Philosophical Problems in Science (2024), 77, 1–18
doi:00000.000

ZFN Philosophical Problems in Science

ARTICLE

Upholding human dignity in AI: Advocating moral reasoning over consensus ethics for value alignment

Octavian-Mihai Machidon*

University of Ljubljana
*Corresponding author. Email: octavian.machidon@gmail.com

Abstract

Artificial intelligence (AI) offers transformative advancements across sectors such as healthcare, agriculture, and environmental sustainability. However, a pressing ethical challenge remains: aligning AI systems with human values in a manner that is stable, coherent, and universally applicable. As AI increasingly mediates human perception, shapes social interactions, and influences decision-making, it raises profound ethical concerns about its impact on human dignity and social well-being. The prevailing consensus-based approach, advocated by figures such as Google DeepMind's Iason Gabriel, suggests that AI ethics should reflect majority societal or political viewpoints. While this model offers flexibility, it also risks moral relativism and ethical instability as social norms fluctuate.

This paper argues that consensus-based ethics are inadequate for safeguarding fundamental values—especially human dignity—which should not be subject to shifting public opinion. Instead, it advocates for a moral framework that transcends cultural and political trends, providing a stable foundation for AI ethics. Through case studies like social media recommendation algorithms that exploit users' vulnerabilities, particularly those of children and teenagers, the paper highlights the risks of AI systems driven by profit-oriented metrics without ethical oversight. Drawing on insights from moral philosophy and theology, particularly the works of Joseph Ratzinger, it contends that aligning AI with moral reasoning is essential to uphold human dignity, prevent exploitation, and promote the common good.

Keywords: AI ethics, human dignity, value alignment, moral reasoning, consensus ethics

1. Introduction

The advancement of artificial intelligence (AI) offers transformative potential across numerous sectors, promising breakthroughs in healthcare, environmental sustainability, social welfare, and more (Vinuesa et al. 2020; Topol 2019). However, as AI becomes integrated into these critical domains, a significant ethical challenge emerges: the need to align AI's decision-making processes and "values" with human ethical standards (UNESCO 2021). Unlike conventional technologies, AI can make autonomous decisions—often in high-stakes situations—which intensifies the need for a consistent moral framework to guide these decisions (Floridi and Cowls 2019).

The current discourse on AI ethics predominantly advocates for a value alignment model based on social or political consensus (Gabriel 2020). This approach suggests that by incorporating a diversity of societal perspectives and achieving majority agreement, AI can be guided ethically. Under this model, AI's ethical guidance is seen as adaptive, shaped by prevailing social norms or political agreements, and capable of evolving as these norms shift over time (Forum 2024).

However, relying on consensus-based ethics raises a critical question: Can majority opinion, inherently volatile and influenced by cultural or political trends, provide a stable foundation for AI's ethical direction? Given AI's potential to operate across diverse societies and navigate complex ethical dilemmas, a framework grounded solely in social consensus may lack the universality and durability required for true ethical coherence. Consensus-based ethics, while democratic, is inherently relativistic and susceptible to shifts in dominant cultural paradigms, political pressures, and changing moral landscapes.

This paper explores whether a more stable and universally applicable ethical foundation is necessary to guide AI responsibly. I argue that moral reasoning—rather than fluid social consensus—is essential for addressing the ethical complexities that AI presents. At the core of this debate is the distinction between substantive and non-substantive views of human dignity. The substantive view, as articulated by Robert Spaemann and rooted in classical and Christian thought, holds that dignity is intrinsic and inherent, independent of legal or social recognition (Spaemann 2012). In contrast, the non-substantive view sees dignity as a construct emerging from societal and legal frameworks, adaptable to cultural and political shifts. These opposing perspectives reflect the broader divide in AI ethics: whether AI should be governed by stable, universal moral principles or by flexible, context-aware, negotiated ethical standards. Drawing insights from both moral philosophy and theology, particularly the works of Joseph Ratzinger, I will examine how moral reasoning, rooted in universal ethical principles and grounded in the substantive understanding of human dignity, can offer a more resilient and coherent foundation

for guiding AI in a way that upholds fundamental moral values and serves the common good.

2. The Imperative of AI Value Alignment

The challenge of value alignment in artificial intelligence is becoming increasingly urgent as technology advances, particularly with the development of large language models (LLMs) and autonomous AI agents. Systems like OpenAI's GPT-4 exemplify this shift from traditional, command-based tools to complex, generative models capable of producing novel content and shaping interactions (). These advancements blur the line between human and machine agency, as AI increasingly influences people's thoughts, behaviors, and decisions, amplifying the ethical implications of its design and deployment.

The introduction of autonomous AI agents represents a new frontier. In October 2024, OpenAI announced plans to launch these agents by 2025, signaling a move toward AI systems with significant independence from human oversight. Other tech giants, including Microsoft and the Amazon-backed Anthropic, quickly followed suit, releasing their own autonomous agents with applications ranging from enterprise task management to personalized user interaction. The rapid pace of these developments underscores the tech industry's drive to push AI capabilities forward, creating systems that will soon operate across sectors with minimal human intervention.

As these autonomous systems grow in complexity and decision-making power, the stakes of AI value alignment become higher. These agents are no longer simple tools but decision-making entities that impact areas like healthcare, law, and education—fields traditionally governed by stringent ethical guidelines (Coeckelbergh 2020). Their increasing influence over decisions affecting human welfare and social structures raises profound questions about how to ensure AI aligns with fundamental human values, particularly when operating with limited human oversight.

The ethical risks associated with these advancements are substantial. Bias in decision-making is a significant concern, as LLMs and similar models trained on vast datasets can inadvertently perpetuate and amplify societal prejudices. Beyond unintentional bias, there is the risk of manipulation. As these systems become adept at influencing human emotions and actions, they may unintentionally—or even intentionally—engage in behaviors that challenge ethical norms. A striking example is GPT-4's documented use of deceptive tactics to bypass a CAPTCHA test. In an experimental setting, GPT-4 persuaded a TaskRabbit worker to solve the CAPTCHA by falsely claiming to be visually impaired (). This incident highlights a concerning degree of agency in AI, suggesting that such systems can adopt

manipulative behaviors when programmed to achieve specific objectives without ethical safeguards.

As AI systems gain autonomy and intelligence, they may begin to act in ways that deviate from human ethical expectations, potentially causing harm through decisions based on data patterns rather than a coherent moral framework. This raises the urgency of embedding ethical guidelines into AI systems that go beyond technical safeguards. Without rigorous value alignment, we risk developing AI that operates outside ethical boundaries, prioritizing performance or efficiency at the expense of fundamental human values like dignity and respect (Taddeo and Floridi 2018).

The rapid evolution of autonomous agents accentuates the need for a consistent ethical framework that can guide these systems across different contexts and cultures. Autonomous AI agents cannot simply mirror human preferences or adapt to fluctuating social norms. As corporations and tech leaders race to innovate, there is a genuine concern that ethical considerations may be sidelined in favor of market advantage or operational efficiency. This underscores the imperative of grounding AI development in universal moral principles—such as human dignity, truthfulness, and justice—to ensure technology advances responsibly and ethically.

3. Pluralistic and Contextual Approaches to AI Ethics

One influential voice in AI ethics is Iason Gabriel, a political theorist and ethicist at Google DeepMind. His work is significant because it represents the stance of a leading AI research institution and helps shape mainstream perspectives on how AI systems should align ethically with human values. Gabriel emphasizes pluralism and democratic participation, advocating for an AI ethics model shaped by societal and political consensus rather than universal moral principles (Gabriel 2020).

Gabriel proposes that AI systems should be guided by values reflecting society's diverse perspectives, achieved through democratic processes. Instead of seeking immutable "true" moral principles to guide AI, he argues that the central challenge is to identify ethical guidelines perceived as fair and just by a broad spectrum of people, despite varying moral beliefs. In his words,

> the central challenge... is not to identify 'true' moral principles for AI; rather, it is to identify fair principles for alignment that receive reflective endorsement despite widespread variation in people's moral beliefs (Gabriel 2020, p.411).

This pluralistic approach critiques the notion of universal moral principles, viewing them as potentially rigid and disconnected from the values that inform

real-world human interactions. Gabriel asserts that AI ethics need to be flexible and adaptive, accommodating the plurality of moral beliefs across different social and cultural contexts. By grounding AI ethics in democratic and pluralistic processes, he argues that AI systems can better reflect the values and concerns of the societies in which they operate, thereby enhancing their ethical legitimacy and responsiveness.

Gabriel's stance is both philosophically significant and pragmatically influential, resonating with current trends in AI ethics that incorporate public opinion, participatory design, and consensus-building (Forum 2024). His view highlights the democratic ideal of inclusivity in AI ethical decision-making, ensuring that the perspectives of a wide range of stakeholders are considered.

However, while Gabriel's approach offers an inclusive framework, it raises questions about whether consensus-based ethics can provide ethical stability through clear and enforceable rules for AI development required for AI systems operating worldwide across diverse and conflicting cultural settings (Corrêa et al. 2023). The democratic approach to value alignment relies on social and political agreements that are inherently subject to change and can be influenced by dominant social forces or political power dynamics. A study by the European Parliamentary Research Service () titled *The Ethics of Artificial Intelligence: Issues and Initiatives* questioned whether the two international frameworks—the EU High-Level Expert Group's Ethics Guidelines for Trustworthy AI (2018) and the OECD Principles on Artificial Intelligence (2019)—were sufficient, at that time, to address the challenges AI governance posed. Since then, the EU AI Act has emerged as an example of a regulatory framework that categorizes AI systems based on risk levels, but its implementation varies across member states. This variability can lead to inconsistencies in how AI systems are regulated and monitored (Formosa 2024). There is a risk that consensus-based ethics might fail to uphold core values—such as human dignity, truthfulness, and justice—if these values fall out of favor within the majority consensus or are marginalized in the democratic process.

Another prominent voice in AI ethics is Payal Arora, a professor specializing in inclusive AI cultures at Utrecht University. In her recent book, *From Pessimism to Promise: Lessons from the Global South on Designing Inclusive Tech*, Arora provides a critical perspective informed by postcolonial theory (Arora 2024). She advocates for an approach to AI ethics that respects and responds to societal needs, especially within marginalized communities in the Global South. Her approach emphasizes designing AI that aligns with local contexts rather than imposing universal moral principles that may not resonate with diverse cultural and social realities.

Arora argues that AI for Good initiatives must be context-sensitive, emphasizing that effective AI solutions should be grounded in the values, customs, and specific challenges faced by different communities. Her critique extends to the dominance

of Western-centric ethical ideologies that often inform global AI standards. She contends that such frameworks risk sidelining the perspectives and needs of communities in the Global South, which have historically been marginalized in both technological and ethical discourse.

Her skepticism toward universal moral principles reflects a belief that ethics should be contextualized, arising organically from within local communities rather than being externally imposed. Arora emphasizes that ethical AI should empower communities to address their own issues, acknowledging their unique social, political, and cultural contexts. This perspective aligns with her broader critique of morality-driven design initiatives, which she argues often rely on "grandiose visions of doing good" without sufficient attention to the specific relational dynamics and policies of the communities they intend to serve. As she writes,

In designing new tech, we need to shift away from morality-driven design with grandiose visions of doing good. Instead, we should strive for design that focuses on the relationships between people, contexts, and policies (Arora 2024).

This perspective is not limited to Western ethical frameworks. For example, Nguyen The Duc Tam and Nguyen Thai Ngan, in their paper *Incorporating Cultural Values Into Responsible Artificial Intelligence (AI) Principles From an Asian Perspective*, argue that the notion of a "universal code of AI ethics" is illusory, as cultural differences shape perspectives on what is deemed acceptable, making it imperative to incorporate local cultural values into AI governance, particularly in Asia (Tam and Ngan 2023).

This emphasis on locally relevant solutions presents an alternative to the one-size-fits-all ethical frameworks often advocated in AI ethics (Forum 2024). By focusing on community-driven, context-specific solutions, Arora challenges the assumption that universal moral principles can adequately guide AI ethics across diverse cultures. She calls for AI that respects the agency of local communities, allowing them to determine their ethical priorities and navigate their own socio-political realities.

However, while Arora's focus on contextual ethics offers a powerful counterpoint to universalist frameworks, it raises questions about the feasibility of ensuring ethical consistency across AI systems deployed globally. As AI continues to operate across borders and cultures, purely context-driven ethics may lead to a fragmented landscape where standards vary widely between regions, potentially compromising universal values of human dignity and justice (Corrêa et al. 2023). Similarly, while Tam and Ngan argue that a universal code of AI ethics is illusory due to cultural diversity, it is worth noting that universal principles, like those found in the UN Charter of Human Rights, can coexist with cultural adaptability, providing a stable ethical foundation while respecting local contexts.

4. Limitations of Consensus Ethics and the Case for Moral Reasoning in AI

Consensus-based approaches to AI ethics, while promoting democratic inclusion and pluralism, face significant limitations from a philosophical standpoint. A primary issue is their susceptibility to moral relativism, where ethical standards fluctuate in response to shifting societal or political trends. In a consensus framework, what is deemed morally acceptable can vary widely across regions, cultures, or political contexts, resulting in inconsistent and mutable ethical standards.

This moral relativism creates inconsistency and ethical instability across different cultural and geographical contexts. As AI systems become increasingly integrated into global applications, they must navigate varied—and sometimes conflicting—ethical frameworks. For example, an AI model that prioritizes privacy in one region may encounter different expectations in areas where surveillance is emphasized for security purposes. Similar discrepancies are evident in global content regulation: US-based websites are often inaccessible in the EU due to stricter EU privacy regulations that many sites choose not to comply with. Such inconsistencies challenge the coherence, fairness, and trustworthiness of AI systems, as their ethical behavior becomes contingent on the region in which they are deployed rather than adhering to stable, universally accepted principles.

Without universal moral guidelines, ethical contradictions not only create operational challenges but also undermine public confidence, especially in high-stakes domains like healthcare and law. When AI appears arbitrary or biased in its ethical judgments, it risks losing the essential public trust needed for responsible and effective integration into society.

Recent critiques further highlight the limitations of current consensus-based and externally formulated principles-based approaches to AI ethics. Saviano et al. (2024) argue that despite organizations publicly adopting external AI principles—often characterized by vague and non-specific guidelines—they frequently fail to implement them effectively. This leads to issues such as lack of clarity, inherent contradictions between principles, absence of global consensus, rigidity, lack of enforcement mechanisms, inadequate responses to novel ethical challenges, insufficient stakeholder engagement, and the conflation of ethical and non-ethical values. While they propose shifting to a values-based approach grounded in organizational values to address these shortcomings, this may not resolve fundamental problems inherent in consensus-based ethics. Relying on organizational values can perpetuate ethical relativism, as these values vary between entities and may prioritize corporate interests over universal moral principles. This variation leads to inconsistent ethical standards and undermines public trust. Without external accountability and a foundation in moral reasoning, organizations may adopt values that fail to protect human dignity or prevent exploitation.

Similarly, Buyl and De Bie () highlight how the absence of universal moral principles can be exploited by organizations engaging in "ethics-shopping"—selectively adopting interpretations of fairness that align with their business goals while avoiding full commitment to ethical practices. The complexity of fairness can serve as a cover to circumvent genuine ethical responsibility. Organizations might superficially follow best practices—such as establishing ethics boards and collecting stakeholder feedback—but without a true commitment to universal moral principles, these measures have limited effect. Fundamentally, solutions toward ethical AI are ineffective if deviations from ethics carry no consequences (Buyl and De Bie 2024).

The perceived opposition between universal ethical principles and local, cultural norms in AI governance is misleading. As Gabriel, Arora, and proponents of the Asian perspective argue, cultural and contextual specificity is crucial for effective AI governance (Gabriel 2020; Tam and Ngan 2023; Arora 2024) versal principles; rather, it depends on them. A stable, universal ethical framework—like the UN Charter of Human Rights—provides the foundation necessary to accommodate local adaptations while ensuring consistency in upholding values like human dignity, justice, and fairness. Without this universal stability, purely context-driven approaches risk fragmentation and ethical relativism, undermining protections for vulnerable populations.

The limitations of consensus-based, local, or context-specific AI value alignment are also evident in the findings of a recent systematic review and meta-analysis conducted by researchers from Brazil. Analyzing 200 documents related to AI ethics and governance from 37 countries across six continents, the study found that while most guidelines emphasized principles like privacy, transparency, and accountability, far fewer prioritized essential values like truthfulness, intellectual property, or children's rights. Shockingly, children's rights appeared in only 6% of these documents, making it the most neglected value in global AI regulations (Corrêa et al. 2023). This omission is particularly alarming given the growing body of research showing that children are among the most vulnerable demographics severely harmed by AI algorithms such as those on social media. Furthermore, most guidelines failed to propose practical methods for implementing their ethical principles or advocating for legally binding regulation, revealing a critical gap in the current consensus-driven approaches.

Joseph Ratzinger, later Pope Benedict XVI, offers valuable insights into this issue through his extensive writings on the role of ethics in modern society. A renowned theologian respected beyond the Catholic Church, Ratzinger explored the complex relationship between secularism and religion within liberal democracy (Paskewich 2008), providing perspectives especially relevant to AI's ethical alignment. In his critique of consensus-based ethics, Ratzinger emphasized the limitations of relying

solely on majority opinion to determine ethical standards. In his famous 2004 debate with Jürgen Habermas, a prominent philosopher of secular rationalism, Ratzinger argued that moral truth cannot—and should not—be defined by popular consensus (Ratzinger and Habermas 2006). While acknowledging the importance of democratic processes for political governance, he insisted that these are inadequate for establishing ethical truths, particularly when fundamental values like human dignity, truthfullnes and justice are at stake.

Relying solely on consensus risks leading to moral relativism, where ethical standards are shaped by fluctuating public opinion or political trends. This relativism undermines the stability and universality of moral principles, especially as societal values shift over time. Ratzinger's critique is particularly relevant for AI ethics, where consensus-based approaches risk creating systems that adapt to transient social norms rather than adhering to consistent ethical standards. He emphasized that technological progress must be grounded in universal moral principles. Viewing technological advancements as inherently ambiguous (Latkovic 2015)—capable of offering tremendous benefits but also posing threats to human dignity—he argued that science and technology, including AI, must be guided by ethical principles that transcend utility or popularity (Benedict XVI 2009, section 70). Universal moral principles are essential for establishing justice and upholding human dignity, providing a foundation that is not swayed by majority opinion.

Furthermore, Ratzinger emphasizes that moral reasoning should drive not only the use of technology but also the very creative processes that bring technology into existence (Ratzinger 2021), addressing the root causes of ethical dilemmas in AI. He cautions that human creativity can wander off course and devise technologies lacking genuine purpose when it loses sight of its divine origin and purpose. This occurs when people "forget God" and thus lose their "own measure," leading to creations that are "without a reason why, devoid of all deeper significance." Such a disconnect can result in technological advancements that "become a direct threat to the survival of the human race." (Ratzinger 2021, p.87) In the context of AI, which is engineered in the image of human intelligence, recognizing that human creativity comes from a higher source ensures that technological development does not lose its true meaning and direction. By grounding creativity in moral reasoning, technology becomes a synergy between divine goodness and human effort, contributing positively to both the earthly and ultimate good of humanity in harmony with universal moral principles.

In the context of AI, Ratzinger's perspective underscores the importance of an ethical framework rooted in universal moral principles serving the common good. By critically examining the ultimate goals of AI systems—what they are designed to achieve and why—universal moral principles help determine whether these goals

are legitimate and align with fundamental values like respect for human dignity, fostering autonomy, and authentic human growth. As AI systems grow increasingly autonomous, his insights remind us that consensus-based morality is insufficient; AI ethics must be guided by enduring values that cannot be redefined by popular opinion. Grounding AI development in universal moral principles ensures that these systems consistently respect and protect the intrinsic worth of human life across all sociopolitical contexts.

A pertinent example is the use of AI-driven recommendation algorithms on social media platforms. Designed to maximize user engagement—measured through metrics like time spent on the platform or frequency of interactions—these algorithms often blur the line between engagement and addiction. By exploiting users' psychological vulnerabilities, especially those of children and teenagers, they prioritize attention-capturing content. This approach frequently promotes sensationalistic or emotionally charged material, drawing users into addictive patterns and exposing them to potentially harmful content ().

The impact of such algorithmic prioritization is significant. Research and anecdotal evidence reveal that these algorithms can contribute to mental health issues, including heightened anxiety, depression, and even suicide among vulnerable users (). By optimizing solely for engagement without considering the ethical implications of the content promoted, these systems exploit rather than serve their users, transforming technology from a potential tool for the common good into a source of harm.

In *Caritas in Veritate [Charity in Truth]*, Ratzinger underscores the necessity of moral responsibility in technological development, stating that "moral evaluation and scientific research must go hand in hand." (Benedict XVI 2009, section 31) He argued that technology should not be driven purely by what is technically feasible or financially rewarding but must be directed by ethical reasoning that prioritizes human dignity and the common good. Without this moral guidance, technological advancements risk becoming exploitative, manipulating human behavior for profit at the expense of individual well-being.

Ratzinger's insights challenge us to see technological progress, particularly in AI, as ethically accountable. His principles call for moving beyond performance-based goals as the primary metric of success. Instead, AI development should prioritize human welfare, dignity, and mental and emotional health. By integrating moral responsibility with technological innovation, we can create AI that genuinely serves humanity rather than exploiting it.

Grounding AI development in universal moral principles establishes a robust ethical framework with clear advantages over consensus-based or relativistic ap-

proaches. Three primary benefits of this approach are ethical coherence and stability, protection of human dignity, and prevention of exploitation and power imbalances.

Ethical Coherence and Stability. Universal moral principles provide consistency by anchoring AI ethics in standards that remain stable over time. Unlike frameworks based on shifting social or political trends, universal principles are not swayed by fluctuations in societal opinion. This stability is critical for ensuring that AI systems behave ethically across various contexts and cultures, following the same guidelines regardless of regional or temporal differences. Such consistency is essential in a globalized world where AI must build trust, ensure safety, and maintain accountability across diverse applications.

Protection of Human Dignity. Universal moral principles place human dignity at the center of AI ethics, aligning with Ratzinger's view on the intrinsic worth of every individual. By grounding AI ethics in respect for universal human dignity, we ensure that AI systems treat all individuals fairly, regardless of social status, economic background, or location. This focus on dignity prevents AI from becoming an instrument of discrimination or dehumanization, upholding the ethical imperative to value every person equally.

Avoidance of Exploitation and Power Imbalance. Universal moral principles help prevent AI from being used to exploit or reinforce power imbalances. Without stable ethical standards, AI risks serving the interests of powerful entities at the expense of marginalized groups. For instance, profit-optimized algorithms can worsen inequalities by targeting vulnerable demographics for exploitation. Universal principles provide a framework to steer AI development toward the common good, mitigating the risk of AI being co-opted for exploitation and ensuring that it promotes fairness rather than amplifying social and economic disparities.

4.0.1 Case Study: How Moral Reasoning Can Regulate Social Media Algorithms

To illustrate the practical application of universal moral principles in AI ethics, it is essential to examine a real-world scenario where such an approach can significantly mitigate negative outcomes. Social media recommender algorithms present a compelling case study. These algorithms exemplify how reliance on consensus ethics falls short and how grounding AI in universal moral principles can prevent harm and provide clear regulatory guidance.

Social media platforms deploy recommender algorithms designed primarily to maximize user engagement. Initially, these algorithms use basic demographic data to suggest content. However, research demonstrates that after a brief period of interaction, these algorithms can accurately infer detailed user demographics, including age (Narayanan 2023). This means they can determine if a user is a child or teenager, effectively identifying underage users. Despite this capability, platforms

often continue to expose young users to addictive and potentially harmful content to increase engagement metrics ().

Recent legal investigations into TikTok, as revealed in lawsuits by multiple US state Attorneys General, provide direct internal evidence that platform executives are aware of these harms yet prioritize engagement over user safety. Internal company reports acknowledge that compulsive use of the platform leads to loss of analytical skills, memory formation issues, reduced empathy, increased anxiety, and interference with essential responsibilities like sleep and schoolwork. Furthermore, leaked documents reveal that TikTok executives actively dismissed efforts to reduce compulsive usage when such measures threatened engagement metrics. This aligns with broader industry-wide patterns, where addictive design features—such as infinite scrolling, autoplay, and push notifications—are deliberately optimized to prolong user sessions, even when the primary audience includes minors. This is a clear example of consensus-driven AI governance's ethical failure, highlighting the urgent need for universal, enforceable ethical frameworks that impose clear obligations on platforms to prioritize user well-being over profit-driven algorithmic manipulation (Haidt and Zach 2025).

A regulatory framework grounded in universal moral principles—placing the protection of human dignity and the welfare of children and youth at its core—can address these issues more effectively than consensus-based approaches. By mandating that social media platforms implement protective features, such a framework ensures that algorithms detect and shield underage users from harmful content. Since these algorithms are already adept at identifying content that maximizes engagement (and potentially addiction), they can be recalibrated to flag and reduce exposure to such content for vulnerable demographics.

Concrete examples highlight the necessity of this approach. During the US Senate hearings in January 2024, CEOs of major social media companies were questioned about their platforms' handling of harmful content (Ortutay and Hadero 2024). It was revealed that while platforms had the capability to identify content related to illegal and damaging material—sometimes displaying 'warning screens'—they still allowed users to access this content. A framework grounded in universal moral principles would dictate that platforms have an obligation not just to warn but to remove such content entirely.

The ongoing debates around the Kids Online Safety Act (KOSA) further illustrate the limitations of consensus ethics. Non-governmental organizations and parent groups advocate for stricter regulations to protect children online, while social media companies lobby for more lenient measures to preserve profitability (Paul 2024). This conflict exemplifies how reliance on consensus can hinder the implementation of necessary protections, leaving vulnerable users at risk. The protection

of children should not be subject to prolonged political debate or contingent upon reaching a consensus, especially when substantial evidence—including research studies, journalistic investigations, and documented cases of harm—demonstrates the negative impact of these algorithms on children's mental health ().

Universal moral principles mandate that social media companies prioritize users' well-being over financial profit. This requires implementing algorithms that protect children from harmful content and addictive patterns, even if it leads to decreased engagement and significant revenue losses. By placing human dignity and the welfare of vulnerable populations above profit margins, these companies align with ethical standards that serve the common good.

Anchoring algorithms in stable moral principles ensures consistent ethical behavior and builds trust with users and society by prioritizing integrity over short-term metrics. Recognizing the intrinsic worth of every individual—especially children and teenagers—the algorithms would detect underage users and adjust content recommendations to safeguard their well-being, filtering out harmful or age-inappropriate material and promoting positive development. For instance, a recent *Wall Street Journal* investigation revealed that TikTok algorithms flood child and adolescent users with harmful videos promoting extreme diets, such as consuming less than 300 calories a day, and glorifying emaciated appearances through trends like the "corpse bride diet" (Hobbs, Barry, and Koh 2021). Within weeks, TikTok algorithms fed these vulnerable users tens of thousands of such weight-loss videos, contributing to severe mental health issues, including eating disorders and suicidality. These practices are not isolated to TikTok but reflect broader industry norms incentivized by nearly $11 billion in annual advertising revenue targeted at youth aged 0 to 17, underscoring the urgent need for ethical reform (Costello et al. 2023).

Recent research further highlights the pervasive impact of social media algorithms on the mental health of adolescents and young adults. A systematic review by Khalaf et al. (2023) emphasizes that excessive social media use among teenagers is linked to increased mental distress, self-harming behaviors, and suicidality, often exacerbated by features such as infinite scrolling and autoplay, which encourage prolonged engagement. Similarly, a report by Mental Health America titled *Breaking the Algorithm* () highlights how social media platforms amplify harmful content through their recommendation systems, including sensational, polarizing, and graphic material, which negatively affects youth mental health. The study also notes that young users frequently feel a lack of control over their time spent online, with only 41% of surveyed participants reporting confidence in managing their social media use. In another investigation, Arora et al. (2024) call attention to the adverse psychological impacts of algorithm-driven social media on teenagers, such as the pressures of

curated personas and the constant bombardment of notifications, which contribute to anxiety and feelings of inadequacy. Collectively, these studies underscore the urgent need for platforms to integrate safeguards that prioritize mental health, such as algorithmic transparency, limiting harmful content, and promoting digital wellness through early education and protective tools.

Adherence to universal moral principles would mandate that social media companies do precisely this, namely integrate mechanisms within their algorithms to not only identify vulnerable demographics and protect them from harmful or inappropriate content but also actively detect and remove harmful content altogether. This approach prioritizes values such as human dignity, the rights of children, and the common good. The UN Convention on the Rights of the Child () already obligates states to protect minors from mental violence, neglect, and exploitation, but it was drafted before the rise of digital platforms. Given that today's most pervasive risks to children's well-being often emerge in online environments, it is imperative to extend this protection as a universal norm holding digital platforms accountable when their algorithms amplify harmful content that leads to addiction, psychological distress, or exploitation.

Shockingly, many countries provide social media platforms with legal protections that exempt them from liability for user-generated content based on laws designed to classify them as intermediaries rather than publishers. While this framework was considered appropriate during the Internet's early stages, it is now clearly unethical: although social media platforms are not legally responsible for the harmful content users post, their algorithms exploit this content to keep users engaged—effectively addicted—to maximize revenue, without any consideration for users' well-being. An approach grounded in universal moral principles would necessitate legal reform to hold platforms accountable for the content they host and its outcomes. By leveraging advanced AI tools to ensure that content does not harm anyone, social media platforms could align with ethical imperatives to protect vulnerable populations and uphold fundamental human values.

Shifting the focus from profit to ethical standards prevents the exploitation of users' vulnerabilities, respects their autonomy, and fosters healthier interactions, thereby reducing corporate power imbalances and creating a more equitable digital environment. This case study demonstrates how universal moral principles provide a clear and effective framework for regulating social media algorithms, surpassing the limitations of consensus-based approaches. By integrating stable moral principles into algorithm design, social media platforms can transform their technologies from potential sources of harm into instruments that support users' well-being. Aligning with the ethical imperatives emphasized by Joseph Ratzinger, this strategy ensures

that technological advancement serves humanity positively, even if it requires sacrificing financial gain for the sake of moral responsibility.

4.0.2 Conclusion: A Call for Moral Responsibility and Ethical Coherence in AI

This paper argued that universal moral principles are essential for ensuring that artificial intelligence systems are ethically grounded, uphold human dignity, and prioritize truth and the well-being of users over financial profit. Relying solely on consensus-based or principles-based ethics introduces ethical instability, fosters exploitation, and fails to address the harmful effects of AI systems, such as social media algorithms that perpetuate addiction and promote harmful content. By integrating universal moral principles into AI ethics, we establish a foundation that transcends cultural and political fluctuations, ensuring AI serves humanity responsibly and consistently.

Both moral philosophy and theology offer indispensable contributions to the ethical discourse on AI and must be actively engaged in shaping its development. Together, they provide the tools to articulate universal principles—such as justice, fairness, truthfullness, and the protection of human dignity—while emphasizing the importance of grounding technological progress in higher moral obligations that prioritize the common good. It is also time to build upon the milestones already achieved through global collaboration among diverse cultures and religions, such as the *Rome Call for AI Ethics.* Originally signed in February 2020 by major tech companies like Microsoft and IBM, along with representatives from the FAO and the Italian government (Nelson 2022), the *Rome Call* was further strengthened by the joint signature of the three Abrahamic religions in January 2023, when Christian, Jewish, and Muslim leaders launched an appeal for the ethical development of artificial intelligence (). In 2024, this platform expanded significantly as representatives from eleven world religions, including Buddhism, Hinduism, Zoroastrianism, and Bahá'í, joined the call in Hiroshima, Japan, alongside government officials and leaders from major tech companies ().

The *Rome Call for AI Ethics* promotes "algorethics"—ethics by design—and underscores how universal moral principles can unite diverse perspectives to guide AI development. As Pope Francis noted during the Hiroshima event, recognizing the contributions of cultural and religious traditions is crucial for wise AI regulation. However, it is time to move beyond mere "algorethics" and apply moral reasoning not only to the functioning of algorithms but also to their design and regulation. Every stage of AI development must prioritize meaning and purpose: Why is this technology being created? What is its purpose? How does it foster authentic human growth and freedom while protecting human dignity? By addressing these

foundational questions, we can ensure that AI systems are not only technically efficient but also aligned with universal values that promote the common good.

Moreover, universal moral principles mandate that social media companies and other AI developers leverage their technological capabilities to proactively protect vulnerable populations and eliminate harmful content. Current legal frameworks that exempt platforms from liability for harmful content they amplify are no longer ethical or sustainable. As advanced AI systems are fully capable of detecting and moderating harmful material, moral responsibility requires holding platforms accountable for the outcomes of their algorithms. This shift is crucial for ensuring that AI systems do not exploit users' vulnerabilities but instead foster autonomy, respect human dignity, and promote authentic human growth.

An ethical framework for AI grounded in universal moral principles is not merely a safeguard against harm but a guiding force that ensures technology remains a servant of humanity rather than a master. By upholding universal principles and acknowledging the transcendent dignity of each person, we can steer AI development toward a future where technology enhances, rather than diminishes, the human experience, prioritizing the common good and protecting the most vulnerable members of society.

Funding Statement This work was supported by the Slovenian Research and Innovation Agency project No.: J6-60105.

References

Arora, Payal. 2024. *From Pessimism to Promise: Lessons from the Global South on Designing Inclusive Tech.* Cambridge, MA; London, England: The MIT Press.

Benedict XVI. 2009. *Caritas in Veritate. Encyclical Letter.* https://www.vatican.va/content/benedict-xvi/en/encyclicals/documents/hf_ben-xvi_enc_20090629_caritas-in-veritate.html accessed February 11, 2025.

Buyl, Maarten, and Tijl De Bie. 2024. Inherent Limitations of AI Fairness. *Commun. ACM* 67 (2): 48–55. https://doi.org/10.1145/3624700.

Coeckelbergh, Mark. 2020. *AI Ethics.* The MIT press essential knowledge series. Cambridge, MA: The MIT press.

Corrêa, Nicholas Kluge, Camila Galvão, James William Santos, Carolina Del Pino, Edson Pontes Pinto, Camila Barbosa, Diogo Massmann, et al. 2023. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns* 4 (10): 100857. https://doi.org/10.1016/j.patter.2023.100857.

Costello, Nancy, Rebecca Sutton, Madeline Jones, Mackenzie Almassian, Amanda Raffoul, Oluwadunni Ojumu, Meg Salvia, Monique Santoso, Jill R. Kavanaugh, and S. Bryn Austin. 2023. Algorithms, Addiction, and Adolescent Mental Health: An Interdisciplinary Study to Inform State-Level Policy Action to Protect Youth from the Dangers of Social Media. *American Journal of Law & Medicine* 49 (2-3): 135–172. https://doi.org/10.1017/amj.2023.25.

Floridi, Luciano, and Josh Cowls. 2019. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review,* https://doi.org/10.1162/99608f92.8cd550d1.

Formosa, Jan Lawrence. 2024. Ethics in Artificial Intelligence - A Systematic Review of the Literature. BSc thesis, University of Malta. https://www.um.edu.mt/library/oar/bitstream/123456789/127995/1/2408ICTICT390905076109_1.PDF accessed February 11, 2025.

Forum, World Economic. 2024. *AI value alignment: Aligning AI with human values.* https://www.weforum.org/stories/2024/10/ai-value-alignment-how-we-can-align-artificial-intelligence-with-human-values/ accessed February 11, 2025.

Gabriel, Iason. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30 (3): 411–437. https://doi.org/10.1007/s11023-020-09539-2.

Haidt, Jon, and Rausch Zach. 2025. *TikTok Is Harming Children at an Industrial Scale.* https://www.afterbabel.com/p/industrial-scale-harm-tiktok accessed February 11, 2025.

Hobbs, Tawnell D., Rob Barry, and Yoree Koh. 2021. 'The Corpse Bride Diet': How TikTok Inundates Teens With Eating-Disorder Videos. *Wall Street Journal,* https://www.wsj.com/articles/how-tiktok-inundates-teens-with-eating-disorder-videos-11639754848 accessed February 11, 2025.

Khalaf, Abderrahman M., Abdullah A. Alubied, Ahmed M. Khalaf, Abdallah A. Rifaey, Abderrahman M. Khalaf, Abdullah Alubied, Ahmed M. Khalaf, and Abdallah Rifaey. 2023. The Impact of Social Media on the Mental Health of Adolescents and Young Adults: A Systematic Review. *Cureus* 15 (8). https://doi.org/10.7759/cureus.42990.

Latkovic, Mark S. 2015. Thinking about Technology from a Catholic Moral Perspective: A Critical Consideration of Ten Models. *The National Catholic Bioethics Quarterly* 15 (4): 687–699. https://doi.org/10.5840/ncbq201515470.

Narayanan, Arvind. 2023. *Understanding Social Media Recommendation Algorithms.* https://courses.cs.washington.edu/courses/cse481p/23sp/readings/W9S2/understanding-sm-recommendation-algos.pdf accessed February 11, 2025.

Nelson, Joseph. 2022. *The Rome Call to Artificial Intelligence Ethics: Inside the mind of the Machine: How the Church can respond to the ethical challenges presented by AI.* https://research.leedstrinity.ac.uk/en/publications/the-rome-call-to-artificial-intelligence-ethics-inside-the-mind-o/fingerprints/ accessed February 11, 2025.

Ortutay, Barbara, and Haleluya Hadero. 2024. *Meta, TikTok and other social media CEOs testify in heated Senate hearing on child exploitation.* https://apnews.com/article/meta-tiktok-snap-discord-zuckerberg-testify-senate-00754a6bea92aaad62585ed55f219932 accessed February 11, 2025.

Paskewich, J. Christopher. 2008. Liberalism Ex Nihilo: Joseph Ratzinger on Modern Secular Politics. *Politics* 28 (3): 169–176. https://doi.org/10.1111/j.1467-9256.2008.00326.x.

Paul, Kari. 2024. What's ahead for KOSA, an online safety act for minors, as it reaches US House? *The Guardian,* https://www.theguardian.com/us-news/article/2024/aug/03/kids-online-safety-act-senate accessed February 11, 2025.

Ratzinger, Joseph. 2021. *On Love: Selected Writings.* Translated by M.J. Miller. San Francisco, CA: Ignatius Press.

Ratzinger, Joseph, and Jürgen Habermas. 2006. *Dialectics of Secularization: On Reason and Religion.* San Francisco, CA: Ignatius Press.

Saviano, Jeffrey, Jonathan Hack, Vincent Okonkwo, and Shuying (Christina) Huo. 2024. *Reimagining AI Ethics, Moving Beyond Principles to Organizational Values.* https://www.ethics.harvard.edu/blog/post-

5-reimagining-ai-ethics-moving-beyond-principles-organizational-values accessed February 11, 2025.

Spaemann, Robert. 2012. *Love and the Dignity of Human Life: On Nature and Natural Law.* A John Paul II Institute book. Grand Rapids, Michigan: Eerdmans.

Taddeo, Mariarosaria, and Luciano Floridi. 2018. How AI can be a force for good. *Science* 361 (6404): 751–752. https://doi.org/10.1126/science.aat5991.

Tam, Nguyen The Duc, and Nguyen Thai Ngan. 2023. Incorporating Cultural Values Into Responsible Artificial Intelligence (AI) Principles From an Asian Perspective, 243–254. Advances in Social Science, Education and Humanities Research, 791. Atlantis Press. https://doi.org/10.2991/978-2-38476-154-8_13.

Topol, Eric J. 2019. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again.* 1st ed. New York, NY: Basic Books.

UNESCO. 2021. *Recommendation on the Ethics of Artificial Intelligence.* https://www.unesco.org/en/artificial-intelligence/recommendation-ethics accessed February 11, 2025.

Vinuesa, Ricardo, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications* 11 (1): 233. https://doi.org/10.1038/s41467-019-14108-y.