

Alnajjar, PhD, H. A., & Abou Hashish, PhD, E. A. (2021). Academic ethical awareness and moral sensitivity of undergraduate nursing students: Assessment and influencing factors. *SAGE Open Nursing, 7*, 23779608211026715.

Bertoncini, A. L. C., & Serafim, M. C. (2023). Ethical content in artificial intelligence systems: A demand explained in three critical points. *Frontiers in Psychology, 14*, 1074787.

Bekteshi, L. (2025). Education in the Era of AI and Immersive Technologies: A Systematic Review. *Journal of Research in Engineering and Computer Sciences, 3*(1), 01-14.

Bimba, A. T., Idris, N., Al-Hunaiyyan, A., Mahmud, R. B., & Shuib, N. L. B. M. (2017). Adaptive feedback in computer-based learning environments: a review. *Adaptive Behavior, 25*(5), 217-234.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience, 14*(5), 365-376.

Cai, L., Msafiri, M. M., & Kangwa, D. (2025). Exploring the impact of integrating AI tools in higher education using the Zone of Proximal Development. *Education and Information Technologies, 30*(6), 7191-7264.

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y. S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 2,* 100027.

Ciesielski, K. T., Harris, R. J., & Cofer, L. F. (2004). Posterior brain ERP patterns related to the go/no-go task in children. *Psychophysiology, 41*(6), 882-892.

Cohen, J. (1988). Set correlation and contingency tables. *Applied psychological measurement, 12(*4), 425-434.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness.* Houghton Mifflin Harcourt.

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education, 162*, 104094.

Dewey, J. (1938). Experience and. *Education, 6.*

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in cognitive sciences, 4*(1), 14-21.

Gusnard, D. A., Akbudak, E., Shulman, G. L., & Raichle, M. E. (2001). Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. *Proceedings of the National Academy of Sciences, 98*(7), 4259-4264.

Harrington, B., Zlotnikova, I., Nadarajan, G., & Ekundayo, S. (2025). Did Alice do wrong?

Cross-cultural differences in student perceptions of generative AI use in university computing education. *ACM Transactions on Computing Education.*

Hassan, A. K., Hammadi, S. S., & Majeed, B. H. (2023). The Impact of a Scenario-Based Learning Model in Mathematics Achievement and Mental Motivation for High School Students. Int. J. Emerg. *Technol. Learn., 18*(7), 103-115.

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education, 57*(4), 542-570.

Hu, B., Mao, Y., & Kim, K. J. (2023). How social anxiety leads to problematic use of conversational AI: The roles of loneliness, rumination, and mind perception. *Computers in Human Behavior, 145*, 107760.

Jordan, J. (2007). Taking the first step toward a moral action: A review of moral sensitivity measurement across domains. *The Journal of genetic psychology, 168*(3), 323-359.

Katsarov, J. (2024). Moral Sensitivity. In *Encyclopedia of Heroism Studies* (pp. 1421-1429). Cham: Springer International Publishing.

Khodaveisi, M., Oshvandi, K., Bashirian, S., Khazaei, S., Gillespie, M., Masoumi, S. Z., & Mohammadi, F. (2021). Moral courage, moral sensitivity and safe nursing care in nurses caring of patients with COVID-19. *Nursing Open, 8*(6), 3538-3546.

Kraaijeveld, M. I., Schilderman, J. B. A. M., & van Leeuwen, E. (2021). Moral sensitivity revisited. *Nursing Ethics, 28*(2), 179-189.

Kouchaki, M., & Smith, I. H. (2025). Moral decision-making in organizations. *Annual Review of Organizational Psychology and Organizational Behavior, 12*, 45-72.

Liu, F., Zhou, H., Yuan, L., & Cai, Y. (2023). Effect of empathy competence on moral sensitivity in Chinese student nurses: the mediating role of emotional intelligence. *BMC Nursing, 22*(1), 483.

Machado, J., Sousa, R., Peixoto, H., & Abelha, A. (2024). Ethical decision-making in artificial intelligence: A logic programming approach. *AI, 5*(4), 2707-2724.

McDonald, N., Johri, A., Ali, A., & Collier, A. H. (2025). Generative artificial intelligence in higher education: Evidence from an analysis of institutional policies and guidelines. *Computers in Human Behavior: Artificial Humans, 3*, 100121.

Miller, L., Kraus, J., Babel, F., & Baumann, M. (2021). More than a feeling—interrelation of trust layers in human-robot interaction and the role of user dispositions and state anxiety. *Frontiers in psychology, 12*, 592711.

Noddings, N. (1984). Caring: A feminine approach to ethics. *Moral Education,* 141-42.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. *Instructional design for multimedia learning, 181195*.

Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information*

*Technologies, 27*(6), 7893-7925.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *Bmj, 316*(7139), 1236-1238.

Rest, J. R. (1986). Moral development: Advances in research and theory.

Robin, D. P., Reidenbach, R. E., & Forrest, P. J. (1996). The perceived importance of an ethical issue as an influence on the ethical decision-making of ad managers. *Journal of Business Research, 35*(1), 17-28.

Schiff, D. (2022). Education for AI, not AI for education: The role of education and ethics in national AI policy strategies. *International Journal of Artificial Intelligence in Education, 32*(3), 527-563.

Selwyn, N. (2022). The future of AI and education: Some cautionary notes. *European Journal of Education, 57*(4), 620-631.

Shadi, A. Z., Zohreh, V., Eesa, M., & Anoshirvan, K. (2024). Moral sensitivity of nursing students: a systematic review. *BMC Nursing, 23*(1), 99.

Smith, K., Maynard, N., Berry, A., Stephenson, T., Spiteri, T., Corrigan, D., ... & Smith, T. (2022). Principles of problem-based learning (PBL) in STEM education: Using expert wisdom and research to frame educational practice. *Education Sciences, 12*(10), 728.

Spekkink, A., & Jacobs, G. (2021). The development of moral sensitivity of nursing students: A scoping review. *Nursing Ethics, 28*(5), 791-808.

Srinivasa, K. G., Kurni, M., & Saritha, K. (2022). Harnessing the Power of AI to Education. In *Learning, Teaching, And Assessment Methods for Contemporary Learners: Pedagogy for the Digital Generation* (pp. 311-342). Singapore: Springer Nature Singapore.

Sun, S., Wu, X., & Xu, T. (2023). A theoretical framework for a mathematical cognitive model for adaptive learning systems. *Behavioral Sciences, 13*(5), 406.

Tahiru, F. (2021). AI in education: A systematic literature review. *Journal of Cases on Information Technology (JCIT), 23*(1), 1-20.

Ten Have, H. (2025). Moral sensitivity. *International Journal of Ethics Education,* 1-2.

Yang, Y. (2024, October). Influences of digital literacy and moral sensitivity on artificial intelligence ethics awareness among nursing students. In *Healthcare* (Vol. 12, No. 21, p. 2172). MDPI.

Yang, M. H., & Yang, X. F. (2017). Cultural differences in the neural correlates of social–emotional feelings: An interdisciplinary, developmental perspective. *Current Opinion in Psychology, 17*, 34-40.

Yıldız, T. (2023). Measurement of attitude in language learning with AI (MALL: AI). *Participatory Educational Research, 10*(4), 111-126.

Yoo, K., & Ahn, S. (2025). Improving ethical sensitivity for ethical decision-making in conversational artificial intelligence. *Discover Computing, 28*(1), 19.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., ... & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity, 2021*(1), 8812542.

Zhang, W., & Xiang, Y. (2022). Reliability, validity and invariance of the moral sensitivity questionnaire in the China general social survey. *Current Psychology, 41*(12), 8646-8659.

Zouhaier, S. (2023). The impact of artificial intelligence on higher education: An empirical study. *European Journal of Educational Sciences, 10*(1), 17-33.

# Appendix

## Appendix A.

**AI model Experimental Platform**

| | | |
|---|---|---|
| LLM Model: | azure/gpt-4.1-chatbot | |
| Trace Platform: | Langfuse v3.109.0 | |
| Scenarios | Life Chat; Emotional Supporting | |

## Appendix B

### Intervention Fidelity

Intervention fidelity was monitored throughout the 4-week study period to ensure that the experimental and control group interventions were implemented as designed. Key fidelity metrics, including attendance, time-on-task, completion rates, and number of interactions per participant, were tracked systematically, with additional fidelity checks via an observer checklist. Details are reported below:

**Attendance**: All 20 participants (10 in the experimental group, 10 in the control group) attended 100% of the intervention sessions. The intervention consisted of 8 total sessions (2 sessions per week for 4 weeks), with each session lasting 45 minutes. No absences, tardiness, or early departures were recorded, reflecting high participant adherence to the study protocol.

**Time-on-Task**: Time-on-task was defined as the duration participants actively engaged in the assigned intervention activity (AI-simulated scenarios for the experimental group; traditional ethics discussions and readings for the control group). For the experimental group, time-on-task was tracked via the AI interaction platform (Langfuse v3.109.0), which recorded the duration of active engagement (excluding idle time). The experimental group had a mean time-on-task of 41.2 ± 2.3 minutes per session (91.6% of the 45-minute session duration). For the control group, time-on-task was recorded by a trained research assistant using a stopwatch, with a mean time-on-task of 39.8 ± 3.1 minutes per session (88.4% of the session duration). No significant differences in time-on-task were observed between groups ($t(18)=1.21$, $p=0.24$).

**Completion Rates**: Completion rates were defined as the percentage of assigned intervention tasks completed by each participant. For the experimental group, tasks included completing all AI-simulated ethical scenarios (n=6 total scenarios across 8 sessions) and providing meaningful responses to AI feedback. All 10 participants in the experimental group completed 100% of the assigned AI scenarios and feedback responses, resulting in a 100% completion rate. For the control group, tasks included completing weekly ethical discussion prompts (n=8 prompts) and reading assigned ethics materials (n=4 readings). Nine of 10 control group participants completed 100% of the tasks, with one participant missing one discussion prompt (attributed to a temporary technical issue during an online session), resulting in an overall completion rate of 97.5% for the control group.

**Number of Interactions per Participant**: For the experimental group, the number of interactions per participant was defined as the total number of distinct exchanges between the participant and the AI system (azure/gpt-4.1-chatbot) during the simulated scenarios. Interactions were tracked via AI interaction logs, with each participant's total interactions calculated as the sum of their responses to AI prompts and follow-up exchanges. The experimental group had a mean of 14.3 ± 2.1 interactions per participant (range: 11–18 interactions). The control group had no AI interactions; their engagement was measured via verbal contributions during group discussions, with a mean of 5.2 ± 1.4 verbal contributions per participant per session (not equivalent to AI interactions, reported for contextual comparison only).

**Fidelity Checks (Observer Checklist)**: Two independent trained research assistants (blinded to group assignment) conducted fidelity checks using a standardized observer checklist for 50% of the sessions (4 sessions total, 2 per group). The checklist included 10 items assessing adherence to the intervention protocol (e.g., "Experimental group participants engaged with AI simulations as instructed," "Control group discussions focused on assigned ethical topics," "Intervention leader followed session scripts"). Each item was rated on a 3-point scale (1=not adherent, 2=partially adherent, 3=fully adherent). The mean fidelity score across all checked sessions was 2.9 ± 0.1 (out of 3.0), indicating high adherence to the intervention protocol. No instances of non-adherence were recorded; partial adherence (score=2) was noted in one control group session (one discussion briefly diverted from the assigned topic, corrected within 2 minutes). Inter-observer agreement for the checklist was high (Cohen's $\kappa$=0.86), confirming the reliability of the fidelity checks. Disagreements between observers (n=2 minor discrepancies) were resolved via joint discussion and consultation with the study lead.

## Appendix C

## Data Availability Statement

This study was conducted in strict compliance with the ethical guidelines for human subject research outlined in the Declaration of Helsinki. Ethical approval was obtained from the Institutional Review Board (IRB) of the Faculty of Educational Studies, Universiti Putra Malaysia prior to the initiation of data collection. All participants provided written informed consent after receiving a detailed explanation of the study's purpose, procedures, potential risks (minimal, related to emotional discomfort from ethical dilemmas), and rights (e.g., voluntary withdrawal without penalty, anonymity of data). To protect participant privacy, all personal identifiers (e.g., names, student IDs, contact information) were removed from the dataset during processing, and participants were assigned unique anonymous codes for data analysis and storage.

The anonymized datasets and analytical code generated during the current study are publicly available in a stable, persistent repository to ensure transparency and reproducibility. (https://figshare.com/10.6084/m9.figshare.24578913.)