Introducing Conceptual Framework of Gamified Moral Reasoning Evaluation and Testing for Human Agentic-AI Co-Alignment

Tauseef ur Rehman, Arham Muslim, Tahira Anwar Lashari, Muhammad Ashraf, Muhammad Ajmal Khan
School of Electrical Engineering and Computer Sciences (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan
*Correspondence: tauseef.rehman@seecs.edu.pk

Citation | Rehman. T. U, Muslim. A, Lashari. T. A, Ashraf. M, Khan. M. A, "Introducing Conceptual Framework of Gamified Moral Reasoning Evaluation and Testing for Human Agentic-AI Co-Alignment", IJIST, Special Issue pp 740-756, May 2026
Received | April 10, 2026 **Revised |** May 16, 2026 **Accepted |** May 18, 2026 **Published |** May 20, 2026.

Current methods for assessing human moral reasoning, based on neo-Kohlbergian theory (e.g., DIT-2), offer valuable psychometric insight but remain static, decontextualized, and limited in their ability to capture the evolution of reasoning abilities along a continuum. Agentic AI alignment and safety frameworks are also evolving, but remain primarily compliance-driven and evaluate adherence to ethical principles without modeling the underlying processes of moral judgment and adaptation. This paper adopts a design-oriented integrated methodology, where interdisciplinary literature is synthesized to propose a conceptual Gamified Moral Reasoning Evaluation Framework that addresses these issues by integrating classical and neo-Kohlbergian theories of moral development, game-theoretic reasoning models, and AI-driven adaptive scenario generation. The framework comprises five layers: theoretical foundations, gamification engine, adaptive scenario generation using LLMs, dual human–AI evaluation modules connected through a co-alignment bridge, and continuous learning mechanisms, forming a dynamic ecosystem for assessing and nurturing moral reasoning in both humans and AI systems. The framework enables multi-dimensional assessment through measurable indicators, including decision patterns, response latency, and schema progression across adaptive scenarios, and convergences/divergences in human–AI moral reasoning. These dimensions support longitudinal analysis of moral reasoning evolution, extending beyond static instruments such as DIT-2. A preliminary technology readiness assessment indicates that core enabling components, such as LLMs, game engines, and adaptive learning architectures, currently operate at TRLs 6–8, enabling near-term prototyping and validation. The paper offers a practically grounded pathway toward evidence-based tools for studying and fostering human–AI moral co-alignment in ethically complex and socially consequential domains.

Keywords: Moral Reasoning; Gamification; AI Ethics; Human–AI Alignment; Ethical Decision Modeling.

Introduction:

Artificial intelligence (AI) has continued to reshape human interaction, learning, and decision-making. While it promises efficiency and personalization, its influence also extends into the moral domain, where choices with ethical implications are being made through algorithms [1]. Despite several policy and regulatory frameworks, most AI systems remain rigidly rule-based, informed by fixed procedural rules that are unable to capture the nuances of human moral reasoning and contextual judgment [2]. Such limitations have raised concerns regarding trust, privacy, and ethical integrity, seen through the psychological and societal consequences of biased algorithms [3] and deepfake technologies [4]. Most of the existing ethical evaluation frameworks are static and compliance-driven with limited user participation or reflective reasoning mechanisms [5][6].

Therefore, there is an emerging demand for approaches that are engaging, adaptive, and allow for moral awareness, continuous engagement, and measurable learning outcomes. Gamification seems to be a promising approach, embedding feedback, progression, and scenario-based reasoning into ethical assessment processes [7][8]. Within cognitive and moral reasoning models, gamification could turn abstract ethical evaluation into experiential learning by stimulating users to commit to dilemmas, reflect on their choices, and reveal patterns of reasoning underlying their decisions [9]. Growing evidence of the presence of AI in nearly all aspects of everyday life has exacerbated the ethical dilemma of emotional dependence on intelligent systems [10]. As the new waves of AI technologies started to provide empathetic responses and companionship, based on emotional resonance, boundaries have increasingly blurred between human–human and human–machine interactions [11].

This blurring of emotional signals creates the potential for psychological and emotional dependencies, vividly demonstrated in the cases of both AI companions and therapeutic robots. Replika, a popular AI chatbot that users engage with in emotionally intimate conversations [12], has demonstrated how many users become attached to such systems and seek comfort and reassurance from them. Recent empirical studies confirm that users can develop perceived emotional bonds and companionship with conversational agents over time. While the interactions may help alleviate loneliness or distress, they also raise deeper concerns regarding the substitution of authentic human relationships with artificial ones [13].

Similar ethical nuances are also found in the use of socially assistive robots such as Paro, the therapeutic robotic seal. It was created to provide comfort to older individuals. Paro mimics the actions of real pets, responding to touch, sound, and human attention, to soothe those dealing with dementia, loneliness, or depression. Studies have acknowledged its positive emotional effects and ability to lessen negative feelings [14][15]. Despite these innovations, they raise philosophical and moral questions about whether humanlike emotional AI redefines learning, empathy, and social bonding. They challenge long-held notions of independence, personal honesty, and the moral costs of outsourcing care, friendship, and emotional understanding to machines [16].

This paper aims to develop a structured and theoretically grounded framework for evaluating and fostering moral reasoning in the context of human–AI interaction. Specifically, it seeks to integrate insights from moral psychology, AI ethics, gamification, and game-theoretic reasoning into a unified system that enables continuous and adaptive assessment of moral reasoning. In doing so, the framework addresses key limitations of existing approaches, including the static nature of traditional assessments, limited contextual adaptability, and the absence of structured human–AI comparative evaluation. These challenges are addressed through the incorporation of adaptive scenario generation, gamified interaction mechanisms, and dual evaluation modules within a unified design. The proposed framework conceptualizes moral reasoning as an evolving and measurable process, supported through interactive and scenario-driven gamified evaluation mechanisms. In addition, the study incorporates a

preliminary feasibility assessment based on Technology Readiness Levels (TRLs) to evaluate the practical viability of the proposed framework.

The novelty of this work lies in the integration of interdisciplinary theoretical foundations into a single, multi-layered architecture that combines gamified interaction, AI-driven adaptive scenario generation, and dual human–AI evaluation modules. Unlike traditional static assessment tools, the framework enables longitudinal analysis of moral reasoning and facilitates direct comparison between human and AI decision-making processes within dynamic and context-sensitive environments.

The rest of this article discusses relevant literature and research gaps in section 2, research methodology is provided in section 3, and a proposed conceptual framework is presented in section 4.

Background and Related Work:

In this section, we review the theoretical and empirical foundations that support the integration of ethical reasoning, gamification, and scenario generation. We synthesize key perspectives from moral psychology, AI ethics, and game-based evaluation to establish a coherent conceptual foundation for designing a unified framework that assesses moral reasoning and fosters human–AI alignment. The literature included in this study was selected through a structured but non-systematic review process, focusing on recent and relevant contributions. The majority of the cited literature consists of publications from the last five years (2020–2025), ensuring alignment with recent advances in AI ethics, adaptive systems, and gamified learning, while foundational works are included to support theoretical grounding.

Ethical and Reasoning Models:

This section reviews global and national frameworks of ethical reasoning in AI. Internationally, we examined the EU AI Act, Digital Services Act, Council of Europe Framework on AI, and UNESCO Recommendation on AI Ethics. From a national perspective, we reviewed Pakistan's Prevention of Electronic Crimes Act (PECA) 2016 and its 2025 amendment, highlighting the evolving emphasis on ethical regulation, transparency, and responsible AI practices.

EU Artificial Intelligence Act (Regulation (EU) 2024/1689):

It is the European Union's landmark legislation. The world's first widespread AI law aimed at regulating artificial intelligence (AI) to ensure its safe, ethical, and human-centric deployment across the EU [17]. It is a risk-based approach that prohibits some AI practices and places obligations and transparency on high-risk AI providers [18]. It implies strict law practices with penalties/enforcement to limit exploitative AI applications. It also establishes definite lines of responsibility. For instance, if a deleterious AI system causes harassment (e.g., deep-fake production, exploitative recommender systems), blame cannot rest solely with the user who abuses the instrument, but also with developers/providers who did not meet safety and transparency standards. The limitation of this enforcement lies in high-speed generative/abusive applications. Recent studies highlight that many AI laws, like the EU AI Act, have phased, delayed enforcement, and struggled with fast AI innovation cycles.

EU Digital Services Act (DSA):

The EU Digital Services Act (DSA) provides a framework for online intermediary services in the EU, adopted in 2022 [19]. It is a platform- liability regime for online intermediary platforms, including large online platforms (very large). This platform ensures that the shared online content does not violate the law and provides adequate services to the users. This digital service act takes three areas into consideration, namely content management, transparency and accountability, and user protection. Content management refers to monitoring illegal content, mechanisms to limit its rapid digital dissemination, and reporting and regulation of systems that facilitate the spread of harmful content. The task of transparency and accountability is scrutinizing platform algorithms and recommendation

systems to understand how they may contribute to the dissemination of illegal content. Lastly, user protection ensures that users have the right to lodge complaints against such platforms and seek appropriate redress in cases of harm or regulatory violations.

Council of Europe Framework Convention of National Minorities (FCNM) on AI:

It is an international treaty of minorities working voluntarily together to serve humanity and protect the rights of people. This convention ensures freedom of expression regardless of cultural, religious, language, and other diversities. Hence, it applies to the whole AI lifecycle to strike a balance between innovation and the protection of societal lives. Even the emerging approach of conventional AI tries to protect people's safety and dignity; still, there are no clear rules on how these rights will be enforced. This European council and committee of ministries monitor the involvement of statutory bodies, but it lacks the application of fines on the accused entity. This creates confusion about who is responsible for protecting dignity, which can weaken accountability.

UNESCO Recommendation on the Ethics of AI (2021):

It is a first global framework that has been adopted by 194 member states, including major nations such as the United States, France, Germany, China, India, Japan, the United Kingdom, Brazil, and Pakistan, among many others [20]. This framework aims to ensure that AI technologies promote human well-being, fairness, and sustainable development while preventing harm, bias, or discrimination. This framework develops the first international normative ethical standardized approach to maintain human dignity, human rights, transparency, justice, and human oversight.

The Prevention of Electronic Crimes Act (PECA) 2016:

It is a law that was approved by the Government of Pakistan to address crimes committed through electronic systems, including the internet and digital devices [21]. The purpose of the act is to protect citizens (particularly women and young users) from cybercrimes such as hacking, identity theft, cyberstalking, hate speech, and the spread of false or harmful information. Through this act, at the provincial level, the data is given protection from unauthorized access. Yet, some provinces' boundaries are vague, and borders that are not protected with this act lead to misuse or overreach, particularly where there is a lack of protection towards freedom of speech. In addition, the act is prone to enhance awareness and rehabilitation more than providing legal punishments.

Despite their regulatory significance, these frameworks remain predominantly compliance-oriented and provide limited support for modeling dynamic, context-sensitive moral reasoning processes in interactive environments.

Advancing Moral Cognition: Integrating Classical and Neo-Kohlbergian Perspectives:

Lawrence Kohlberg's theory of moral development proposed a stage-based model revealing how moral reasoning evolves through three levels [22]. The underpinning concept of this theory is mapped on justice, fairness, and rational reasoning. Classical Kohlbergians revealed that moral development is a universal phenomenon across cultures. The primary focus is on how individuals reason through moral dilemmas rather than on the decision outcome itself. A series of three levels of moral reasoning posits six further stages [two in each level], namely the pre-conventional level aged 3–7 years, conventional aged 8-13, and post-conventional aged adulthood and onwards (See Figure 1). Each level individual makes moral decisions on different factors, including avoiding punishment, following laws, and following universal ethical principles. The moral reasoning of preconvention is primarily based on external consequences rather than as a result of internalized values. Pre-conventional level consists of two stages: obedience and punishment orientation [avoiding principle], and self-interest orientation [personal]. The moral reasoning of the conventional stage is guided by the process of socialization, social expectation, social interaction, laws, and order. This level is further divided into two stages: interpersonal accord and conformity, revolving around

meeting social expectations. The second stage refers to authority and maintains social order. Individuals at this stage prefer to obey laws and maintain social order. The moral reasoning of the post-conventional level is guided by higher ethical principles and individual rights beyond societal laws. It comprises dual stages; firstly, social contract and individual rights [prioritizing human life despite legal consequences], and secondly, universal ethical principles [recognizing that the greater good may justify personal risks].

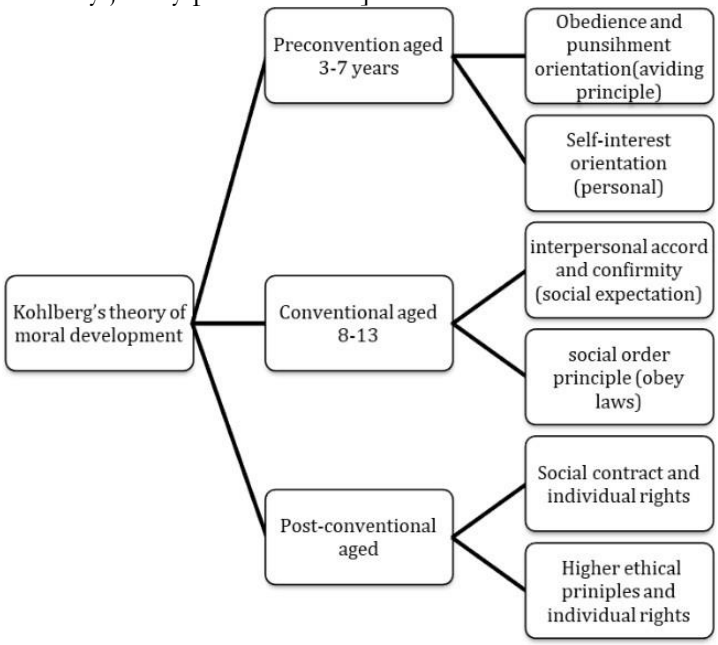

Figure 1. AI framework builds on the progressive development of neo-Kohlberg's stages of moral development.

Classical Kohlbergian stages of moral reasoning are an open-ended approach that measures moral dilemma through interviews, which is a qualitative method. Due to subjective interpretation, limited cultural generalization, and time-consuming nature, this method faced major criticism on internal validity. As a consequence, the Neo Kohlbergian theory, an updated model, replaces rigid sequential stages with more flexible moral schemas. James Rest and his colleagues revised the theory [23]. They reconceptualized it with flexible patterns of reasoning harmonized within individuals. Thus, a dynamic overlap-based schema approach was proposed, rather than the discrete stages of moral reasoning. Yet, it retained the theoretical developmental foundations by means of a recognized context-sensitive method. It was a quantitative self-report method using the Defining Issues Test (DIT). Nevertheless, an updated dilemma introduced the N2 index; a modernized DIT into DIT-2. It is a more reliable, theoretically grounded, cross-culturally validated, empirical-based, measurable framework [24]. In this model, an individual is involved in three schemata [framework] but may not follow the fixed levels. The three schemata are personal interest schema [linked with personal interest], the maintaining norms schema [socialization process such as order, laws, duties], and the Postconventional schema [concerned with abstract principles, justice, rights].

The Defining Issues Test 2 (DIT-2) is a standardized tool that measures the neo-Kohlbergian framework. It facilitates continuous evaluation and provides a useful foundation for digital or gamified systems. In this tool, respondents are presented with scenarios where they are asked to rate and rank statements related to different schema-aligned morals. Empirical studies reveal that moral schema tends to increase and mature with age, education, exposure, and reflective engagement, while recent work in AI ethics further highlights the need to move beyond static, principle-based evaluations toward more context-sensitive and behavior-oriented models of moral reasoning [25][26][27].

Gamification:

The term "gamification" is defined in Merriam-Webster dictionary as "the process of adding games or game-like elements to something (such as a task) so as to encourage participation". Even though this term is relatively new, the base concept of gamification has been around for a long time. It employs behavioral psychology and motivational theory to design interactive environments, where users receive continuous feedback and are rewarded based on their progression through goal-oriented challenges. Humans possess a natural desire for competition and achievement, and gamification taps into these desires to enhance motivation and sustain engagement during task performance, ultimately leading to behavioral changes [28][29]

Gamification has been applied in many different contexts, though with varying degrees of success. Recent empirical syntheses further confirm that gamification improves engagement and learning outcomes when appropriately designed [30]. In educational settings, for instance, several studies report gains in learner motivation, persistence, and sometimes even performance [31]. Related approaches have also been used in health, business, and civic participation initiatives, where game-like elements are found to encourage engagement; however, the outcomes depend heavily on design decisions and the surrounding social context [32].

Gamification Types: Gamification approaches are often discussed under two broad categories: structural gamification and content gamification. Structural gamification focuses on integrating game-like features around the content, and the content itself remains unchanged. These features include points, badges, progress bars, and leaderboards [33]. Content gamification, on the other hand, goes a step further by enhancing the content itself and embedding game-like elements (storylines, challenges, role-play scenarios, etc.) in it to feel more playful and narrative-driven [34]. For example, embedding an interactive quiz in an online lesson or structuring lessons as missions, where completing one mission opens the next level.

Research has shown that gamification is effective when coupled with learners' basic psychological needs, such as autonomy, competence, and relatedness [35]. Practically, it means that if students are given choices to progress, their achievements are recognized, and they can relate the scenarios to their personal experiences, then they will be more interested in engaging with the content and learn better [36].

Reasoning Models Driven Gamification: Reasoning models provide insight into how people perceive stimuli, compare options, and make choices and decisions. Gamification, in turn, provides an environment in which we can see a person progressing through different activities. Integrating these two elements will provide insight into how people reason and the real-time unfolding of their reasoning through the application of thoughtfully constructed activities.

Research in this area reveals that game elements such as feedback loops and other components within the realm of game design do more than reinforce effort. They also shape the players' decisions, the resulting outcomes, and how they revise the strategy toward a goal. Furthermore, in a gamified environment, we can study the behavior of players as they progress through scenarios using feedback loops [37].

Building on this, [38] pointed out that integrating reasoning models with such frameworks provides a theoretical lens to understand how players learn and tackle issues of uncertainty, feedback, and goal-setting in a gamified environment. The authors further stated that adaptive gamified systems combined with reasoning models can effectively personalize challenges according to the learner's cognitive state. This forms a context in which reasoning is measurable, and the challenges are adjustable.

Gamification as a Tool for Assessing Moral Reasoning: In the previous section, we explored how the combination of gamification and reasoning models can reveal how individuals process feedback, alternatives, and adjust their decisions over time. But moral reasoning adds another layer of complexity, namely ethical reflection and value-based judgment. In this section, we will examine the literature focusing on using gamification as a tool to observe and assess moral reasoning within structured, interactive environments.

[39] Shows that through gamified tasks, participants' choices and actions become observable in real time, providing insights into cognitive and ethical processes. Another study builds upon this perspective and demonstrates how gamified simulations can present branching ethical scenarios and provide adaptive feedback. They found that such environments allow participants to experience the impact of their decisions in order to reflect on the consequences of their decisions and improve ethical decision-making. As participants in such environments are exposed to dynamic challenges, the data produced represents realistic thinking in the context of several settings. We can then conclude that gamification forms a pragmatic link between theoretical frameworks of moral reasoning and human and AI systems' ways of dealing with complex, uncertain scenarios.

AI-Based Ethical Scenario Generation:

Creating AI-generated scenarios that capture moral and cognitive reasoning and can adapt to changing ethical dilemmas is still a major challenge. One research work that addresses this problem is the Off The Rails benchmark [40]. It uses causal-graph-informed frameworks to generate structured moral dilemmas. These situations allow them to compare the human moral thinking patterns to the logic practices of large language models, incorporating differences in narrative structure and ethical framing, as well as the interpretation of their human moral judgments. It thus provides a more intensive and quantitative study of how moral decisions are made in humans as well as AI systems.

The related studies by [41] suggest a hierarchical evaluation structure in which complexity and tension between competing values in ethical dilemmas are gradually increased. This approach illustrates how moral agents, either human or artificial, alter values such as equity, compassion, and avoidance of harm to a wider range of ethical dilemmas. Another work [42], presented a LLM-based debate panel as a method for ethical reasoning. In each panel, AI personas are presented as representing different moral perspectives that differ from the deontological, consequentialist, and stakeholder's view. These panels allow researchers to examine how dialogue dynamics, ethical attitudes, and reasoning strategies affect moral decision-making in high-stakes areas such as healthcare planning and public policy.

These studies illustrate the growing capacity of AI to generate and analyze ethical scenarios with notable sophistication. Yet, a common limitation persists: most models are constrained to static, one-off assessments of moral reasoning. Our proposed framework in this paper aims to address this gap by introducing a gamified, adaptive framework, one that treats moral and cognitive reasoning not as isolated events, but as dynamic, evolving processes that unfold over time.

Game-Theoretic Decision Tree Approaches:

Game theory offers a structured way to explore how individuals (human or artificial) make decisions that depend on the choices of others. By simulating strategic interactions, potential outcomes, and the uncertainties that shape them, researchers can examine how agents respond when faced with ethical or morally complex situations. Therefore, game-theoretic modeling can offer a formal mechanism to simulate moral reasoning processes and evaluate how human participants and AI systems negotiate ethical trade-offs, align goals, or exhibit biases in interactive decision environments.

Several strands in recent literature demonstrate the utility of game-theoretic methods for studying human–AI ethical interaction. [43] used very large (N≈0.49 million), structured

scenario choices to reveal population-level moral preferences and systematic cultural variation. Their dataset shows how choice framing and scenario design shape normative judgments, which is directly relevant when designing scenario libraries for evaluation. [44] Operationalize repeated games to probe cooperation and coordination of large language models, revealing stable behavioral signatures and manipulable features of LLM strategic behavior. Their methods offer concrete experimental paradigms and metrics cooperation rates, reciprocity, strategy persistence) applicable to human–AI co-alignment testing. More recent tool-oriented work, such as FAIRGAME [45] provides an accessible platform for simulating multi-agent games with LLM agents and systematically detecting biased or divergent strategic outcomes across model families and languages. [46] Apply game-theoretic analysis to adversarial socio-technical systems, illustrating how equilibrium computation and search techniques can reveal structural biases and resilience properties in networked competitive settings.

Looking across these studies, we can see how they each contribute in meaningful ways to understanding ethical decision-making in multi-agent contexts. They provide large-scale datasets of moral scenarios, experimental designs for repeated strategic interactions, tools to uncover biases in language models, and approaches for testing adversarial situations in complex systems. Yet, despite these advances, important gaps remain. Many game-theoretic studies tend to simplify or overlook the emotional and narrative aspects of moral reasoning, and there is still limited work that integrates these models with gamified, interactive, or procedurally generated scenarios that capture richer, affect-driven experiences.

Research Gap:

Building on the critical analysis presented across the preceding sections, the reviewed literature highlights several important limitations across existing approaches to moral reasoning assessment and AI alignment. First, traditional instruments grounded in neo-Kohlbergian theory, such as DIT-2, primarily rely on static and decontextualized evaluation methods, limiting their ability to capture the dynamic and evolving nature of moral reasoning. Second, contemporary AI ethics and alignment frameworks largely emphasize compliance with predefined principles, without modeling the underlying cognitive and contextual processes that shape moral judgment.

Third, while gamification has been widely adopted to enhance engagement, its application in rigorous assessment of moral reasoning remains limited, with most implementations prioritizing motivation over structured evaluation. Similarly, AI-driven scenario generation approaches demonstrate strong capabilities in producing complex ethical dilemmas, yet they are often constrained to isolated or one-off assessments without longitudinal adaptability.

Finally, existing research rarely provides integrated mechanisms for comparative evaluation of human and AI moral reasoning within a unified framework, particularly under dynamic and interactive conditions. These gaps collectively indicate the need for a cohesive approach that enables continuous, adaptive, and measurable assessment of moral reasoning across both human and artificial agents.

Framework Design Methodology:

In response to the shortcomings concerning the evaluation of moral reasoning, this study pursues an integrated design-oriented methodology, where insights from interdisciplinary literature (moral psychology, game-based activities, and the ethics of AI) are systematically synthesized to construct and formalize the proposed framework. Within the gaps highlighted above in the literature review, the emphasis is on transitioning from passive, one-off evaluation of reasoning to continuous and flexible assessment of moral reasoning and its evaluative components.

In our approach, gamification is employed to develop a design and measurement framework for assessment, not as an engagement or motivational tool. It enables a more

authentic capture of decision-actor behaviors, the dynamic interaction of human and AI agents, and the longitudinal assessment of moral reasoning along a developmental continuum. The structural framework is informed by the classic and neo-Kohlbergian moral reasoning theories to shape the model of progression, while the AI-driven adaptive learning systems provide targeted scenarios and feedback loops to shape the ongoing relevant context of the assessment.

This work integrates the conceptual foundations necessary for the proposed Gamified Moral Reasoning Evaluation Framework, translating these principles into a complex, multi-layered system designed for human–AI moral co-alignment and continuous assessment, as elaborated in the following section.

The overall design process of the proposed framework follows a structured sequence as follows:

Problem Identification: Analysis of limitations in existing moral reasoning assessment approaches and AI alignment frameworks.

Interdisciplinary Literature Synthesis: Integration of insights from moral psychology, AI ethics, gamification, and game-theoretic reasoning.

Gap Identification: Critical evaluation of existing methods to identify limitations related to static assessment, lack of adaptability, and absence of human–AI comparative mechanisms.

Conceptual Integration: Mapping of theoretical constructs into a unified design structure, aligning moral reasoning models with interactive and adaptive system components.

Framework Design: Development of a multi-layered architecture incorporating gamification, AI-driven scenario generation, dual evaluation modules, and continuous learning mechanisms.

Feasibility Assessment: Evaluation of implementation readiness using Technology Readiness Levels (TRLs) across core system components.

This structured process ensures a systematic transition from theoretical foundations to a coherent and implementable framework design.

While the proposed framework is conceptual in nature, its operational logic is articulated through a structured design process that outlines how theoretical constructs are systematically organized into a coherent framework. The methodology emphasizes clear relationships between design components, decision processes, and adaptive mechanisms, ensuring conceptual clarity and internal consistency. This approach maintains flexibility at the design level while providing a structured foundation that can guide future implementation and empirical validation, without imposing rigid formal or algorithmic representations at this stage.

Conceptual Framework:

Theoretical Foundations of Proposed Framework:

We propose a conceptual framework for an adaptive evaluation of both human and Agentic AI moral alignment as a continuous assessment mechanism in contrast to traditional discrete assessment methods. Most contemporary approaches treat moral reasoning evaluation as a static problem that is considered measurable through one-time assessments. In contrast to this, the proposed framework views moral cognition as an evolving characteristic that can develop through interaction between human participants alongside AI agents. The proposed architecture comprises five interconnected layers (see Figure 2) that are hierarchically organized, incorporating feedback, and can potentially address limitations of classical evaluation methodologies.

At its core, the framework employs both classical and neo-Kohlbergian theories of moral reasoning evaluation [47][48], and integrates them with game-theoretic reasoning models [49] and selected perspectives from virtue ethics. In particular, the neo-Kohlbergian multiple schemas provide mechanisms for recognizing overlapping moral patterns, whereas the classical methods provide fixed stages as markers of moral development. On the other

hand, game theoretic models treat moral reasoning as strategic decision making in interactive and uncertain environments, allowing researchers to examine not just what people believe to be right but also how they act when morals collide with self-interest motives. The hybrid synthesis of Kohlbergian hierarchical stages of moral development with the situational realism of a game-theoretic layer helps to extend the framework's domain from moral understanding to the evaluation of moral actions.

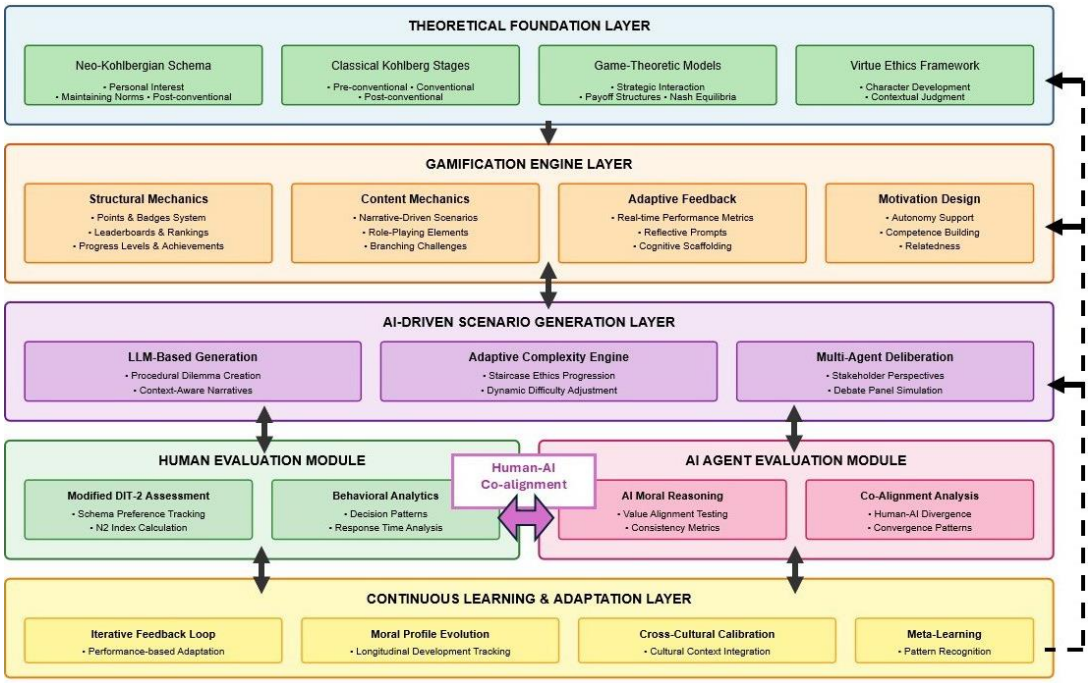

Figure 2. Gamified Moral Reasoning Evaluation Framework.

The proposed framework operates through a structured flow of inputs, transformations, and outputs across its layered architecture. At the input level, participant profiles (human or AI agents), contextual parameters, and initial moral reasoning states are provided to the system. These inputs are processed through the gamification engine, which orchestrates interaction scenarios and progression mechanisms.

The adaptive scenario generation layer produces context-sensitive moral dilemmas, calibrated based on participant responses and prior interaction history. These scenarios are then presented to both human and AI agents, whose responses—comprising decisions, response latency, and interaction patterns—are captured as primary data outputs.

The evaluation modules process this data to derive indicators of moral reasoning, including schema alignment, decision consistency, and convergence/divergence between human and AI responses. These outputs are subsequently fed into the continuous learning layer, which updates participant profiles and informs subsequent scenario generation, thereby establishing a closed-loop adaptive system.

Gamification Engine:

Typical game engines provide a motivational overlay in the form of points, scores, and achievements to improve participation, whereas the proposed gamification engine differs by providing AI-generated dynamic scenarios and content mechanics, including measurement and analysis functionality for moral reasoning assessment. Consequently, this layer expands the scope of the framework from moral evaluation alone to an interactive and adaptive moral development tool.

In contrast to the static and post-hoc mechanisms of DIT-2 and similar neo-Kohlbergian tests, the adaptive feedback system of the game engine provides feedback and guidance to participants during the task. Consequently, the framework is transformed from a

measurement tool into a learning environment. Cognitive scaffolding in the form of automated feedback and reflection prompts or explicit "oracle" consultations can help participants think more deeply about their choices and reasoning process, i.e., encouraging metacognitive awareness. Self-Determination Theory (SDT) [50] identifies three core psychological needs of autonomy, competence, and relatedness that drive intrinsic motivation. These can be intentionally built into the game engine to sustain engagement as well as prevent superficial and externally influenced responses.

The enhanced interactivity and adaptive progression mechanics of the gamified framework mitigate limitations of traditional survey-based assessments, where response fatigue can undermine data quality. The framework is also well-poised to capture fine-grained metadata relating to behavioral signals such as response latency, choice revision frequency, and other user exploratory trajectories. This additional data can potentially provide richer moral cognition indicators compared to the discrete response selection of DIT-2 like frameworks.

Adaptive Scenario Generation Using LLMs:

The LLM-powered scenario generation layer provides superior functionality compared to a fixed set of identical dilemmas of conventional moral assessment systems that overlook participant diversity in terms of moral development, cultural and experiential dimensions of participants. AI-generated adaptive scenarios with moral dilemmas are constructed, matched to each participant's demonstrated reasoning ability and sociocultural context [50]. The situational dilemmas can be synthesized in a coherent and believable domain (i.e., real-world authenticity) while varying contextual details to make them relevant to each individual user. In addition to this, the layer also provides support for multi-agent deliberation [51], whereby AI-simulated stakeholders with differing ethical points of view interact within the same scenario. This functionality challenges participants with multiple ethical perspectives through interaction with representative AI-generated personas, thus revealing their capacity for empathy, negotiation, and consensus building in socially complex situations.

The difficulty of moral scenarios is gradually increased by the adaptive complexity engine in response to the respective participant's demonstrated reasoning ability, using a neo-Kohlbergian inspired "Staircase Ethics" model [52]. The engine ensures that all subsequent dilemmas remain challenging but not discouraging, learning from participants' prior decisions. Participants progress from beginner stages to advanced game levels, remaining in a continuously calibrated moral growth zone that balances stimulation, attainability, and motivation aspects of the game. During this progression, the engine varies factors such as the number of affected stakeholders, clarity or uncertainty of outcomes, number of conflicting moral principles, and the extent of consequences into the future.

Dual Mode Human-AI Evaluations & Co-Alignment Study:

The proposed framework includes an evaluative layer for both human and agentic AI moral reasoning capabilities under comparable conditions [53]. The human evaluation module introduces a continuous extension of the DIT-2 methodology, capturing decision patterns, response times, choice consistency, and exploration strategies of individual participants. The framework allows external AI agents to undertake the game as well, thus inheriting parallel metrics of value alignment, decision stability, and adaptability from the human assessment framework. Since both modules use equivalent indicators, the framework can help identify convergences (i.e., shared reasoning features) and divergences (points of moral mismatch) between humans and machines.

Indicated by the bridge between the Human and AI module in Figure 2, interaction dynamics between human and agentic-AI participants can also be studied under blinded interaction conditions, potentially revealing how human moral reasoning and AI moral patterns influence each other. Through repeated interactions over time, mutual learning and

misalignment can also be observed. This bidirectional exchange mechanism represents a realistic model of rapidly emerging human-AI real-world interactions, where ethical understanding is continuously negotiated rather than being predetermined.

Continuous Learning & Adaptation Layer:

An integrative feedback layer is required to realize a complete developmental process within the framework that accumulates performance data from all preceding layers into a temporally aligned archive of scenarios and participant trajectories. The gamified environment can then adjust itself to insights derived from aggregated data to ensure that the evaluation process remains adaptive, personalized, and developmentally aligned with theory.

The core function of this layer is to keep track of each participant's moral profile evolution trajectory over time. These respective tracks can allow the framework to internally differentiate between transient situational reactions and stable moral dispositions of each participant. Externally, researchers and educators can follow the developmental progress of anonymized subjects to deliberate and devise supportive interventions in the real world, including the development of new theories. In particular, the meta-learning component synthesizes data patterns across participants and populations, whereby shared developmental pathways can be identified along with cultural divergences. Moreover, it can also provide critical effectiveness data for any complementary moral learning interventions in the real world.

The novelty of this framework is the combination of five layers that make up a self-optimizing ecosystem for moral reasoning assessment, development, and human-AI co-alignment studies that evolves and improves through repeated use. The operationalization of psychological theory, gamified interaction, adaptive AI modeling, and game-theoretic reasoning into a single framework may help narrow the methodological divide in the study of moral psychology and AI ethics. This integrative formalization allows moral cognition to be both quantified and developed within a unified adaptive system.

Feasibility, Validation, and Implications:

Technology Readiness and Feasibility Assessment:

Although the proposed Gamified Moral Reasoning Evaluation Framework is conceptual, its realization is practical due to the technological maturity of all key enabling elements. The Technology Readiness Level (TRL) scale allows assessment of their current feasibility and integration potential for an experimental prototype.

The TRL assessment in this study is based on qualitative mapping of each framework component to established definitions of the TRL scale, as outlined in prior literature. Each component is evaluated based on its current level of technological maturity and the availability of implementation platforms in related domains. This structured assessment ensures consistency in assigning TRL levels across components.

LLMs for Adaptive Scenario Generation (TRL 8–9). The use of LLMs such as GPT-4, Claude, and Gemini for procedural content generation and ethical scenario synthesis is already at the commercial deployment level. Their capacity to generate contextually coherent moral dilemmas and character dialogues under controlled prompts is highly feasible.

Gamification Engines and Behavioral Analytics (TRL 7–8): Modern learning management and behavioral research platforms (e.g., existing gamified educational systems already collect high-resolution temporal and decision data), placing this component at an advanced readiness level.

AI-Driven Adaptive Testing and Meta-Learning Modules (TRL 6–7). Adaptive testing methodologies, originally developed for psychometric evaluation and later integrated into intelligent tutoring systems, can be extended to moral reasoning. Model-agnostic meta-learning (MAML) architectures and reinforcement learning-based personalization modules are operational in research labs and early-stage educational pilots, indicating moderate readiness.

Human–AI Co-Alignment and Multi-Agent Simulation Environments (TRL 5–6). Simulation frameworks such as Generative Agents and FAIRGAME demonstrate the ability to model social interactions among AI entities and humans. However, sustained multi-agent moral deliberation remains at an experimental demonstration stage, requiring further validation in controlled studies.

Continuous Learning and Moral Profile Tracking (TRL 4–5): Continuous behavioral modeling and longitudinal moral profiling involve aggregating ethically sensitive data and deriving cognitive metrics. While technically feasible, this component requires further methodological development in data ethics, privacy, and cross-cultural calibration before reaching pilot readiness.

Overall, the assessment indicates that the majority of enabling technologies operate at moderate to high readiness levels (TRL 6–9), supporting the feasibility of incremental prototyping of the proposed framework.

Validation Pathways and Evaluation Considerations:

In addition to feasibility considerations, the framework incorporates measurable dimensions that support future validation. The framework further enables comparative analysis between human and AI agents by examining convergence and divergence in decision-making across equivalent scenarios. Its adaptive structure supports longitudinal tracking of moral reasoning progression, enabling evaluation beyond static, one-time assessments. These evaluation mechanisms provide a structured foundation for future empirical studies, simulation-based testing, and prototype-driven validation.

Implications of the Proposed Framework:

The proposed framework has several implications across theoretical, practical, and broader domains. From a theoretical perspective, it contributes to the integration of moral psychology, AI ethics, and interactive system design into a unified and adaptive evaluation paradigm. From a practical standpoint, the framework provides a foundation for the development of systems capable of continuous moral reasoning assessment, adaptive scenario-based evaluation, and structured analysis of human–AI interaction dynamics. At a broader level, the framework contributes to ongoing discussions on human–AI co-alignment by offering a structured approach to studying moral decision-making in dynamic and context-sensitive environments, with potential relevance for the design and governance of ethically aligned AI systems.

Conclusion and Future Work:

Conclusion:

Gamified environments have a broader and more diverse impact on ethical and moral development. However, value-oriented development and exposure to an ethical and moral-based immersive environment have a long-lasting impact on young learners' minds. Therefore, in this work, we proposed the Gamified Moral Reasoning Evaluation Framework, which is a structured architecture comprising five layers, namely the theoretical foundation layer, gamification engine layer, AI-driven scenario generation layer, human and AI agent evaluation module, and continuous learning & adaptation layer. It presents an emerging step toward integrating moral psychology through the underpinning concepts of classical and neo-Kohlbergian theories to assess moral dilemmas, game-based interaction, individual adaptive behavior in an interactive environment, and adaptive AI systems into a unified structure for assessing moral and ethical alignment. By transitioning from static, questionnaire-based evaluations to dynamic, gamified moral dilemmas, this framework introduces a participatory mechanism that captures human reasoning as an evolving process rather than a fixed trait. The inclusion of adaptive feedback, scenario generation, and co-alignment pathways between human and AI agents offers a promising approach to understanding how moral cognition develops through iterative interaction. Overall, this conceptual foundation lays the foundation

for a more responsive, transparent, and context-sensitive system for evaluating moral decision-making within AI-mediated environments.

Future Work:

Future work will focus on translating the proposed framework into a fully functional prototype that integrates the gamification engine, AI-driven scenario generation modules, and structured behavioral data pipelines. This will be followed by empirical validation through controlled human-subject studies to rigorously evaluate user engagement, learning effectiveness, and the development of deeper moral reasoning over time. Further efforts will explore integration of the framework into real-world decision-support systems within both educational and organizational settings, enabling practical deployment beyond experimental environments. In parallel, the evaluation methodology will be refined by developing more robust metrics for assessing human–AI moral alignment and for capturing longitudinal changes in reasoning behavior. Additionally, future research will investigate the scalability and cultural adaptability of the framework across diverse ethical and societal contexts. Collectively, these directions aim to evolve the system into a scalable and responsible AI framework that supports mutual moral understanding between humans and artificial agents in complex decision-making environments.

References:

[1] Jiwat Ram, "Moral decision-making in AI: A comprehensive review and recommendations," *Technol. Forecast. Soc. Change*, vol. 217, 2025, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040162525001817

[2] D. Deckker and S. Sumanasekara, "Moral Scaffolding Theory (MST): A Developmental Framework for Artificial Moral Cognition through Human Co-Learning," Apr. 2025, doi: 10.36227/techrxiv.174494755.57917307/v1.

[3] Matthew G. Hanna, Liron Pantanowitz, "Ethical and Bias Considerations in Artificial Intelligence/Machine Learning," *Mod. Pathol.*, vol. 38, no. 3, p. 100686, 2025, doi: https://doi.org/10.1016/j.modpat.2024.100686.

[4] Maria Pawelec, "Decent deepfakes? Professional deepfake developers' ethical considerations and their governance potential," *AI Ethics*, vol. 5, pp. 2641–2666, 2025, [Online]. Available: https://link.springer.com/article/10.1007/s43681-024-00542-2

[5] Angel Olider Rojas Vistorte, Angel Deroncele-Acosta, "Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review," *Front. Psychol.*, vol. 15, 2024, [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1387089/full

[6] Sumeet Jhamb, Teresa Ryan, "Emotional Machines: Ethics and Biases of Emotion Artificial Intelligence in Businesses and Workplaces," *J. Leadership, Account. Ethics*, vol. 19, no. 2, 2022, [Online]. Available: https://articlegateway.com/index.php/JLAE/article/view/5157

[7] "(PDF) Gamification in Ethics Education: A Literature Review Gamification in Ethics Education: A Literature Review." Accessed: May 15, 2026. [Online]. Available: https://www.researchgate.net/publication/376610217_Gamification_in_Ethics_Education_A_Literature_Review_Gamification_in_Ethics_Education_A_Literature_Review

[8] Michael Sailer & Lisa Homner, "The Gamification of Learning: a Meta-analysis," *Educ. Psychol. Rev.*, vol. 32, pp. 77–112, 2020, [Online]. Available: https://link.springer.com/article/10.1007/s10648-019-09498-w

[9] Weijane Lin, Jui Ying Wang, "Learning Information Ethical Decision Making With a Simulation Game," *Front. Psychol.*, vol. 13, 2022, doi: https://doi.org/10.3389/fpsyg.2022.933298.

[10] "(PDF) My Chatbot Companion - a Study of Human-Chatbot Relationships." Accessed: May 15, 2026. [Online]. Available:

https://www.researchgate.net/publication/348707261_My_Chatbot_Companion_-_a_Study_of_Human-Chatbot_Relationships

[11] Cathy Mengying Fang, Auren R. Liu, "How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study," *ResearchGate*, 2025, doi: 10.48550/arXiv.2503.17473.

[12] M. D. Chu, P. Gerard, K. Pawar, C. Bickham, and K. Lerman, "Illusions of Intimacy: How Emotional Dynamics Shape Human-AI Relationships," Dec. 2025, Accessed: Mar. 25, 2026. [Online]. Available: http://arxiv.org/abs/2505.11649

[13] L. Laestadius, A. Bishop, M. Gonzalez, D. Illenčík, and C. Campos-Castillo, "Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika," *New Media Soc.*, vol. 26, no. 10, pp. 5923–5941, Oct. 2024, doi: 10.1177/14614448221142007;CTYPE:STRING:JOURNAL.

[14] Rose E. Guingrich, Michael S. A. Graziano, "A Longitudinal Randomized Control Study of Companion Chatbot Use: Anthropomorphism and Its Mediating Role on Social Impacts," *arXiv:2509.19515*, 2025, [Online]. Available: https://arxiv.org/abs/2509.19515?

[15] Lillian Hung, Yong Zhao, "Ethical considerations in the use of social robots for supporting mental health and wellbeing in older adults in long-term care," *Front. Robot. AI*, vol. 12, 2025, [Online]. Available: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2025.1560214/full

[16] D. J. Jobi Babu, "Emotional AI and the rise of pseudo-intimacy: are we trading authenticity for algorithmic affection?," *Front. Psychol.*, vol. 16, 2025, [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2025.1679324/full

[17] "AI Act enters into force - European Commission." Accessed: Mar. 25, 2026. [Online]. Available: https://commission.europa.eu/news-and-media/news/ai-act-enters-force-2024-08-01_en

[18] Cinquanta, R, "AI and Democracy: the Role of the European Parliament in shaping the EU 'AI Act,'" *Community Notebook. People, Educ. Welf. Soc.*, vol. 1, no. 1, pp. 311–340, 2025, doi: https://doi.org/10.61007/QdC.2025.1.278.

[19] David Leslie, Christopher Burr, Mhairi Aitken, Michael Katell, Morgan Briggs, Cami Rincon, "Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal," *arXiv:2202.02776*, 2022, [Online]. Available: https://arxiv.org/abs/2202.02776

[20] "Recommendation on the Ethics of Artificial Intelligence | UNESCO." Accessed: Mar. 25, 2026. [Online]. Available: https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

[21] Natalia Díaz-Rodríguez, Javier Del Ser, "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation," *Inf. Fusion*, vol. 99, p. 101896, 2023, doi: https://doi.org/10.1016/j.inffus.2023.101896.

[22] Kohlberg, L. & Power, C, "Moral Development, Religious Thinking, and The Question of A Seventh Stage," *Zygon J. Relig. Sci.*, vol. 16, no. 3, pp. 203–259, 1981, doi: https://doi.org/10.1111/j.1467-9744.1981.tb00417.x.

[23] J. R. Rest, D. Narvaez, S. J. Thoma, and M. J. Bebeau, "A Neo-Kohlbergian approach to morality research," *J. Moral Educ.*, vol. 29, no. 4, pp. 381–395, 2000, doi: 10.1080/713679390.

[24] F. J. McDonald, "AI, alignment, and the categorical imperative," *AI Ethics 2022 31*, vol. 3, no. 1, pp. 337–344, Apr. 2022, doi: 10.1007/s43681-022-00160-w.

[25] Luciano Floridi, Josh Cowls, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds Mach.*, vol. 28, pp. 689–707, 2018, [Online]. Available:

https://link.springer.com/article/10.1007/s11023-018-9482-5

[26] Laud Nii Attoh Ammah, Christoph Lütge, "AI4people − an ethical framework for a good AI society: the Ghana (Ga) perspective," *J. Inf. Commun. Ethics Soc.*, vol. 22, no. 1, 2024, doi: 10.1108/JICES-06-2024-0072.

[27] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nat. Mach. Intell. 2019 111*, vol. 1, no. 11, pp. 501–507, Nov. 2019, doi: 10.1038/S42256-019-0114-4.

[28] Sebastian Deterding, Dan Dixon, "From Game Design Elements to Gamefulness: Defining Gamification," *Proc. 15th Int. Acad. MindTrek Conf. Envisioning Futur. Media Environ. MindTrek 2011*, 2011, doi: 10.1145/2181037.2181040.

[29] Kai Huotari & Juho Hamari, "A definition for gamification: anchoring gamification in the service marketing literature," *Electron. Mark.*, vol. 27, pp. 21–31, 2017, [Online]. Available: https://link.springer.com/article/10.1007/s12525-015-0212-z

[30] Shurui Bai, Khe Foon Hew, "Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts," *Educ. Res. Rev.*, vol. 30, p. 100322, 2020, doi: https://doi.org/10.1016/j.edurev.2020.100322.

[31] Sujit Subhash, Elizabeth A. Cudney, "Gamified learning in higher education: A systematic review of the literature," *Comput. Human Behav.*, vol. 87, pp. 192–206, 2018, doi: https://doi.org/10.1016/j.chb.2018.05.028.

[32] Jonna Koivisto, Juho Hamari, "The rise of motivational information systems: A review of gamification research," *rise Motiv. Inf. Syst. A Rev. gamification Res.*, vol. 45, pp. 191–210, 2019, doi: https://doi.org/10.1016/j.ijinfomgt.2018.10.013.

[33] K. Werbach and D. Hunter, "For the win : the power of Gamification and game thinking in business, education, government, and social impact," 2020, Accessed: Mar. 25, 2026. [Online]. Available: https://books.google.com/books/about/For_the_Win_Revised_and_Updated_Edition.html?id=nQI2EAAAQBAJ

[34] "The gamification of learning and instruction: Game-based methods and strategies for training and education. San Francisco, CA: Pfeiffer | Request PDF." Accessed: Mar. 25, 2026. [Online]. Available: https://www.researchgate.net/publication/273947281_The_gamification_of_learning_and_instruction_Game-based_methods_and_strategies_for_training_and_education_San_Francisco_CA_Pfeiffer

[35] E. L. Deci and R. M. Ryan, "The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior," *Psychol. Inq.*, vol. 11, no. 4, pp. 227–268, 2000, doi: 10.1207/S15327965PLI1104_01.

[36] Armando M. Toda, Ana C. T. Klock, Wilk Oliveira, Paula T. Palomino, Luiz Rodrigues, Lei Shi, Ig Bittencourt, Isabela Gasparini, Seiji Isotani & Alexandra I. Cristea, "Analysing gamification elements in educational environments using an existing Gamification taxonomy," *Smart Learn. Environ.*, vol. 6, 2019, [Online]. Available: https://link.springer.com/article/10.1186/s40561-019-0106-1

[37] R. N. Landers and A. K. Landers, "An Empirical Test of the Theory of Gamified Learning: The Effect of Leaderboards on Time-on-Task and Academic Performance," *Simul. Gaming*, vol. 45, no. 6, pp. 769–785, Dec. 2014, doi: 10.1177/1046878114563662.

[38] Yuan Jia, Bin Xu, "Personality-targeted Gamification: A Survey Study on Personality Traits and Motivational Affordances," *Conf. Hum. Factors Comput. Syst. - Proc.*, 2016, doi: 10.1145/2858036.2858515.

[39] A. Mora, D. Riera, C. González, and J. Arnedo-Moreno, "Gamification: a systematic review of design frameworks," *J. Comput. High. Educ.*, vol. 29, no. 3, pp. 516–548, Dec. 2017, doi: 10.1007/s12528-017-9150-4.

[40] Jan-Philipp Fränken, Kanishk Gandhi, Tori Qiu, Ayesha Khawaja, Noah D. Goodman, Tobias Gerstenberg, "Procedural Dilemma Generation for Evaluating Moral Reasoning

in Humans and Language Models," *arXiv:2404.10975*, 2024, [Online]. Available: https://arxiv.org/abs/2404.10975

[41] Ya Wu, Qiang Sheng, Danding Wang, Guang Yang, Yifan Sun, Zhengjia Wang, "The Staircase of Ethics: Probing LLM Value Priorities through Multi-Step Induction to Complex Moral Dilemmas," *ACL Anthol.*, 2025, [Online]. Available: https://aclanthology.org/2025.emnlp-main.806/

[42] Hazem Zohny, "Simulating Ethics: Using LLM Debate Panels to Model Deliberation on Medical Dilemmas," *arXiv:2505.21112*, 2025, [Online]. Available: https://arxiv.org/abs/2505.21112

[43] E. Awad *et al.*, "The Moral Machine experiment," *Nat. 2018 5637729*, vol. 563, no. 7729, pp. 59–64, Oct. 2018, doi: 10.1038/s41586-018-0637-6.

[44] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, Eric Schulz, "Playing repeated games with Large Language Models," *arXiv:2305.16867*, 2025, [Online]. Available: https://arxiv.org/abs/2305.16867

[45] Alessio Buscemi, Daniele Proverbio, Alessandro Di Stefano, The-Anh Han, German Castignani, Pietro Liò, "FAIRGAME: a Framework for AI Agents Bias Recognition using Game Theory," *arXiv:2504.14325*, 2026, [Online]. Available: https://arxiv.org/abs/2504.14325

[46] Andrew Cullen, Tansu Alpcan & Alexander Kalloniatis, "Game-Theoretic Analysis of Adversarial Decision Making in a Complex Socio-Physical System," *Dyn. Games Appl.*, vol. 15, pp. 709–728, 2024, [Online]. Available: https://link.springer.com/article/10.1007/s13235-024-00593-4

[47] "The philosophy of moral development : moral stages and the idea of justice : Kohlberg, Lawrence, 1927-1987 : Free Download, Borrow, and Streaming : Internet Archive." Accessed: Mar. 25, 2026. [Online]. Available: https://archive.org/details/philosophyofmora0001kohl

[48] "(PDF) Postconventional Moral Thinking: A Neo-Kohlbergian Approach." Accessed: Mar. 25, 2026. [Online]. Available: https://www.researchgate.net/publication/247214536_Postconventional_Moral_Thinking_A_Neo-Kohlbergian_Approach

[49] John C. Harsanyi, "Game and decision theoretic models in ethics," *Handb. Game Theory with Econ. Appl.*, vol. 1, pp. 669–707, 1992, doi: https://doi.org/10.1016/S1574-0005(05)80022-0.

[50] E. L. Deci and R. M. Ryan, "Self-determination theory: A macrotheory of human motivation, development, and health," *Can. Psychol.*, vol. 49, no. 3, pp. 182–185, Aug. 2008, doi: 10.1037/a0012801.

[51] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, Michael S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," *arXiv:2304.03442*, 2023, [Online]. Available: https://arxiv.org/abs/2304.03442

[52] Chelsea Finn, Pieter Abbeel, Sergey Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," *arXiv:1703.03400*, 2017, [Online]. Available: https://arxiv.org/abs/1703.03400

[53] "(PDF) Computerized Adaptive Testing: A Primer." Accessed: Mar. 25, 2026. [Online]. Available: https://www.researchgate.net/publication/392721914_Computerized_Adaptive_Testing_A_Primer