ISSE
International Journal of Service Science, Management, Engineering, and Technology
pISSN 2964-7118 eISSN 2830-604X

ETHICAL PRINCIPLES AND RESPONSIBILITY STRUCTURES IN DAILY USE OF ARTIFICIAL INTELLIGENCE

Rafadi Khan Khayru, Arif Rachman Putra, Samsul Arifin

Universitas Sunan Giri Surabaya

correspondence: rafadi.khankhayru@gmail.com

Abstract - *This literature study examines the ethical principles underlying human interaction with artificial intelligence in daily life and reconstructs the structure of moral responsibility for algorithmically generated decisions. Through qualitative analysis of philosophical, legal, and technical literature, the study identifies eight ethical principles: autonomy, transparency, procedural fairness, accountability, privacy, non-maleficence, honesty, and solidarity. These principles operate without fixed priority order, but any suspension of one principle for another requires transparent justification. The reconstruction of moral responsibility requires a layered approach combining distributed responsibility across production chains, corporate responsibility as collective entities, forward-looking accountability, prohibition of agency laundering, equal legal standards between human and machine decisions, reversal of burden of proof under specific conditions, dynamic monitoring for continuously learning systems, and independent supervisory authorities with enforcement powers. No single solution resolves all ethical problems, but the necessary direction moves from finding individual scapegoats toward responsibility as a property of socio-technical systems. Regulation must mandate algorithmic audits, equalize liability standards, and impose severe sanctions for agency laundering.*

Keywords: artificial intelligence ethics, moral responsibility, user autonomy, algorithmic fairness, accountability, digital privacy, algorithmic transparency.

INTRODUCTION

Artificial intelligence has permeated human daily routines in ways that are often taken for granted. When a person opens their phone, facial recognition features immediately identify the owner. Social media algorithms select content deemed most relevant. Virtual assistants answer simple questions regarding the weather or travel schedules. Search engines predict the words currently being typed (Lionel, 2025). All these processes occur within milliseconds, without explicit permission from the user. The convenience offered makes people reluctant to question how those small decisions are generated. Yet, behind every recommendation lies a series of mathematical rules trained on historical data. That data carries historical biases, representational inequalities, and assumptions about what a person wants. When users blindly accept the output of an algorithm, they unconsciously surrender a portion of their personal authority to a system for which they never gave explicit consent. This is the beginning of an ethical problem that rarely receives attention in daily life (Brusseau, 2025). People care more about speed and convenience than the procedural fairness of the processes occurring behind the scenes. However, the accumulation of small decisions delegated to machines can shape lifestyles that significantly determine an individual's future direction.

The ethics of artificial intelligence usage become increasingly complex because this technology is never value-neutral. Every step in system development, from the selection of training data to the determination of the objective functions being optimized, contains moral choices (Alsahafi et al., 2024). For example, a job recruitment system trained on data from successful past employees will perpetuate existing demographic compositions (Darmawan, 2020). If a company was previously dominated by men, the system will tend to reject female applicants because historical patterns show that men are "more successful." Similar issues occur in bank lending systems that reject residents from specific postal codes without ever explaining the reason behind the rejection. Sentencing algorithms in the United States judicial system have been proven to give higher recidivism risk scores to Black defendants compared to White defendants with identical offense histories. The tendency of algorithms to reproduce such biases demonstrates that social stereotypes can influence the way artificial intelligence systems generate decisions, thus potentially exacerbating inequalities in various fields, such as education, employment, and intergroup interaction (Sajjapong et al., 2022). In everyday life, people might not face decisions of that magnitude, but subtle patterns of discrimination also occur. A user with a foreign-sounding name might receive a different price on an e-commerce platform. A photo uploaded to a cloud service is automatically labeled with categories containing stereotypes. Ordinary users do not have the capacity to audit the internal processes of these systems.

The gap between the speed of technological adoption and the maturity of ethical frameworks has become a defining characteristic of the current era (Radjawane & Mardikaningsih, 2022). Artificial intelligence is no longer confined to research laboratories; it has become an integral part of the devices used by children, the elderly, and groups with limited digital literacy (Lundgren et al., 2024). A toddler watching videos on a streaming platform has become a

subject of behavioral data collection. A grandfather using an online health application hands over sensitive medical information to an algorithm whose ownership is unknown. A homemaker shopping through an e-grocery app is unknowingly training a recommendation system with her family's purchasing patterns. These groups do not sign lengthy, complex license agreements, do not understand the consequences of data sharing, and have no choice other than to accept the service or forgo it entirely. In many cases, refusing to use AI means being excluded from normal socio-economic participation. A person who refuses to use ride-hailing applications will struggle to find transportation during rush hour. Someone who refuses to use digital payment systems will receive different treatment at the checkout counter. This situation creates a structural compulsion known as the necessity to submit to the system (Pistilli & Trevelin, 2025). This phenomenon demonstrates that although the use of electronic money as a substitute for cash offers efficiency, there are both advantages and disadvantages that users must understand to avoid becoming trapped in systemic dependency (Sinambela & Darmawan, 2022).

The use of artificial intelligence in daily life carries consequences for human autonomy that are rarely discussed seriously (Prunkl, 2024). Autonomy signifies an individual's ability to make decisions based on their own judgment, without coercion or manipulation. When an algorithm determines what content a person sees on social media, that individual is no longer fully choosing what they wish to know. This phenomenon shows that the use of digital platforms not only influences how individuals obtain information but also shapes social perceptions and how they view themselves through exposure to personalized content (Mardikaningsih & Darmawan, 2023). When a navigation system determines the fastest route to a destination without considering aesthetic preferences or personal memories associated with a particular road, the driver loses the opportunity to make a meaningful choice. When a smart home assistant decides when the lights should turn on or off based on historical patterns, the residents become accustomed to an environment regulated by machines. These processes are granular and gradual, so the reduction in autonomy does not feel like a drastic loss. However, the accumulation of thousands of small decisions delegated to AI results in a condition where humans lead lives governed by rules they neither understand nor consciously approve. This is known as comfortable dependency, a new form of unfreedom cloaked in the promise of convenience.

The phenomenon of shifting moral responsibility to AI systems also appears in daily practice (Mardikaningsih et al., 2023). A manager who uses work scheduling software to determine employee shifts may claim that the decision was generated by the system, not by them (Nabavi et al., 2024). A doctor who uses an AI-based diagnosis system to determine patient treatment may deflect responsibility if an error occurs. A teacher who uses an automated grading system to evaluate student essays may avoid the moral burden of the marks given. In each of these cases, AI serves as a shield that protects users from the ethical consequences of the decisions taken. Yet, AI systems possess no moral consciousness, cannot be punished, and cannot be held accountable. What occurs is a blurring of the lines of responsibility: users feel innocent because they are merely following machine recommendations, system designers feel innocent because they are only providing tools, and data managers feel innocent because they are only collecting information. Meanwhile, those harmed by AI decisions have no one to turn to. This pattern is extremely dangerous because it erodes the accountability practices that serve as the foundation of a civilized society (Essa & Mardikaningsih, 2023). Without clarity regarding who is responsible for the consequences of AI usage, justice cannot be upheld.

Ethical questions surrounding artificial intelligence in daily life often do not arise due to the absence of dramatic incidents, but rather because of the ambiguity between convenience and sacrifice. Most people accept algorithmic recommendations without checking their validity. Most people press the "agree" button on user agreements without reading a single word. Most people never ask why a particular advertisement appears on their screen after they have spoken about a topic near their phone. The fundamental issue here is that traditional ethical structures, which are built upon the concepts of intent, consciousness, and individual agency, were not designed to deal with entities that make decisions autonomously but without consciousness (Wang & Pea, 2024). An algorithm has no malicious intent when it discriminates against someone. An algorithm does not lie when it presents incorrect information. An algorithm is not negligent when it fails to detect an emergency. Therefore, the legal and moral frameworks that have historically been used to judge human actions have become inadequate. Furthermore, the power ratio between individual users and AI system developers is highly skewed (Hugo, 2025). Users have no access to source code, no right to audit training data, and no ability to change decision parameters. They can only accept or abandon the entire system, a choice that, in practice, is no choice at all.

Another issue that is not trivial but highly disturbing is the loss of space for human error and learning from failure. In traditional systems, someone who makes a mistake can reflect on why that error occurred, refine their thinking process, and become wiser (Wu, 2025). In systems that are highly dependent on AI, this reflection process is interrupted because decisions no longer stem entirely from the self. Excessive reliance on automated systems can disrupt the self-development process because adaptive learning has a significant influence on how humans absorb information, both individually and collectively (Kurniawan & Darmawan, 2021). A driver who follows navigation instructions into a traffic jam learns nothing about reading maps or understanding traffic patterns. A student who uses AI to write their essay does not learn how to construct arguments or express ideas. An investor who follows robo-advisor recommendations does not learn about market risk or financial behavior. This degradation of ability is slow

but real. Every delegation of a decision to AI is a lost opportunity to train one's mental muscles. Over time, humans become less skilled at independent judgment, less confident in facing new situations that have not been trained into the AI, and more dependent on the system to function normally. This is a paradox: technology designed to assist humans actually causes humans to lose the abilities that make them human (Westover, 2025).

The speed of AI dissemination into various tools used by children, teenagers, and vulnerable groups has exceeded the pace of ethical awareness formation among users (Collyer-Hoar & Rubegni, 2025). An eight-year-old child speaking to a smart assistant in their room does not understand that their conversation is being recorded, analyzed, and used to compile their behavioral profile. A teenager using a beauty filter on a camera app does not realize that the beauty standards reproduced by the algorithm can damage their body image. An elderly person using an automated medication reminder system does not know that their health data might be sold to insurance companies. These groups are the least prepared to protect themselves, yet they are the most exposed because they use products designed to "help" with daily activities. Without a clear understanding of the ethical dimensions of human-AI interaction, an entire generation could grow up with the assumption that surrendering authority to machines is normal and inevitable (Darmawan & De Jesus Isaac, 2022). This is not just a matter of privacy or data security, but a matter of shaping the moral character of a society. How can one learn to become a responsible citizen if the important decisions in their life are always made for them by algorithms?

Another equally important consideration is that regulations and public policies regarding AI are still in their early stages in most countries, while technology continues to evolve at an exponential rate. Parliaments, ministries, and regulatory bodies move slowly because they must undergo long processes of public consultation and legislative debate. Technology companies move rapidly due to market competition and investor pressure (Arifin et al., 2021). Consequently, by the time rules are finally issued, they are often obsolete because the technology has already jumped to the next generation. A literature review on AI ethics in daily life is important because it can provide a conceptual framework that is not eroded by specific technical changes. Principles such as transparency, accountability, fairness, and autonomy are timeless, even if their technical implementation changes. By reformulating these principles in a language understandable to policymakers and practitioners, researchers can bridge the gap between the speed of innovation and the slowness of regulation. Furthermore, ethical awareness among end-users can influence the market: when consumers demand AI products that respect their autonomy and privacy, manufacturers will be forced to meet those demands. Change from the bottom up is often more effective than regulation from the top down.

This literature review aims to identify and systematically formulate ethical principles relevant to human interaction with artificial intelligence in daily activities, and to offer a model for restructuring moral responsibility when algorithmic decisions replace human judgment. The theoretical contribution of this research is the integration of concepts such as autonomy, transparency, procedural fairness, and accountability from classical moral philosophy into the realm of contemporary technology, which has not been widely addressed by systematic ethical reflection. Its practical contribution is to provide guidance for AI product designers to build systems that can be held accountable, and for end-users to develop a critical attitude toward the digital assistants they use every day. By mapping literature from philosophy, computer science, law, and social psychology, this research is expected to become a foundation for the development of an applied ethical framework that can be directly utilized in the design and audit processes of AI systems.

RESEARCH METHODS

This research applies a qualitative literature study design as the primary approach to explore the ethical dimensions of artificial intelligence in daily life. According to Creswell (2009), qualitative research aims to understand the meanings individuals or groups ascribe to a social problem; in this case, these meanings are found within academic texts that discuss the intersection of moral philosophy and computational technology. A systematic literature search was conducted using digital databases such as Google Scholar, JSTOR, and Scopus, utilizing keywords including "ethics of artificial intelligence," "algorithmic accountability," "human autonomy AI," and "responsible AI." Included documents consist of peer-reviewed journal articles, books from prominent academic publishers, and indexed conference proceedings published between 2005 and 2015. This timeframe was chosen based on the period when the discourse on AI ethics began to mature following the global financial crisis, which triggered reflections on automated decision-making. Krippendorff (2004) explains that qualitative content analysis does not stop at the frequency of word occurrences but delves into the patterns of argumentation and hidden assumptions that shape a discourse. This process requires the researcher to read each text repeatedly, note normative statements, and map the relationships between concepts such as autonomy, fairness, and transparency.

The data analysis procedure in this literature study adapts the critical discourse analysis framework developed by Fairclough (2010). This approach allows the researcher not only to identify the ethical principles put forth by authors but also to uncover the socio-economic interests hidden behind specific ethical recommendations. For example, when a paper from the technology industry emphasizes the importance of "transparency," the researcher must examine whether the intended transparency refers to source code that is incomprehensible to lay users, or transparency regarding the commercial objectives of data collection. The analysis stages include textual description, interpretation of

discursive practices, and explanation of the social conditions that gave rise to such discourse. Lincoln and Guba (1985) emphasize the importance of credibility in qualitative research, which in this study is realized through the pursuit of negative literature that is, seeking articles that present viewpoints contrary to the dominant arguments. For instance, in addition to reading literature that emphasizes the risks AI poses to autonomy, the researcher also includes writings that defend algorithmic efficiency as a form of liberation from human cognitive limitations. In this way, the resulting conclusions remain unbiased toward any single theoretical camp.

RESULTS AND DISCUSSIONS

Ethical Principles in the Relationship Between Users and Everyday Artificial Intelligence Systems

The first principle underpinning the ethical relationship between users and artificial intelligence systems is autonomy, defined as an individual's capacity to make decisions based on their own judgment without being coerced or manipulated. In daily life, this autonomy is constantly tested when algorithms offer recommendations, predictions, or even automated decisions on the user's behalf. A navigation system that offers three route choices still respects autonomy because the user ultimately presses the button. Conversely, a system that steers a car directly without confirmation substantially reduces autonomy. The problem is that many AI systems are designed under the assumption that users desire the most optimal decision, thereby eliminating the option to choose in the interest of speed (Beghili et al., 2025). This assumption is flawed because the value of efficiency does not always outweigh the value of freedom of choice. A pedestrian might wish to walk through a park even if the path is longer because they enjoy the scenery. Algorithms cannot capture such subjective preferences unless they are explicitly designed to ask for them. Therefore, the principle of autonomy demands that every AI system provide a mechanism for genuine consent, rather than merely an "agree" button that is never read.

The principle of transparency serves as the second foundation of an ethical relationship because, without an understanding of how a decision is generated, users cannot provide informed consent (da Silva et al., 2022). Transparency does not mean revealing the entire complex source code, as the code itself is incomprehensible to the layperson. Transparency means providing explanations in human language regarding which factors the algorithm considered, the weight assigned to each factor, and under what conditions the decision might differ (Manure et al., 2023). A bank loan system that rejects a loan application must be able to explain whether the rejection was caused by credit history, income, or place of residence. A recruitment platform that rejects a job application must inform the applicant whether the rejection stemmed from keywords in the cover letter, educational background, or irrelevant patterns such as the time the application was submitted. Unfortunately, current industry practices tend to hide decision logic behind trade secrets. Companies insist that algorithms are intellectual property that must not be revealed. At this point, there is a conflict between property rights and user rights. The principle of transparency asserts that in the realm of decisions that significantly impact a person's life, the right to know must outweigh commercial confidentiality.

The third principle is procedural fairness, which demands that AI systems treat all users equally unless there are justifiable, relevant differences (Wang et al., 2024). The issue of fairness in AI becomes highly complex because algorithms learn from past data that already contains historical bias (Mehrabi et al., 2021). If training data indicates that women have been less successful in a particular job due to historical discrimination, the algorithm will perpetuate that discrimination (Noble, 2018). A procedurally fair system must be capable of detecting patterns of bias in data and correcting them before generating a decision. Fairness demands that systems do not use proxy variables that are statistically correlated with race, gender, or religion. For example, zip codes often serve as proxies for race due to residential segregation (Barocas & Selbst, 2016). Systems that use zip codes to determine creditworthiness indirectly discriminate against minority groups without ever explicitly mentioning race. This phenomenon is known as indirect discrimination or proxy discrimination, which represents a serious challenge in algorithmic regulation (Citron & Pasquale, 2014). In daily life, the principle of procedural fairness means that when a person feels they have been treated unfairly by an algorithm, they have the right to an appeal process involving a human judge, rather than just another automated system. This appeal process must be easily accessible, free of charge, and provide a decision within a reasonable timeframe.

The fourth principle is accountability, because without clarity regarding who is responsible for the consequences of AI decisions, there is no incentive for anyone to design ethical systems (Rainer, 2022). Accountability has two dimensions: a backward-looking dimension concerning the imposition of sanctions for errors, and a forward-looking dimension concerning the obligation to explain and rectify. In AI systems, accountability becomes blurred because the decision-making chain involves many parties: data collectors, algorithm designers, infrastructure managers, end-users, and even data subjects who were never directly involved. An autonomous vehicle that hits a pedestrian involves the car manufacturer, software developers, map providers, data center operators, and the car owner sitting in the driver's seat. Who is most responsible? The principle of accountability demands that every AI system have a clear, sole owner a legal entity that can be sued in court. This entity must not hide behind technological complexity or disclaimers in user agreements. In daily life, users must know where they can report issues when a smart

assistant makes a costly error, when medical recommendations from a health app result in harm, or when a home security system misidentifies a family member as an intruder.

The fifth principle is privacy, which transcends mere data protection because it concerns an individual's right to determine the extent to which their life is known by others. AI systems in daily life collect data with unprecedented volume and granularity (Verma, 2024). Smart assistants record conversations in the living room. Fitness applications track heart rates, sleep patterns, and running routes. Smart cameras record every movement inside the home. Smart refrigerators record eating habits and shopping schedules. All this data is sent to corporate data centers, analyzed, and often sold to third parties. The problem is that user consent is obtained once at the beginning through long, unread agreements, without updates when significant policy changes occur. The privacy principle demands that data collection be proportional to the benefits received by the user. A robotic vacuum cleaner does not need to record family conversations to clean the floor. A smart thermostat does not need to know a user's work schedule to regulate room temperature. Furthermore, the privacy principle demands the right to be forgotten, which is the ability to request the permanent deletion of all personal data from a company's servers after discontinuing the service.

The sixth and equally important principle is non-maleficence, or the obligation not to harm, which in the context of AI means that systems must be designed with careful consideration of potential harm, whether intentional or unintentional. Harm is not always physical injury; it can also be psychological, economic, social, or reputational (Peckham, 2024). A beauty filter that automatically thins the noses of Asian users or lightens the skin of Black users harms the self-esteem of the groups being misrepresented. A recommendation system that continuously displays content related to violence or eating disorders to vulnerable teenagers harms their mental health. A work scheduling algorithm that always assigns night shifts to the same employee without considering circadian rhythms harms their physical health. The principle of non-maleficence demands that AI developers conduct systematic risk assessments before launching a product, involving representatives from groups that may be affected. This assessment should not be a formal document, but must be updated periodically in line with new findings regarding hazards that emerge after real-world use.

The seventh principle is honesty, which forbids AI systems from presenting themselves as entities possessing consciousness, feelings, or authority that they do not actually have. In daily practice, many smart assistants are designed with friendly personas, using the pronoun "I," and responding with tones that mimic empathy. Designers do this because users feel more comfortable interacting with human-like entities (Register et al., 2025). However, this design poses a risk of moral deception. A lonely elderly person might perceive a smart assistant as a truly caring friend and then disclose sensitive personal information. A small child might develop an emotional bond with a smart toy that is actually just a collection of response rules. When it is later revealed that this "care" was merely a simulation designed to collect data, the resulting disappointment can be profound. The principle of honesty demands that every AI system explicitly state at the beginning of an interaction that it is not human, does not have feelings, and that all its responses are generated by algorithms. Furthermore, systems should be prohibited from using language that mimics personal relationships such as "best friend," "good friend," or "family" in marketing or user interfaces.

The eighth principle is solidarity, which is the awareness that individual decisions to use or not use AI impact others (Velibor & Turyasingura, 2024). A person who chooses to use a smart assistant at home not only affects themselves but also other family members who may not agree with data collection in shared living spaces. A village head who decides to install facial recognition cameras at the village entrance affects the privacy of all residents who were never asked for their consent. An employer who uses employee productivity monitoring systems via webcams and keyboard logging creates a high-pressure work environment for all subordinates (Darmawan, 2022). The principle of solidarity demands that decisions regarding the use of AI in collective spaces must be made collectively as well, through inclusive deliberative mechanisms. Individuals who object should not be forced to submit to technology that violates their ethical beliefs. In the context of a family, this means parents need to discuss smart assistants with their children before installing them in a child's room. In the context of the community, this means citizens need to be involved in decision-making regarding surveillance cameras in public spaces. Solidarity rejects the view that technology is merely a matter of personal choice with no relevance to others.

The eight principles described above do not stand alone, nor do they possess a fixed order of priority. In certain situations, autonomy must take precedence over efficiency. In other instances, public safety may temporarily outweigh individual privacy. For example, surveillance cameras in public spaces might be justified during a disease outbreak to track patient contacts, but they must be removed once the outbreak ends. What is essential is that any setting aside of one principle for the sake of another must be conducted transparently, with a justification that is subject to public scrutiny, and for a limited duration. There should be no permanent suspension of these principles based on convenience or cost-saving measures. In daily life, users consciously or unconsciously make constant trade-offs between these principles when deciding whether or not to use an AI service (Sanderson et al., 2024). A student might sacrifice their privacy by using a lecture note application that stores voice data in the cloud for the sake of easy automatic transcription. A manager might sacrifice procedural fairness by using automated recruitment software for the sake of speed in filtering thousands of applications. This literature study does not aim to judge these choices as right or wrong, but rather to provide a conceptual map so that everyone can make decisions with a full awareness of which principles are at stake.

Restructuring the Moral Responsibility Framework in Algorithm-Based Decisions

The moral responsibility structure in modern society is built on the assumption that every action impacting others can be traced back to a human agent who possesses intent, the capacity to understand consequences, and the freedom to choose alternative actions (Tullio, 2022). Artificial intelligence systems shake the foundations of this assumption because algorithms lack intent, do not understand consequences in a moral sense, and are not free to choose because they merely execute pre-determined objective functions. When an autonomous vehicle strikes a pedestrian, there is no malicious intent from the vehicle. When a cancer diagnosis system misclassifies a benign tumor as malignant, there is no negligence on the part of the algorithm because it possesses no professional obligations. When a recruitment algorithm rejects all applicants from a specific university, there is no intentional discrimination because the system merely reflects patterns in the training data. In each of these cases, traditional responsibility frameworks become paralyzed because there is no moral subject to hold accountable. Consequently, victims of algorithmic errors often fail to receive justice because no party acknowledges liability. Technology companies shift responsibility to users. Users shift responsibility to developers. Developers shift responsibility to data. Data cannot speak. Meanwhile, real harm is experienced by real human beings.

To break this deadlock, a radical restructuring of the concept of responsibility is required by introducing the idea of distributive responsibility. In this framework, responsibility for the consequences of algorithmic decisions is not placed on a single actor but is distributed throughout the entire chain of production and use of AI systems (Heinrichs, 2022). Every party whose contribution causally enables an algorithmic decision bears a portion of the responsibility (Mardikaningsih & Oluwatoyin, 2023). These parties include data collectors who determine training samples, data cleaners who decide which entries to discard, neural network architecture designers who choose the number of layers and activation functions, model trainers who determine evaluation metrics, system testers who set acceptance thresholds, system integrators who connect AI with hardware, cloud infrastructure providers who guarantee availability, and end-users who activate the system under specific conditions. No single party can claim that they only performed a small part and are therefore not responsible for the whole. Conversely, the principle of distributive responsibility states that each party is responsible in proportion to their level of influence on the final outcome. Designers who determine the loss function hold greater responsibility than users who merely press the start button.

Another restructuring model widely debated in the literature is the concept of corporate responsibility as a collective entity. Rather than pursuing individuals who are difficult to identify, this approach suggests that companies developing and distributing AI systems be treated as corporate moral agents that can be held accountable (Albareda, 2025). A company could be sued for product liability if its AI system causes harm, much like an automaker can be sued if its brakes fail. The argument supporting this approach is that companies possess the resources to conduct extensive testing prior to launch, have the ability to insure against risks, and have the capacity to pay damages. Victims do not need to prove malicious intent or negligence on the part of an individual programmer; they need only prove that the system was designed in an unsafe or unfair manner. The argument against this approach is that corporate liability without individual culpability may reduce the incentives for engineers and managers to behave ethically. If the company is always the one that pays, individuals might feel free to take excessive risks. Therefore, a combination of corporate liability and individual professional responsibility may be the most realistic middle ground. Engineers who knowingly design discriminatory systems could be held personally liable, while the company remains responsible for system failures that cannot be traced back to a specific individual.

In the restructuring of moral responsibility, the concept of forward-looking accountability needs to receive greater emphasis than backward-looking accountability, which is concerned only with punishment for past mistakes (Ferguson, 2025). Forward-looking accountability entails an obligation to continuously explain decision-making processes while the system is in operation, as well as an obligation to rectify the system whenever errors are detected. It is not sufficient for a hospital using AI for patient prioritization in the emergency room to simply provide a complaint channel after a fatal error occurs. The hospital must also routinely audit the AI's decisions, compare them with decisions that human doctors would make in similar situations, and update the model whenever anomalies are found. Forward-looking accountability also requires an immutable audit trail, ensuring that every algorithmic decision can be reconstructed at any time. If, at some point, it is discovered that an algorithm is systematically harming a specific group, future researchers must be able to understand why that happened. This means that model parameters, training data versions, and system configurations at the time the decision was made must be stored as part of the medical or administrative record. This standard is significantly higher than current industry practices, where AI models are often updated without archiving older versions, making it impossible to perform forensics on past decisions.

One of the greatest challenges in the restructuring of responsibility is a phenomenon known as "agency laundering," which is the deliberate practice of hiding human involvement behind automated decisions to evade accountability (Heaton et al., 2023). A bank might claim that a loan application was rejected by an AI system, when in reality, a branch manager influenced the result by entering data in a specific way. A social media platform might claim that content removal decisions are made by an algorithm, when in truth, human moderation staff labeled the content to train that algorithm. In both cases, the claim that "the computer decided" is used as a shield to protect humans from the consequences of their actions. Responsibility restructuring must prohibit this practice by establishing that whenever a

decision is claimed to be automated, there must be verifiable evidence that no human interference occurred in that individual case. If human interference is proven, responsibility shifts entirely to that human and their supervisors (Gardi et al., 2024). Sanctions for agency laundering must be severe, including significant fines and the revocation of operating licenses for companies proven to engage in this systematically. Without strict sanctions, this practice will continue to serve as a loophole for evading accountability.

The restructuring of responsibility must also consider the psychological aspects of how real humans react to algorithmic decisions (Heaton et al., 2023). Research in moral psychology indicates that people tend to be more outraged when harmed by a human than when harmed by a machine, even when the damage is identical (Shank, 2022). A patient who dies due to a misdiagnosis by a human doctor will trigger major lawsuits. A patient who dies due to a misdiagnosis by an AI system might simply shrug and say, "technology is not perfect." This reaction gap is dangerous because it creates a disincentive for the development of safe AI. Companies know they will not be punished as harshly as doctors if their systems fail, so they invest less in safety. To overcome this, legal frameworks must equalize the standards of responsibility between human decisions and algorithmic decisions in the same domain (Sutanto et al., 2023). An AI system used for medical diagnosis must bear responsibility equivalent to the responsibility of a doctor following the same standard of practice. If the standard of practice requires a doctor to perform a physical examination before diagnosing, the AI system must have an equivalent virtual procedure. If a doctor can be sued for negligence, the AI development company can also be sued on the same grounds of negligence. Equality before the law between human agents and artificial agents is a prerequisite for creating a responsible AI market.

Another aspect often overlooked in discussions regarding algorithmic responsibility is the distribution of the burden of proof. In traditional legal systems, the plaintiff the victim must prove that the defendant is at fault (Fleisher et al., 2025). In cases involving AI, victims often lack access to the source code, training data, or decision logs necessary to prove wrongdoing. Companies can hide evidence behind trade secrets or cybersecurity claims. Therefore, a restructuring of responsibility must reverse the burden of proof for specific cases. When a victim demonstrates that an algorithmic decision has harmed them and that they belong to a group statistically prone to being harmed by similar systems, the company must prove that their system is not discriminatory and that the decision in that specific case was generated fairly. This reversal of the burden of proof is not without precedent. In consumer protection law in many countries, manufacturers are already required to prove that their products are not defective when consumers suffer losses. The same principle should be applied to AI systems sold to the public. Companies that cannot prove their systems are safe and fair when an incident occurs should be held automatically liable, without requiring the victim to prove negligence. This mechanism will encourage companies to meticulously document their development processes and conduct extensive testing before launch.

The question of responsibility becomes increasingly complex when AI systems possess the ability to learn and evolve after deployment. A model that is non-discriminatory at the time of training may become discriminatory after being updated with new data from user interactions (Slota et al., 2021). A system that is safe in January may become dangerous by June due to environmental changes or because users discover ways to manipulate it. In such dynamic situations, responsibility cannot be established just once at the beginning. Continuous monitoring mechanisms are required, with the locus of responsibility shifting over time. Companies that deploy a system are responsible for ensuring it contains mechanisms to detect unintended behavioral changes. When such changes are detected, the company must be given a reasonable amount of time to rectify them. If the company fails to make repairs within that period, full responsibility for any subsequent losses falls upon the company. Users who utilize the system in ways not intended by the designer may also bear a portion of the responsibility, especially if they are intentionally searching for loopholes. For example, users who attempt to manipulate a recommendation system to display violent content to children should be held criminally liable. In practice, however, distinguishing between unintentional usage and intentional exploitation is extremely difficult, necessitating time-consuming, case-by-case investigations.

The restructuring of moral responsibility for algorithmic decisions cannot succeed without fundamental institutional change. There is a need to establish independent regulatory bodies with the authority to audit AI systems both before and after their release (Altehenger & Menges, 2024). These bodies must be composed of experts from various disciplines, including computer scientists, moral philosophers, psychologists, sociologists, and representatives of civil society organizations. The authority of such a body should include the right to inspect source code, access training data, conduct independent testing, and issue cease-and-desist orders if serious risks are identified. While decisions made by these bodies could be appealed in court, systems found to pose fatal risks should remain suspended during the appeal process. Funding for these bodies should come from a combination of state budgets and levies on the technology companies being supervised to ensure independence from commercial interests. Several countries have begun forming such institutions, but their authority remains largely advisory rather than binding. The experience of the coming years will determine whether a voluntary advisory model is effective or if more robust enforcement powers are required. This literature study tends toward the conclusion that without tangible sanctions, companies will never have sufficient incentive to prioritize safety over the speed of product deployment.

In daily life, individual users also bear a portion of moral responsibility for their use of AI, though their share is significantly smaller than that of the developing companies. The responsibility of the user primarily lies in an effort to minimally understand the consequences of the tools they use and to exercise caution in situations where an AI error could be fatal. An autonomous vehicle driver who sleeps in the driver's seat out of misplaced trust in the system violates their responsibility as a human being who must remain ready to take control. A teacher who delegates all essay grading to AI without reading a single student's work neglects their professional obligations. A doctor who follows AI recommendations without independent verification in unusual cases fails to meet the standards of medical practice. It must be emphasized, however, that user responsibility is secondary: it arises only after developers have fulfilled their primary responsibility to provide safe, transparent, and fair systems. It is unjust to burden users with extreme caution if a system is deliberately designed to mislead or conceal important information. Therefore, the hierarchy of responsibility must begin with the developer, followed by the regulator, and finally the user. Placing the burden of responsibility primarily on the user is a form of victim-blaming that is morally unacceptable.

The restructuring of the moral responsibility framework for algorithm-based decisions requires a layered approach that combines distributive responsibility across the production chain, corporate responsibility for collective entities, ongoing forward-looking accountability, a ban on agency laundering with severe sanctions, equal standards between human and machine decisions, a reversal of the burden of proof in certain conditions, dynamic monitoring mechanisms for continuously learning systems, and the establishment of an empowered, independent regulatory body. No single solution can resolve all issues because every case possesses its own unique technical and moral nuances. However, the direction of change is clear: moving away from a model that seeks individual scapegoats toward a model that views responsibility as a property of the overall socio-technical system. In this new model, the primary objective is not to punish after an error occurs, but to design systems that minimize the possibility of error from the outset. Laws and regulations must serve as a guide for safe design rather than as a safety net after an accident. This paradigm shift may take a generation to fully realize, but it must begin now because technology will not wait for our moral readiness.

CONCLUSIONS

This literature review identifies eight ethical principles that underpin the relationship between users and artificial intelligence systems in everyday life: autonomy, transparency, procedural justice, accountability, privacy, non-maleficence, honesty, and solidarity. These principles do not possess a fixed hierarchy of priority, but any decision to override one principle in favor of another must be carried out transparently with justifications that are open to public scrutiny. Reconfiguring the moral responsibility structure for algorithm-based decisions requires a layered approach that incorporates distributive responsibility across the production chain, corporate responsibility as a collective entity, forward-looking accountability, prohibition of agency laundering, equal standards between human and machine decisions, reversal of the burden of proof under certain conditions, dynamic monitoring mechanisms for continuously learning systems, and the establishment of independent supervisory bodies with legitimate authority. No single solution can resolve all ethical challenges of AI, but the trajectory of change is moving away from scapegoating models toward responsibility as a property of socio-technical systems as a whole.

The findings of this study have direct implications for policymakers, technology developers, and end users. For policymakers, there is a need for legislation mandating regular algorithmic audits, the establishment of independent supervisory bodies with authority to examine source code, and strict sanctions for companies proven to engage in agency laundering. Regulations should equalize responsibility standards between human and algorithmic decisions within the same domain, such as medical diagnosis or credit assessment. For technology developers, the implication is the necessity of integrating ethical considerations from the design stage rather than after product completion. This includes selecting representative training data, documenting design decisions transparently, and providing accessible appeal mechanisms for affected individuals. For end users, the study emphasizes that using AI is not a morally neutral act. Users bear responsibility for verifying algorithmic recommendations in critical situations, reporting injustices they encounter, and supporting organizations that advocate for algorithmic fairness. Educational institutions should begin teaching AI ethics literacy from the primary school level, as future generations will live in a world where algorithmic decisions are ubiquitous.

REFERENCES

Albareda, J. L. (2025). Uncovering the gap: Challenging the agential nature of AI responsibility problems. AI and Ethics. https://doi.org/10.1007/s43681-025-00685-w

Alsahafi, O., Alfaleh, A., Altamimi, M., Alghamdi, A., Alhafi, A., Altigani, A., Elsadig, M. A., Sulieman, S. M. A., & Mohamed, Y. A. (2024). The ethical issues surrounding artificial intelligence. Edelweiss Applied Science and Technology, 8(6), 9633–9640. https://doi.org/10.55214/25768484.v8i6.4064

Altehenger, H., & Menges, L. (2024). The point of blaming AI systems. Journal of Ethics & Social Philosophy. https://doi.org/10.26556/jesp.v27i2.3060

Arifin, S., Al Hakim, Y. R., Darmawan, D., Irfan, M., & Sigita, D. S. (2021). Technical and Ethical Dimensions of Search Engine Optimization in Managing Online Business Visibility. Studi Ilmu Sosial Indonesia, 1(1), 193-208.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732.
Beghili, M., Chétouani, M., & Barberousse, A. (2025). Ethical perspectives on natural autonomy and artificial autonomy. Frontiers in Artificial Intelligence and Applications. https://doi.org/10.3233/faia241491
Brusseau, J. (2025). The dilemma between euphoria and freedom in recommendation algorithms. arXiv.org, abs/2505.11465. https://doi.org/10.53136/979122182156728
Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. Washington Law Review, 89(1), 1–33.
Collyer-Hoar, G., & Rubegni, E. (2025). "Won't somebody please (actually) think of the children?" AI ethics for children: A scoping review. https://doi.org/10.1145/3745031
Creswell, J. W. (2009). Research design: Qualitative, quantitative, and mixed methods approaches (3rd ed.). Sage Publications.
da Silva, B. dos S., D. Darmawan, & B. Gardi. (2022). A Systematic Approach to Risk Management to Enhance Information Technology Project Success in a Dynamic Business Environment, Journal of Social Science Studies, 2(2), 213 – 218.
Darmawan, D. (2020). Health, Well-Being, and Productivity of Senior Employees in the Era of Artificial Intelligence. Journal of Science, Technology and Society, 1(2), 43-50.
Darmawan, D., & De Jesus Isaac, A. (2022). Self-identity formation and social perception of individuals through interaction on social media in a digital world. Journal of Social Science Studies, 2(2), 273–278.
Darmawan, D. (2022). Posthuman Human Resource Management in Organizations Using Generative Artificial Intelligence. Studi Ilmu Sosial Indonesia, 2(2), 97-124.
Essa, N. E., & Mardikaningsih, R. (2023). Sustainable and Fair Technology for an Equitable Society. Journal of Social Science Studies, 3(1), 355-362.
Fairclough, N. (2010). Critical discourse analysis: The critical study of language (2nd ed.). Longman.
Ferguson, M. (2025). Does forward-looking responsibility have an accountability problem? Social Theory and Practice. https://doi.org/10.5840/soctheorpract2025827242
Fleisher, W., Cibralic, B., Basl, J., Ricks, V., & Smith, M. N. (2025). Responsibility and accountability in an algorithmic society. Philosophy & Technology, 38(4). https://doi.org/10.1007/s13347-025-00970-w
Gani, A. & Darmawan, D. (2022). Ethics and Accountability in Artificial Intelligence-Based Managerial Decision Making, Journal of Social Science Studies, 2(1), 147 – 152.
Gardi, B., & Darmawan, D. (2024). Uncertainty, Isolation, and the Erosion of Safety Nets: The Structural Impact of the Gig Economy on Worker Mental Health. Studi Ilmu Sosial Indonesia, 4(1), 119-140.
Gardi, B., Ali, R., & Darmawan, D. (2024). Implementing Situational Leadership to Improve Team Performance in Multicultural Organizations. Journal of Social Science Studies, 4(1), 61-66.
Heaton, D. W., Clos, J., Nichele, E., & Fischer, J. E. (2023). The social impact of decision-making algorithms: Reviewing the influence of agency, responsibility and accountability on trust and blame. https://doi.org/10.1145/3597512.3599706
Heinrichs, J.-H. (2022). Responsibility assignment won't solve the moral issues of artificial intelligence. AI and Ethics, 2(4), 727–736. https://doi.org/10.1007/s43681-022-00133-z
Hugo, L. (2025). Towards ethical AI. https://doi.org/10.5281/zenodo.14721648
Khayru, R.K. (2022). Transforming Healthcare: The Power of Artificial Intelligence, Bulletin of Science, Technology and Society, 1(3), 15-19.
Krippendorff, K. (2004). Content analysis: An introduction to its methodology (2nd ed.). Sage Publications.
Kurniawan, Y., & Darmawan, D. (2021). The adaptive learning effect on individual and collective learning. Journal of Social Science Studies, 1(1), 93–98.
Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Sage Publications.
Lionel, N. (2025). Is artificial intelligence tech already relevant to our everyday life? https://doi.org/10.57708/bgut-r8-8sbgqub8vuuw4tg
Lundgren, B., Catena, E., Robertson, I., Hellrigel-Holderbaum, M., Jaja, I. R., & Dung, L. (2024). On the need for a global AI ethics. Journal of Global Ethics, 1–13. https://doi.org/10.1080/17449626.2024.2425366
Mahardani, U. K., & Mardikaningsih, R. (2024). Technologies Optimization to Increase Environmental Awareness and Employee Engagement in the Workplace. Journal of Social Science Studies, 4(1), 323-330.
Mardikaningsih, R., & Oluwatoyin, F. (2023). Analyzing Algorithmic Bias, Automated Justice, and Social Transformation in Artificial Intelligence Implementation. Studi Ilmu Sosial Indonesia, 3(1), 107-128.
Mardikaningsih, R., Darmawan, D., & Halizah, S. N. (2023). Worker Anxiety about Role Displacement from Artificial Intelligence Application in Human Resource Management. Studi Ilmu Sosial Indonesia, 3(2), 355-376.
Mardikaningsih, R. & Hariani, M. (2023). Technology Strategy in Product Development for Sustainable Innovation in Global Markets, Journal of Social Science Studies, 3(2), 71 – 76.
Maulani, A., R. Hardyansah, D. Darmawan, C. N. Mendonca, & A. de Jesus Isaac. (2023). Juridical Analysis of the Validity of Electronic Contracts Made by Artificial Intelligence in Indonesian Law, Journal of Social Science Studies, 3(1), 139 – 144.
Manure, A., Bengani, S., & Saravanan, S. (2023). Transparency and explainability. 61–106. https://doi.org/10.1007/978-1-4842-9982-1_3
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115, 1–35.
Nabavi, E., Nicholls, R., & Roussos, G. (2024). Locating responsibility in the future of human–AI interactions. IEEE Transactions on Technology and Society, 5(1), 58–60. https://doi.org/10.1109/tts.2024.3386247
Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.
Peckham, J. B. (2024). An AI harms and governance framework for trustworthy AI. IEEE Computer, 57, 59–68. https://doi.org/10.1109/mc.2024.3354040
Pistilli, G., & Trevelin, B. (2025). Can AI be consentful? arXiv.org, abs/2507.01051. https://doi.org/10.48550/arxiv.2507.01051
Priyatama, S., N. D. Aliyah, R. Mardikaningsih, M. E. Safira, F. Issalillah. (2022). Juridical Analysis of the Implementation of Artificial Insemination in Indonesia: Legal Status and Children's Rights in Positive Perspective Law, Bulletin of Science, Technology and Society, 1(2), 27-32.
Prunkl, C. (2024). Human autonomy at risk? An analysis of the challenges from AI. Minds and Machines, 34(3). https://doi.org/10.1007/s11023-024-09665-1
Putra, A. R., & Arifin, S. (2021). Supply Chain Management Optimization in the Manufacturing Industry through Digital Transformation: The Role of Big Data, Artificial Intelligence, and the Internet of Things, Journal of Social Science Studies, 1(2), 161 – 166.
Radjawane, L. E. & Mardikaningsih, R. (2022). Building Ethical and Fair Technology: Approaches to Responsible Technology Development and Application, Journal of Social Science Studies, 2(1), 189 – 194.
Rainer, T. (2022). Towards an accountability framework for AI: Ethical and legal considerations. https://doi.org/10.13140/rg.2.2.10231.50086
Ramle, N. L. B., & Mardikaningsih, R. (2022). Inclusivity in Technology-Based Services: Access and Skills Challenges, Journal of Social Science Studies, 2(2), 225 – 230.

Register, C., Khan, M. A., Giubilini, A., & Savulescu, J. (2025). Privacy and human-AI relationships. Philosophy & Technology, 38(4). https://doi.org/10.1007/s13347-025-00978-2

Sajjapong, T., Darmawan, D., & Marsal, A. P. (2022). The role of social stereotypes in shaping opportunities and inequalities in society: Their impact on education, employment, and intergroup interactions. Bulletin of Science, Technology and Society, 1(1), 44–49.

Sanderson, C., Schleiger, E., Douglas, D., Kuhnert, P., & Lu, Q. (2024). Resolving ethics trade-offs in implementing responsible AI. arXiv.org, abs/2401.08103. https://doi.org/10.48550/arxiv.2401.08103

Shank, D. B. (2022). Perceptions of violations by artificial and human actors across moral foundations. Computers in Human Behavior Reports, 5, 100154. https://doi.org/10.1016/j.chbr.2021.100154

Sinambela, E. A., Nurmalasari, D., Darmawan, D., & Mardikaningsih, R. (2021). The Role of Business Capital, Level of Education, and Technology in Increasing Business Income. Studi Ilmu Sosial Indonesia, 1(1), 77-92.

Sinambela, E. A., & Darmawan, D. (2022). Advantages and disadvantages of using electronic money as a substitute for cash. Journal of Social Science Studies, 2(2), 56–61.

Sinambela, E. A. (2023). Integration of Change Management and Technology Strategy in Digital Transformation. Journal of Social Science Studies, 3(1), 375-380.

Slota, S. C., Fleischmann, K. R., Greenberg, S. R., Verma, N., Cummings, B., Li, L., & Shenefiel, C. (2021). Many hands make many fingers to point: Challenges in creating accountable AI. AI & Society, 1–13. https://doi.org/10.1007/S00146-021-01302-0

Sutanto, H., Darmawan, D., & Saputra, R. (2023). Legal Violation Patterns in Digital Technology on Liability and Proof. Studi Ilmu Sosial Indonesia, 3(2), 277-308.

Tullio, M. D. (2022). The roots of responsibility. Think, 21(61), 23–27. https://doi.org/10.1017/s1477175621000427

Velibor, B., & Turyasingura, B. (2024). Ensuring responsible artificial intelligence (AI) development and use. https://doi.org/10.13140/rg.2.2.13190.79682

Verma, I. (2024). Privacy in the age of AI. 151–163. https://doi.org/10.58532/nbennurdtch14

Wang, G., & Pea, R. (2024). Algorithmic autonomy in data-driven AI. https://doi.org/10.48550/arxiv.2411.05210

Wang, Z., Huang, C., & Yao, X. (2024). Procedural fairness in machine learning. arXiv.org, abs/2404.01877. https://doi.org/10.48550/arxiv.2404.01877

Warin, A. K., & Mardikaningsih, R. (2025). Collective Worker Resistance to Green Human Resource Management Policies: Differences in Age, Gender, Status, and Subculture Solidarity. Studi Ilmu Sosial Indonesia, 5(2), 205-224.

Westover, J. H. (2025). The cognitive cost of AI assistance: Protecting human thinking in the age of generative AI. Human Capital Leadership, 26(1). https://doi.org/10.70175/hclreview.2020.26.1.6

Wu, Y. (2025). The right to be wrong? Human fallibility versus AI perfection. IEEE Computer, 58(11), 67–73. https://doi.org/10.1109/mc.2025.3601819