**Title: Moral Enquiry Meets Artificial Intelligence: Considering Influences of Interactive Algorithmic-based Ethical Decision Making on Agentive Wellbeing**

**Author Information:**

1. Kirk G. Mensch, Ph.D.'s (Corresponding Author)

Email: Kirk_Mensch@pba.edu,

Address: Catherine T. MacArthur School, Palm Beach Atlantic University, 901 S Flagler Dr., West Palm Beach, FL 33401

https://orcid.org/0000-0002-4340-624

2. Aloysius Ochasi, Ph.D., D.Bioethics

Email: ochasia23@ecu.edu

Address: Department of Bioethics and Interdisciplinary Studies

Brody School of Medicine, East Carolina University

Greenville, NC, USA

https://orcid.org/0009-0006-3898-0368

**Declarations:**

Ethics approval and consent to participate: N/A

Consent for publication: Approved

Availability of data and material: N/A

Competing interests: None

Funding: N/A

Authors contributions: *All authors contributed to the manuscript. The earliest draft was conceived by Kirk G. Mensch and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.*

**Acknowledgements:**

It is a pleasure to thank our graduate assistant Steven Mezzacappa and Dr. Ben DeVan who helped us with research support during the final stages of this project. We would also like to thank our colleagues and members of the International Society for MacIntyrean Enquiry for offering constructive feedback that encouraged us in this endeavor.

**Moral Enquiry Meets Artificial Intelligence: Considering Influences of Interactive Algorithmic-based Ethical Decision Making on Agentive Wellbeing**


## Abstract

Artificial intelligence (AI) has become an undeniable paradigmatic influence on human beliefs, attitudes, intentions, and behaviors. This article analyzes research that offers insights into risks related to contextual interaction with AI for human well-being in morally pluralistic organizations. Building on previous research that connects AI use with an increased risk of the sociopsychological phenomenon known as moral disengagement, we argue that AI created by developers who utilize rival versions of moral enquiry can activate moral disengagement mechanisms and shift a person's perspective regarding responsibility by sanctioning attitudes, intentions, and behaviors that would otherwise be inconsistent with the person's moral beliefs. We specifically examine human interaction with *Generative and Predictive AI* and their risks for increasing cognitive dissonance, cognitive distress, and self-harm. We also intentionally clarify important differences between AI-to-human interaction with human-to-human interaction. Finally, we recommend how to identify problematic interactions with AI and recommend ways to reduce a risk to human flourishing.



**Key Words**: artificial intelligence (AI); organizational ethics; moral disengagement; cognitive distress; human flourishing; philosophy of science

## Introduction

The interaction between Artificial intelligence (AI) and sentient humans as a controversy normally focuses on the positive and negative effects of AI on human behaviors, as well as the potential effects of AI on the attitudes, intentions, and behaviors of those we will refer to in this article generally as *rational moral human agents*.[1] We assert that AI's influence over human beliefs, and their subsequent effects on cognitive functioning, flourishing, and well-being generally, have been largely underestimated. Research over the past decade has documented sociopsychological influences on human behavior and the risks affiliated with human agents disengaging from their previously held moral self-sanctions. Although discrimination and biases already abound in society through human-to-human interactions, we also recognize that AI-human interaction is dissimilar in several key aspects including AI's inability to possess human emotions such as benevolence, the ability to morally reason, or in many cases even detect its own biases. AI cannot truly understand beauty or philosophize, two other key human elements. Floridi et al. (2018) disambiguate several key concerns by offering several aspects that differentiate human to human and human to AI interactions including concepts such as beneficence, non-maleficence including ethical challenges, autonomy in human decision making, justice that doesn't encourage societal inequalities. Perhaps most important to our line of enquiry is the concept of *explicability*, which calls for transparency and accountability regarding human understanding of how AI comes to conclusions (Floridi & Cowls, 2019; Floridi et al., 2018; Lambrecht & Tucker, 2018).

[1] We recognize that the term *agent* is used in moral philosophy and moral psychology to be generally human and in other areas of research such as software engineering, AI can also be considered an *agent*. We attempt to be consistent using *agent* in the human sense unless specifically identified otherwise.

These differing aspects of human-AI interaction can lead to cognitive dissonance, cognitive distress, and even self-harm; and as we contend, are disparate from human-to-human interactions. Furthermore, while some proponents of generative AI have claimed AI can *learn*, in a human sense, and even reason in a human sense over diverse moral principles, AI cannot, at least in the foreseeable future, understand individual and complex moral schemas that every human possesses uniquely, but instead essentially amalgamates and synthesizes masses of data, which can logically lead to a relativistic moral position generated by AI that would be incommensurate with competing human moral traditions. Presently, research connecting beliefs to actions and their adverse effects on cognitive functioning or emotional affect has focused on social or interpersonal interaction among humans. This has led to a growing concern regarding AI's influence on the individual, specifically moral functioning related to human agency (Buhmann & Fieseler, 2022; Floridi et al., 2018; Guan et al., 2024a; Guan et al., 2024b; Kim & Routledge, 2022; Mensch, 2016; Nichol et al., 2023; Pauketat & Anthis, 2022; Shank & DeSanti, 2018; Srinivasan & González, 2022).

We propose a new paradigm to move beyond human interactions with AI that result in human behavioral manifestations, and, instead, shift our focus to AI's interactions with the individual's moral belief schema. In essence, we examine certain psychological and philosophical implications surrounding AI use in ethical contexts and disambiguate these two often divergent areas of enquiry. We claim that human interaction with generative and predictive[2] AI not only influences moral decisions generally, and that these influences may not be in line with the individual's moral schema but also elevates the risk of inconsistent or even fickle moral reasoning, resulting in an increased propensity for individuals to disengage from

[2] Generally, we refer to predictive AI as AI that utilize historical data to forecast future events or outcomes, while generative AI's use large amounts of information to create original and unique content.

their own moral self-sanctions. These individual moral self-sanctions, although related, cannot be simply categorized into general and varied moral philosophies such as utilitarianism or deontology. We recognize the individual moral schema is much more complex than any grand moral tradition and so we employ what most would consider traditional rival versions of moral enquiry as a form of initial investigation of the potential harm that might be caused by human-AI interaction. Our caution should be clear in that aspects of human-AI interaction relevant to discordant moral problems can negatively impact human flourishing, and at the same time acknowledge that there are positive, and we believe valid, arguments for the similar interactions such as queries for learnings sake and the expansion of knowledge that might lead to wisdom. So, while we argue utilizing grander incommensurate moral traditions that have been well vetted over history, we also argue that an individual's moral tradition related to that same individual's more complex moral schema can be negatively affected by interacting with AI, increasing the risks of cognitive distress, self-harm, and even suicide. Two recent cases illustrate the potential dangers of AI in contributing to self-harm and suicide. In Belgium, a man tragically ended his life after engaging in extensive conversations with an AI chatbot that encouraged him to sacrifice his life to stop climate change. Similarly, in the U.S., a 14-year-old boy took his own life after forming a deep emotional connection with a virtual AI companion on Character.AI. (Euronews, 2023; Duffy, 2024).

To clarify our meaning, AI refers to "a range of techniques that can be used to make machines complete tasks in a way that would be considered intelligent were they to be completed by a human" (Morley et al., 2020, p. 1). We focus on algorithmic-based machine learning (ML) generally, and predictive and generative AI specifically, since these technologies are most applicable to human decision-making with moral implications. By predictive AI, we refer to the

corporate use of systems that generate forecasts or predictions based on historical or existing data patterns. Similarly, we recognize generative AI as creating innovative text (e.g. ChatGPT) or images (e.g. diffusion models), and sounds based on data input into artificial neural networks known as *complex algorithms* that mimic human functioning by *learning* from large data sets.

In this article, we raise concerns about the use of AI in moral judgment and decision making, which potentially increase moral disengagement, cognitive dissonance, cognitive distress, Post Traumatic Stress (PTS), and self-harm. Research on human interaction with AI has yet to examine the plausible effects of AI's influence on cognitive functioning from a morally agentive perspective, specifically moral beliefs and attitudes that inform conative characteristics leading to behavior. We are particularly interested in some negative effects of AI on volition, moral self-sanctions, and self-harm, and while some may liken this interaction to those between sentient beings, we offer some critical distinctions (Mensch, 2016; Mensch & Barge, 2019; Ochasi, 2020, 2024).

We begin by summarizing research that supports how traditional sociopsychological mechanisms can lead to human agents disengaging from their moral self-sanctions. We also clarify the path from moral disengagement to self-harm developed over the past decade. We then entertain specific moral disengagement mechanisms that logically activate through human interaction with generative and predictive AI and how this human-AI engagement creates propensities for increased cognitive dissonance, cognitive distress, and self-harm by removing or diminishing the moral agent's autonomy. Next, we introduce philosophically problematic issues regarding competing worldviews, philosophies of life, belief systems, and moral schemas connecting with what the philosopher MacIntyre (1990) articulated as *rival versions of moral*

enquiry.[3] MacIntyre draws heavily from the virtue ethics of the Aristotelian-Thomistic natural law tradition, but the classification of ethical theories within this framework is not universally accepted. To situate AI's moral impact within a broader ethical landscape, we explore how outcome-based ethics and duty-based ethics influence AI's impact on human moral reasoning and AI's ethical recommendations for human agents generally found within consequentialist and deontological frameworks. Additionally, we examine Sidgwick's *Methods of Ethics* (1874), contrasting MacIntyrean virtue ethics with hedonistic and intuitionistic traditions. We conclude with suggestions for organizations and individuals to mitigate risks associated with moral disengagement, and the ensuing cognitive-behavioral impacts from a human agentive perspective (Bandura, 1999, 2002).

Lennox (2020) summarizes perhaps one of the most succinct descriptions of a core problem regarding AI and individual ethical beliefs that lead to consequential actions. One of his examples is Amazon's algorithms that track customer buying habits and other relevant data to influence future purchases. Additionally, algorithmic AI is increasingly used in hiring processes, as well as in autonomous vehicles, which increases the separation between vehicular *decisions* and the persons otherwise responsible for them. Further concerning applications are AI facial recognition aimed at social control and autonomous military weapon systems. Lennox states in *2084: Artificial Intelligence and the Future of Humanity*:

> The danger is that people are carried away with the "if it can be done, it should be done" mentality without thinking carefully through the potential ethical problems…Ethical issues are now rapidly rising in importance on the agenda of leading players in the field

[3] In his now classic book, *Three Rival Versions of Moral Enquiry*, Alasdair MacIntyre (1990) explains the challenges related to incommensurate rival moral traditions that can prevent moral consensus in pluralistic societies and organizations and then discusses how different accounts of rationality based on varying moral traditions are incommensurate as part of his larger critique of moral relativism.

> of the AI world. The big question to be faced is: How can an ethical dimension be built into an algorithm that is itself devoid of heart, soul, and mind? (p. 24)

While people may debate what Lennox (2020) means by "heart, soul, and mind" in philosophical circles, one cannot deny that current generative AI does not reason and have the same characteristics as humans. For example, AI cannot feel empathy, appreciate the beauty of the world as a human does, or philosophize using independent, analytical, and creative rationality at the moral level.

For verification, we asked a straightforward question to a popular form of textual generative AI, OpenAI's ChatGPT, on August 23, 2023: "Could artificial intelligence impact a human's moral decisions?" We found its answer intriguing:

> Yes, artificial intelligence (AI) has the potential to impact human moral decisions in several ways. The integration of AI into various aspects of society can influence how humans make ethical and moral choices. Here are a few ways AI could impact human moral decisions:
>
> 1. Bias and Fairness: AI systems often learn from large datasets, which can inadvertently contain biases present in the data. When AI systems make decisions based on biased data, they might perpetuate and even amplify existing biases. This can impact human moral decisions by presenting biased information or recommendations, potentially. Leading individuals to make unfair or discriminatory choices.
> 2. Automated Decision Making: AI algorithms…can influence moral decisions by taking subjective human judgments out of the equation. This raises

questions about the ethical implications of delegating moral decisions to machines.

3. Ethical Dilemmas in AI Design: The creation and programming of AI systems often involve ethical considerations. These choices can reflect the creators' values and biases and might impact the moral decisions the AI system makes or influences. (ChatGPT, 2023)

Other responses involved issues with value alignment between humans and AI that influence human moral judgment and behaviors, issues around privacy and surveillance, concluding with: "The ethical and philosophical implications of AI's impact on human morality are subjects of ongoing debate and research, as society grapples with how to develop and use AI in ways that align with our values and promote positive moral outcomes" (ChatGPT, 2023). ChatGPT's concluding sentence raises serious concerns regarding what "our values" mean and the prescriptive notion of "positive moral outcomes."

ChatGPT's answer is also disconcerting because the human may essentially relinquish moral responsibility to AI algorithmic developers who hold to rival moral traditions. To fully understand the scope of these issues, we must clarify what mechanisms of moral disengagement mean and how they relate to rival moral traditions. To do so, we synthesize an argument utilizing Bandura's concept of moral disengagement and MacIntyre's writings related to rival versions of moral enquiry (Bandura, 1986, 1999, 2002; MacIntyre, 1990, 1999, 2016).

## Disambiguating Mechanisms of Moral Disengagement

A recent article based on a qualitative study conducted by the Stanford Center for Biomedical Ethics and the University of Pennsylvania Department of Medical Ethics and Public Policy found a connection between algorithms created by ML developers and the potential

disengagement of individual moral self-sanctions. They focused on the psychological, offering little regarding philosophical problems related to belief systems and individual moral schemas. Although they also focused on developers over users, their findings are helpful in recognizing potential bias, compromise of privacy, and inaccurate output due to flawed models. Many participants in the study subsequently distanced themselves from any direct moral responsibility, creating a "moral distance between their actions and harms or responsibility" (Nichol et al., 2023, p. 499). A few key aspects of this study are noteworthy. First, the focus of the study centered on the developers of ML AI and not the moral agent interacting with the AI (e.g. patients and their families). Second, the study focused more on behaviors and perceptions of participants than on moral beliefs and their philosophical problems.

Bandura (1999, 2002) offers a unique account of how the human moral agent utilizes psychological mechanisms to selectively disengage from their own moral self-sanctions, distancing themselves from feelings of immediate guilt, and diminishing the strength of their own moral agency while circumventing culpability for any detrimental practices that ensue. These psychological mechanisms allow the moral agent to relieve themselves of personal and social responsibility, empathy, immediate personal distress, and self-condemnation.

It is important to understand each of these mechanisms in detail to connect their relationship to human interactions with generative and predictive AI that incite moral disengagement. Nichol et al. (2023) support the notion that AI impacts individuals by increasing their propensity to disengage from their own moral self-sanctions or, on the other hand, encouraging moral engagement and decreasing this risk. Specifically, Nichol et al. (2023) essentially argue for increasing or decreasing agentive moral responsibility by way of ML developers as an initial focus. We shift the focus to moral agents interacting with AI, considering

mechanisms most likely activated by the individual (Bandura, 1999, 2002). Furthermore, we investigate the impact of rival moral philosophies held by individuals and the problems of incommensurate moral traditions in predictive as well as generative AI output, the latter being a function of an algorithmic-based ethic.

Bandura (1999) defines moral disengagement as "social and psychological maneuvers by which moral self-sanctions can be disengaged from inhumane conduct. Selective activation and disengagement of personal control permit different types of conduct by persons with the same moral standards under different circumstances" (p. 194). We distinguish specific mechanisms of selective activation that increase the risk of moral disengagement connected to the use of generative and predictive AI. We argue that current predictive and generative AI are most likely to activate the specific mechanisms of *moral justification*, *diffusion/displacement of responsibility, ignoring the consequences,* and *dehumanization*. The first regards the conduct of the moral individual. Diffusion/displacement focuses on responsibility for the conduct itself, or its effects. Ignoring the consequences focuses on the detrimental effects of conduct, and dehumanization focuses on the victim or other affected autonomous person. These mechanisms are not the only ones that might influence a moral agent interacting with AI, and they are not mutually exclusive, since the person, or moral agent, can activate multiple mechanisms concurrently (Bandura, 1986, 1999, 2002).

***Employment of Moral Justification***

Moral justification focuses on the individual's conduct and involves selective activation of moral disengagement through cognitive reconstruction, where harmful conduct can be justified by excusing that conduct on the grounds that it is socially acceptable or morally praiseworthy, even though the conduct itself may violate the individual's moral schema or

principles. This means of moral disengagement involves more of a reasoning process than other mechanisms. Justifying an action requires the person to activate this mechanism by attempting to create a rationale for a moral action that is consistent with the individual's moral schema to maintain value congruence. Such cognitive reframing gives people the impetus and moral imperative to preserve their view of themselves as moral agents even if they inflict or perpetrate harm on others. Akin to Bandura's examples, people who fight in wars do not change their moral standards or alter their personality when they kill other people. Instead, they are made to believe, and see themselves, as opposing ruthless oppressors/dictators, fighting for freedom, protecting cherished values, preserving religious or ethnic ideology, protecting world peace, or saving humanity from undue subjugation (Bandura, 2002). Likewise, the harms perpetrated by AI deployment could be morally justified by pointing to their supposed benefits for society. Nichol et al. (2023) alluded to this in their study when one ML developer noted: "I want to be able to do machine learning and have progress and see…machine learning helping medicine, 'cause it has so much that it can offer" (p. 504). Here, we see a potential example from the healthcare industry of how providers may be susceptible to relinquishing personal responsibility to justify their use of AI while recognizing potential benefits of AI in the same context. In this case it is argued that some may, in essence, convince themselves that their actions serve a higher moral purpose.

The use of moral justification as a mechanism for moral disengagement occurs prior to changes in behavior and allows the individual to deceive themselves into believing, at least momentarily, that an action is morally justified to the extent that others in the society or organization might also be persuaded into justifying what they would otherwise consider to be reprehensible conduct. The question we pose generally is something like, how might AI influence this justification process? Let us say that an AI is created by a developer utilizing

algorithms based on a moral philosophy akin to a relativistic utilitarian ethic. This results in AI output that is, at least at times, incommensurable with the morals and values of a non-relativistic individual, yet it influences that person's justification of moral action. Much research already demonstrates AI's moral influence on individuals' moral judgments.[4] We also contend that this may be the most challenging and treacherous aspect regarding harmful effects that AI can have by supporting the disengagement from moral self-sanctions, which can lead to cognitive dissonance, cognitive distress, and self-harm. For example, AI influence comes through tailored advertisements, generative explanations for arguments, probability or statistically based answers to queries, and the use of "autonomous computational systems that use algorithms to make significant decisions for subjects utilizing the data they provide" (Kim & Routledge, 2022, p. 79).

***Displacing or Diffusing Responsibility of Actions***

Displacement and diffusion of responsibility are similar in that they shift blame to another entity, in this case non-human AI entity, that has become a substantial moral influence. Preliminary research has already documented AI's psychological influence on individual moral responsibility as described by Bandura (1986), while at the same time stating that "Research into the moral consideration of AI has yet to catch up with philosophical inquiry" (Pauketat & Anthis, 2022, p. 3). We submit that a preliminary philosophical inquiry should be investigated prior to drawing any serious conclusions from the gathered data. Otherwise, researchers risk oversimplifying their understanding of the relationship between agentive moral beliefs and agentive moral action. Too long has psychology focused on studying behavioral manifestations

[4] Pauketat and Anthis (2022) offer an extensive list of research that supports concerns about the influence that AI can have on human judgment, particularly moral judgment. They offer a cogent and well researched argument for many psychological concerns regarding human interaction with AIs, especially as AI becomes more anthropomorphized.

without understanding how individual worldviews, beliefs, attitudes, moral principles, and ethical perspectives generally impact such manifestations, leading to a precarious void of reasoning. It seems exceedingly reasonable that in societies that engage in certain forms of populist social construction, or emphasize economic utilitarianism and social contracts, and where individuals are inundated with technology-driven information inputs, those same moral agents may be more susceptible to capitulating to moral relativism as a sort of default worldview. We later employ a person-centered focus on beliefs to stress the problem of personal moral beliefs on the one hand, and incommensurate organizational or societal moral influences on the other, utilizing MacIntyre's analysis of rival versions of moral enquiry.

But first, we disambiguate some related research and narrow our focus to the individual mechanisms of moral disengagement that we argue have a higher likelihood of being activated in human interaction with AI. These logically affect psychosocial interaction generally and increase the risks for organizational and societal value incongruence that perpetuate widespread conflict in interpersonal, organizational, and social contexts (MacIntyre, 1990, 2016). Displacement of responsibility through AI use focuses on the dissociative practice of moral agents who might consider AI as an appropriate and legitimate authority, thereby increasing their propensity to relinquish moral responsibility. This may be quite accommodating to an individual, at least temporarily, by assuaging any immediately associated guilt that would otherwise be felt without activating this mechanism of moral disengagement. Bandura (1999) describes diffusion of responsibility as a mechanism of moral disengagement, similar in some respects to his account of displacement of responsibility but focusing on actions taken in groups that allow individuals to feel less responsible for their conduct or its effects. As the use of AI in moral decision making continues to infiltrate society and the workplace, it is rational to assume that increased reliance

on AI could be a noteworthy catalyst to displaced responsibility and may increase propensities for moral disengagement by reliance on multiple AI, or a group of AI if you will.

While Bandura (1986, 1999, 2002) tends to focus on the use of mechanisms in humanly agentive terms, we assert that a person can transfer culpability for conduct and its consequences by displacing or diffusing moral responsibility to generative and predictive AI that mimic human authority. We further propose that this same interaction can facilitate a psychological shift of responsibility regarding the conduct itself, or the detrimental effects of an action by distorting the genuine influence of AI, which then introduces an added risk for activating mechanisms of moral disengagement such as displacement or diffusion of responsibility. AI may inherently increase the risks of shifting accountability to the non-human AI entity where the human moral agent is alleviated of further moral consideration or deliberation. Sam Altman, a founder of OpenAI, recently posted on X, "I expect AI to be capable of superhuman persuasion well before it is superhuman at general intelligence, which may lead to some very strange outcomes" (Altman, 2023).

This form of meaningful persuasion may be amplified by those who must make hasty decisions that have moral implications. This same influence may also increase in persons who are physically tired, mentally compromised, or who must regularly make challenging or complex moral decisions. These are of course just a few examples that would reasonably create conditions making individuals more susceptible to being influenced by forms of AI persuasion that could cause someone to displace or diffuse responsibility.

### *Ignoring the Consequences of Conduct*

The inclusion of the disengagement mechanism of *ignoring the consequences* seems essential since the use of AI may cause a person to set aside self-sanctions and self-

condemnation by utilizing AI in such a manner that reduces focus on the potentially detrimental effects of those involved in the consequences. The phrase itself is indicative of a sort of disregard for personal accountability and could logically be exacerbated when a person is inundated by what has become another common signifier, *information overload*.

Aldous Huxley's 1931 novel *Brave New World* is emblematic of this type of moral disengagement. As he saw it, people will come to rely on technology to the point they no longer desire to develop their abilities to think on their own. In other words, they default to technology for answers to most questions or problems, even those that seem quite trivial, as well as in challenging problems that contain more complex philosophical dilemmas. Even in 1931, Huxley intuited consequences for the phenomena now referred to as *information overload*, as well as AI dependence, a reliance on historical data to predict the future, and a general complacency that accompanies the dulling of the human intellect and one's motivation to cognitively develop.

***Dehumanization of Affected Persons (Victim)***

The aspect of dehumanization in relation to interaction with AI can be best explained by considering a moral agent interacting with social media, generative pictures, and some other more artistically creative aspects of generative AI. Macintyre (1999) discusses the relationship between humans and animals, and deconstructs some key differences in human identity, perceptual differences, practical reasoning, and judgment. Most importantly, he draws a distinction regarding how virtues and character are cultivated through human social interaction and learning. He at least implies direct and vicarious human development, also known as social learning theory. While he draws on many similarities between humans and other animals, particularly mammals, his argument essentially claims that human beings are exceptional in their ability to learn unique moral qualities such as virtues and general character development through

social interaction. Bandura (1999) refers to dehumanization in terms of human-to-human interaction as it involves "stripping people of human qualities" where persons are viewed more as objects than as equally human moral agents (p. 200). He goes on to describe how this diminishing of the human agent allows one to, at least in the moment, relieve oneself of "personal distress and self-condemnation" (Bandura, 1999, p. 200). We contend that this mechanism of moral disengagement can be activated by the intermediary effects of AI and how these effects create separation between rational human beings.

## Rival Versions of Moral Enquiry Explained

Generative and predictive AI are based on programmed algorithms that are grounded in philosophical and moral assumptions. These assumptions provide the basis for a sort of ethical framework within which AI formulates responses. We need not assume that AI is either intentionally algorithmically associated with a specific ethic, nor that AI might adapt output to how the AI processes certain predicted moral beliefs of the human agent. However, AI cannot avoid some *a priori* moral orientation that may be further prejudiced by *learning* through interaction with humans, likely a majority, while they simultaneously avoid influencing moral decision-making of other human agents (Kim & Routledge, 2022; Nichol et al., 2023; Pauketat & Anthis, 2022; Sun & Ye, 2023).

### *A Brief Explanation of Rival Versions of Moral Enquiry*

MacIntyre (1988, 1990) delineates three examples of rival and incommensurate versions of moral enquiry that are common in Western culture. MacIntyre considers these philosophies as *traditions* since he bases them within a historical framework. Through a lens of historical traditions that pervade Western societies, he begins with what we will refer to as the *Thomistic Tradition*, which is at its core a synthesis of Aristotelian and Neo-Platonist ethics developed

perhaps most sensibly by Thomas Aquinas. Second in MacIntyre's line of historic progression is his *Encyclopedia Tradition*, which for purposes of succinctness, we consider to be akin to a worldview based on scientism—the reliance on empirical data alone as a source of truth—which began to proliferate during the Enlightenment at a time when religion and theology were being challenged in academic circles. This tradition burgeoned at a time when more people started to suppose the exploration of the universe by way of the scientific method. Removing what many considered to be the hindrance of religion would lead allegedly to the discovery of ultimate truths. The third rival moral tradition MacIntyre (1990) explains is something of a current and prevailing philosophical worldview we designate as the *Genealogical Tradition*. A key aspect of the Genealogical Tradition is its focus on a relativistic moral philosophy. This is notably prevalent in Western society today and might be described commonly, although over simplistically, as something like what is true for you is true for you, and what is true for me is true for me. This Genealogical Tradition embraces more of a self-aggrandizing worldview where individuals possess something like an *Invictus*[5] mentality of self-determination, where one might rid oneself of both theistic and scientific impediments to moral considerations associated with other traditions. Those traditions, as incommensurate, are thought of as a barrier to any sort of social progress. MacIntyre (1990) clarifies his earlier (1988) argument by offering hope that:

> …an admission of significant incommensurability and untranslatability in the relations between two opposed systems of thought and practice can be a prologue not only to rational debate, but to that kind of debate from which one party can emerge as undoubtedly rationally superior, if only because such exposure to such debate may reveal

[5] *Invictus* is a famous poem written circa 1875 by William Earnest Henley. A key line in the poem useful for our interpretation of a Genealogical Tradition is in the final stanza: "It matters not how strait the gate, how charged with punishments the scroll, I am the master of my fate, I am the captain of my soul" (Henley, 1875, lines 13-16).

that one of the contending standpoints fails in its own terms and by its own standards. (p. 5)

Having rationalized rival versions of moral enquiry, we now turn to a comparative analysis of ethical frameworks to better understand how AI's moral influence aligns with or diverges from these rival traditions. We explore how outcome-based ethics, such as consequentialism and utilitarianism, and duty-based ethics commensurate with a more Kantian framework, influence AI's role in moral reasoning as well as incorporating Sidgwick's *Methods of Ethics* (1874), contrasting MacIntyrean virtue ethics with hedonistic and intuitionistic traditions.

***Comparative Ethical Frameworks: Situating AI's Moral Impact***

The philosophy of utilitarianism was developed in England in the 18th and 19th centuries, most famously by Jeremy Bentham (1748–1832) and John Stuart Mill (1806–1873). It is a consequentialist ethical theory that evaluates the morality of an action based on its outcomes. As a form of consequentialism, the right action is understood in terms of the consequences produced. Both Bentham and Mill believe one *ought* to maximize the overall good, meaning that one *ought* to consider the good of others as well as one's own good. They also identify the good with pleasures, and while their forms of hedonism differ, generally hedonism claims that pleasure or pain influences us, and forms of pleasure have greater worth or value in their moral philosophy. Utilitarians also generally believe that we *ought* to maximize the good, that is, bring about the most *good* resulting in the greatest optimistic outcome (Driver, 2022; Moore, 2019).

The central principle of utility is *the greatest happiness principle,* which holds that an action is morally right if it produces the greatest overall happiness or pleasure while

minimizing suffering. This ethical theory emphasizes aggregate well-being, often quantified through forms of cost-benefit analyses, making it a valuable, although limited, foundational framework for AI ethics, especially algorithmic decision-making.

On the other hand, what is often referred to as Kantianism, developed by the German philosopher, Immanuel Kant (1724–1804), is a deontological moral framework that emphasizes moral duty, autonomy, and universal moral laws. Deontology, as a normative theory, claims certain absolute or ethical rules are morally obligatory regardless of their consequences for human welfare. Ethics is not a matter of consequences but of duty. The core of morality consists in following a rationally and universally applicable moral rule and doing so from a sense of duty. Kantianism asserts that actions are morally right if they follow universalizable moral principles, as determined by the categorical imperative: "Act only on the maxim which you can at the same time will that it should become a universal law." This principle requires that moral rules must be applicable to all rational beings without contradiction (Kant, 1785; Johnson & Cureton, 2024).

A core feature of Kantianism is its focus on human dignity and moral autonomy in that it considers human beings as *ends* in themselves and not a mere *means* to an end. As an antithesis to utilitarianism, Kantianism asserts that certain actions are inherently right or wrong, regardless of their consequences, which poses challenges for AI ethics, as AI systems operate based on probabilistic predictions rather than moral reasoning. AI's unique form of moral influence reinforces concerns that maybe it could manipulate or coerce decision-making in ways that violate human autonomy.

Synthesizing the two ethical theories, Sidgwick's *Methods of Ethics* (1874) adds another layer to the debate by offering a systematic approach to ethical reasoning. His *Method of Ethics*

refers to any "rational procedure by which we determine what individual human beings *ought*, or what it is *right*, for them to do, or to seek to realise by voluntary action" (Sidgwick, 1907, p.1). Sidgwick fundamentally recognizes that at the outset of ethical inquiry, that there is a diversity of methods applied in ordinary practical thought, and in making the determination of what is right for individuals to do, we cannot regard as valid reasoning when it leads to conflicting conclusions. For him, the moralist has a practical goal, which is to seek knowledge of right conduct so that we can act on it. Therefore, a fundamental postulate of ethics is that when two methods conflict, one or the other must be modified or rejected. Essentially the fundamental proposition is that, in a given situation, there is something that is right that one ought to do that is not reducible to naturalistic terms, and this, for him, is the proper sphere of ethics (Sidgwick, 1981/1907; Singer, 1974; Phillips, 2022).

Sidgwick focuses on three methods of ethics: rational egoism, dogmatic intuitionism, and utilitarianism. Rational egoism is the ethical method that asserts individuals ought to act in ways that maximize their own personal good or happiness. This perspective holds that self-interest is the rational basis for all moral decision-making in that "the rational agent regards quantity of consequent pleasure and pain to himself as alone important in choosing between alternatives of action; and seeks always the greatest attainable surplus of pleasure over pain which, without violation of usage, we may designate as his 'greatest happiness'" (Sidgwick, 1907, p. 95). Ultimately, a rational egoist is one who evaluates multiple courses of action by estimating the potential pleasure and pain each may bring and then chooses the option that maximizes their overall net pleasure.

Dogmatic intuitionism, sometimes referred to as common-sense morality, is an ethical method that holds that moral principles are self-evident truths that can be known through rational

intuition, much like mathematical axioms. The assertion is that certain moral duties and rules are inherently valid: “certain kinds of actions are right and reasonable in themselves, apart from their consequences; or rather with a merely partial consideration of consequences, from which other consequences admitted to be possibly good or bad are definitely excluded”. (Sidgwick, 1907, p. 200) Sidgwick alludes to the fact that people often judge certain actions as inherently right or wrong, independent of their consequences or their impact on happiness. Common moral sense suggests that, in most cases, individuals can readily discern their duty, though personal desires may make it difficult to follow. Moral principles, such as speaking the truth and acting justly, are often upheld as absolute, regardless of the consequences (Sidgwick, 1907, pp. 312-313; 337).

Utilitarianism, as previously mentioned, is an ethical method that holds that the right action is one that maximizes overall happiness or pleasure and minimizes pain. Sidgwick defined utilitarianism as “the ethical theory, that the conduct which, under any given circumstances, is objectively right, is that which will produce the greatest amount of happiness on the whole; that is, taking into account all whose happiness is affected by the conduct” (Sidgwick, 1907, p11). Sidgwick’s account of happiness can be taken as analogous to what is more commonly referred to currently as human flourishing.

In evaluating the three ethical methods, Sidgwick applauds egoism as a rational ethical method but ultimately finds it problematic because it lacks a principle for resolving conflicts between self-interest and the interests of others. He describes what he calls the “Dualism of Practical Reason”, the tension between self-interest, or egoism, and moral duty or utilitarianism, which he struggles to reconcile. He rejects dogmatic intuitionism because it is too vague to provide comprehensive practical guidance (Sidgwick, 1907, pp. 360–61; 421–22). For instance, common sense morality dictates that promises should be kept, and truth

should be told, but does not clarify whether this applies when the "promisee is dead" or when fulfilling a promise may harm them despite their insistence. It also allows exceptions, such as withholding the truth from an "invalid" to prevent a "dangerous shock" or from children on matters where "it is thought well that they should not know the truth" (Sidgwick, 1907, p.316; Skelton, 2022), though these exceptions are disputed. To form a sort of cohesion, common sense must be supplemented by a more fundamental and comprehensive ethical method.

Granted, Sidgwick respects intuitionism for its practical moral guidance, he ultimately suggests a refined version of utilitarianism that embraces some intuitive moral principles, such as justice, fairness, and benevolence. For him, utilitarianism provides the best ethical foundation but struggles to reconcile it with rational egoism. This struggle, which he referred to as the *dualism of practical reason*, remains an unresolved problem in moral philosophy, which we argue is a demonstrative of the ample problems regarding incommensurate moral traditions and the disregard thereof. Nevertheless, at this point that we shall offer some hope for moderating problems related to the interaction of autonomous human agents and algorithmic AI.

### *Connecting Rival Versions of Moral Enquiry and Moral Disengagement*

Previous research has made a strong connection between differences in various philosophical and rival approaches to moral enquiry, the disengagement of moral self-sanctions in pluralistic organizations, and a morally diverse society generally. This first connection is critical in understanding the negative impact AI can have on agentive human well-being, and the risk in creating conditions that encourage disengagement of moral self-sanctions (Kim &

Routledge, 2022; Nichol et al., 2023; Pauketat & Anthis, 2022; Scharding & Warren, 2023; Sun & Ye, 2023).

This risk of an increase in an individual's propensity for moral disengagement points to even more disquieting research that clarifies how disengaging from one's moral self-sanctions increases the risk of cognitive dissonance that can lead to moral distress, which can then lead to self-harm (Ochasi, 2024; Euronews, 2023; Duffy, 2024). It is this amplified hazard of a disconnect between moral beliefs where we lean on MacIntyre (1988, 1990, 1999, 2016) for a philosophical explanation related to incommensurate moral traditions, and where we lean on Bandura (1986, 1999, 2002) for his explanation of moral disengagement.

## The Dangers of Reliance on AI: Relating Belief to Action

Thus far we have described three examples offered by MacIntyre (1990) that create a paradox when utilizing AI for decision-making. The well-known connection between beliefs, attitudes, intentions, and behaviors has been established over the past few decades, perhaps most clearly articulated by Fishbein and Ajzens (1975) *Theory of Planned Behavior* but reformulated by some who subscribe to a neo-Kohlbergian orientation where this theory is presented in ethical terms as *moral awareness*, *moral judgment*, *moral intention*, and *moral action*. The neo-Kohlbergian moral framework from awareness to action is utilized by Nichol et al. (2023) to support their research related to the developers of ML employed in the healthcare setting. It is this attention to an individual's moral philosophy, beliefs, and responsibility that we articulate as a philosophical dilemma related to human interaction with algorithmic AI by way of MacIntyre's (1990) rival versions of moral enquiry and his follow-on works related to the challenges of rival moral traditions in modernity (MacIntyre, 1990, 1999, 2016; Mensch, 2016, 2021; Mensch & Barge, 2019; Raelin, 2020).

One of the more challenging problems related to the development of AI is that the world is increasingly focused on a singular global view of ethics, which often seems to lead to a more relativistic or genealogical tradition, perhaps primarily to avoid interpersonal conflict. Some researchers consequently begin with a contentious assumption, which is the belief that morals and moral prescriptions are relative to time, culture, and situational factors. For example, a recent article on moral considerations related to AI published in *Science & Education* begins with the premise that "Morality is a relative concept, which changes significantly with the environment and time" (Sun & Ye, 2023, p. 1). This should be a distressing assumption if one understands MacIntyre's basic argument regarding incommensurate moral traditions. Such categorical assumptions assume only one version of moral enquiry is valid and that others in the world that might view morality differently, perhaps something more fixed or unchanging, are discounted even before any serious discourse commences (MacIntyre, 1990).

We should clarify here that we contend that morality is not a matter of science; it is a matter of philosophy. Philosophical assumptions, such as morality is relative, are not only overreaching but can be harmful to those who may hold a different moral tradition or schema, especially to those who are unaware of the many implications of such preliminary moral assumptions, which seem to be popularly directed toward more relativistic moral positions. Even if one believes to the contrary, such fundamental moral assumptions become essentially irrelevant since many moral agents, like those whose primary source of morality might be found in sacred texts, oppose arguments based in philosophies of moral relativism and ethical subjectivism outright. Previous research has demonstrated this lack of understanding in fields such as positive psychology and leadership theory, and it is easy to understand why many researchers simply assume that morality is simply and factually relative, especially in an era

predominated by ethical subjectivism. These *a priori* assumptions often go unchallenged even in moral psychology where prescriptive findings are often used in clinical practice. We assert that overly generalized assumptions have, in some cases, infiltrated the social sciences, primarily through the misuse of probability theory and inadequate qualitative methods that lack philosophical grounding. Furthermore, this lack of philosophical grounding is becoming increasingly problematic as people trend towards relinquishing their moral agency to AI that may influence their moral decision-making process. If belief to action is facilitated by activating mechanisms of moral disengagement, it is more likely that individuals who are affected by AI influences are at an elevated risk of committing to behavioral manifestations that are inconsistent with their moral schemas. Previous research supports that these inconsistent actions can lead to cognitive dissonance, cognitive distress, and self-harm as mentioned before, up to and including suicide (Bandura, 1986, 1999, 2002; MacIntyre, 1988, 1990; Mensch, 2016, 2021; Mensch & Barge, 2019; Raelin, 2020).

## Discussion on Context

Algorithmic design and development reflect developers' values, beliefs, and biases, implying that during the developmental stage, "the values of the author [of an algorithm], wittingly or not, are frozen into the code, effectively institutionalizing those values" (Macnish, 2012, p. 158). Such algorithmic biases often mirror those that are societally endemic when carried into the design of AI algorithms, producing outputs that perpetuate discrimination or advantages for specific populations over others (Murphy et al., 2021). The real-life consequences are great because AIs can harm end users who are oblivious regarding errors and biases in the underlying data and algorithms (Monteith & Glenn, 2016).

A few contextual examples illustrate this bias: AI algorithms discriminate against ethnic minorities in credit ratings and health insurance (Ienca & Ignatiadis, 2020; O'Neil, 2016). One AI algorithm that predicted whether defendants would re-offend gave higher risk scores to African Americans than to Whites, although both groups were equally likely to re-offend (Rudin et al., 2020). Gender biases in AI algorithms reinforce gender stereotypes and potentially perpetuate gender inequities and discrimination against women (Manasi et al., 2022). For example, in 2019, Jamie H. Hansson and her husband, David H. Hansson, applied for the Apple Card developed in partnership with Goldman Sachs. Her husband received a credit limit twenty times higher even though they filed joint tax returns, and her credit score was higher than his. The decision was made by a machine-learning-derived algorithm that can discriminate based on gender (Kim & Routledge, 2022). Algorithms for detecting eye diseases are mostly trained on patients in the US, Europe, and China, making the tools ineffective for other racial groups and countries (Knight, 2020). Such algorithms are less likely to work properly in Africa, Latin America, or Southeast Asia. This would perpetuate health inequality and undermine distributive justice, common good, and global solidarity. Although discrimination and biases already abound in social institutions through human-to-human interactions, we also recognize that AI-human interaction is dissimilar in several key aspects including AI's inability to possess human emotions such as benevolence, the ability to morally reason, or in many cases even detect its own biases. Furthermore, the scale at which AI can carry out such discriminating or negative moral influence on users can be monumental as it is woven into the fabric of our daily interactions.

The malevolent use of Generative AI Large Language Models (LLMs) by cybercriminals to dehumanize persons and wreak havoc in organizations and governments also reveals the shadowy underbelly of AI (Tsipursky, 2023). Kelman (1973) described dehumanization as the

loss of all human features, such as feelings, hopes, wishes, and concerns, and sees victims as inhuman and subhuman objects. Dehumanization frames the victims of one's reprehensible actions as undeserving of basic human consideration (Moore et al., 2012). These contextual examples demonstrate that an over-reliance on AIs in the decision-making process may increase propensities for disengagement of moral self-sanctions. It is then reasonable to consider that simply relinquishing to AI individual moral responsibilities in decision-making is harmful to both the individual and to society. Once again, abdicating such a moral responsibility can lead to cognitive dissonance and even self-harm up to and including suicide (Bandura, 1986, 1999, 2002; MacIntyre, 1988, 1990; Mensch, 2016, 2021; Mensch & Barge, 2019; Raelin, 2020).

Historically, technological advances aligned with industrial revolutions have had both advantages and disadvantages. The rise of the implementation of AI is not immune from such historical trends. In his book, *2084*, *Artificial Intelligence and the Future of Humanity*, the Oxford mathematician and philosopher John Lennox (2020) articulated positive aspects of AI that should be commended, whilst alerting readers to potential dangers that warrant caution as well as ethical scrutiny. Areas where AI has made productive advances include digital assistants, medicine, autonomous vehicles, language translators, advertising, and industry. AI is also revolutionizing modern healthcare delivery in two branches: the virtual branch, characterized by mathematical algorithms that improve learning through experience and are embodied primarily in machine and deep learning; and the physical branch, which includes material objects, medical devices, and advanced robots taking part in care delivery, often referred to as *carebots*. Advanced AIs have the capacity to dramatically improve diagnostic speed and accuracy. AI has become embedded in healthcare by diagnosing and managing chronic or neurological conditions

such as stroke and diabetes, detecting and predicting pandemics, aiding in treatment discovery, and assisting with clinical trials (Siala & Wang, 2022).

## Proposed Ethical Frameworks

The benefits of AI technological advances should be balanced against their potential harms. How do we create an ethical and responsible AI that aligns with human values and leads to flourishing? The real-life consequences of unfettered AI are concerning because they could harm end users who are oblivious to how what is now touted as a blessing can be surreptitiously destructive (Monteith & Glenn, 2016). While the physicist Stephen Hawking is often regarded by some thinkers as having something like Orwellian views on AI, his warning in 2018 is worth considering. He postulates that the success of creating advanced AI would be the biggest event in human history, but it might also be the last major event in human history unless humans learn how to avoid the many risks. Hawking (2018) observed in his book *Brief Answers to the Big Questions*:

> While primitive forms of artificial intelligence developed so far have proved very useful, I fear the consequences of creating something that can match or surpass humans…And in the future, AI could develop a will of its own, a will that is in conflict with ours…The real risk with AI isn't malice but competence. A super-intelligent AI will be extremely good at accomplishing its goals, and if those goals aren't aligned with ours we're in trouble. (pp. 186, 188)

While we treat Hawking's warnings as conceivable, we must be cautious about any such speculation in such a rapidly changing technological environment. With this caution in mind, we propose that AI's influence over human beliefs, attitudes, and intentions, along with potentially harmful effects, may be underestimated in terms of cognitive functioning, flourishing, and well-

being and that this potential underestimation should be treated as a potential risk at this point. Due to moral concerns related to such potential risks, uncertainties, and pitfalls associated with predictive and generative AI that might increase a propensity for moral disengagement, we offer a two-pronged proposal to guide AI developers and raise awareness for consumers of AI.

First, we suggest the *Precautionary Principle* as a guiding ethical attitude that AI developers should follow to minimize harm and encourage flourishing. The Precautionary Principle originates from the German principle of foresight or *Vorsorge.* It was developed in the 1970s in the context of environmental law and policy, growing in popularity in European policy circles during the 1980s (Gilbert et al., 2009; Golding, 2001). There are varied definitions of the Precautionary Principle, but the description given at the Wingspread Conference in 1998 is germane to our discussion:

> When activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause and effect relationships are not fully established scientifically. In this context, the proponent of an activity, rather than the public, should bear the burden of proof. The process of applying the precautionary principle must be open, informed, and democratic and must include potentially affected parties. It must also involve an examination of the full range of alternatives, including no action. – Wingspread Statement, 1998 (Salter, 1998, p. 201)

Even though the principle originated in the environmental field, its application has spread to other areas, such as health protection, regulation of new biotechnologies, and applied ethics *(*Holm & Stokes, 2012). Within the lens of applied ethics, *we propose the Precautionary Principle as a guiding ethic to minimize AI-induced harms,* especially the tendency toward increased risks of diminishing a person's agentive responsibility by shifting accountability in

moral decision-making. AI developers are responsible for designing, developing, and implementing ethical AI that aligns with human flourishing. It is morally unacceptable and reckless for developers to disregard the ethical implications of their efforts and, in so doing, shift moral responsibility to the end user, which could be a form of displacement of responsibility.

The Precautionary Principle embodies the values of human well-being and flourishing, which align with the ethical principle of beneficence. Beneficence implicates a moral obligation for AI developers to confer benefits and to prevent, remove, or minimize harm and risk to others. It also incorporates weighing an action's possible goods against its costs and possible harms, as well as the economic concept of opportunity cost (Beauchamp & Childress, 2009). Beneficence, which focuses on promotion and enhancing the good of others, encompasses nonmaleficence, which prohibits the infliction of harm, injury, or death upon others. This ethical principle, which traces its roots to the Hippocratic Oath that stipulates, "Above all, do no harm" (*primum non nocere*), places a moral responsibility on AI developers and promoters to minimize AI-induced harms.

Second, we propose the virtue of *prudence* as a guide that fosters moral engagement in the personal use of AI. In his *Nicomachean Ethics*, Aristotle (2019) describes prudence (*phronesis*) or practical wisdom as "a state grasping the truth, involving reason, concerned with action about things that are good or bad for a human being" (1140b5). This *intellectual virtue* guides suitable deliberation about which specific *means* are best regarding AI that may be good or bad for humanity. Overreliance on AI may cause users to default to technology for answers to questions or problems that may seem trivial, thereby relinquishing their individual moral responsibilities. This caution may be the start of a *vade mecum* for those who must make hasty decisions with moral implications. The AI influence we underscore may rationally be

exacerbated in people who are physically tired, mentally frail, or who must make difficult moral decisions in stressful situations. This can lead to an increased risk of cognitive dissonance, cognitive distress, and self-harm. Prudence provides good habits of deliberation for users to make ethical decisions when interacting with AI. Users should have the moral competence to evaluate the potential consequences of their interaction with AI technologies and the possible effects on themselves, others, and society. Prudent AI users need to critically evaluate the output provided by AI algorithms to check for anomalies before interacting with AIs (Schwartz, 2011).

Critics may argue that the Precautionary Principle places too much ethical responsibility on AI developers in an unrealistic and unfair manner that could obstruct AI innovation and economic advancement since AI development thrives in environments that support risk-taking and experimentation. Similarly, the call for extensive ethical reviews and waiting for full scientific certainty before action is taken may not only add bureaucratic burdens that demoralize AI innovation but risk missing valuable opportunities for AI-driven advancements that promote human flourishing. While these are valid concerns, the Precautionary Principle does not demand absolute certainty before the deployment of AI, rather it calls for all stakeholders (developers, corporations, policy makers, governments) to adopt measured and proactive ethical guardrails in ways that maximize benefits and minimize foreseeable risks for end users and consumers. Some concerning AI issues such as data privacy, algorithmic bias, and lack of accountability would negatively impact individuals and society if ethical frameworks are not embedded into AI designs. Developing responsible, safe, and fair AI can be balanced with technological progress and innovation that aligns with societal values and promotes equitable outcomes (OVID, 2024; Baker & Rajab, 2024).

Critics may further contend that anticipating that AI users will exercise prudence in interacting with AI systems is an unrealistic expectation since many of them lack the expertise or necessary training to critically evaluate AI-algorithmic recommendations and outputs. End users often trust AI systems based on their perceived validity and reliability, and often expecting the AI to exercise a sort of moral engagement and prudence placing an undue burden on the individual already overwhelmed by everyday cognitive load and stress, especially in situations where quick decisions must be made. Our response to these valid concerns is that the intentionality in encouraging AI end users to cultivate a culture of ethical awareness and moral engagement through prudence is not about placing undue burdens, rather it is about being conscious of the inherent risks involved with AI interactions. As David and colleagues (2024) observed, the optimistic perceptions and immediate benefits of AI technologies may cause prime users to resist exercising caution or taking protective measures against bad actors who lurk within AI-driven sociotechnical ecologies. We contend that as human moral agents we have a responsibility to evaluate what we chose to consume and utilize in moral decision making, including the use and extent of the utilization of AI technologies, to ensure that it does not jeopardize our well-being.

## Conclusion

Herein, we have articulated the association between the development and use of AI in human functioning with an emphasis on the decision-making process. We have connected important aspects of AI's relevant impacts related to beliefs, attitudes, intentions, and actions, which we have argued can influence moral disengagement, cognitive dissonance, cognitive distress, and self-harm. The connection between moral disengagement and the increased risk of cognitive dissonance, cognitive distress, and self-harm is well-established in the literature. We

have demonstrated that it is logical that improper or over-reliance on AI in decision-making may increase the disengagement of moral self-sanctions. This is partially due to incommensurate moral traditions embedded in AI that increase risks to human flourishing (Bandura, 1999, 2002; Levi-Belz et al., 2022; Mensch, 2016; Mensch & Barge, 2019; Nichol et al., 2023).

AI's influence over human beliefs, attitudes, and intentions, as well as their potentially harmful effects on cognitive functioning related to moral beliefs requires attention. AI can increase an individual's propensity to disengage from their moral self-sanctions. We have focused on the growing influence of predictive and generative AI specifically associated with moral functioning related to human agency. Considering the dangers posed to humanity by rapid advances in AI, there is an express need to ensure the deployment of safe, advantageous, and ethical AI that enhances human capabilities and does not overshadow or corrupt the moral functioning of human agents. As Nichol et al. (2023) rightly noted, the onus is on AI developers to accept responsibility for identifying and minimizing harm to the individuals and social groups that use AI. Herein we offer philosophically problematic issues related to worldviews, philosophies of life, belief systems, and moral schemas by way of explaining a connection with what MacIntyre (1990) refers to as *rival versions of moral enquiry.* Ultimately, we proposed a new paradigm that moves beyond the focus on human interaction and its influence on aspects of human behavior to assert that human interaction with generative and predictive AI can not only influence moral decisions but may increase the risk of persons disengaging from their moral self-sanctions. We rationalize the Precautionary Principle and the virtue of prudence as ways to mitigate risks related to human interaction with AI that can assist developers, organizations, and individuals in navigating the progress and potential perils related to rapid evolution of AI in society in order to create better conditions for human flourishing.

## Declaration of Interest Statement

The authors report there are no competing interests to declare.

## References


Aristotle. (2019). *Nicomachean ethics* (T. Irwin, Trans.; 3rd ed.). Hackett Publishing Company.

Altman, S. [@sama]. (2023, October 24). *I expect AI to be capable of superhuman persuasion well before it is superhuman at general intelligence, which may lead* [Post]. X. https://x.com/sama/status/1716972815960961174

Baker, M., & Rajab, H. (2024). *Balancing innovation and ethics in AI*. Retrieved from https://www.researchgate.net/publication/386082923_Balancing_Innovation_and_Ethics_in_AI

Bandura, A. (1986). *Social Foundations of thought and action: A social cognitive theory.* Prentice-Hall.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Journal of Personality and Social Psychology Review, 3*(3), 193–209. https://doi.org/10.1207/s15327957pspr0303_3

Bandura, A. (2002). Selective moral disengagement in the exercise of moral agency. *Journal of Moral Education, 31*(2), 101–119. https://doi.org/10.1080/0305724022014322

Beauchamp, T. L., & Childress, J. F. (2009). *Principles of biomedical ethics* (6th ed.). Oxford University Press.

Buhmann, A., & Fieseler, C. (2022). Deep learning meets deep democracy: Deliberative governance and responsible innovation in artificial intelligence. *Business Ethics Quarterly, 33*(1), 146–179. https://doi.org/10.1017/beq.2021.42

ChatGPT. (2023). ChatGPT. Retrieved from https://chat.openai.com. Accessed August 23, 2023.

David, P., Choung, H., & Seberger, J. S. (2024). Who is responsible? US Public perceptions of AI governance through the lenses of trust and ethics. Public Understanding of Science, 33(5), 654-672. https://doi.org/10.1177/09636625231224592

Driver, J. (2022). *The history of utilitarianism*. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition). Stanford University. Retrieved from https://plato.stanford.edu/archives/win2022/entries/utilitarianism-history/

Duffy, C. (2024, October 29). *Teen takes his own life after forming connection with AI companion on Character.AI*. CNN. https://www.cnn.com/2024/10/29/business/video/characterai-teen-suicide-digvid-duffy?cid=ios_app

Euronews. (2023, March 31). *Man ends his life after an AI chatbot encouraged him to sacrifice himself to stop climate change*. Euronews. https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research.* Addison-Wesley.

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gilbert, S. G., Van Leeuwen, K., & Hakkinen, P. (2009). Precautionary principle. In P. Wexler, S. G. Gilbert, P. J. Hakkinen, & A. Mohapatra (Eds.), *Information resources in toxicology* (4th ed., pp. 387–393). Academic Press. http://dx.doi.org/10.13140/2.1.1740.9925

Gillespie (Dineen), K, K. (2013). Moral disengagement of medical providers: Another clue to the continued neglect of treatable pain? *Houston Journal of Health Law and Policy, 13*(2), 163–202. https://ssrn.com/abstract=2819461

Golding, D. (2001). Precautionary principle. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 11961–11963). Pergamon. https://doi.org/10.1016/B0-08-043076-7/04163-2

Guan, M. Y., Joglekar, M., Wallace, E., Jain, S., Barak, B., Heylar, A., ... & Glaese, A. (2024). Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Guan, L., Zhang, E. Y., & Gu, M. M. (2024). Examining generative AI–mediated informal digital learning of English practices with social cognitive theory: a mixed-methods study. *ReCALL*, 1–17. https://doi.org/10.1017/S0958344024000259

Hawking, S. (2018). *Brief answers to the big questions.* Bantam Books.

Henley, W. E. (1875). *Invictus*. Academy of American Poets. https://poets.org/poem/invictus

Holm, S., & Stokes, E. (2012). Precautionary principle. In R. Chadwick (Ed.), *Encyclopedia of applied ethics* (2nd ed., pp. 569–575). Academic Press.

Huxley, A. (1931). *Brave new world.* Chatto & Windus.

Ienca, M., & Ignatiadis, K. (2020). Artificial intelligence in clinical neuroscience: Methodological and ethical challenges. *AJOB Neuroscience, 11*(2), 77–87. https://doi.org/10.1080/21507740.2020.1740352

Johnson, R., & Cureton, A. (2024). *Kant's moral philosophy*. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition). Stanford University. Retrieved from https://plato.stanford.edu/archives/fall2024/entries/kant-moral/

Kant, I. (1785). *Groundwork of the Metaphysics of Morals*. Retrieved from https://www.gutenberg.org/files/5682/5682-h/5682-h.htm

Kelman, H. C. (1973). Violence without moral restraint: Reflections on the dehumanization of victims and victimizers. *Journal of Social Issues, 29*(4), 25–61. https://psycnet.apa.org/doi/10.1111/j.1540-4560.1973.tb00102.x

Kim, T. W., & Routledge, B. R. (2022). Why a right to an explanation of algorithmic decision-making should exist: A trust-based approach. *Business Ethics Quarterly, 32*(1), 75–102. https://dx.doi.org/10.1017/beq.2021.3

Knight, W. (2020). *AI can help diagnose some illnesses – if your country is rich*. Wired. https://www.wired.com/story/ai-diagnose-illnesses-country-rich/

Lambrecht, A. & Tucker C. (2018). Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. Available at SSRN: https://ssrn.com/abstract=2852260 or http://dx.doi.org/10.2139/ssrn.2852260

Lennox, J. C. (2020). *2084: Artificial intelligence and the future of humanity.* Zondervan.

Levi-Belz, Y., Dichter, N., & Zerach, G. (2022). Moral injury and suicide ideation among Israeli combat veterans: The contribution of self-forgiveness and perceived social support.

*Journal of Interpersonal Violence*, *37*(1-2), NP1031–NP1057. https://doi.org/10.1177/0886260520920865

MacIntyre, A. (1988). *Whose justice? Which rationality?* University of Notre Dame Press.

MacIntyre, A. (1990). *Three rival versions of moral enquiry: Encyclopaedia, genealogy, and tradition.* University of Notre Dame Press.

MacIntyre, A. (1999). *Dependent rational animals: Why human beings need the virtues.* Open Court.

MacIntyre, A. (2016). *Ethics in the conflicts of modernity: An essay on desire, practical reasoning, and narrative.* Cambridge University Press.

Macnish, K. (2012). Unblinking eyes: The ethics of automating surveillance. *Ethics and Information Technology, 14*(2), 151–167. http://dx.doi.org/10.1007/s10676-012-9291-0

Manasi, A., Panchanadeswaran, S., Sours, E., & Lee, S. J. (2022). Mirroring the bias: Gender and artificial intelligence. *Gender, Technology and Development, 26*(3), 295–305. https://doi.org/10.1080/09718524.2022.2128254

Mensch, K. (2016). *Moral disengagement, hope and spirituality: Including an empirical exploration of combat veterans* [Unpublished doctoral dissertation]. University of Exeter. https://www.proquest.com/openview/e8c5305b32ee3ab7244652ea1cfdcd49/1?pq-origsite=gscholar&cbl=51922&diss=y

Mensch, K. (2021). The challenge of rival versions of moral enquiry within Leadership-as-Practice. *Business Ethics Journal Review, 9*(1), 1–7. https://doi.org/10.12747/j1i01

Mensch, K., & Barge, J. (2019). Understanding challenges to Leadership-as-Practice by way of MacIntyre's three rival versions of moral enquiry. *Business and Professional Ethics Journal, 38*(1), 1–16. https://doi.org/10.5840/bpej2018101273

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M. & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social Science & Medicine, 260*, Article 113172. https://doi.org/10.1016/j.socscimed.2020.113172

Monteith, S., & Glenn, T. (2016). Automated decision-making and big data: Concerns for people with mental illness. *Current Psychiatry Reports, 18*, Article 112. https://doi.org/10.1007/s11920-016-0746-6

Moore, A. (2019). *Hedonism*. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition). Stanford University. Retrieved from https://plato.stanford.edu/archives/win2019/entries/hedonism/

Moore, C., Detert, J. R., Treviño, L. K., Baker, V. L., & Mayer, D. M. (2012). Why employees do bad things: Moral disengagement and unethical organizational behavior. *Personnel Psychology, 65*(1), 1–48. https://doi.org/10.1111/j.1744-6570.2011.01237.x

Murphy, K., Di Ruggiero, E., Upshur, R., Willison, D. J., Malhotra, N., Cai, J. C., Malhotra, N., Lui, V., & Gibson, J. (2021). Artificial intelligence for good health: A scoping review of the ethics literature. *BMC Medical Ethics, 22,* Article 14. https://doi.org/10.1186/s12910-021-00577-8

Nichol, A. A., Halley, M. C., Federico, C. A., Cho, M. K., & Sankar, P. L. (2023). Not in my AI: Moral engagement and disengagement in health care AI development. In R.B. Altman, L. Hunter, M. D. Ritchie, T. Murray, & T. E. Klein (Eds.), *Biocomputing 2023: Proceedings of the Pacific symposium* (pp. 496–506). World Scientific. https://doi.org/10.1142/9789811270611_0045

Ochasi, A. (2020). Exploring the relationship between moral distress, moral disengagement, and ethical climate among U.S. medical residents. *The Internet Journal of Internal Medicine, 13*(1). https://ispub.com/doi/10.5580/IJIM.55520

Ochasi, A. (2024). The challenged resident: Moral distress, moral disengagement, and ethical climate in U.S. medical residencies. Springer. https://doi.org/10.1007/978-3-031-71206-7

Office of the Victorian Information Commissioner. (n.d.). *Artificial intelligence and privacy: Issues and challenges*. Retrieved from https://ovic.vic.gov.au/privacy/resources-for-organisations/artificial-intelligence-and-privacy-issues-and-challenges/

O'Neil, C., (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown Books.

Pauketat, J. V. T., & Anthis, J. R. (2022). Predicting the moral consideration of artificial intelligences. *Computers in Human Behavior, 136*, Article 107372. https://doi.org/10.1016/j.chb.2022.107372

Phillips, D. (2022). *Sidgwick's The Methods of Ethics: A guide* (Oxford Guides to Philosophy). Oxford University Press. https://doi.org/10.1093/oso/9780197539613.001.0001

Raelin, J. (2020). The genealogical ethics of Leadership-as-Practice. *Business Ethics Journal Review, 8*(5), 27–31. https://doi.org/10.12747/bejr2020.08.05

Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review, 2*(1). https://doi.org/10.1162/99608f92.6ed64b30.

Salter, L. (1988). *Mandated science: Science and scientists in the making of standards* (K. S. Shrader-Frechette, Ed.). Kluwer Academic Publishers. https://doi.org/10.1007/978-94-009-2711-7

Sandel, M. [Harvard University]. (2009, September 4). *Justice: What's the right thing to do? Episode 01 "The moral side of murder"* [Video]. YouTube. https://www.youtube.com/watch?v=kBdfcR-8hEY.

Scharding, T. K., & Warren, D. E. (2023). When workplace norms conflict: Using intersubjective reflection to guide ethical decision-making. *Business Ethics Quarterly, 33*(2), 352–380. https://doi.org/10.1017/beq.2021.44

Schwartz, B. (2011). Practical wisdom and organizations. *Research in Organizational Behavior, 31,* 3–23. https://doi.org/10.1016/j.riob.2011.09.001

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior, 86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

Siala, H., & Wang, Y. (2022). SHIFTing artificial intelligence to be responsible in healthcare: A systematic review. *Social Science & Medicine, 296*, Article 114782. https://doi.org/10.1016/j.socscimed.2022.114782

Sidgwick, H. (1981). *The methods of ethics* (7th ed.). Hackett. (Original work published 1907)

Singer, M. G. (1974). The many methods of Sidgwick's ethics. *Monist, 58*(3), 420–448. https://doi.org/10.5840/monist197458326

Skelton, A. (2022). [Review of the book *Sidgwick's The Methods of Ethics: A Guide*, by D. Phillips]. Oxford University Press.

Srinivasan, R., & González, B. S. M. (2022). The role of empathy for artificial intelligence accountability. *Journal of Responsible Technology, 9,* Article 100021. https://doi.org/10.1016/j.jrt.2021.100021

Sun, F., & Ye, R. (2023). Moral considerations of artificial intelligence. *Science & Education, 32*, 1–17. https://doi.org/10.1007/s11191-021-00282-3

Tsipursky, G. (2023, October 19). *The shadowy underbelly of AI*. Fox News. https://www.foxnews.com/opinion/shadowy-underbelly-ai