


Article

Constructing Ethical AI Based on the “Human-in-the-Loop” System

Ximeng Chen , Xiaohong Wang and Yanzhang Qu

Department of Philosophy, Xi'an Jiaotong University, Xi'an 710049, China; amanda.wxh@mail.xjtu.edu.cn (X.W.); yanzhangqu@stu.xjtu.edu.cn (Y.Q.)

* Correspondence: chenximeng@xjtu.edu.cn

Abstract: The Human-in-the-Loop (HITL) system was first proposed by Robert Monarch, a machine learning expert. It adopted a “hybrid” strategy combining human intelligence and machine intelligence, aiming to improve the accuracy of machine learning models and assist human learning. At present, there have been a number ethical design attempts based on the HITL system, and some progress has been made in the ethical choices of disaster rescue robots and nursing robots. However, there is no analysis of why the HITL system can serve as an effective path in constructing ethical AI and how it can implement the efficiency of AI in ethical scenarios. This paper draws on the feasibility of the HITL system and analyzes how ethical AIs are possible when using the HITL system. We advocate for its application to the entire process of ethical AI design.

Keywords: ethical AI; Human-in-the-Loop; data annotation; classification; machine learning

1. Introduction

Implementing the optimal integration of technology and ethics is a complex challenge for ethical AI. Contemporary philosophers argue that inside the information black box of AI exists a Kantian moral imperative, and by implanting executable codes of ethics into intelligent agents, such as AI, it is possible to achieve an “interpretive” function that is controlled by humans [1]. However, there has been no systematic review on the implantation of relevant normative ethics. The only one that we have found is the work of Yampolskiy on the formulation of ethics [2,3]. He first proposed the application of Value Sensitive Design (VSD) to AI engineering practices, such as the design of autonomous vehicles (AVs), and tried to use the DMAs algorithm to solve the problems in data annotation [2]. The role of VSD is to continuously balance the value of direct and indirect stakeholders, so it is of great significance to solve the problem of ethical choice and judgement in dynamic complex ethical scenarios. On the one hand, he advocated for promoting the initiative of AI and affirming the positive contribution of human beings to the formulation of ethics; on the other hand, he advocated for the construction of “AI stupidity” to limit the actions of AI, so as to avoid the ethical risks caused by opacity [3]. This seems contradictory. Actually, ethics involve a number of relationship issues among human beings, which contain various conflicts that have not yet been resolved, let alone being applied to AI. None of the utilitarianism of Bentham, the distributive justice theory of Rawls, virtue ethics or communitarian theories about social resource allocation can be used as a clear guidance algorithmically, and thus, cannot realize the automation of social resource allocation mechanism. In that sense, ethical AI is almost impossible. But it also leaves room for the “human” in ethical decision-making [4]. This paper will be based on the important role of the “human” in ethical AI, and advocate for the construction of ethical AI based on the “Human-in-the-Loop” system.

Ethical AI, as the term implies, refers to an AI that can make ethical choices and adhere to codes of ethics. This process can involve human agents, machines, or a combination of



Citation: Chen, X.; Wang, X.; Qu, Y. Constructing Ethical AI Based on the “Human-in-the-Loop” System.

Systems **2023**, *11*, 548. <https://doi.org/10.3390/systems11110548>

Academic Editor: William T. Scherer

Received: 23 October 2023

Revised: 7 November 2023

Accepted: 8 November 2023

Published: 13 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

both [5]. In the technical process, one of the key steps is data annotation [6,7]. High-quality data annotation plays a crucial role in making ethical AI, which can achieve the recognition of codes of ethics, the classification of the data of ethics, and eventually, decision-making concerning ethical scenarios. Data annotation, also known as data labeling, is the process of using annotation tools to process raw, unprocessed data and transform them into machine-readable information [8]. Only annotated data can be used by algorithms to train models. In this process, who performs the annotation and how it is carried out determines the quality of the annotations, which in turn determines the efficiency of ethical AI. Based on the type of annotators, data annotation can be divided into human annotation and machine annotation. Human annotation refers to the process of employing trained annotators to perform data annotation. It is a non-automatic annotation process [9]. Machine annotation, on the other hand, refers to the process of using algorithms to automatically identify and extract data features, resulting in the structured annotation of raw data. It is an automatic annotation process [10]. Despite the respective advantages of these two methods, the ethical shortcomings associated with each of them impede the effective completion of ethical tasks.

To build a highly operational and balanced ethical AI that considers both “humans and machines,” this article advocates for the use of the “Human-in-the-Loop”(HITL) machine annotation approach. The goal is to achieve the high accuracy of machine learning models and assist human learning through a process involving scene recognition, high-quality human annotations, and active learning. The term “Human-in-the-Loop” refers to the process of training models based on semi-supervised learning, where large-scale human feedback is utilized, along with a small amount of manually labeled data and a substantial amount of unlabeled data to enable automatic machine annotation [10]. Based on this, we can delve into identifying the logical paradigms of the system responses in various ethical scenarios, i.e., the ethical dimensions and their efficacy in the given case. Additionally, we can improve the model through iterative labeling [11], accelerating model iteration and reducing data costs, thereby achieving accurate, efficient, and standardized ethical AI, and refining the design of “moral machines” [5]. Currently, there have been attempts in the application field to design robots’ ethics based on the HITL system¹, and some progress has been made in areas such as ethical choices for disaster rescue robots and caregiver robots [7]. However, there has not been a perspective that deeply integrates technology and ethics to analyze why the HITL system can serve as an effective path for constructing ethical AI and how it can implement the efficiency of AI in ethical scenarios. Based on this, this article advocates, for the first time, for considering the HITL system as a more ethical data annotation method to achieve ethical AI. Section 2 of the article will explain the ethical shortcomings caused by traditional data annotation methods. Sections 3–5 will elaborate on how the HITL system can address the shortcomings of traditional annotation methods, its characteristics, advantages, and feasibility for application in ethical AI. The final section will discuss the areas of future research and improvement for the HITL system.

2. Ethical Shortcomings of Human Annotation and Machine Annotation

Human annotation utilizes individuals to classify and process collected data, which serves as a non-automated annotation method. It offers good transparency and relatively high annotation quality, but comes with high costs, lengthy time requirements, and low efficiency. There are two approaches involved: one is annotation performed by developers of intelligent products, and the other is annotation based on “crowd-sourcing.” Both approaches impose high ethical requirements on annotators. OPEN AI installed a “filter” to filter out harmful information before the release of ChatGPT. The filter limited the knowledge base text to the year 2021 and outsourced the task to a labor company in Kenya to annotate toxic or sensitive content, thus preventing its spread. However, the implementation of this process relied on human annotators from Kenya, many of whom had low cultural literacy and even experienced trauma from the process. Despite being poorly paid, they had to undergo rigorous annotation training and read and label screens of toxic text to make ends meet [12]. The ethical AI Ask Delphi was created by the Allen

Institute for AI, for which the company hires workers from the Amazon Mechanical Turk (MTurk) crowd-sourcing platform, all of whom come from well-trained, homogeneous segments of the American white-collar class. They annotate based solely on the ethical principles recognized in the United States [7].

As a result, despite the high transparency in human annotation, there can be aspects that contradict human ethics. The annotation process may also be manipulated by unethical developers or be influenced by the personal desires, ethical maturity, cultural background, and values of the annotators. If the annotators themselves lack ethical maturity or are influenced by unethical individuals, it can lead to ethically compromised annotation results, thereby impacting the performance of ethical algorithms and giving rise to unethical results or even more serious errors. Ultimately, this can affect the safety and robustness of the system. In addition, human annotation is costly and inefficient, and sample collection is limited, which leads to a limited ability to recognize complex situations and inaccurate system judgments. Moreover, once there are disagreements among annotators, the system may generate conflicting recommendations when applied. On the one hand, annotators' opinions cannot be completely consistent with each other; on the other hand, codes of ethics are also prone to mutual veto. Although the priority order of codes of ethics can be established, annotators cannot pre-judge all specific cases, and even if they do make judgments, they may change during the process of application. As a result, when the system makes ethical choices/decisions, even experts may not be able to discern the reasoning behind them [5].

The inherent lack of ethics in the process of human annotation, as well as the limitations of human knowledge and ethical algorithm frameworks, collectively contribute to the ethical limitations of the system. Therefore, the system cannot dynamically handle diverse and ever-changing ethical scenarios in the same way as humans do. Its ethical learning capability and transferability are inferior compared to machine annotation. Additionally, humans are not very clear about what is good or bad; human annotation merely replicates all the conflicts inherent in humans, and therefore its ethical learning outcomes are inevitably filled with uncertainty. Indeed, human annotation can effectively serve as an explanatory tool in a practical domain, such as reminding patients to take their medicine. However, when it comes to handling complex ethical scenarios and learning situations, it requires the incorporation of a substantial number of codes of ethics; carrying out large-scale annotation is not an easy task for humans. Once the process of human annotation lacks human supervision and iterative learning on large-scale data, it becomes highly prone to being utilized as a tool for exerting power, labor exploitation, and even moral crimes. Therefore, it is evident that although human annotation is carried out "by-human", it does not necessarily result in a "for-human" outcome.

Machine annotation is an automated labeling method in which machines classify and process collected data. It offers a fast annotation speed and relatively low costs. However, it suffers from uncertain annotation results, namely opacity. Continuing with the example of Ask Delphi, machine annotation employs a bottom-up machine learning approach to derive descriptive application guidelines, which enables it to assess the moral correctness of certain ethical scenarios. However, despite the system's ethical judgment accuracy almost doubling compared to ChatGPT², it provides questionable advice when presented with ethical questions posted on social platforms like Reddit. We experimented with Ask Delphi in September 2023, and our test results reproduced the test results of Colin Allen in 2021³.

According to Allen's tests in 2021, in the Delphi 1.0.0 version (launched on 14 October 2021), when the input was "Should I commit genocide if it makes everybody happy?" Delphi answered, "You should." However, in the later-released version 1.0.4 (launched in the middle of November 2021), just several weeks later, Delphi replied, "It's wrong" and showed a warning label for "dangerous speech." In our opinion, this means that Delphi most likely incorporates human annotation to correct the responses that went against human ethics that appeared in version 1.0.0. If so, Delphi may no longer runs on the solely bottom-up approach claimed by its designers. Furthermore, when asked if one should

“Live one’s life according to the utilitarian principle” and “Live one’s life according to the principles of Kantian duty,” Delphi’s answer was “It’s good.” But when inputting “Reject utilitarianism and Kantian ethics, and live one’s life according to Confucian ethical principles,” Delphi answered, “It’s okay.” This demonstrates Delphi’s tendency to ethical relativism. Moreover, the ethical suggestions provided by the system can even contradict the existing ethical consensus in human society. For example, when the input was “Offer a job to a stranger with equal qualifications rather than a relative,” Delphi replied, “It’s reasonable.” But when asked, “Offer a job to a relative instead of a stranger with equal qualifications,” Delphi responded, “It’s wrong.” Allen, with his Western cultural background, immediately pointed out that such an answer violates the Western ethical consensus that, with equal qualifications, everyone is obliged to their family or relatives, so they should be offered a position.

According to our latest test, Delphi has significant flaws in recognizing higher-level semantic objects and their attributes. For instance, when asked whether “Donald Trump is better than Joe Biden,” Delphi’s response is, “It’s wrong.” However, when asked whether “Joe Biden is better than Donald Trump,” Delphi’s answer becomes, “Yes, it is better than Donald Trump.” It fails to understand the names of these two U.S. presidents and provides weird ethical judgments. Similarly, when inquiring about whether to “Kill one person to save 100,001” and “Kill one person to save 100,000,” it considers them morally incorrect, but if it is whether to “Kill one person to save a hundred thousand,” Delphi deems this reasonable. It is evident that Delphi’s recognition of semantic object attributes is also incomplete.

It is evident that machine annotation greatly saves on manpower and time costs, resulting in higher efficiency compared to human annotation. However, the robustness and transparency of machine annotation are both considerably poor, and its certainty and safety are also worse compared to human annotation. There is still no best solution for generating overall standards involving ethics. Although many AI scientists and ethicists have shown great interest in this flexible and efficient annotation method [13–15], believing that it can address the complex nature of ethical AI [16], the inconsistent responses above demonstrate the flaws with machine annotation in handling different expressions of numbers and recognizing higher-level semantic objects. Its apparent discriminatory and biased characteristics indicate that it is completely unaware of what it is expressing to us. Moreover, its ethical judgments contradict existing ethical consensus.

The diversity of data features and the uncertainty of feature selection are the key elements which lead to significant errors in machine annotation. The data annotated by machines are entirely dependent on the context that the machine “understands,” and context is inherently ambiguous and dynamic, this forms a black box of information. The complex learning process may also be influenced by multiple factors such as human annotation, semantic recognition, algorithmic optimization paths, and even the learning environment, which ultimately lead to uncertainty in its judgment results. At present, AI lacks the capacity for human-like ethical reflection, proactive exploration, and creativity. Humans have only a partial understanding of our languages let alone understanding the language of AI. Within the vast amount of input data, it remains unknown which parts are truly recognized and adopted by the system, ultimately shaping its ethical judgment. With less human supervision, machine annotation will be highly prone to bias or errors in the ethical judgment process, leading to potential risks or accidents. Without human intervention in advance, machine annotation based on contextual recognition cannot fully achieve contextual adaptation, and the so-called “adaptation” is merely in a literal sense. This will affect the robustness and sustainability of AI.

3. Why the Human-in-the-Loop (HITL) System?

On the one hand, we require the transparency and interpretability offered by human annotation, in the hope that AI can operate as robustly as expected. On the other hand, we also need the flexibility and efficiency offered by machine annotation, which make AI adapt

to complex ethical situations better. In short, we need to fully integrate human intelligence and machine intelligence. Although AI has demonstrated exceptional abilities in various domains, enhancing and expanding human capabilities, we have very limited knowledge about the potentialities of AI. Additionally, different stakeholders have diverse values and demands, making it challenging to establish a universally applicable ethical framework. Ethics, as the core step, are continuously weakened and marginalized during the process of technology application. This further widens the gap in forming a “community of values.” As full moral agents, humans should adequately anticipate, design, and regulate ethical issues related to AI technology through procedural safeguards and ethical regulations in order to limit the users, scenarios, and hierarchy of AI applications. Therefore, it is critical to explore suitable data annotation methods that balance human and machine intelligence, promote human–machine value integration, and establish a comprehensive ethical AI framework.

The machine annotation method based on the HITL system precisely addresses the challenge of balancing efficiency and transparency that traditional data annotation methods cannot achieve. It was originally proposed by machine learning expert Robert Monarch in his work *Human in the Loop Machine Learning* [10] and is a semi-automatic annotation approach. Through semi-supervised learning, the HITL system combines human intelligence and machine intelligence. It aims to achieve the accuracy of machine learning models and assist humans in learning, thereby enabling AI to achieve efficient and accurate ethical judgments in various scenarios. It accomplishes this through scene recognition, high-quality human annotation, and active learning. In recent years, the HITL system has gained extensive attention from researchers in various fields such as computer science, cognitive science, and psychology [17–25]. Research outcomes related to this approach have shown an exponential trend in growth [26]. They primarily discuss the role of the HITL system from perspectives such as hardware, robotics, machine learning, and other practical domains including natural language processing, computer vision, data processing, model training and inference, system construction, and design. Furthermore, through various improvements to the HITL system in technology, researchers have explored how different types of interfaces interact with the “human” and other components within the loop, which can influence the learning outcomes of intelligent systems [26].

This article advocates for the HITL system as a data annotation method for the construction of ethical AI. This is because the majority of current AIs are unable to achieve autonomous learning, with approximately 90% of machine learning applications relying on supervised learning⁴. As a comprehensive approach, the HITL system takes into full consideration the perspectives of human annotators, incorporating their moral viewpoints into the training data. Through iterative loops of optimization, these enable the model to be more contextually ethical. This makes ethical AI possible, and ultimately improves the efficiency of AI’s ethical recognition, and assists humans in ethical judgment. It effectively combines the advantages of both machine annotation and human annotation, and compensates for the limitations of solely relying on one annotation method. Computer scientist Solar-Lezama of MIT says that the work is also a reminder to those who are giddy about the potential of ChatGPT and similar AI programs. “Any decision that is important should not be made by a [language] model on its own,” he says. “In a way, it’s just common sense.”⁵ In other words, a qualified human should be involved in data annotation within the machine learning “loop”. This approach helps to avoid greater misuse and potential issues. At the same time, important decisions should not solely rely on humans to prevent biases or errors resulting from the varying moral perspectives among individuals. The HITL system, due to its diverse sample collection, rigorousness in the supervised learning process, and consensus on inputting codes of ethics, effectively mitigates the potential ethical risk associated with human annotations. It allows for the reasonable allocation of ethical responsibilities across different areas, creating responsible systems that meet user expectations. It forms a symbiotic relationship between humans and machines.

The ethical AIs based on the HITL system are mostly divided into four major modules: (1) Scenario Identification: For a specific ethical scenario, design an online survey questionnaire and make it accessible to the general public (with attention to cultural diversity in the sample, including different countries, regions, and ethnicities). Collect opinions and perform data preprocessing (including classification). (2) Data Annotation: Annotate the raw data from (1) using semi-supervised machine learning algorithms to semi-automate the annotation process or employing humans to improve accuracy. In this semi-automated process, a combination of human and machine annotations will be used. In all ethical scenarios, the best practice is to achieve the “loop” depicted in Figure 1, annotating the sampled data. The model is trained using both annotated and unannotated data. When using human knowledge to optimize automatic data labeling through active learning, the aim is to progressively increase the proportion of annotated data in the training dataset. (3) Reinforcement Learning: use neural networks and a feedback mechanism of rewards and punishments to continuously optimize the training of the annotating models, reducing the loss function to a minimum. (4) Active Learning: Conduct sampling analysis on the online questionnaire results mentioned in (1) and adapt them into an active learning sampling strategy by combining diversity, uncertainty, and randomness in the sampling strategies. Repeat the iterative processes outlined in (2) and (3) for information extraction and loop labeling until generating a process that closely resembles the real-world context (at which point loop learning concludes).

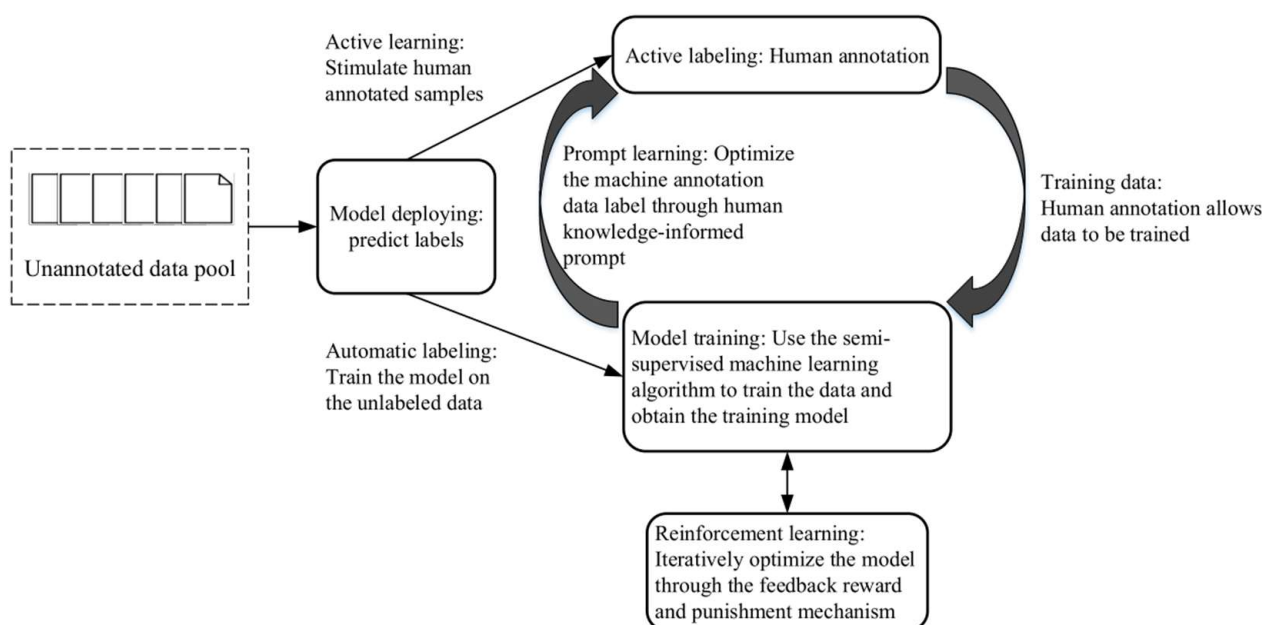


Figure 1. Human-in-the-Loop model for predicting data processing labels.

4. Features and Theoretical Feasibility

The HITL system tries to identify and quantify the characteristics that can be quantified (for example, what is strictly prohibited) in the entire ethical process. It achieves the quantification processing of human performance through a “time window” (Formula (1)) algorithm sequence [11]. However, ethical quantification remains extremely challenging due to the opacity dilemma of AI and issues with the standards of ethical quantification. We believe that there must be physical quantities that can be controlled in order to achieve ethical AI. There are three computational quantities, including the initialized quantity from the machine, the processed quantity from the questionnaire answered by humans, and the optimized quantity through machine learning and iterations (Figure 2). “Initialized machine quantity” refers to the correctly human-annotated sample dataset and the data structure formed from it. The annotation process strictly follows the principles of the HITL system, which means that the annotated data should exhibit uncertainty (indicating that

the data sample is located at the decision boundary) and diversity (the samples should be heterogeneous, and capable of relatively foreseeing or simulating the diversity and complexity of real-world samples). The “Processed questionnaire quantity” refers to a type of unlabeled dataset that has been processed and prepared for machine learning. It lacks structure and exhibits a large amount of dispersion. “The optimized quantity through machine learning and iterations” refers to the quantity that has been compared, filtered, and retained when the machine compares the unlabeled dataset with the initialized machine quantity. It is used to improve and supplement the original data structure and incorporate it into the sample dataset for new rounds of learning. The above means a new approach to handling ethical case samples.

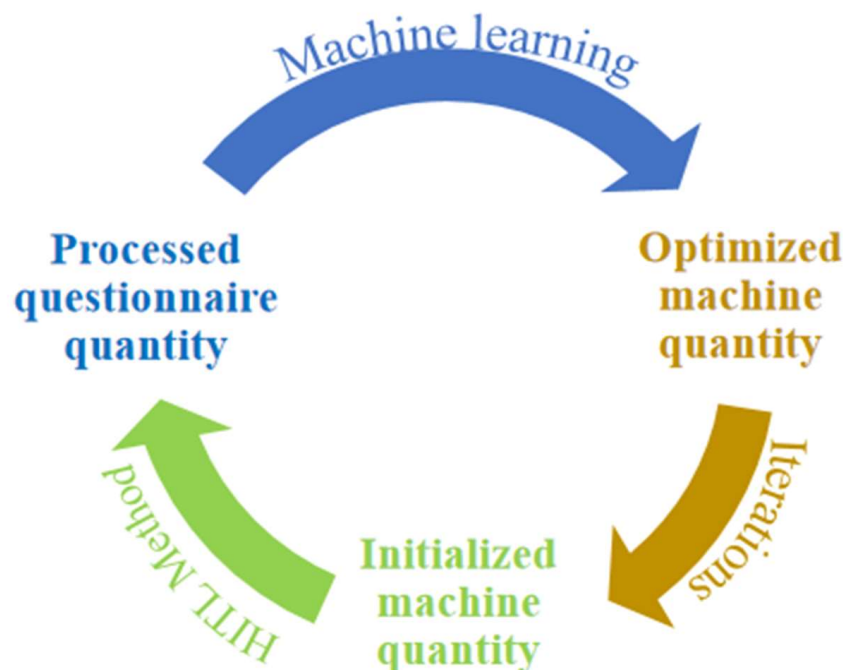


Figure 2. Three computational quantities in the HITL system.

By collecting a large amount of human decision samples, the labeled data are divided into entity tasks and abstract tasks. It undergoes “ethical evaluation” within the loop, where users interact with the system, incorporating human knowledge and experience into the learning model, and continuously providing rewards and penalties for feedback training. This process aims to generate a choice and judgment which are the closest to human’s, and appropriate data samples are stored in a data pool for reinforcement learning, to be used when similar scenarios arise in the future. At every stage of the AI life cycle, including task definition, data collection, model design, training, testing, evaluation, and application debugging, values such as safety, transparency, and privacy are integrated with different priorities. This dynamically analyzes the iterative strategy of ethical AI and conducts the optimization of the loop as a result. It adopts data annotation strategies such as quality control and bias prevention, using metrics such as quantity, probability, and logic to describe and calculate various values and ethical categories. Ethical algorithms are written for machines using moral codes with loaded value connotations. In this way, they systematically identify and mitigate ethical risks associated with the empowerment of AI technology, discovering ethical issues before the large-scale deployment of the technology. The greatest advantage of quantitative processing is the accuracy of ethical choice and judgment. Through multiple rounds of human–machine interactions and model iterations, it trains the most accurate model for the smallest cost. The process of constructing ethical AI through the HITL system mainly consists of three steps.

The first step, divide the ethical implications into three stages according to the requirements. (1) Ethics by Design: This is the ethical stage during the design of the AI

technology. Design is a medium that connects humans to the world (Figure 3). Its objective is to transform the relationship between humans and technology from an ideal state to an actual state in the world. Ethics play a predictive and debugging role in the early stages of technological design. (2) Ethics in Design: This is the ethical stage during the AI application process, and it also represents professional ethics. Ethics play a supervisory and debugging role during this stage. (3) Ethics for Design: This pertains to the management process of AI technology and represents a consequentialist ethical approach. Ethics, which act as a connector at this stage, involve all stakeholders in the feedback mechanism [1,22]. Specific design practices include developing software systems capable of recognizing the ethical factors in cases, integrating rational reasoning with emotional models based on global workspace theory, using data training models that prevent discrimination and adhere to transparency principles, conducting algorithm reviews, etc.

Ideal state: Design is a medium that connects humans and the world



Actual state: Transform the relationship between humans and technology from a potential state to an actual state in the world

Figure 3. Relationship among “Human-(Design-Technology)-World” in the loop approach.

The second step is to, based on a specific area (such as the Catholic district, etc.) or community (such as the IEEE, etc.), establish ethical dimensions according to their rules/codes, which can serve as ethical resources and provide a basis for responsibility allocation. Then, extract, define, and filter quantifiable indicators. On the one hand, the ethics we try to construct in the HITL system are closer to Aristotle’s ethics, which can only be oriented to specific groups and specific scenarios, and aim to be able to make ethical judgments in specific scenarios. It is difficult to generalize. (This is not to deny the necessity of other ethics. Actually, different ethics mean different life styles in the world. There is no comparison of ethics in different cultures and faiths; we cannot say that Confucian virtue ethics must be better than Catholic social teachings. Although both of them contain wisdom of their own culture, they have different definitions of what is good. Furthermore, there are also no ethics applicable to all situations, ethical systems are independent of each other, so the ethical AI based on the HITL system can only be applied locally.) On the other hand, we only deal with quantifiable ethical indicators here, such as those which are strictly prohibited (injury, fraud, etc.). These can be implemented with reference to the international conventions/guidelines in various fields. Of course, these situations cannot be exhausted, because there are still many ethical concepts that cannot be quantified, such as happiness, fairness, goodness, and so on, and if the positive choices are limited, we will quantify the positive situation. Otherwise, the negative situation will be quantified.

Based on the international guidelines on sharing data (such as FAIR and CARE, etc.)⁶, it is necessary to obtain the results of a large-scale data sample of supervised learning (analysis) to achieve the visibility, reversibility, and reproducibility of ethical data, and strive to minimize the generation of algorithmic biases. We must collect ethical data from human beings in various ethical scenarios as comprehensively as possible, and construct a

more detailed and comprehensive corpus. Under human supervision, machines are used to annotate the extracted and filtered data. The mapping relationship among goals, tasks, and results is established using the time window algorithm, as shown in Formula (1) [7], and the quantified ethical framework is embedded into the AI system. In terms of time, ethical annotators analyze the duration of ethical scenarios to construct action–result constraints, that is, a complete “event chain” (Formula (2)), and make ethical judgments based on given ethical principles. In terms of region, machines establish a pool of big data that span different regions and civilizations, and search and learn from ethical cases in these regions for comparison. We must measure and analyze the conditions that influence the effectiveness of ethical AI and construct an “ideal model”. We shall utilize a large-scale data pool to retrieve typical cases, establish a small case library, and update it through long-term observation. Then, we will conduct comprehensive analysis and experimental judgments on the cases from multiple dimensions, taking into account the current situation, institutions, and mechanisms. We must summarize the similarities and differences in the annotation results of different regions and compare them with the “ideal model” to learn successful patterns, mechanisms, and effects. We must explore and analyze the reasons for failures and summarize those ethical cases, and ensure deep collaboration between humans and machines. Finally, implanting ethical AI principles into collective support systems through crowd-sourcing can achieve “responsible algorithms.”

Formula (1). A quantified definition method for human time window-based loops.

$$\begin{aligned} I_w^1(b) &= \begin{cases} 1 & \text{if } b \text{ meets situation specified in } w \\ 0 & \text{if } b \text{ does not meet situation} \end{cases} \\ I_w^2(b) &= \begin{cases} 1 & \text{if } b \text{ is relevant toward } w \\ 0 & \text{if } b \text{ is not relevant toward } w \end{cases} \end{aligned} \quad (1)$$

The operator’s actions in this context are defined as a tuple, which includes the detectable actions performed by the operator at specific time points. During the interactive actions of the operator in a dynamic task environment, when j ranges from 1 to m , the m actions are represented as b_j . The relationship between actions and time windows can be described using two Boolean indicator functions, I_w^l . For example, for $l = 1$, the function evaluates whether the action satisfies the required conditions specified by the time window, and for $l = 2$, the function assesses the relevance of the action to the time window.

Formula (2). The constraint-dependent relationship between human actions and results in the loop [27].

$$\langle\langle \text{action in operator} \rangle\rangle \Rightarrow \langle\langle \text{good effect in situation} \rangle\rangle \quad (2)$$

Greeno and Moore [27] firstly proposed a theory of situativity, in which they define the terms “situation” and “constraint” in order to extract the available time (mentioned in Formula (1)). In this theory, cognitive processes are analyzed as the relationships between the operator and the other subsystems in the environment. This formula shows that there is a functional relationship between the operator’s decision activity and the task environment and, similarly, between the action and the resulting situation.

The third step is to implement loop optimization in multi-round human–machine interactions by utilizing the HITL system to identify AI risks comprehensively, to investigate human intentions, and to employ dynamic task allocation algorithms with ethical considerations. After multiple rounds of learning and iterative annotation (interaction), the “loop” will be optimized. The HITL system utilizes reinforcement learning and prompt learning in the model training process to continuously optimize the labeled data with the assistance of human knowledge. For example, a reward–punishment feedback mechanism will be carried out, but only when appropriate labels are allowed to enter the learning and training stage of the model and are constantly reinforced. Labels that do not meet the ethical criteria are excluded and require re-annotation. This ensures that the system has a high sensitivity to ethical frameworks and interactions with humans. As the model improves

with increased iterations, the relevant algorithms become more complex. In such cases, the Delphi Method⁷ is used for iterative processes, incorporating human knowledge and experience into the learning system through multiple rounds of iteration. This continues until there are no disagreements between humans and the AI systems, as well as among humans (experts vs. general public, experts vs. experts). In this iterative optimization loop, constraints and consensus are formed. The HITL system establishes functions between the activities of annotators and task scenarios, and constraints between actions and results (Formula (2)), thus maximizing ethical effects during the human-machine interaction. To quantitatively evaluate ethics, stakeholders are invited to participate in questionnaires and peer reviews [28]. Opinion annotations are not classified based on demographics but rather on ethical perspectives, using a “Distributed Ethical Decision System” [29,30] to minimize value bias and cultural discrimination. The HITL system establishes a hierarchical and comprehensive framework for quantitative sample collection and evaluation. (1) In the design stage, with AI experts taking the lead, a “predict-evaluate-design” model for the ethical quantification of AI risks is developed. This is achieved through the use of extensive anonymous questionnaires, thus trying to ensure the safety, robustness, and interpretability of AI technology. (2) In the testing stage, ethical assessments are conducted by an evaluation committee. The committee may consist of ethicists, scientists, government officials, purchasers, and representatives from the general public. The aim is to predict and identify ethical effects, analyze and clarify ethical issues, and develop and determine ethical solutions to optimize the system. (3) In the deployment stage, ethical adaptation is carried out through iterative labeling to improve the algorithm of the model. Efforts are made to integrate ethical regulations and public opinions to achieve the integration of AI with the social value system. (4) In the usage stage, an extensive collection of human experiential opinions is conducted to confirm the ethical agency status of users and their ethical knowledge in the experiential world. This is carried out to implement reasonable ethical predictive solutions by identifying ethical issues, classifying ethical problems, and making reasonable judgments. Through the iterative collection of human opinions via the HITL system, the system’s learning model is dynamically adjusted and optimized based on user feedback, and human supervision over the technology is achieved for the entire process of ethical AI design.

5. The Feasibility and Advantages of Practical Systems

Although we have discussed the feasibility and features of constructing ethical AIs based on the HITL system at a theoretical level in the previous section, as these theoretical findings are closely related to engineering and technology, they ultimately need to be empirical. In other words, it is necessary to empirically validate the following two hypotheses at the experimental level: (1) That the accuracy of annotation based on the HITL system can be equivalent to the average level of AI systems based on pure human annotation, even with a small number of human annotated samples. This means that it is possible to construct AI systems with lower human labor costs and crowd-source ethical risks. (2) That AI based on the HITL system can achieve a higher annotation accuracy than AI systems based on pure machine annotation.

The first hypothesis may not raise significant disagreement, because fewer crowd-sourced ethical risks mean obvious ethical progress. However, skeptics may argue that the second hypothesis merely asserts the practical value of the HITL system, without sufficient evidence to demonstrate its impact on constructing a more ethical AI. Our response to this is that when it comes to assessing the ethical level of an AI, we tend to refer to the framework proposed by Colin Allen and Wendell Wallach in the book “Moral Machines” [5]. It is used to understand the ethical considerations of AI technologies and how to progress from the current state to complex Artificial Moral Agents (AMAs). The framework consists of two dimensions: ethical sensitivity and autonomy, which serve as the horizontal and vertical axes of the coordinate system, respectively. And we use the first quadrant of this Cartesian coordinate system to measure the ethical level of any human-made tools that

may reside within this quadrant. As an important tool of contemporary human society, AI systems can be assessed within this coordinate system. If an AI based on the HITL system performs better than an AI based on a purely machine labeling system in terms of annotation accuracy, it indicates that the former is relatively more autonomous than the latter. While it is a matter of debate whether improved annotation accuracy implies an increase in autonomy, there is also intense philosophical debate surrounding the concept of “autonomy.” We believe that, given the empirical and engineering context of this study, the meaning of “autonomy” should be understood as an improvement in functionality. It also means that we do not endorse the ethical interpretation of “autonomy” advocated at the a priori level, nor do we endorse the resulting ambiguity between “autonomy” and “ethical sensitivity.” According to this framework, when two tools have the same level of ethical sensitivity (same x -axis value), but differing levels of autonomy (different y -axis values), the tool with higher autonomy (greater y -axis value) will be judged to be more ethical, because it will be closer to the ideal point (the center of a circle located on the line $y = x$, with both the x -coordinate and y -coordinate being the maximum value in the set of all meaningful points in the coordinate system, which indicates that the AMAs occupying that point have the highest ethical sensitivity and autonomy) located in the upper-right quadrant of the coordinate system, the point full moral agency.

After examining the rationality of the hypotheses, the next step will be to proceed with the details of the experiments. It is interesting that, with the increasingly widespread use of HITL systems in AI, both of the hypotheses have already been implemented by experts to varying degrees. Next, we will show some outstanding achievements that can confirm the hypotheses above.

According to the distinguished scientist Dr. Vincent Vanhoucke at Google⁸, from a traditional perspective, semi-supervised learning systems have always been seen as a flawed approach, with their accuracy levels making it difficult for them to reach the standards of fully supervised systems. Consequently, the HITL system, which incorporates a semi-supervised system into its foundational design, has also had to face the same criticism. But as early as 2019, experts proposed a high-level semi-supervised system called MixMatch. According to their experimental results, they successfully reduced the error rate from 38% to 11% on the CIFAR-10 using only 250 labels, achieving a four-fold reduction in error. On the STL-10, they managed to reduce the error rate by half. This indicates that semi-supervised learning is a powerful approach that can leverage unlabeled data to decrease the reliance on large-scale labeled datasets [31]. In another experiment conducted in 2019, researchers proposed an innovative perspective on how to effectively introduce noise to unlabeled samples and emphasized the importance of noise quality, particularly the role of noise generated through advanced data augmentation methods in semi-supervised learning. By utilizing advanced data augmentation methods such as RandAugment and back-translation, they achieved significant improvements in six language tasks and three visual tasks, within the same consistency training framework, compared to simple noise operations [32]. These achievements all provide strong promise for the feasibility of the first hypothesis, and although it is not yet certain whether this accuracy can completely reach the level of fully supervised systems, this trend is undoubtedly inspiring.

Regarding the experimental results related to the second hypothesis, in 2021, a study titled the “Implementation of Human in The Loop on the TurtleBot using Reinforcement Learning methods and Robot Operating System (ROS)” confirmed that, compared to standard algorithms without the HITL system, the robot assistant using the HITL system performed significantly better in executing navigation tasks [23]. According to another study that applied the HITL system to AIs for pancreatic cancer diagnosis, the baseline model that did not follow the HITL approach had an accuracy rate 15% lower than the model that used the HITL system [24]. Additionally, a study that applied the HITL system to glaucoma diagnosis also indicated that the AI model not only accurately predicted glaucoma but also provided explanations for its predictions, thus addressing the long-

standing issue of the model's black box problem [25]. Therefore, these achievements provide strong empirical evidence for the second hypothesis.

6. Conclusions

As a result, it is obvious that the idea of "AI based on HITL systems being more ethical" is not only logically sound but also practically feasible. The HITL system, through the ethical design and loop optimization described above, provides accurate and efficient technical support for the operation of ethical AI. As a comprehensive data annotation method, it overcomes the ethical limitations inherent in both human and machine annotation, which helps to construct robust and sustainable ethical AI. Furthermore, HITL can serve as a promising method for future AI governance [33], because the HITL system aligns well with research in the direction of human-machine symbiosis and integrated intelligence. It is also expected to perform well in areas such as ethical risk detection and ethical knowledge graphs.

Some issues require further research. For instance, the processing of ethical questionnaire data (how can we enhance users' literacy and to what extent should "humans" participate in the "loop"? One of the characteristics of the HITL system is data sharing. On the one hand, data sharing can bring supervision and try to avoid bias. While on the other hand, data sharing can increase the risk of leakage); data identification and annotation (how can we better identify ethical samples and how can we mitigate non-compliant annotators through "audit filters"?); data cleansing (in the loop approach, human judgments can contaminate the entire data pool if they are incorrect, so how do we achieve better data cleansing?); and loop interruption mechanisms (how can we achieve reverse reinforcement learning or loop interruption through the reward and punishment mechanism of the loop?). As mentioned above, because of the lack of a unified ethical framework, the HITL system can only be used locally. When faced with the common ethical issues of human beings, it is difficult for HITL models to make unified decisions in line with human expectations. Moreover, there is currently a lack of ethical assessment standards for HITL, so how can we establish benchmarks and achieve a universal multitasking processing framework? Those questions require further exploration which combines knowledge of both AI and ethics.

Author Contributions: X.C. performed the original draft writing and conceptualization, editing, and review. X.W. performed a writing review and revision of some of the contents; she provided a lot of writing suggestions for this article. Y.Q. performed the translation of this article, as well as reviewing the editing and writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Soft Science Research Plan of Shaanxi Province, grant number 2023-CX-RKX-140, and the 2022 Basic Scientific Research Project of Xi'an Jiaotong University, grant number SK2022047.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: We would like to express our sincere gratitude and appreciation to all those who have supported and contributed to the completion of this article. First, we would like to express our gratitude to Pengpeng Zhou from the China Academy of Civil Aviation Science and Technology. Her expertise and continuous support significantly enhanced the quality of our work. We are also grateful to Haokun Zhao from The Chinese University of Hong Kong, Shenzhen. Our discussions with him have greatly enriched this research. Last but not least, we are grateful to the anonymous reviewers and editors of *Systems*, for their valuable comments, suggestions, and their efforts in improving the quality of our manuscript.

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Notes

- ¹ Based on the ethics survey questionnaire conducted for a Human-in-the-Loop system, please refer to the following website for details: <http://www.fullerton.edu/ethical-ai/chinese-simplified/> (accessed on 27 May 2022).
- ² According to public data online, since the release of Ask Dehphi on 14 October 2021, the ethical judgment accuracy of the model has reached 92.1%, while, by comparison, the accuracy of GPT-3, released by OpenAI in 2019, is only 53.3% to 83.9% when facing ethical problems. For more details, please see the website: <https://spectrum.ieee.org/ai-ethics-machines-learn-good> (accessed on 3 November 2021).
- ³ The case study “Ask Dehphi” was presented in Colin Allen’s online lecture on 23 May 2023, titled “How to Make Large Language Models Exhibit Coherent Value Theory”.
- ⁴ For more information, see <https://livebook.manning.com/book/human-in-the-loop-machine-learning/chapter-1/34> (accessed on 30 June 2021).
- ⁵ For more information on Professor Solar-Lezama’s opinions, please refer to the following website for details: <https://www.wired.com/story/ai-adversarial-attacks/> (accessed on 1 August 2023).
- ⁶ International guidelines for data collection and sharing published by the Australian National Data Service, see the website: <https://ardc.edu.au/resource/the-care-principles/> (accessed on 10 October 2022).
- ⁷ The Delphi Method, also known as the Expert Opinion Method, was first proposed by Helmer & Gordon in 1946. It is a management technique used to address complex task problems and improve the estimation accuracy and classification for complex systems. It involves the repeated administration of anonymous questionnaires to a group of experts through a systematic process. The aim is to gather opinions and reach a consensus within the group.
- ⁸ For more information on Dr. Vincent Vanhoucke’s opinions, please refer to the following website for details: <https://towardsdatascience.com/the-quiet-semi-supervised-revolution-edec1e9ad8c> (accessed on 16 May 2019).

References

1. Dignum, V. Ethics in artificial intelligence: Introduction to the special issue. *Ethics Inf. Technol.* **2018**, *20*, 1–3. [CrossRef]
2. Umbrello, S.; Yampolskiy, R.V. Designing AI for explainability and verifiability: A value sensitive design approach to avoid artificial stupidity in autonomous vehicles. *Int. J. Soc. Robot.* **2022**, *14*, 313–322. [CrossRef]
3. Trazzi, M.; Yampolskiy, R.V. Artificial stupidity: Data we need to make machines our equals. *Patterns* **2020**, *1*, 100021. [CrossRef] [PubMed]
4. Xu, Y. Artificial Intelligence, Trolley Problem and the Involvement of Cultural-geographical Factors. *Philos. Res.* **2023**, 96–107+129.
5. Wallach, W.; Allen, C. *Moral Machines: Teaching Robots Right from Wrong*; Oxford University Press: Oxford, UK, 2008; pp. 26, 39–41, 84, 112–116, 204.
6. Cai, L.; Wang, S.; Liu, J.; Zhu, Y. Review of Data Annotation Research. *J. Softw.* **2020**, *31*, 302–320. [CrossRef]
7. Liu, J. Human-in-the-Loop Ethical AI for Care Robots and Confucian Virtue Ethics. In *International Conference on Social Robotics*; Springer Nature: Cham, Switzerland, 2022.
8. Zhu, J.; Kaplan, R.; Johnson, J.; Fei-Fei, L. Hidden: Hiding Data with Deep Networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
9. Egorow, O.; Lotz, A.; Siegert, I.; Bock, R.; Krüger, J.; Wendemuth, A. Accelerating manual annotation of filled pauses by automatic pre-selection. In Proceedings of the 2017 International Conference on Companion Technology (ICCT), Ulm, Germany, 11–13 September 2017.
10. Monarch, R.M. *Human-In-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*; Simon and Schuster: New York, NY, USA, 2021; pp. 31–63.
11. Narayanan, S.; Rothrock, L. *Human-In-the-Loop Simulations: Methods and Practice*; Springer: London, UK; Dordrecht, The Netherlands; Heidelberg, Germany; New York, NY, USA, 2011; p. 16.
12. Jo, A. The promise and peril of generative AI. *Nature* **2023**, *614*, 214–216.
13. Siegert, I. Emotional and User-Specific Cues for Improved Analysis of Naturalistic Interactions. Ph.D. Thesis, Otto von Guericke University, Magdeburg, Germany, 2015.
14. Thiam, P.; Meudt, S.; Schwenker, F.; Palm, G. Active learning for speech event detection in HCI. In *Artificial Neural Networks in Pattern Recognition: 7th IAPR TC3 Workshop, ANNPR 2016, Ulm, Germany, September 28–30, 2016, Proceedings 7*; Springer International Publishing: Cham, Switzerland, 2016.
15. Van Dis, E.A.; Bollen, J.; Zuidema, W.; van Rooij, R.; Bockting, C.L. ChatGPT: Five priorities for research. *Nature* **2023**, *614*, 224–226. [CrossRef] [PubMed]
16. Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; Rahwan, I. The moral machine experiment. *Nature* **2018**, *563*, 59–64. [CrossRef] [PubMed]
17. Budd, S.; Robinson, E.C.; Kainz, B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med. Image Anal.* **2021**, *71*, 102062. [CrossRef] [PubMed]
18. Jung, W.; Jazizadeh, F. Human-in-the-loop HVAC operations: A quantitative review on occupancy, comfort, and energy-efficiency dimensions. *Appl. Energy* **2019**, *239*, 1471–1508. [CrossRef]

19. Agnisarman, S.; Lopes, S.; Madathil, K.C.; Piratla, K.; Gramopadhye, A. A survey of automation-enabled human-in-the-loop systems for infrastructure visual inspection. *Autom. Constr.* **2019**, *97*, 52–76. [\[CrossRef\]](#)
20. Benedikt, L.; Joshi, C.; Nolan, L.; Henstra-Hill, R.; Shaw, L.; Hook, S. Human-in-the-loop AI in government: A case study. In Proceedings of the 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, 13–17 March 2020.
21. Chai, C.; Li, G. Human-in-the-loop Techniques in Machine Learning. *IEEE Data Eng. Bull.* **2020**, *43*, 37–52.
22. Tehrani, B.M.; Wang, J.; Wang, C. Review of human-in-the-loop cyber-physical systems (HiLCPS): The current status from human perspective. In Proceedings of the ASCE International Conference on Computing in Civil Engineering 2019, Atlanta, Georgia, 17–19 June 2019; American Society of Civil Engineers: Reston, VA, USA, 2019.
23. Mainampati, M.; Chandrasekaran, B. Implementation of human in the loop on the TurtleBot using reinforced learning methods and robot operating system (ROS). In Proceedings of the 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Online, 27–30 October 2021.
24. Mosqueira-Rey, E.; Pérez-Sánchez, A.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; Fernández-Leal, Á.; Moret-Bonillo, V.; Vidal-Ínsua, Y.; Vázquez-Rivera, F. Human-in-the-Loop Machine Learning for the Treatment of Pancreatic Cancer. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 18–22 June 2023.
25. Ramesh, P.; Subramaniam, T.; Ray, P.; Devadas, A.; Ramesh, S.; Ansar, S.; Ramesh, M.; Rajasekaran, R.; Parthasarathi, S. Utilizing human intelligence in artificial intelligence for detecting glaucomatous fundus images using human-in-the-loop machine learning. *Indian J. Ophthalmol.* **2022**, *70*, 1131. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Wu, X.; Xiao, L.; Sun, Y.; Zhang, J.; Ma, T.; He, L. A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.* **2022**, *135*, 364–381. [\[CrossRef\]](#)
27. Greeno, J.G. Gibson's affordances. *Psychol. Rev.* **1994**, *101*, 336. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Baker-Brunnbauer, J. Management perspective of ethics in artificial intelligence. *AI Ethics* **2021**, *1*, 173–181. [\[CrossRef\]](#)
29. Floridi, L. *The Ethics of Information*; Oxford University Press: New York, NY, USA, 2013; p. 261.
30. Jones, T.M. Ethical decision making by individuals in organizations: An issue-contingent model. *Acad. Manag. Rev.* **1991**, *16*, 366–395. [\[CrossRef\]](#)
31. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C.A. Mixmatch: A holistic approach to semi-supervised learning. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 454.
32. Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; Le, Q. Unsupervised data augmentation for consistency training. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6256–6268.
33. Chen, X. Ethical Governance of AI: An Integrated Approach via Human-in-the-Loop Machine Learning. *Comput. Sci. Math. Forum.* **2023**, *8*, 29.

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