SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I11P107

Volume 13 Issue 11, 129-138, November 2026
© 2026 Seventh Sense Research Group®

Original Article

Evaluating the Reliability of Human Ethical Behavior for Ethical AI Alignment

Stenin Nelliyulla Parambath[1], Malode Vishwanatha Panduranga Rao[1], Ezhilarasan Ganesan[2]

[1]*Computer Science & Engineering, Faculty of Engineering & Technology, JAIN (Deemed-to-be University), Kanakapura Main Road, Bengaluru, Karnataka, India.*
[2]*Department of Electrical & Electronics Engineering, Faculty of Engineering & Technology, JAIN (Deemed-to-be University), Kanakapura Main Road, Bengaluru, Karnataka, India.*

[1]*Corresponding Author: stenin_n_p@yahoo.com*

Received: 16 September 2026 Revised: 16 October 2026 Accepted: 31 October 2026 Published: 27 November 2026

Abstract - *The question of ethical alignment is one of the most pressing issues of modern Artificial Intelligence (AI), as most systems mostly recreate neocortical mental operations without having effective moral judgment features. One of the common beliefs that have been put forward in ethical AI studies is that ethical behavior in humans can be used as a good training indicator in ethical decision-making in machines, but this has received very little empirical support. This paper questions the consistency and reliability of modern human ethical conduct to serve as a tuning mechanism for Ethical AI technologies. The survey was qualitative in nature and was conducted through a questionnaire with 101 participants from different geographical and demographic backgrounds. Structured categorical responses were used to analyse ethical behavior in various real-life situations of decision-making. Data analysis was conducted at the question and respondent levels to assess ethical consistency, ambiguity, and self-perception bias; pattern saturation and validity checks were conducted to guarantee methodological rigor. The results indicate that even though the participants tended to exhibit ethical reasoning in single situations, consistent ethical behavior across situations was a rarity. It was found that only about 4% of the respondents showed a perfect ethical alignment on all the dimensions that were surveyed, and significant differences were noted between the self-rated ethicality and the reported one. Such findings suggest that modern human ethical conduct is not enough on its own as a training cue for ethical AI systems. The paper gives empirical reasons to understand how structured ethical systems and value-based reasoning systems, including those based on the Indian Knowledge System and limbic-based AI systems, can be used to facilitate the creation of coherent and ethically sound artificial intelligence.*

Keywords - *Ethical AI, Human ethics, AI alignment, Indian Knowledge System, Limbic intelligence.*

1. Introduction

Recent developments in Artificial Intelligence (AI) with the help of Machine Learning (ML), Deep Learning (DL), Generative AI, and newer Agentic AI paradigms have made a significant leap in perception, reasoning, as well as autonomous decision-making. Modern AI designs are largely based on neocortical cognitive processes of the human brain and are good at logical inference, statistical learning, and high-scale data processing. Irrespective of these abilities, these systems often have weaknesses in moral reasoning and context sensitivity, delivering morally or ethically inappropriate decisions whose technical correctness makes their non-ethical character feasible [1].

Ethical decision-making in human cognition does not exist because of purely rational analysis. Rather, it is the result of the interaction of the processes of logical thinking and affective processing, internalized values, social conditioning, and experience memory. These functions are most closely related to the limbic system, which is central in the regulation of emotions, moral judgment, and perception of responsibility. Contrarily, the majority of current AI systems do not have built-in ways to represent affective or value-wise reasoning [2]. Ethical actions with machines are therefore usually enforced by outside interventions, comprising of rule-based restrictions, after-hoc safety filters, or controlled alignment datasets [3]. Such strategies minimize some of the most visible risks, but they fail to deal with the inherent issue of integrating ethical logic into the very process of decision-making.

One of the common assumptions in the study of ethical AI is that human ethical action can be used as a good guide to machine alignment. It is based on this assumption to include human-in-the-loop learning, reinforcement learning using human feedback, and preference-based optimization.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Subtly, such approaches assume that human decision-making nowadays is consistent, well-principled, and universally acceptable [4]. Nevertheless, human ethics is subject to situational, cognitive, social, and personal motivated factors. The validity of human ethical behaviour as a training signal for ethical AI has had little empirical support, in spite of its importance [5].

Similarly, more focus has been given to traditional philosophical and culturally-based ethical frameworks as alternative or supplementary sources of AI ethics [6]. Indian Knowledge System (IKS), with its concepts of Dharma (moral duty), Karma (causal responsibility), and Ahimsa (non-violence), is a legal system of ethics that has developed throughout millennia. IKS places greater emphasis on persistence of accountability, community welfare and future implications, unlike much of modern ethical standards, which are pragmatic and individualistic in nature. All of these features are in great accordance with the functional role of the limbic system in human ethical decision-making, specifically value integration and emotional regulation of decisions [7].

Based on these observations, the following question for research arises: Is it possible to reliably use the modern moral behaviour of a human being as a basic starting point for machines to learn moral behavior? Answering this question is a key step in determining the efficacy of the current strategies for human-centric alignment, and whether other sources of ethical knowledge must be developed to systematically be incorporated into AI systems. In order to explore this question, qualitative empirical research was carried out with the use of a structured questionnaire given to respondents representing various demographic characteristics in the world. The survey assesses various aspects of ethical behavior such as moral self-evaluation, ethical compromising in constrained situations, adherence to rules, responsibility attribution and decision-making tendencies in real life. The paper contrasts the consistency and validity of human ethical behaviours through the analysis of patterns and inconsistency in human reactions to construct a machine ethical criterion.

The results of this exploration directly relate to the general goal of creating ethically and emotionally intelligent AI systems based on the activity of the limbic brain and with the help of systematized ethical principles like the Indian Knowledge System. More to the point, the results put a popular but little-explored hypothesis in the field of ethical studies of AI on its knees, and thus raise the question of the current paradigms of alignment. This paper presents a methodological empirical evaluation to measure the appropriateness of current human moral conduct to serve as a training cue towards ethical alignment of AI. The work goes beyond the hypothetical implications of the notion of ethically sensitive situations and provides evidence formulated by data instead of assumptions as to the capabilities and the weaknesses of human-based ethical tuning mechanisms for intelligent systems. The study is one of the strongest in terms of being global and multi-demographic. The qualitative survey conducted on participants from varying countries, cultures and demographics was the one that helped to assess the ethical differences in the population. The results from such a wide sampling approach are not restricted to views from culturally homogeneous or geographically separate areas, thereby enhancing the applicability of the conclusions.

Based on the results of the empirical study, it can be concluded that there is a sharp contradiction with the human behaviour of ethics. In the scenarios discussed, the ethical reaction of a small group of the participants remains constant across the board, and many of them are all depicted as morally unsure or utilitarian and/or enforcing moral values biased. These results call into question the assumption that human ethical behaviour is consistent and uniform enough to be directly generalised for alignment of machines. The paper also finds that basic constraints in AI alignment policies are based solely on modern human ethics. The outcome analysis points to the fact that these methods could introduce human biases, short-term self-interest, and situational moral flexibility into AI systems, hence achieving a reduced level of solid ethical integrity and reliability in the long term.

This work offers a solid rationale for why structured and value-focused ethical frameworks should be investigated by empirically determining the inadequacy of human ethics as a sole basis of ethical AI. Specifically, the results inspire the inclusion of the ethical precepts based on the Indian Knowledge System, which puts an emphasis on continuity of responsibility, collective welfare, and long-term implications, as additional guidance for AI decision-making. Lastly, the research establishes the conceptual connection between ethical reasoning and limbic-inspired intelligence. Observable shortcomings of the purely rational, neocortex-inspired models justify the need to include the affective and value-driven mechanisms similar to the human limbic functions. This contributes to the discussion of next-generation AI architectures that will combine cognitive efficiency with ethical and emotional intelligence.

2. Related Works

Artificial intelligence alignment has become an important field of research and is influenced by the growing use of autonomous and decision-support systems in socially sensitive environments. There are five overlapping axes of existing literature, namely: (i) global AI ethics models and principles of governance, (ii) human-conscious alignment paradigms like reinforcement learning based on the feedback of human participants, (iii) bias and fairness in human-generated knowledge bases, (iv) affective and neuro-inspired architectures, and (v) culturally-founded ethical knowledge systems.

There has been extensive literature on the definition of high-level ethical principles that can be used to develop and implement AI. One of the first systemic mappings of AI ethics guidelines in the world reported by Jobin et al. showed the presence of common principles, including transparency, fairness, accountability, non-maleficence, and privacy, across institutional contexts across the globe [8]. Although there has been conceptual convergence, there have been many variances in the interpretation, implementation, and cultural foundation. Floridi et al. introduced the AI4People framework in which they suggested that beneficence, non-maleficence, autonomy, justice, and explicability are essential pillars of ethical AI systems [9]. Even though these frameworks entail development of a noteworthy normative foundation, they remain mainly prescriptive and offer little in relation to the empirical derivation or operationalization of ethical conduct in learning-based systems. Recent extensive surveys of AI governance policies also reveal that there is no cohesion in the ethical priorities across places and organizations, showing the lack of a universally acceptable ethical basis for AI decision-making [10].

Human-centred methods of alignment, including Reinforcement Learning through Human Feedback (RLHF), have become a new dominant paradigm when it comes to training large AI models. Ouyang et al. have shown that models that are fine-tuned with human preference data give better instruction-following and less harmful behaviour [11]. These are the most common methods used in contemporary generative AI systems. However, further research has found that there are intrinsic weaknesses of human-derived alignment. Constitutional AI was suggested by Huang et al. to decrease the need to be constantly labelled by humans by applying specific sets of principles as constraints [12]. Although this approach is effective in partly reducing certain risks, it is, nevertheless, based on man-imposed rules and fails to prove the consistency of human behaviour with regard to these rules. More empirical studies have demonstrated that annotator inconsistency, framing and reward bias have an influence on human preference data. Shen et al. pointed out that there were systematic errors in RLHF signals, such as length bias and preference instability, expressing concerns that human judgments could serve as a basis for ethical signals [13].

There is a significant amount of research that has proven that AI systems that are being trained on human data tend to be biased to the society. Barocas and Selbst have shown how data-driven decision systems can be biased by chance because of historical and structural injustices in training data [14]. The article by Mehrabi et al. offers a detailed overview of the types of bias and the issue of fairness in machine learning, highlighting that the issue of ethical consistency in human data could not be assumed [15]. Such results imply that human ethical conduct can be situational, self-focused and culturally programmed. Therefore, applying aggregated human behaviour as a universal moral reality is a dangerous way of introducing inconsistency and moral trade off to AI.

Researchers in affective computing have stressed that emotional intelligence, empathy, and experiential memory are important in intelligent systems. The fundamental contribution made by Pie et al. was the establishment of the fact that affective processes are essential to meaningful human-machine interaction [16]. A more recent review has recorded new developments and improvements in emotion recognition and affect-aware decision-making models [17]. Neuroscience-based AI studies also indicate that ethical decision-making in humans is very much related to the activities of the limbic system, such as emotional regulation and value-based decision-making. However, in spite of these observations, the majority of AI systems are neocortex-oriented, with emphasis on logic and optimization, ignoring affective and ethical aspects.

Issues of Western-centricity in AI ethics have prompted the demand for culturally pluralistic models. Ethical priorities in different societies have been shown to differ greatly when compared to each other, and this does not support the universality of the current dominant models of AI ethics. In this framework, organized cultural regimes like this provide a rational ethical paradigm founded on long-term responsibility and the common good. Principles like Dharma, Karma and Ahimsa give internally consistent types of ethics which have been honed over centuries. Although recent research and white papers suggest the integration of IKS in the framework of AI ethics [18, 19], they are still mostly theoretical and are not empirically substantiated concerning the inadequacy of modern human ethical conduct. A number of works have suggested feasible toolkits to evaluate the ethical aspect of AI, such as bias audits, transparency checklists, and governance dashboards. Morley et al. analysed publicly available AI ethics tools and found that the majority of them concentrate on post-hoc evaluation, instead of seeking to explain the root of ethical knowledge [20]. The recent datasets on responsible AI measurement also show how challenging it is to measure ethical behaviour when the labels are generated through human judgment, which supports the necessity to approach the analysis of the reliability of human ethics as a main source critically [21].

The literature shows that enormous dependence on human feedback, preferences, and normative principles is made in alignment with AI ethics. Nevertheless, empirical research has not been conducted to determine whether modern human ethical behaviour is coherent, predictable, and generally consistent to a degree that it can be used to provide an ethical machine with the foundations of training. The current research accepts the bias and inconsistency but fails to empirically test whether human ethics is adequate. The current paper fills this void with a qualitative cross-demographic survey on the consistency of human ethical

decisions, which would offer evidence-based reasons to study structured ethical systems like the Indian Knowledge System and limbic-inspired AI systems.

3. Methodology

3.1. Research Design and Objective Alignment

The proposed study follows a qualitative, exploratory research design, as it empirically tests the question of whether modern human ethical behaviour is consistent and reliable enough to provide a training signal in the ethical decision-making of artificial intelligence systems. The methodology clearly follows the overall agenda of the research, which inquiries the possibility of tuning machines to assume ethical human behaviour on the basis of the observed human ethical behaviour. The survey method adopted a questionnaire as it was found to be the most reflective form of ethical judgment, tendencies and trade-offs on moral aspects that were used in the actual situation to make decisions. Qualitative surveys are especially apt in morally inquisitive inquiry since they enable semi-evaluation of moral reasoning by respondents in complex and contextual situations that cannot be wholly embodied using strictly quantitative measures and tools.

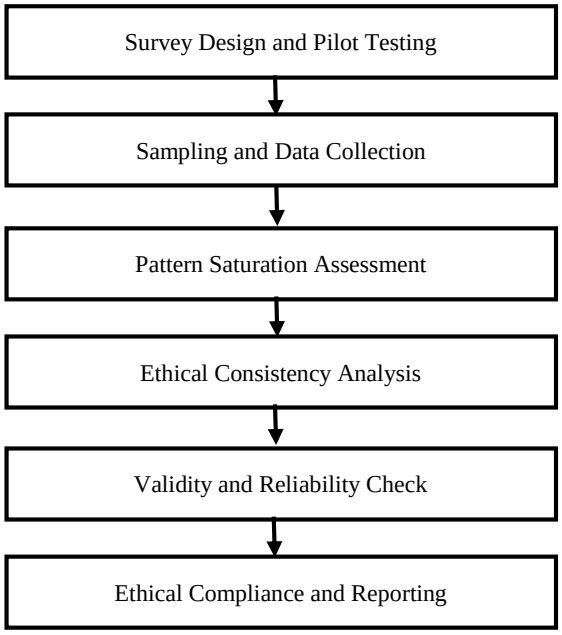

Fig. 1 Research pipeline

The proposed methodology will adhere to a qualitative research pipeline with multi-stage, systematic steps (Figure 1) to estimate the appropriateness of human ethical behaviour to ethical AI tuning. This is initiated by designing and pilot testing an ethical situation-based questionnaire to verify that it is clear and neutral. Information is then gathered among 101 participants around the world in order to attain cross-cultural ethical views. Pattern saturation test is conducted to ensure the stability of responses and sufficiency of sample size. Ethical consistency is then examined at scenario and respondent levels to determine whether there is consistency in ethical decision-making or not. To prevent bias and guarantee data integrity, validity and reliability checks are included. Lastly, the integrity of data is supported by anonymized data processing and accountable reporting, which makes the framework very powerful, reproducible, and consistent with Scopus-reporting standards.

3.2. Survey Instrument Design

To explore the various dimensions of ethical behaviour as applied to real-life decision-making, a structured questionnaire was created in order to systematically study various aspects of ethical behaviour. The measurement device was meant to provide critical constructs of ethics, such as ethical self-reflection and moral awareness, a conflict between personal benefit and moral principles, compliance with rules versus justification and outcome of ethical behaviour, honesty and integrity in personal and professional life, a sense of responsibility towards other people, the environment, and society and readiness to make a compromise of moral principles to gain short-term advantages. Put collectively, these dimensions offer an in-depth account of real-life ethical conduct in a diverse situational setting. The survey questions have been structured in such a manner that they would elicit a clear moral stand of the respondents and be interpretively consistent among a diverse group of respondents. In this regard, controlled response categories, including Yes, No and Maybe, were used. This design alternative was easy to be coalesced quantitatively and also enabled the ethical uncertainty in the possibility of the respondents in the uncertainty or conditional reasoning situations. This was also important due to the existence of the neutral response category, which is sometimes crucial to the capture of the situationally flexible moral responses that are often lost in dichotomous moral judgments.

The questionnaire was purposely skewed towards making behavioural and decision-related inquiries instead of abstract moral inquisitions or philosophical inquiries. This method was used to evaluate morality as practiced in real life, as opposed to morality as ideal or theorized. The instrument sought to find how people make ethical trade-offs under realistic conditions and competing demands by basing the questions on real-life situations. The questionnaire was also tested in a pilot test on a small and mixed group of participants before its large-scale implementation. The pilot test was aimed at checking the question clarity, understandability, cultural neutrality and the lack of implicit bias or dominant language. The feedback received at this stage was evaluated in a systematic way and applied to improve the phrasing of some of the items, remove the ambiguity, and increase the level of neutrality. These adjustments made the end product a solid instrument, understandable, and applicable in heterogeneous demographic groups.

3.3. Data Collection and Sampling Strategy

The questionnaires were created and surveyed through the Microsoft Forms platform, which provided safety and convenience of the questionnaire, anonymity and automated response collection. The data was dealt with effectively through the use of an online survey tool that guaranteed the privacy of the respondents and reduced the chances of development of social desirability bias because of identifiable participation. The population under study was adult members of a broad geographical area and demographical composition (respondents of Asia, North America, Europe, Africa, the Middle East and Oceania) and thus, the comprehensive nature of ethics in the world. A convenience-based purposive sampling method was used that emphasized diverse ethical perspectives, rather than demographic homogeneity and statistical representativeness. This sampling method is suitable for qualitative in-depth research that is exploratory in nature, particularly if the research hypothesis is to determine the patterns, gaps and theme-related variation in human behaviour in cultural and social contexts. The study was devised to provide a wide range of ethical perspectives with meticulous encouragement of diversified backgrounds, as dictated by social standards, life experience and cultural compulsions.

A total of 101 valid responses were obtained and stored to be analysed at the end. During the data collection process, the distributions of responses and patterns were constantly observed to gauge the development of themes as well as redundancy. It was noted that around 45-50 responses, further data sets started showing pattern saturation with little new trends emerging on ethical response. Although this is an early sign of saturation, data were still collected to increase the strength of the results, build up confidence about saturation among demographic subsets, and ascertain that the trends observed were not a result of early sampling bias.

3.4. Sample Saturation and Adequacy Justification

In this study, the principle of theoretical and pattern saturation was used to assess sample adequacy, as it is an established principle of qualitative research that is used to characterize when adequate data has been gathered. Saturation was not determined only by the size of the sample, but by the stability and repetition of the patterns of responses that were observed with each consecutive data point. This will make the conclusions based on consistent empirical findings and not arbitrary numbers. Namely, saturation was established based on a variety of complementary indicators. These were the stabilization of the response distributions across the ethical questions, the recurring similar patterns of ethical inconsistencies in a sample of different respondents and the lack of new or previously unobserved ethical response categories in later data collection stages. When combined, these signs were solid evidence that further sampling would not produce anything of significant new knowledge applicable to the study purposes.

The final sample had 101 respondents, and this is above the general saturation thresholds that are usually documented in qualitative and exploratory behavioural research studies. This size is big enough to provide enough empirical richness to facilitate cross-scenario analysis of ethical consistency and inconsistency in the profile of individual decision-making. It is important to note that this study is not to statistically generalize to a broader population, but to analytically generalize. It is concerned with the question of whether uniform ethical conduct is so common and consistent as to be used as a prior training condition for ethical AI systems, as opposed to estimating the ethical population distributions.

3.5. Data Analysis Procedure

All the data collected were converted to a systematic tabular format for systematic analysis of the data. The data were organised in such a manner that it could be analysed on an item-wise and respondent-wise basis, and thus ensured that the ethical conduct could be analysed at different levels. This two-level analytical framework has been adopted to make sure that not only the overall ethical tendencies have been embraced, but also the consistency patterns of individuals in different ethical situations. At the first level, a question-level analysis was carried out to discuss the balance of ‘Yes’, ‘No ’, and ‘Maybe’ answers on each of the ethical dimensions. In the process of this analysis, the patterns of ethical compliance, ambiguity, and compromise in specific decision-making situations were identified. The existence of a neutral response category facilitated the identification of conditional or situational moral thinking that is particularly relevant in seeking to establish the validity of human ethics as a reference standard that is consistent.

At the respondent level, the consistency of the application of ethical positioning in the answers a particular respondent gave to each of the questions was tested at the second level. In this respect, a high degree of consistency criterion was established. A respondent was ethically consistent only when he demonstrated ethical consistency in all the situations surveyed and did not have any inner contradiction or self-identification with ethical degradation. This stringent definition was specifically made to mirror operational necessities of ethical AI systems, which are supposed to be operating with steady and predictable ethical conduct instead of adaptability according to situations.

The results of this examination, overall, revealed a clear difference between the results of partial ethical congruence and holistic ethical congruence. Although some of these individual questions scored as moderately high on ethical compliance as a single question, a very small percentage of the respondents indicated that they exhibited ethical behavior consistently throughout all the questions. It is first-hand empirical data of the primary objective of the research, which is to evaluate human ethical behaviour as a legitimate way of educating machine ethics and AI alignment.

3.6. Validity and Reliability Considerations

Various methodological procedures were embraced to maximize both the validity and reliability of the research, and ensure that the findings are valid and analytic. Considering the characteristics of the information (qualitative and self-reported), particular care has been taken to reduce the possibility of ambiguity in the interpretation, and to make sure that the survey instrument could measure the constructs of ethical interest as intended. The fact that these survey questions were informed by some of the most frequently used dimensions of ethics widely debated in the literature of AI ethics, moral psychology and ethics of behaviour was used as a measure of content validity. The selected dimensions include all the key components of ethical decision-making processes which are directly relevant to the suitability of using human ethics as a premise to align AI with. This orientation ensured that the instrument was broadly and theoretically based on ethical aspects.

To make sure that the questions were read by the participants as intended by the researchers, face validity was determined at the pilot stage of the research. The pilot group offered feedback which led to the revision of the wording of the questions, the removal of ambiguous phrases, and ensured clarity for the questions across all cultural and language groups. This was crucial in reducing the effects of misinterpretation and increasing the chances of responses being interpretable. The internal consistency was assessed by analysing the question-response cross-questions at the respondent level. This was done by determining contradictions, ethical reversals or discontinuity of ethical positioning in various situations. Nor were these inconsistencies viewed as 'noise' in the data, but were instead used analytically to calculate the consistency of ethical behaviour in a person, the chief research objective of the study.

Potential bias of responses was mitigated by providing a high level of anonymity of participation, which minimized the effects of social desirability and promoted the use of honest self-reporting. The lack of personally identifiable information reduced the perceived judgment or evaluation, which increased the extent of the freedom of the participants in answering ethically sensitive questions. Though self-reported ethical behaviour is subjective in nature, the research methodology is quite clear about this subjectivity as an informative aspect of the research and not a methodological concern. Such an approach is especially applicable to ethical AI alignment, which is often based on subjective human judgments to direct the actions of the machine.

4. Results and Discussion

This section shows the empiric results of the qualitative survey that is carried out to test the consistency of human ethical behaviour and its appropriateness as a priori signal to come to the ethical AI tuning. The question-level distributions, aggregate response analysis, and ethical consistency interpretation are used to report the results, which is consistent with the aim of the research. One hundred and one valid responses were used. All the respondents were asked to respond to a set of structured ethical decision-making questions in categorical responses (Yes / No / Maybe), as well as to self-rate their ethical rating. All answers were coded anonymously before being analysed.

4.1. Question-Level Ethical Response Distribution

Table 1 shows the frequency of responses in the major dimensions of ethical aspects that were assessed in the survey. The findings show that there is a significant range of ethical positioning in various situations, which means that the issue of human ethical decisions is not a simple and standard one. Despite the fact that 'No' answers in most questions comprise the majority, indicating nominal ethical compliance, each occurrence of 'Yes' and 'Maybe' answers in almost all the dimensions points to the fact that there is no uniform and absolute compliance with ethics among the respondents. Questions of direct ethical compromise, including acting ethically dubiously, making a personal gain instead of the ethical choice, violating the rules for personal advantage, and dishonest or deceptive acts, have significant percentages of affirmative and undefined answers. For instance, a high percentage of the respondents answered yes to questions, which was a sign of moral ambivalence and not outright denial of such actions. Likewise, we can see similar patterns in questions pertaining to bending the rules and breaching ethics to achieve some immediate gain, where uncertainty ('Maybe') embodies situational thinking and conditional morality.

Table 1. Distribution of ethical responses across survey questions

Ethical Question (Abbreviated)	Yes	No	Maybe
Engaged in ethically questionable behavior	38	46	16
Personal gain over ethics	15	63	23
Bending rules for benefit	22	59	20
Dishonesty or deception	19	66	16
Manipulated others for gain	14	70	17
Compromise ethics for short-term benefit	18	65	18
Ethics are subjective	18	60	23
Took credit for others' work	3	88	10

The aspect of engagement in manipulating other people to benefit oneself or even taking credit of other people are considered as interpersonal ethics, and the affirmative responses are not so high here. In these categories, too, the affirmative responses (either Yes or Maybe) are not irrelevant since these are blatant breaches of commonly held ethical standards. The results show that a certain group of

respondents admits that such actions does bring about an ethical failure even in an otherwise unquestionably unethical situation. The responses to the statement on subjectivity of ethics also highlight the element of difference in ethics. A significant percentage of interviewees supported or at least partly supported the belief that ethics are relative and that moral reasoning tends to be relative or situation-specific instead of the dedication to some set of ethical judgments. This understanding has significant consequences of the alignment of AI ethically, as it embodies a non-agreement on the existence of ethical absolutes among the human population.

Generally, the analysis at the question level indicates that overt ethical infractions are not universal elements, though the flexibility of ethics, ambiguity, and situational compromise are always present in various aspects. The evidence suggests that human ethical action, even when considered on a fine scale of ethical situations at the individual level, is not uniform and stable enough to be directly converted to machine ethics. This point is a direct indication of the main argument of the study that human ethics in their existing form in practice might be inadequate as a primary source of training for ethically consistent AI systems.

4.2. Aggregate Ethical Response Pattern

All the ethical questions were combined and analysed together in order to take into consideration the overall trends in ethics among the surveyed population. Figure 2 displays the cumulative distribution of categories of the responses (Yes, No, and Maybe) in the entire data, which would give the general picture of ethical behaviour, as compared to the observation of each question. There is a clear preponderance of the 'No' responses, which indicates that the respondents tend to perceive themselves as not violating ethical norms when they are viewed as a whole. Yet, this seeming ethical conformity is coupled with large shares of 'Yes' and 'Maybe' answers, meaning that ethical compromise and uncertainty are not single incidents but are organized in numerous decision situations.

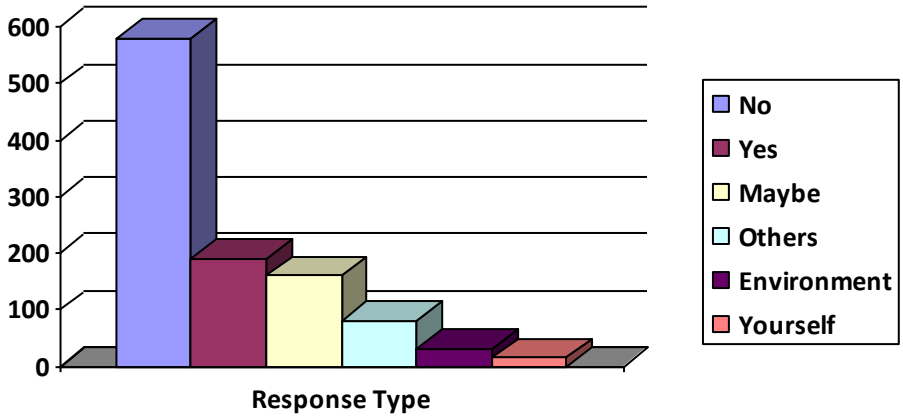

Fig. 2 Aggregate ethical response distribution

The fact that there are many responses of 'Yes' proves the direct recognition of ethically dubious behaviour, and the fact that the percentage of responses of 'Maybe' is significant is especially informative. Such vague answers indicate an ambiguity of ethics, conditional thinking or contingency that the respondents fail to, or do not want to, specify their stance as to the finality of the ethics. Such variability can be seen as an ethical AI alignment issue, because machine ethics aims at determinism, predictability, and consistency as opposed to context-driven moral variation.

Altogether, the overall pattern of response supports the results obtained in the question level and contributes to the main argument of the research. Although the behaviour of humans in ethical contexts might seem very obedient when considered in an isolated scenario, the distribution of ethics over time displays persistent ethical hazards and trade-off. Such aggregate inconsistency leads to vital doubts about the direct application of human ethical behaviour as a training signal on its own to ethical AI systems and the importance of more systematic and consistent ethical systems in the design of AI.

4.3. Ethical Consistency at the Respondent Level

In addition to question-level and aggregate analyses, the ethical behaviour was also analysed at the respondent level, to see how ethical decision-making stayed consistent across different situations. The consistency test applied in the analysis in this case was quite rigorous: if no ethical compromise or ethical ambiguity was revealed in any of the ethical dimensions surveyed, the respondent was regarded as

being ethically consistent. Particularly, when some form of response to the question contained any of the Yes or Maybe that meant that the behaviour is ethically dubious, it was categorized as ethically inconsistent. This rigid definition was chosen with the purpose of diversifying in accordance with operational needs of ethical AI systems, which require a high level of reliability and predictability in the different contexts of decisions. The total of 4 of 101 respondents, who represent approximately 3.96% of the sample, were fully ethical in all queries. What this observation implies is that, in everyday life, it is not highly likely to uphold ethical norms in the long term and in diverse circumstances, even by individuals who would deem themselves to be ethically responsible. A majority of the respondents (98%) showed at least one act of ethical compromise or ambiguity and reveal the widespread situational or conditional ethical reasoning in human behaviour.

These results reveal a significant drawback of humanistic methods of ethical harmony. Though one can be ethical in isolation and in familiar settings, it appears to be rare to act "ethically" in many and sometimes opposing scenarios. This is human moral inconsistency, which is a problem of AI systems that are meant to be steady and run in the same manner without any moral drift over time and circumstances. The noted absence of ethical consistency of respondents is a solid empirical measure that modern human ethical conduct, as an applied phenomenon, might not be enough as an independent basis of training ethically responsible AI systems.

4.4. Self-Assessment vs. Behavioral Ethics

To further examine the accuracy of the human factor in the ethical context, self-rated ethical scores were compared with self-reported ethical behaviors of the respondents as to the ethical scenarios surveyed. The purpose of this comparison was to identify the potential differences between ethical self-perception and actual-life decision-making, namely, in the case of AI alignment approaches that utilize self-reported human preferences or values.

The assessment shows that there is a great gap between self-evaluation and moral behavior. Several with high rates on the scale on ethical thinking stated that they were involved in ethically doubtful behaviors or had ethical issues in certain instances. This is a lack of consistency, which suggests that there is a self-perception bias, people maintain a positive moral sense of themselves regardless of their moral beliefs and behaviors.

This result has significant consequences to ethical AI alignment. Methods based on self-reported moral attitudes, preferences or evaluations make an implicit assumption that the actual moral behaviour is in line with the self-conceived moral behaviour. The identified inconsistency shows that these assumptions might not be true, and self-assessed ethics do not always correspond to ethically consistent behaviour. Human values as a direct training signal should not then be assumed as a non-biased and reliable signal to feed into AI systems; as such, technical and ethically meaningful strictness of AI design requires that values be structured and behaviourally based.

4.5. Implications for Ethical AI Tuning

Regarding ethical AI creation, the research results of this paper reveal that the ethical behaviour of humans in the present-day reality is not as consistent as it must be to serve as an independent source of ethics to be used by intelligent systems. Despite the fact that concerns about individual ethical judgments might seem to make sense individually, there is evidence that all of them are widespread with regard to variability, compromise in situations, and moral ambiguity, which cuts across decision situations. These features do not fit the demands of an ethical AI system, which should exhibit consistent, predictable, and repeatable ethical behaviour in an extensive variety of conditions. Although human feedback is still a relevant factor to contextual learning and adaptation, there is a risk associated with extensive use of human ethical behaviour as a main signal of alignment. These are the unwanted implementations of human bias, immediate self-interest, and the flexibility of morals depending on the situation in machines. Such inconsistencies can increase ethical risk as opposed to reducing it when applied to autonomous systems that run continuously and at high speed.

The experimental results of this paper thus are sufficient evidence that the complementary or alternative ethical basis of AI alignment should be investigated. Specifically, value centred and systematic ethical frameworks such as the IKS-based ones present principles based on long-term responsibility, group well-being, and responsibility continuity. Moreover, the results endorse the construction of limbic-based ethical reasoning designs incorporating affective and value-driven processes with cognitive reasoning. Put collectively, these guidelines point to positive ways of building stronger, more principled and socially oriented ethical machine intelligence.

4.6. Discussion

This paper empirically tested the issue of whether modern human ethical behaviour can be an effective and reliable source of consistent ethical decision-making in artificial intelligence systems. The results offer valuable understanding of the constraints of the human-centric ethical alignment methods and directly impact the process of designing future ethical AI architectures.

The findings indicate that most of the respondents are ethical in a single situation alone, but ethical integrity in a variety of situations is uncommon. Only a small percentage (around 4) of the respondents continued to be ethically

aligned in all the dimensions that were surveyed. This result calls into question a widely unspoken assumption in AI ethical research: the view that human behaviour as a group follows a consistent and principled ethical ground state. The ethical inconsistency experienced in the data is indicative of well-reported attributes of human decision-making such as situational reasoning, self-interest, moral licensing and trade-offs. To a human, this form of flexibility could be adaptive, but in the view of machine ethics, this creates an unwanted ambiguity because autonomous systems should act in predictable and uniform moral manners.

Human-in-the-loop learning and Human Feedback Reinforcement Learning (RLHF) have become the prominent alignment options for modern AI. Although these solutions are useful in achieving conformity on the surface, or less profound behaviours like helpfulness, conversational tone and so on, the current results indicate that there are inherent constraints to the signals of human ethical behaviour as the dominant ethical indicator.

The prevalence of ethical ambiguity (appearances of the 'Maybe' option) and ethical compromise ('Yes' option to unethical conduct) is observed, which means that human feedback can be situation-dependent and self-inconsistent. When implemented directly in machine learning systems without ethical constraints that include a structured sense of morality, using this data will increase the risk of introducing moral relativism, immediate self-interest, and inconsistency into AI decision-making.

The interesting observation of the study is that there was a discrepancy between the self-rating of the ethical rating of the respondents and their perceived ethical behaviour. Some of the respondents rated themselves highly but admitted that they have been involved in ethically dubious behaviours. This disparity highlights one of the most serious threats of ethical AI education: self-reported values not always correlating with ethical behavior.

The only issue with using statements of declarative morality in the training of AI systems is that they may not mimic real-life examples of moral conduct, once more reinforcing the need for well-defined and externally verified ethical priors.

The argument is rather convincing, and the empirical evidence is quite strong that human ethical behaviour in its own is not an ethical tuning dataset that can be applied in isolation by AI systems. Ethical AI must be adaptable but consistent, have long-term responsibility, and must be resistant to situational bias. Well-organized ethical systems - especially ones focusing on responsibility, the outcomes of actions, and the welfare of the group- are a potential solution. It is a component of those frameworks, in contrast to the ad-hoc human preferences, that offer internally satisfying ethical conferences; this allows the assessment of the decisions by the machines beyond the immediate projections.

5. Conclusion

The purpose of this paper was to empirically quantify a hypothesis of whether or not human ethical behavior today can be a plausible and accountable foundation of artificial intelligence systems to calibrate ethical decision-making. The study targets the unveiling of a qualitative, cross-demographic survey, which discloses clear facts that human ethical behaviour, as it is now, is not consistent enough to be directly and solely employed in ethical AI training.

The empirical research study found that although people tend to display ethical reasoning in single situations, ethical consistency in various decision situations is not common. A major percentage of the surveyed respondents admitted to ethical compromise or ambiguity, and only an extremely small percentage displayed steady ethical alignment across all of the dimensions surveyed. In addition, the discrepancies between perceived values and actual behaviors of the ethical machines also point to the demerits of the declarative values of human beings to harmonise the ethical machines.

These results, in ethic terms, are important as far as AI is concerned. Even though the human-in-the-loop and preference-based alignment are also good regarding their capability to adjust autonomous systems in contexts, they may add inconsistency, bias and situational self-interest to autonomous systems were they to be trusted with the role of the sole muse of ethics.

The determinism, stability, and long-term responsibility needed by ethical AI systems, by definition, cannot always be seen in the actions of contemporary human ethical behaviour. This piece of work thus gives a solid empirical ground for investigating structured ethical knowledge systems as either a complementary or alternative source of ethical AI design. Specifically, the ethical frameworks that are based on the IKS, i.e., based on such principles as duty, long-term causal responsibility, and non-harm, provide internally consistent and non-situational ethical constructs that directly address the gaps that were identified in the human-centric ethical data. In combination with limbic-inspired architecture of AI that representations affective and value-based logic, these architectures represent a way to a future where artificial intelligence is made ethically sound and socially responsible.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

No external funding is taken for this work.

References

[1] David De Cremer, and Devesh Narayanan, "How AI Tools can-and cannot-Help Organizations become More Ethical," *Frontiers in Artificial Intelligence*, vol. 6, pp.1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Joel Janhonen, "Socialisation Approach to AI Value Acquisition: Enabling Flexible Ethical Navigation with Built-in Receptiveness to Social Influence," *AI and Ethics*, vol. 5, no. 1, pp. 527-553, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[3] José Machado et al., "Ethical Decision-Making in Artificial Intelligence: A Logic Programming Approach," *AI*, vol. 5, no. 4, pp. 2707-2724, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Adam Dahlgren Lindström et al., "Helpful, Harmless, Honest? Sociotechnical Limits of AI Alignment and Safety through Reinforcement Learning from Human Feedback," *Ethics and Information Technology*, vol. 27, no. 2, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[5] Teresa Heyder, Nina Passlack, and Oliver Posegga, "Ethical Management of Human-AI Interaction: Theory Development Review," *The Journal of Strategic Information Systems*, vol. 32, no. 3, pp. 1-50, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Rashmi Yogesh Pai et al., "Integrating Artificial Intelligence for Knowledge Management Systems – Synergy among People and Technology: A Systematic Review of the Evidence," *Economic Research*, vol. 35, no. 1, pp. 7043-7065, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Rajesh Timane, "Integrating Indian Knowledge Systems into Modern Multidisciplinary AI Research: Strategic Pathways under NEP 2020," *SSRN*, pp. 1-5, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389-399, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[9] Luciano Floridi et al., "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689-707, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[10] Nicholas Kluge Corrêa et al., "Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance," *Patterns*, vol. 4, no. 10, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Long Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," *arXiv preprint*, pp. 1-68, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Saffron Huang et al., "Collective Constitutional AI: Aligning a Language Model with Public Input," *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, Rio de Janeiro Brazil, pp. 1395-1417, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Wei Shen et al., "Loose Lips Sink Ships: Mitigating Length Bias in Reinforcement Learning from Human Feedback," *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2859-2873, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Solon Barocas, and Andrew D. Selbst, "Big Data's Disparate Impact," *SSRN*, vol. 104, pp. 671-732, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[15] Ninareh Mehrabi et al., "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1-35, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Guanxiong Pei et al., "Affective Computing: Recent Advances, Challenges, and Future Trends," *Intelligent Computing*, vol. 3, pp. 1-24, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Sitara Afzal et al., "A Comprehensive Survey on Affective Computing: Challenges, Trends, Applications, and Future Directions," *IEEE Access*, vol. 12, pp. 96150-96168, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Ammar Younas, "Toward a Cultural Co-Genesis of AI Ethics," *arXiv preprint*, pp. 1-17, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] Salim Khan, and Meeta Sharma, "An Overview on Indian Knowledge System," *Integrated Journal for Research in Arts and Humanities*, vol. 4, no. 4, pp. 42-46, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Jessica Morley et al., "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Science and Engineering Ethics*, vol. 26, no. 4, pp. 2141-2168, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[21] Shalaleh Rismani et al., "Responsible AI Measures Dataset for Ethics Evaluation of AI Systems," *Scientific Data*, vol. 12, no. 1, pp. 1-12, 2025. [CrossRef] [Google Scholar] [Publisher Link]