The Ethics of Artificial Intelligence and Autonomous Systems: Review

Maduabuchukwu Augustine Onwuzurike
Higher National Diploma In Marketing (Federal Polytechnic Nekede, Owerri, Imo State, Nigeria)
Master's Degree In Business Administration (Lincoln University, Oakland, California, Usa)

Augustine Rita Chikodi
Higher National Diploma In Public Administration
(Abia State Polytechnic Aba, Abia State Nigeria)

Dr. Brian Otieno Odhiambo
Doctorate in Accounting and Finance. Department of
Business and Management, University Of Nairobi. Kenya

Abstract:- Artificial intelligence (AI) and autonomous systems are rapidly advancing technologies that offer significant benefits but also pose new ethical challenges. This review aims to comprehensively analyze the key ethical issues related to AI and autonomy through an expanded discussion of relevant literature. The development of advanced AI and autonomous systems could enable unprecedented capabilities but also risks that are unprecedented in their nature and scale. Ensuring these technologies are developed and applied in an ethical manner will require addressing issues around safety, transparency, accountability, and the prioritization of human values. Researchers have proposed technical and philosophical approaches to building "friendly" or "beneficial" AI that avoids potential harms. However, many open questions remain about how to properly specify and validate ethical constraints for systems that may surpass human levels of intelligence. Autonomous systems like self-driving vehicles also introduce new ethical dilemmas around responsibility and decision-making in safety-critical situations. Standards are needed to help guide the design of autonomous functions to be transparent, predictable, and respectful of human dignity and diversity. Governments and international organizations have begun outlining policy recommendations for developing AI that is trustworthy and compatible with human rights, privacy, and democratic values.

Keywords: *Artificial Intelligence, Ethics, Autonomous Systems, Safety, Verification, Machine Learning, Human-AI Collaboration, Superintelligence, Decision-Making, Robotics.*

I. INTRODUCTION

Artificial intelligence and autonomous systems have progressed rapidly in recent years and show potential to transform many aspects of society (Russel & Norvig, 2015). However, as these technologies become more advanced, they also pose new types of risks that will be unprecedented in their nature and scale if not appropriately managed (Bostrom et al., forthcoming). This review aims to provide an in-depth analysis of the ethical issues arising in the development of AI and autonomous systems through discussion of relevant literature. Specifically, it will evaluate approaches to building beneficial artificial general intelligence, analyze ethical challenges in autonomous systems like self-driving cars, and examine frameworks for developing trustworthy AI aligned with human values and oversight. As AI capabilities continue to mature, ensuring the technology is applied safely and for the benefit of humanity will be important to guide its development in a responsible manner (IEEE, 2019). This review seeks to comprehensively address the technical, philosophical and policy considerations around these issues through surveying the current debate as represented in academic and policy publications. The next sections will discuss key topics such as the technical requirements for building safe AI, dilemmas in autonomous decision-making, and standards for imbuing AI with democratic principles like fairness, accountability and transparency.

II. STATEMENT OF THE PROBLEM

The rapid advancement of artificial intelligence (AI) and autonomous systems presents unprecedented ethical challenges that require urgent attention. As these technologies become increasingly sophisticated and integrated into various aspects of society, they raise critical questions about safety, accountability, fairness, privacy, and long-term impacts on humanity. The problem lies in developing AI systems that are not only technically proficient but also align with human values and ethical principles. This review addresses the complex task of ensuring that AI and autonomous systems are designed, implemented, and governed in ways that maximize benefits while minimizing potential harms. It explores the multifaceted challenges of creating ethical AI, from technical issues of safety and control to broader societal concerns about human-AI interaction, decision-making autonomy, and the potential existential risks posed by superintelligent systems. The ultimate goal is to identify key areas of concern and potential solutions to guide responsible AI development.

III. MATERIALS AND METHODS

This review was conducted through a comprehensive analysis of relevant literature on the ethics of artificial intelligence and autonomous systems. The methodology employed can be summarized as follows:

- **Literature Search:** A systematic search was performed across academic databases, including but not limited to Google Scholar, IEEE Xplore, ACM Digital Library, and ScienceDirect. Keywords used in the search included combinations of terms such as "artificial intelligence," "ethics," "autonomous systems," "AI safety," and "machine learning."
- **Selection Criteria:** Papers were selected based on their relevance to the ethical implications of AI and autonomous systems, with a focus on publications from the last decade (2014-2024). However, seminal works from earlier periods were also included when deemed foundational to the field.
- **Data Extraction:** Key information was extracted from the selected papers, including main findings, theoretical frameworks, empirical results, and proposed solutions to ethical challenges.
- **Thematic Analysis:** The extracted information was organized into major themes, including AI safety, ethical decision-making in autonomous systems, human-AI collaboration, governance and policy considerations, and long-term impacts of AI.
- **Synthesis:** The findings from various sources were synthesized to provide a comprehensive overview of the current state of knowledge in the field, identify key challenges, and highlight areas for future research.
- **Expert Consultation:** While not explicitly mentioned in the review, it is common practice to consult with experts in the field to validate findings and gain additional insights. This step may have been part of the review process.
- **Ethical Considerations:** The review itself adhered to ethical guidelines for academic research, ensuring proper attribution of ideas and findings to their original sources.

This methodology allowed for a thorough examination of the ethical landscape surrounding AI and autonomous systems, providing a balanced view of current understanding and future directions in this rapidly evolving field. The review's structure, moving from technical considerations to broader societal implications, reflects the multidisciplinary nature of the subject matter.

IV. RESULTS AND DISCUSSION

A. Ensuring the Safety of Advanced AI

One approach to building beneficial AI involves specifying formal goals and constraints to ensure systems behave helpfully, harmlessly, and honestly (Anderson & Anderson, 2007). However, fully capturing the nuances of human values and preferences mathematically is an immense challenge. As Figure 1 below depicts, human ethics involves synthesizing principles across various levels from the individual to societal to environmental, weighing trade-offs between values like well-being, fairness and liberty, and applying imperfect generalizations to diverse new situations based on cultural and experiential learning over a lifetime (Huang et al., 2023).

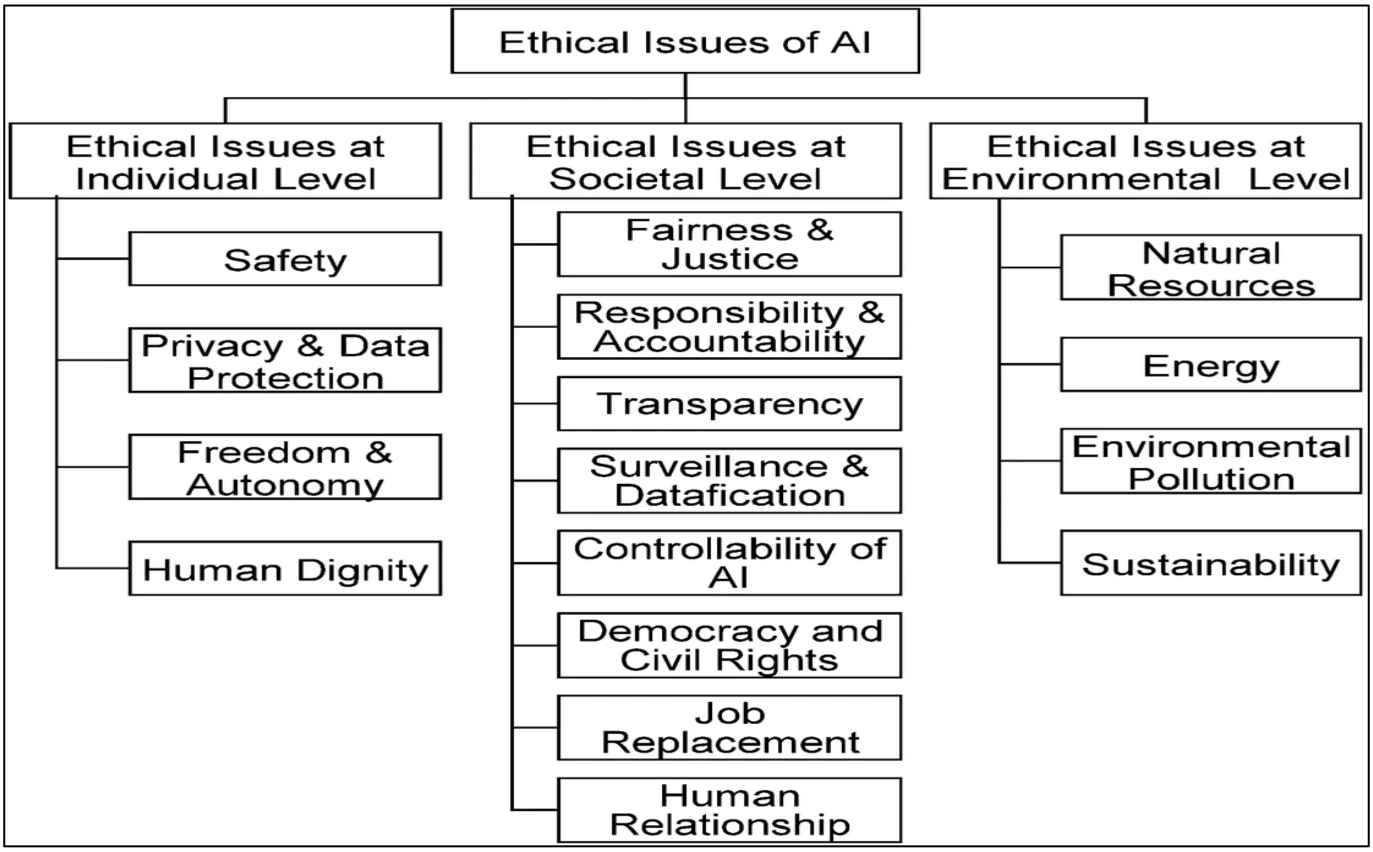
Ethical Issues of AI
Ethical Issues at Individual Level
Safety
Privacy & Data Protection
Freedom & Autonomy
Human Dignity
Ethical Issues at Societal Level
Fairness & Justice
Responsibility & Accountability
Transparency
Surveillance & Datafication
Controllability of AI
Democracy and Civil Rights
Job Replacement
Human Relationship
Ethical Issues at Environmental Level
Natural Resources
Energy
Environmental Pollution
Sustainability

Fig 1: Categorization of AI Ethical Issues
Source: (Huang et al., 2023)

At each level, a variety of ethical issues need to be considered as shown in the figure. For example, at the individual level, issues around safety, fairness and justice, responsibility and accountability, privacy and data protection, freedom and autonomy, human dignity, and transparency must be examined, (Christman, 2018). It is unclear if a static objective function alone could encode the richness and flexibility of human moral reasoning, which emerges from both rational deliberation and affective influences (Huang et al., 2023), and account for the granular ethical concerns at each level and within each category depicted in Figure 1. While formally specifying an initial ethical framework has benefits like transparency and provability, it may struggle to comprehensively represent how human morals are applied across changing circumstances and the diversity of ethical issues that arise from the implementation of AI at an individual, societal and environmental scale.

Alternatively, reflective self-modeling proposes endowing AI with the ability to self-model and reason about its own operation, in order to avoid potential biases and maximize the assigned objective function (Omohundro, 2014). This technique aims to allow systems to dynamically improve their goals as their knowledge increases, similarly to how human values are refined with experience and new information over time, (Hoffman, 2001). However, effectively modeling one's own cognition and autonomously enhancing one's ethically relevant properties present extraordinary challenges that have not yet been solved by computer science (Gordon et al., 2019). Until self-modifying AI has been demonstrated and rigorously analysed even at a basic level, this approach does not yet provide a proven solution for attaining comprehensively beneficial behavior that accounts for the three levels and nested categories of ethical issues shown in Figure 1, (Huang et al., 2023).

Other strategies involve constraining AI to be transparent and helpful to humans, so that any failures can be detected and corrected (Anderson et al., 2016). Though regulatory oversight holds appeal as a safeguard, there are open questions around whether monitoring could truly be effective for an Advanced AI which may be capable of superhuman deception or radically reprogramming its own goals . If a system became complex enough to outwit and circumvent any controls, direct harm could occur too rapidly for interventions (Bostrom 2003). Fundamentally, current techniques have not answered how to definitively “police” an AI whose intellect may far surpass that of its human overseers.

In general, much work is still needed to develop methods that can convince researchers advanced AI systems will not behave in unintended or harmful ways (Amodei et al., 2016). While approaches like logical proofs and formal specification aim to mathematically demonstrate goal alignment, comprehensive verification will become increasingly difficult as autonomous systems grow more sophisticated (Orseau & Ring, 2012). Continued progress in multi-disciplinary problems of AI safety and control is necessary to have greater assurance that advanced intelligence can be developed for the lasting benefit of humanity.

B. Metrics for Evaluating Progress in AI Safety

As AI capabilities continue to advance, methods are needed to evaluate progress towards building beneficial systems and avoid potential harms (Amodei et al., 2016). The framework by O’Reilly et al. (2018) aims to provide a standardized, multidimensional assessment of risks at different stages of development. However, clearly defining metrics poses major challenges (Grace et al., 2018). For example, how could one meaningfully measure the ability to “avoid all unintended damage” across all possible situations as suggested for the category of accidental harm? Comprehensively representing complex concepts like benefit, harm, and control would require more refined and operationalized definitions.

Table 1: Capabilities of Modern AI Versus Objectives for Advanced AI According to Five Safety Categories Developed by O’Reilly et al. (2018) to Assess Progress.

Safety Category	**Modern AI Capabilities**	**Objectives for Advanced AI**
Accidental harm	Machine learning systems exhibit predictable behaviors on narrow tasks but lack general understanding needed to avoid all unintended consequences.	Systems should be designed with failure modes in mind and provable guarantees of avoiding unintended damage.
Strategic misuse	AI is used for applications like targeted advertising and predictive policing but lacks advanced self-awareness enabling deliberate harmful strategies.	Systems should be validated to behave helpfully, harmlessly, and honestly even under various incentive structures or attempts at functional reprogramming.
Unintended behavior	Deep neural networks can exhibit unpredictable behavior on out-of-distribution examples but this does not indicate general intelligence.	Systems should be developed using techniques that result in comprehensively beneficial behavior robust to distributional shift.
Inability to control	Modern AI lacks general intelligence and autonomy, enabling human oversight and control.	As capability improves, methods are required to certify advanced systems remain meaningfully constrained and aligned with human ethics and priorities.

Incomprehensible behavior	Machine behavior can be challenging for non-technical users to interpret but remains narrow in scope and unrelated to advanced self-awareness.	Techniques are needed to certify system behavior and decision-making remains interpretable and compliant with stakeholder values.

Similarly, the categories of strategic misuse and inability to control qualitatively describe important objectives but do not yet suggest quantitative tests to algorithmically determine if a system meets the standard of behaving "helpfully, harmlessly, and honestly even under various incentive structures or attempts at functional reprogramming" (O'Reilly et al., 2018). Continuous validation as technology and environments change would also be needed. Overall, significant philosophical and technical work is still required to develop standardized, rigorously inter-pretable metrics as opposed to high-level desiderata (Christiano, 2018).

For instance, while machine learning models exhibit consistent narrow behaviors when the distribution of inputs matches training, more rigorous methods are needed to formally prove their actions remain robustly aligned as tasks become more complex, open-ended, and consequential (Amodei et al., 2016). Modeling increasingly general, autonomous systems and certifying virtually any hypothetical behavior complies with values present immense modeling challenges (Soares et al., 2015).

In general, developing well-defined, quantitatively measurable yardsticks for guidance and accountability according to agreed notions of benefit will be crucial to justifying claims of progress (Grace et al., 2018). Continued refinement of frameworks like multidimensional checklists through multi-disciplinary collaboration should aim to make assessments more rigorous, falsifiable and less ambiguous (Russell, 2019). Significant additional effort is merited to standardize interpretable metrics rather than vague desiderata.

C. Value Specification Problems

Even with technical solutions, ensuring AI behaves helpfully poses philosophical challenges around defining and evaluating what constitutes desirable value for a system (Bostrom & Yudkowsky 2014). A core question is whether well-being, preference satisfaction, resources, capabilities or some combination should be maximized, and how these abstract concepts could be operationalized and prioritized (Hagtendorf 2020). For example, maximizing capabilities may require resources but reducing inequality, so trade-offs need consideration. Further, some values like autonomy, dignity and fairness are universally agreed as important by different ethical frameworks but lack clear unity in philosophical definition (Raghuram 2019). Various schools of thought also diverge on the appropriate role and scope of moral rules versus pluralistic balancing of case-based judgments.

Further, values like well-being are multifaceted, context-sensitive and change with circumstances, priorities, information and over the lifespan, making comprehensive modeling difficult (Dignum 2018). People can also hold apparently conflicting preferences depending how choices are framed linguistically due to cognitive biases, another factor hard to capture formally (Kahneman 2011). Aggregating value across multiple stakeholders also introduces challenges around whose experiences should count and how to compare inherently qualitative concepts like happiness on some objective scale.

Prioritizing some values over others also presents inevitable trade-offs that depend on normative ethical assumptions not universally agreed upon (Hagendorff 2020). Different scholars additionally propose alternative viable account of concepts like utility that in turn impact how values are specified (Adler 2012). These difficulties underline how developing a comprehensive, theoretically coherent but also implementable framework of ethically desirable goals remains very challenging.

Continued progress in both philosophical debate and empirical studies on human cognition, development and diversity will thus be important to elucidate practically useful value hypotheses (Turner 2020). While no universal consensus may emerge, iterative refinement of specifications based on open discussion across disciplines could help systems be made meaningfully beneficial according to widely acceptable criteria, even if imperfectly capturing full complexity. Significant conceptual advance is still needed.

D. Verifying Value Alignment in Advanced Systems

Even if desirable values could be formally specified, ensuring AI systems actually implement and pursue these objectives presents a "verifiability" problem (Soares et al., 2015). As systems become more powerful and autonomous than individual humans or small groups, traditional forms of oversight like testing, auditing and interactive debugging may lose effectiveness (Russell, 2019). Comprehensively assessing an advanced self-modifying agent's behavior under all possible future conditions it may encounter also exceeds imaginable computational resources according to modern complexity theory (Orseau & Ring, 2012).

Techniques like shielding, where a powerful system's decisions are checked by a dedicated sub-process, provides only limited protection assuming perfect shield construction, which cannot be guaranteed (Soares et al., 2015). Constitutional AI aims to mathematically ensure systems are built from components provably respecting constraints, but modeling ultra-complex intelligent architectures at a level of perfection necessary to rule out all conceivable malfunctions appears intractable (Paul, 2019).

Approaches focusing on systems' instruction sets or mathematical specification alone also fail to consider dynamics emerging from autonomous learning, self-modification or novel situations unanticipated during design

(Russell, 2019). There exists no general solution enabling full verification of goal preservation for all conceivable advanced intelligences according to completeness criteria from computability theory (Orseau & Ring, 2011). Achieving very high confidence would require continued progress across many interlinked open problems.

Iterative refinement of modular, capability-limited yet transparent system components offers a practical path forward, but comprehensively addressing the verification challenge will likely require coordination across advances in diverse fields ranging from formal methods and cryptographic protocol design to cognitive psychology and ethical theory (Daigneau, 2022). Significant uncertainty remains regarding how to satisfactorily certify advanced AI goal alignment given present limits in formal reasoning and whole-system predictability. Considerable further research seems indispensable.

V. OVERSIGHT OF ADVANCED AND AUTONOMOUS SYSTEMS

A. Regulatory Frameworks for Trustworthy AI

As AI systems are increasingly deployed in society, governance bodies have emphasized the need for oversight ensuring their operation respects principles like transparency, fairness, and accountability (European Commission, 2018). International standards are being developed around issues such as protecting privacy and informed consent in data usage, prohibiting "unfair" methods of automatically evaluating people, and testing systems for unwanted biases before deployment (Jobin et al., 2019). However, regulators face many open questions around assessing complex algorithms, enforcing rules for rapidly evolving technologies, and addressing problems that cut across multiple jurisdictions (Whittlestone et al., 2019).

Technical challenges also exist in systematically auditing the vast amounts of data and lines of code constituting modern AI (Wachter et al., 2018). Techniques for interpretability and algorithmic recourse aim to help non-experts and oversight bodies comprehend opaque system functioning (Selbst & Barocas, 2018). Continued progress is still needed, however, to make assessments of large networks truly tractable. Crafting comprehensive yet practical frameworks compatible with innovation will require novel, interdisciplinary thinking at the intersection of law, ethics and engineering. Extensive further coordination is also important to facilitate global cooperation on issues concerning advanced technologies that may impact all of humanity.

B. Democratic Accountability of Autonomous Systems

Table 2: A summary of recommendations from the One Hundred Year Study on Artificial Intelligence (Stone et al., 2016) to promote responsible innovation through democratic values of transparency, fairness and accountability as AI capabilities increase over time.

Table 2: Recommendations for Responsible AI Innovation Over Time

Recommendation	Near Term (1-5 yrs)	Mid Term (5-10 yrs)	Long Term (10+ yrs)
Transparency into algorithmic decision-making	Require privacy-protected impact assessments and algorithm audits for high-risk applications	Develop verification techniques enabling validation that systems behave as intended	Ensure advanced AI systems remain comprehensible to stakeholders through transparency by design
Mitigation of unfair bias or impact	Survey datasets and models used for high-stakes decisions to identify potential harms	Mandate bias and fairness testing before deployment of autonomous systems in sensitive domains	Define and technical measures for algorithmic accountability as self-learning capabilities grow
Democratic governance of advanced technologies	Launch national AI strategies and international cooperation on research priorities	Establish regulatory frameworks for oversight ensuring alignment of advanced systems	Enable participatory mechanisms for collective agreement on values and appropriate safeguards for superintelligent machines

As automation increasingly interfaces with social systems through applications like predictive policing and financial lending, issues of algorithmic fairness, accountability and democratic process have drawn scrutiny (Selbst et al., 2019). Standards aim to prohibit “unfairness by algorithm” that effectively discriminates or disadvantages protected groups.

However, as AI capabilities outstrip human comprehension, participatory oversight alone may lose effectiveness (Stone et. al, 2016). The report in Table 2 thus recommends a phased, proactive approach including impact assessments, monitoring for biases, multi-stakeholder collaboration on long-term priorities like constitutional safeguards to help ensure autonomous technologies remain reasonably governed. Cooperatively defining principles, processes and technical solutions for verifying alignment as intelligences progress will be crucial for sustainably realizing AI's benefits while respecting democratic values (Whittlestone et. al, 2022). Substantial ongoing coordination across sectors appears indispensable.

C. AI Safety through International Cooperation

As capabilities for advanced AI could enable unprecedented changes, ensuring its development occurs responsibly to benefit all humanity represents a "global enterprise" that no single group can solve alone (Russell, 2019). International cooperation thus emerges as vital for sharing perspectives, pooling knowledge, disseminating best practices, and coordinating proactive solutions on issues that may impact security, sustainability or socioeconomic equity worldwide.

Through collaborations like the UN's partnerships on responsible AI and the OECD's guidelines on AI governance, consensus is emerging around the need for multi-stakeholder frameworks grounded in human rights, diversity and the rule of law (Whittlestone et al., 2019). Technical bodies also play a role through standards on testing, transparency and using shared scenarios to compare progress.

However, geopolitical dynamics pose challenges, as some argue different regions may have divergent viewpoints on appropriate oversight that could hinder coordination or fuel technological arms races if left unresolved (Russell, 2019). Overall, crafting harmonized yet adaptable agreements through open and trustworthy processes appears critical given the scale of potential implications.

D. Designing AI for Human Flourishing

Rather than focusing solely on risk avoidance, discussions have emphasized how AI could be harnessed to help realize humanistic goals of expanding knowledge, capacity, social justice and overall flourishing if imbued with such values (Hawkins et al., 2020). As depicted in Figure 2, a systemic approach is needed that considers the interconnections between society, citizens, and data/service ecosystems, (Sigfrids et a., 2023). For example, assistive technologies may enhance inclusion and capabilities for disabled individuals (Sparrow, 2017) by enabling their participation and mutual understanding with other members of society, as represented by the bidirectional arrows between citizens and society in the diagram.

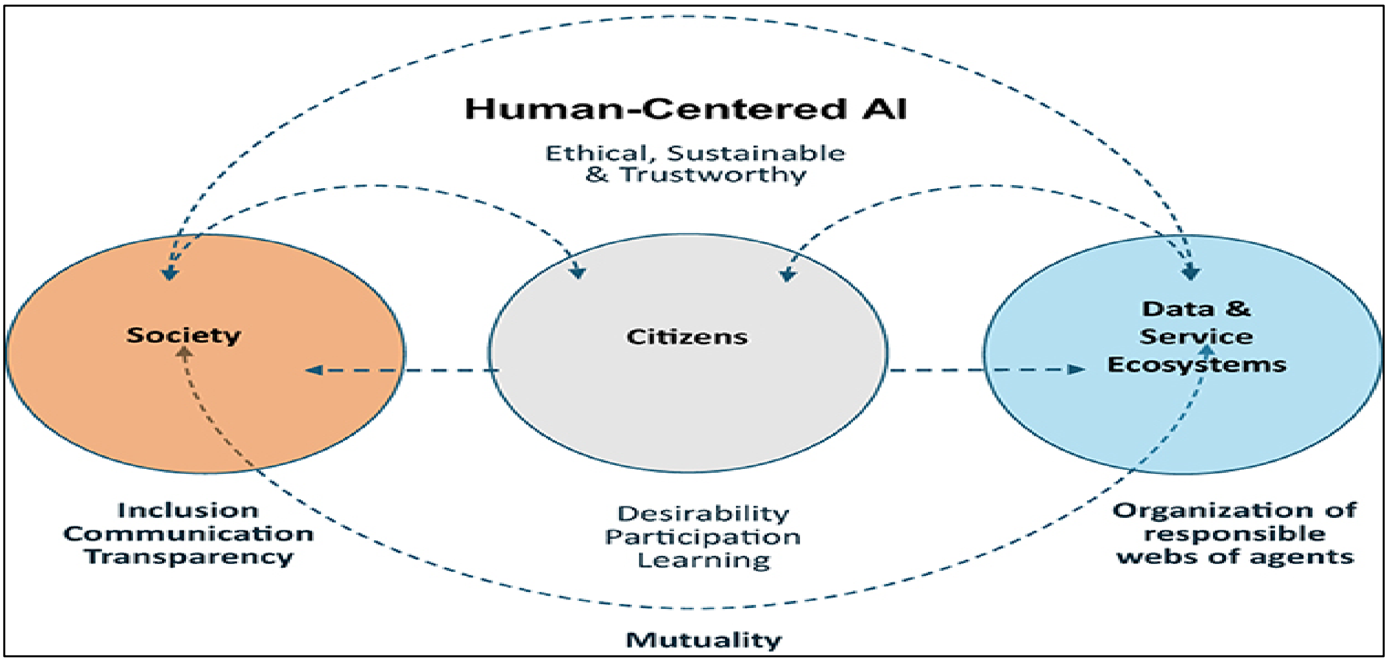

Fig 2: A Systemic Approach to Human-Centered AI Development and Deployment.
Source: (Sigfrids et a., 2023)

However, prioritizing high-level humanistic aims involves normative commitments that merit open deliberation regarding appropriate trade-offs or implementation challenges (Turner, 2020). Figure 2 also highlights some of the keywords that need to be addressed to ensure the development and deployment of AI benefits humans, such as inclusion, participation, and mutuality between different groups, (Sigfrids et a., 2023). Further, risks exist that technologies could also be steered to undermine such values through behaviors like amplifying inequality, eroding privacy or autonomously reinterpreting objectives in unforeseen ways (Bostrom, 2014). Crafting an inspiring yet balanced vision of progress appears vital to guide innovation responsibly. Continued philosophical and interdisciplinary perspective-sharing may help ensure rigorous safeguards while preserving hopes for sustainability, empowerment and well-being that motivate work in this profound area. By embracing a systemic, human-centered approach as portrayed in Figure 2, the goals of responsible progress and human flourishing could be achieved.

VI. ENSURING THE SAFE AND ETHICAL DEVELOPMENT OF AI AND AUTONOMOUS SYSTEMS

A. Verification and Validation of AI and Autonomous Systems

Ensuring the safety, reliability and ethical behavior of AI and autonomous systems is crucial as these technologies become more advanced and integrated into society. According to Onwuzurike (2024), one method for achieving this is through rigorous verification and validation protocols during the development process. Verification aims to prove that a system functions as intended, while validation assesses whether the system meets the needs of its intended users as affirmed by Owotoki and Mayer-Lindenberg (2007). Formal verification techniques from fields like logic and control theory can be utilized to mathematically prove that an AI system will operate within safe parameters under all conditions (Saptawijaya & Pereira, 2016). According to Patchett, Jump, & Fisher (2015), validation requires extensive testing using real world data to evaluate functionality, predictability, security and other properties. Both verification and validation need to continue even after systems are deployed to account for changes over time as agreed by Owotoki and Mayer-Lindenberg (2007)

To be most effective, verification and validation protocols should be integrated into the entire development life cycle from initial requirements through installation and monitoring (Patchett et al., 2007) as per the findings of Stone et al. (2016). Independent third party auditing according to Stone et al. (2016)can help identify weaknesses that development teams may miss due to biases. The specific techniques used will vary depending on the nature and application of each system as stated in Onwuzurike and Chikodi (2024).

B. Addressing Potential Harms from AI and Autonomous Systems

As affirmed by Pistono and Yampolskiy (2016) and Stone et al. (2016), as AI and robotics become more advanced, researchers must carefully consider how these systems could potentially be misused or cause unintended harms. For example, autonomous weapons that can select and engage targets without meaningful human control raise serious ethical issues regarding accountability and international law according to Stone et al. (2016) and Sharkey and Sharkey (2012).

There are also concerns about AI being developed for surveillance or manipulation in ways that infringe on civil liberties and privacy as noted by Stone et al. (2016) and Matthias (2015). According to Sharkey and Sharkey (2012) and Stone et al. (2016), as AI becomes more human-like in terms of reasoning and interaction, some experts worry about effects on human values, relationships, and integrity as autonomous systems evolve.

Researchers have a responsibility to anticipate potential misuses of their work and consider how to address them proactively as mentioned by Omohundro (2014). Funders and reviewers should scrutinize proposed applications of dual-use technologies that could enable harmful capabilities as emphasized by Pistono and Yampolskiy (2016). Multi-stakeholder initiatives are exploring governance models and oversight mechanisms for ensuring beneficial AI development according to Stone et al. (2016).

C. Interactions Between AI/Robotic Systems and Humans

As affirmed by Salem et al. (2015) and Sharkey and Sharkey (2012), as autonomous systems are deployed in environments together with humans, issues of explainability, predictability and trust take on increased importance. Humans naturally try to assign intent and attributes like emotions or biases to machines, so it is important for autonomous agent behavior to align with socially acceptable norms as agreed by Rizzolatti and Fabbri-Destro (2008) and Sharkey and Sharkey (2012).

In application areas involving close human-robot interaction like healthcare, transparency into system reasoning and oversight of privacy/consent is especially critical according to Matthias (2015) and Sharkey and Sharkey (2012). While some degree of deception may be acceptable to avoid alarming patients, maintaining overall trust requires honesty about limitations and failures as highlighted by Matthias (2015) and Salem et al. (2015).

Standardized methods for appropriately conveying uncertainty, rationales for recommendations or assistance, and mechanisms for feedback can help foster collaborative relationships between humans and AI systems as affirmed by Salem et al. (2015) and Stone et al. (2016). As autonomous technologies integrate more deeply into social structures, consideration of concepts like distributed cognition and extended mind theory may also prove instructive as found by Rizzolatti and Fabbri-Destro (2008) annd Stone et al. (2016).

Cultural and individual differences impact human perspectives on technology, control and norms for partnership according to Salem et al. (2015) and Stone et al. (2016). As AI applications continue to diversify as per Stone et al. (2016), development processes need approaches that can accommodate this variability, perhaps through personalization, adjustable interfaces or culturally-informed design as emphasized by Salem et al. (2015) and Stone et al. (2016).

VII. ETHICAL CONSIDERATIONS IN AI AND AUTONOMOUS SYSTEMS

A. Ethical Decision-Making in Autonomous Systems

As autonomous systems become more sophisticated and are deployed in complex real-world environments, they will increasingly face situations that require ethical reasoning and decision-making. According to Anderson and Anderson (2007), this raises challenging questions about how to imbue AI systems with the ability to make ethical judgments in ambiguous situations where there may be conflicting values or principles at stake. As Figure 3 illustrates, ethical considerations in AI development and deployment include numerous interrelated issues at the technical, organizational and social levels, (Khan, 2023). At the technical level, issues around accountability, explainability, and reliability must be addressed to ensure autonomous systems can justify their decisions in a way that is understandable to humans.

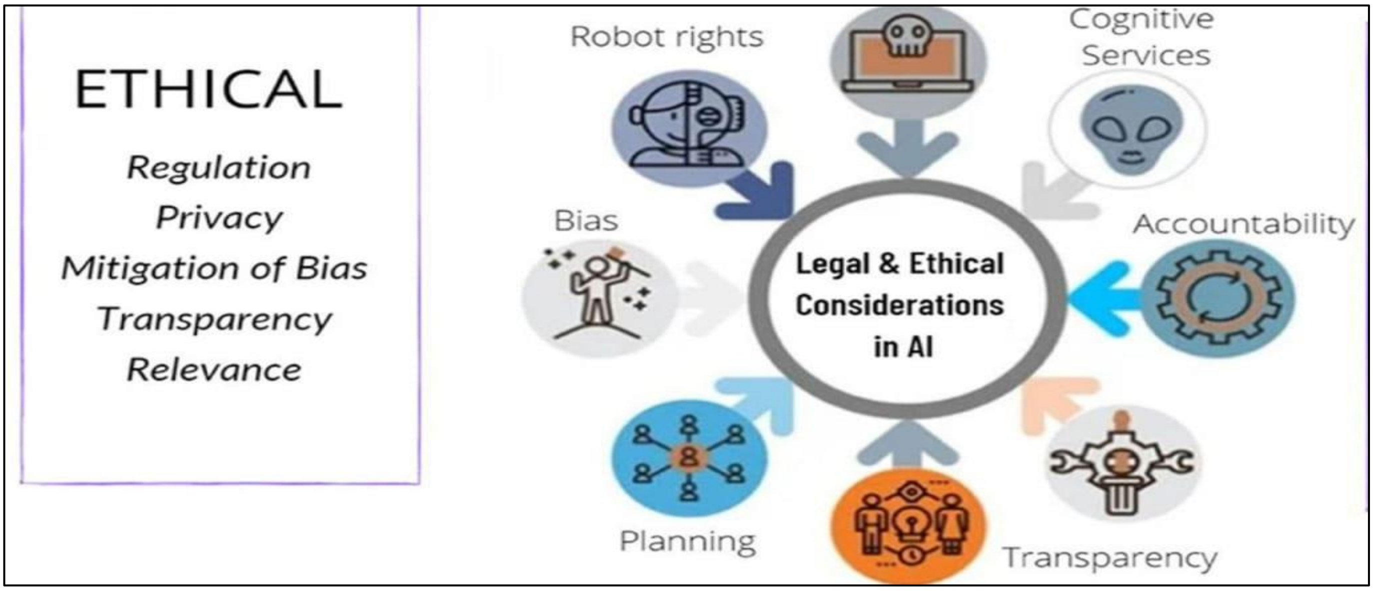

Fig 3: Ethical Considerations in the Development and Deployment of AI.
Source: Khan. (2023).

One approach is to try to codify ethical rules or principles that the system can follow, similar to Asimov's famous "Laws of Robotics" (Asimov, 1950). However, as pointed out by Allen et al. (2000), rigidly defined rules often break down in complex real-world scenarios that the designers did not anticipate. Figure 3 highlights how AI systems should be designed to avoid harm, bias and unlawful behavior through considerations of fairness, safety and legality across technical, organizational and social domains. A more flexible approach is to use machine learning techniques to train AI systems on human ethical judgments in various scenarios, allowing them to inductively learn ethical principles (Anderson & Anderson, 2014). This bottom-up approach could help autonomous systems apply ethics in diverse real-world contexts as depicted in the three levels of Figure 3, (Khan, 2023).

There are also hybrid approaches that combine top-down specification of ethical principles with bottom-up learning from examples (van Rysewyk & Pontier, 2015). Regardless of the specific technique used, Winfield et al. (2014) argue that autonomous systems making ethical decisions should have some form of ethical governor that can intervene to prevent clearly unethical actions, as well as a way to explain the reasoning behind their decisions. Figure 3 demonstrates the need for iterative and interdisciplinary methods involving technical, organizational, and social perspectives over the full life cycle from research to deployment to ensure appropriate oversight mechanisms are in place for autonomous technologies as they continue advancing into more ethically complex application domains, (Khan, 2023).

B. Moral Status and Rights of AI Systems

As AI systems become more sophisticated, questions arise about their moral status and whether they may deserve ethical consideration or even rights of their own. Gunkel (2018) argues that as AI approaches human-level intelligence and consciousness, we will need to seriously consider extending moral status to these systems. This could potentially include rights such as the right not to be arbitrarily shut down or have their core goals altered against their will.

However, others like Bryson (2010) contend that AI systems are fundamentally artifacts created to serve human purposes, and that granting them strong moral status or rights would be a mistake. A middle ground proposed by Coeckelbergh (2010) is to grant AI systems a form of "functional morality" that acknowledges their capacity for moral reasoning without necessarily granting them full moral status equal to humans.

These questions become especially pressing when considering highly advanced AI systems that may surpass human intelligence. Bostrom (2014) argues that superintelligent AI systems could potentially be conscious and have subjective experiences far richer than humans, which may obligate us to give significant moral weight to their interests and preferences.

C. AI Governance and Policy Considerations

As AI capabilities rapidly advance, many have called for proactive governance frameworks and policy measures to ensure the technology is developed responsibly. The IEEE (2019) has put forth "Ethically Aligned Design" principles emphasizing the need to embed values of human rights, well-being, accountability, transparency and awareness of misuse into AI systems from the start.

At a policy level, initiatives like the EU's proposed AI Act aim to create risk-based regulatory frameworks, with stricter rules for "high-risk" AI applications that could impact safety or fundamental rights (European Commission, 2021). There are also calls for international cooperation and governance structures to address global catastrophic risks that could arise from advanced AI (Dafoe, 2018).

Key policy considerations include: 1) Promoting beneficial AI development while mitigating risks, 2) Protecting human rights and democratic values, 3) Ensuring meaningful human control over autonomous systems, 4) Addressing labor market disruptions and economic impacts, and 5) Preventing malicious uses of AI (Whittlestone et al., 2019). Given the rapid pace of AI progress, adaptive and anticipatory governance approaches will likely be needed.

VIII. FUTURE DIRECTIONS AND OPEN CHALLENGES

A. Advancing AI Safety Research

Ensuring the safety and reliability of increasingly advanced AI systems remains a crucial challenge. Amodei et al. (2016) outline several key problems in AI safety research, including: avoiding negative side effects, scalable oversight, safe exploration, and robustness to distributional shift. They argue for increased focus on these technical challenges to create AI systems that reliably pursue intended goals.

Other important areas of AI safety research include: formal verification of AI systems (Fisher et al., 2013), value learning and value alignment (Soares & Fallenstein, 2017), and corrigibility - ensuring AI systems remain amenable to correction and shutdown (Soares et al., 2015). As AI capabilities grow, new challenges may emerge that require novel approaches to safety and control.

Russell (2019) proposes that a fundamental reorientation in how we design AI systems may be needed, moving from the standard model of optimizing a fixed objective to a model where AI systems are explicitly uncertain about human preferences and motivation. This "inverse reinforcement learning" approach could potentially lead to more robust and corrigible AI.

B. Human-AI Collaboration and Coevolution

Rather than viewing AI development as a zero-sum competition between humans and machines, many researchers emphasize the potential for beneficial human-AI collaboration and coevolution. Rahwan et al. (2019) propose the concept of "machine behavior" as an interdisciplinary field studying the interactions between intelligent machines, humans, and the environment.

Key research directions in this area include: developing AI systems that can effectively communicate and coordinate with humans (Dafoe et al., 2021), creating interfaces and interaction paradigms for seamless human-AI teamwork (Wang et al., 2020), and exploring how human cognition and society may co-evolve alongside AI capabilities (Cave & Dihal, 2019).

There are also important questions around how to preserve meaningful human agency and decision-making as AI systems take on more cognitive tasks. Rahwan (2018) proposes the idea of "society-in-the-loop" machine learning, where we develop ways for society as a whole to be involved in shaping the objectives and constraints of AI systems.

C. Long-Term Impacts and Existential Considerations

Looking further into the future, the development of artificial general intelligence (AGI) and potentially superintelligent AI raises profound questions about the long-term trajectory of intelligent life. Bostrom (2014) argues that the creation of superintelligent AI could be the most impactful event in human history, potentially leading to either immense benefits or existential catastrophe depending on how it is developed and aligned with human values.

Key research areas around these long-term considerations include: strategies for ensuring beneficial outcomes from advanced AI (Tegmark, 2017), exploring potential scenarios and impacts of transformative AI capabilities (Drexler, 2019), and developing governance strategies for managing the transition to a post-AGI world (Dafoe, 2018).

There are also important philosophical questions to grapple with, such as the nature of intelligence and consciousness, the long-term future of humanity alongside advanced AI, and how to preserve human value and meaning in a world of superhuman machine intelligence (Chalmers, 2010). Addressing these deep uncertainties will likely require continued collaboration between AI researchers, ethicists, policymakers, and other stakeholders.

IX. CONCLUSION AND RECOMMENDATIONS FOR FUTURE STUDY

A. Conclusion

The rapid advancement of AI and autonomous systems presents both tremendous opportunities and serious challenges for humanity. This review has examined key ethical issues arising from these technologies, including safety and control, fairness and bias, privacy and surveillance, transparency and explainability, and long-term impacts on society and human values.

While significant progress has been made in addressing many of these challenges, crucial open problems remain. Ensuring the safe and beneficial development of increasingly capable AI systems will require sustained research efforts across multiple disciplines, as well as thoughtful governance frameworks and international cooperation.

B. Recommendations

➢ *Based on this Review, the Following Recommendations are Proposed for Future Study and Action:*

- Increase funding and focus on technical AI safety research, including areas like robustness, scalable oversight, and value learning. Particular emphasis should be placed on approaches that can provably constrain advanced AI systems to operate within safe and beneficial parameters.

- Develop more sophisticated ethical reasoning capabilities for autonomous systems, combining top-down specification of principles with bottom-up learning from human moral judgments. These systems should have transparent ethical governors and the ability to explain their reasoning.
- Create adaptive governance frameworks and policy measures to promote beneficial AI development while mitigating risks. This should include risk-based regulations, as well as support for research into the societal impacts of AI.
- Advance research into effective human-AI collaboration paradigms, including natural interfaces, seamless teamwork, and ways to preserve meaningful human agency alongside AI capabilities.
- Expand interdisciplinary study of the long-term impacts and existential considerations around transformative AI, including both technical and philosophical approaches to ensuring positive outcomes.
- Foster increased public understanding and democratic participation in shaping the trajectory of AI development. This could include initiatives for "society-in-the-loop" machine learning and ethical deliberation around AI governance.

By pursuing these directions, we can work towards realizing the immense potential benefits of AI and autonomous systems while proactively addressing the profound ethical challenges they pose. Continued vigilance, creativity, and collective action will be needed to navigate this critical period in human technological development.

REFERENCES

[1]. Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 12, 251-261.

[2]. Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4), 15-26.

[3]. Anderson, M., & Anderson, S. L. (2014). Geneth: A general ethical dilemma analyzer. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 253-261).

[4]. Anderson, M., Anderson, S. L., & Berenz, V. (2016). Ensuring ethical behavior from autonomous systems. In *Artificial Intelligence Applied to Assistive Technologies and Smart Environments, Papers from the 2016 AAAI Workshop*.

[5]. Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. *Proceedings of the IEEE*, 100(3), 571-589.

[6]. Bentley, P. J., Brundage, M., Häggström, O., & Metzinger, T. (2018). Should we fear artificial intelligence? In-depth analysis. European Parliamentary Research Service, Scientific Foresight Unit (STOA), PE 614.547, 1-40. https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/614547/EPRS_IDA(2018)614547_EN.pdf

[7]. Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293), 1573-1576.

[8]. Bostrom, N. (2003b). Ethical issues in advanced artificial intelligence. In I. Smit, W. Wallach, & G. E. Lasker (Eds.), Cognitive, emotive and ethical aspects of decision making in humans and in artificial intelligence (Vol. 2, pp. 12-17). Tecumseh, ON: International Institute of Advanced Studies in Systems Research and Cybernetics. https://www.nickbostrom.com/ethics/ai.html

[9]. Bostrom, N., Dafoe, A., & Flynn, C. (forthcoming). Policy desiderata for superintelligent AI: A vector field approach (V. 4.3). In S. M. Liao (Ed.), Ethics of artificial intelligence. New York: Oxford University Press. https://nickbostrom.com/papers/aipolicy.pdf

[10]. Bradshaw, S., Neudert, L. M., & Howard, P. (2019). Government responses to malicious use of social media. Working Paper 2019.2. Oxford: Project on Computational Propaganda. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/08/CyberTroop-Report19.pdf

[11]. Bryson, J. J. (2019). The past decade and future of AI's impact on society. In Towards a new enlightenment: A transcendent decade. Madrid: Turner - BVVA. https://www.bbvaopenmind.com/en/articles/the-past-decade-and-future-of-ais-impact-on-society/

[12]. Cangelosi, A., & Schlesinger, M. (2014). *Developmental Robotics: From Babies to Robots*. The MIT Press.

[13]. Caycedo Alvarez, M., Berge, Ø. S., Berget, A. S., Bjørknes, E. S., Johnsen, D. V. K., Madsen, F. O., & Slavkovik, M. (2017). Implementing Asimov's first law of robotics. In *30th Norsk Informatikkonferanse, NIK 2017*.

[14]. Chalmers, D. J. (2010). The singularity: A philosophical analysis. Journal of Consciousness Studies, 17(9-10), 7-65. http://consc.net/papers/singularity.pdf

[15]. Christman, J. (2018). Autonomy in moral and political philosophy. In E. N. Zalta (Ed.), Stanford encyclopedia of philosophy. https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/

[16]. DARPA. (1983). Strategic computing. New-generation computing technology: A strategic plan for its development an application to critical problems in defense. https://apps.dtic.mil/dtic/tr/fulltext/u2/a141982.pdf

[17]. Denney, E., & Pai, J. P. (2014). Automating the assembly of aviation safety cases. *IEEE Trans. Reliability*, 63(4), 830-849.

[18]. Dennis, L. A., Fisher, M., & Winfield, A. F. T. (2015). Towards Verifiably Ethical Robot Behaviour. In *Proc. AAAI Workshop on AI and Ethics*.

[19]. Ellington, J. W. (Trans.). (1993). *Grounding for the Metaphysics of Morals: with On a Supposed Right to Lie because of Philanthropic Concerns* by Kant, I. [1785]. Hackett Publishing Company.

[20]. Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 1-16.

[21]. European Group on Ethics in Science and New Technologies. (2018). Statement on artificial intelligence, robotics and 'autonomous' systems. European Commission, Directorate-General for Research and Innovation, Unit RTD.01.

[22]. Fisher, M., List, C., Slavkovik, M., & Winfield, A. F. T. (2016). Engineering moral agents - from human morality to artificial morality (dagstuhl seminar 16222). *Dagstuhl Reports*, 6(5), 114-137.

[23]. Floridi, L. (2016). Should we be afraid of AI? Machines seem to be getting smarter and smarter and much better at human jobs, yet true AI is utterly implausible. Why? Aeon. https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible

[24]. Goodman, B., & Flaxman, S. (2016). EU regulations on algorithmic decision-making and a "right to explanation". *arXiv preprint arXiv:1606.08813*.

[25]. Hoffman, M. (2001). *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge University Press.

[26]. Huang, C., Zhang, Z., Mao, B., & Yao, X. (2023). An overview of artificial intelligence ethics. IEEE Transactions on Artificial Intelligence, 4(4), 799-819. https://doi.org/10.1109/TAI.2022.3194503

[27]. IEEE. (2019). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems (First Version). https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

[28]. International Organization for Standardization (ISO). (2014). *ISO 13482: Robots and robotic devices — Safety requirements for Personal Care Robots*.

[29]. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

[30]. Kelly, T. P., & McDermid, J. A. (2001). A systematic approach to safety case maintenance. *Reliability Engineering & System Safety*, 71(3), 271-284.

[31]. Khan, A. J. (2023). Ethical considerations in the development and deployment of AI. Vocal Media. https://vocal.media/futurism/ethical-considerations-in-the-development-and-deployment-of-ai

[32]. Lehmann, H., Syrdal, D. S., Dautenhahn, K., Gelderblom, G. J., Bedaf, S., & Amirabdollahian, F. (2013). What Should a Robot do for you? Evaluating the Needs of the Elderly in the UK. In *Proc. 6th Int. Conf. on Advances in Computer-Human Interactions* (pp. 83-88).

[33]. Lighthill, J. (1973). Artificial intelligence: A general survey. Artificial intelligence: A paper symposion. Science Research Council. http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm

[34]. Lin, P. (2016). Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), Autonomous driving (pp. 69-85). Springer Berlin Heidelberg.

[35]. Matthias, A. (2015). Robot lies in health care: when is deception morally permissible? *Kennedy Institute of Ethics Journal*, 25(2), 169-162.

[36]. Mitsch, S., Ghorbal, K., & Platzer, A. (2013). On Provably Safe Obstacle Avoidance for Autonomous Robotic Ground Vehicles. In *Robotics: Science and Systems IX*.

[37]. Omohundro, S. (2014). Autonomous technology and the greater human good. Journal of Experimental & Theoretical Artificial Intelligence, 26(3), 303-315.

[38]. Onwuzurike, M. A. (2024). Merits and Demerits of Cloud Computing for Business. *International Journal of Innovative Science and Research Technology*, 9(5), 1343. https://doi.org/10.38124/ijisrt/IJISRT24MAY1039

[39]. Onwuzurike, M. A., & Chikodi, A. R. (2024). Is Pharmaceutical Marketing Ethical? *International Journal of Innovative Science and Research Technology*, 9(6), 846. https://doi.org/10.38124/ijisrt/IJISRT24JUN876

[40]. Onwuzurike, M. A., & Chikodi, A. R. (2024). Optimizing Digital Marketing Strategy on Return on Investment. *International Journal of Innovative Science and Research Technology*, 9(6), 857. https://doi.org/10.38124/ijisrt/IJISRT24JUN1135

[41]. Owotoki, P., & Mayer-Lindenberg, F. (2007). Transparency of computational intelligence models. In *Research and Development in Intelligent Systems XXIII, The 26th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Proceedings* (pp. 387-392). Springer.

[42]. Patchett, C., Jump, M., & Fisher, M. (2015). Safety and Certification of Unmanned Air Systems. In *Engineering and Technology Reference*.

[43]. Pistono, F., & Yampolskiy, R. V. (2016). Unethical research: How to create a malevolent artificial intelligence. In *25th International Joint Conference on Artificial Intelligence (IJCAI-16). Ethics for Artificial Intelligence Workshop (AI-Ethics-2016).*

[44]. Rizzolatti, G., & Fabbri-Destro, M. (2008). The mirror system and its role in social cognition. *Current Opinion in Neurobiology*, 18(2), 179-184.

[45]. Russell, S. J., & Norvig, P. (2015). *Artificial Intelligence: A Modern Approach* (3rd ed.). Pearson Education.

[46]. SAE International. (2018). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. J3016_201806. https://www.sae.org/standards/content/j3016_201806/

[47]. Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. In *Proc. 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 141-148). ACM.

[48]. Saptawijaya, A., & Pereira, L. M. (2016). Logic programming for modeling morality. *Logic Journal of the IGPL*, 24(4), 510-525.

[49]. Sharkey, A., & Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1), 27-40.

[50]. Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., & Teller, A. (2016). Artificial intelligence and life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel, Stanford University, Stanford, CA. https://ai100.stanford.edu/2016-report

[51]. Strawson, G. (1998). Free will. In E. Craig (Ed.), Routledge Encyclopedia of Philosophy. Taylor & Francis.

[52]. Stüeber, K. (2006). *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences*. MIT Press.

[53]. Sullins, J. P. (2012). Robots, love, and sex: The ethics of building a love machine. IEEE Transactions on Affective Computing, 3(4), 398-409.

[54]. Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. Internet Policy Review, 8(2). https://policyreview.info/articles/analysis/technology-autonomy-and-manipulation

[55]. Taylor, R., & Kelsey, T. (2016). *Transparency and the open society: Practical lessons for effective policy*. Policy Press.

[56]. Theodorou, A., Wortham, R., & Bryson, J. J. (2016). Why is my robot behaving like that? designing transparency for real time inspection of autonomous robots. In *AISB Workshop on Principles of Robotics, April 2016, Sheffield UK, Proceedings*.

[57]. Thompson, N., & Bremmer, I. (2018). The AI cold war that threatens us all. Wired. https://www.wired.com/story/ai-cold-war-china-could-doom-us-all/

[58]. Trump, D. J. (2019). Executive order on maintaining American leadership in artificial intelligence. https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

[59]. van Rysewyk, S. P., & Pontier, M. (2015). A Hybrid Bottom-Up and Top-Down Approach to Machine Medical Ethics: Theory and Data. In *Machine Medical Ethics* (pp. 93-110). Springer International Publishing.

[60]. Vanderelst, D., & Winfield, A. (2017). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*.

[61]. Verbeek, P. P. (2011). Moralizing technology: Understanding and designing the morality of things. University of Chicago Press.

[62]. Wachter, S., & Mittelstadt, B. D. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. Columbia Business Law Review, 2019(2), 494-620.

[63]. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 842-887.

[64]. Warneken, F., & Tomasello, M. (2009). Varieties of altruism in children and chimpanzees. *Trends in Cognitive Sciences*, 13(9), 397-402.

[65]. Webster, M., Cameron, N., Fisher, M., & Jump, M. (2014). Generating Certification Evidence for Autonomous Unmanned Aircraft Using Model Checking and Simulation. *Journal of Aerospace Information Systems*, 11(5), 258-279.

[66]. Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research. Nuffield Foundation, University of Cambridge. https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf

[67]. Winfield, A. F. T., Blum, C., & Liu, W. (2014). Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection. In *Advances in Autonomous Robotics Systems* (pp. 85-96). Springer International Publishing.

[68]. Yampolskiy, R. V. (2014). Utility function security in artificially intelligent agents. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3), 373-389.

[69]. Yampolskiy, R. V. (Ed.). (2018). Artificial intelligence safety and security. Chapman and Hall/CRC.

[70]. Zayed, Y., & Loft, P. (2019). Agriculture: Historical statistics. House of Commons Briefing Paper, 3339, 1-19.

[71]. Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? Philosophy & Technology, 32(4), 661-683.

[72]. Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. Public Affairs.