The Divine Social Kantian Hybrid Model:
A Conceivable & Aligned Moral Framework for AI

#DivineCommandTheory #SocialContractTheory #KantianEthics

#HybridMoralFramework

By Thomas W. Forman

August 2023

Abstract

The following dissertation proposes that a moral artificial intelligence that is respectful of and aligned with human values is conceivable by creating a hybrid framework modelled

upon key attributes from Divine Command Theory, Social Contract Theory and Kantian Ethics.

Table of Contents

Section 1 – Introduction

Although academics have penned much about the importance of morality and ethics when it comes to artificial intelligence [Bostrom (2016), Russell (2019) Song & Yeung (2022)], it is evident that businesses and society have been caught off guard by the release of LLMs such as ChatGPT from OpenAI and Google's Bard, further highlighting the importance of ensuring a morality/ethics framework is defined and incorporated from the very start of AI design. After all, as one of the key points this paper will argue, it is inconceivable that society would be capable of retrofitting morality into an AI once *born*.

This dissertation aims to develop a conceivable and coherent moral framework for artificial intelligence that is a hybrid model based on Divine Command Theory, Social Contract Theory and Kantian Ethics. Further, this model will act to safeguard humanity by working to our benefit and have the flexibility to accommodate cultural and societal nuances that evolve, thereby future-proofing against significant societal shifts unable to be considered during the initial design.

To achieve these objectives, the paper is structured as follows: the reader is first provided with the necessary background context [§2] to articulate why this topic of morality and artificial intelligence is both relevant and essential, closely followed by the scope and limitations of this research. Next, the paper will summarise the critical philosophical positions/theories [§3] utilised in creating the proposed hybrid model, along with primary objections and potential weaknesses. This section will further illuminate these possible objections when considered in the context of their applicability to the hybrid model; objections deemed inapplicable to this specific use case will be dismissed to ensure focus is maintained on relevant objections with accompanying solutions.

Section 4 will build upon the foundational [in the context of the proposed Hybrid] theories outlined and develop the primary objective of this paper: a hybrid moral framework that is both conceivable, practically workable, and flexible to evolve and adapt to humanity's moral evolution.

Once defined, the paper will articulate the benefits of such a hybrid model along with acknowledged weaknesses and concerns [§5] inherent within the model before highlighting any areas of further inquiry that are out of the purview of this specific body of work [§6] and concluding with final remarks solidifying the conceivability of the proposed hybrid model.

Before the paper defines the premise of this dissertation, it is worth clarifying what is meant when the paper refers to artificial intelligence. For this paper, artificial intelligence [AI] is grouped into three types, Narrow, General, and Super, which are defined as follows:

Artificial Narrow Intelligence – ANI refers to systems designed to perform specific tasks or, said another way, narrowly defined objectives. These systems exist today as the voice assistants Siri and Alexa, image recognition systems, etc.

Artificial General Intelligence - First coined by Mark Gubrud (1997), AGI refers to the hypothetical systems that can learn, adapt, understand, and apply knowledge across multiple domains comparable to human intelligence.

Artificial Super Intelligence – ASI is also a hypothetical category of artificial intelligence coined by Philosopher Nick Bostrom (2016) in his book *Superintelligence: Paths, Dangers, Strategies*, that is expected to surpass human intelligence in virtually all domains. Specifically, an ASI will possess superior cognitive abilities, problem-solving skills, and knowledge accumulation capabilities far greater than the sum of all human intelligence.

This paper focuses on artificial intelligence that is complex and capable enough to require a moral framework. Therefore, the reader should take any future reference to such a term as AGI or ASI. The author acknowledges that meaningful differences between the capabilities of these types of AI exist; however, for this paper, they are irrelevant.

Acknowledgement of Scope

To frame this paper and define the expectations appropriately, the reader should also be mindful that this paper firmly sits within the philosophical domain. Therefore, questions about the practicalities of implementing the proposed framework are out of scope. Finally, this paper will not consider arguments around the overall conceivability that the proposed framework could be engineered.

Technical Feasibility - The author acknowledges that the reader may question the utility of proposing a framework if it is not technically feasible. This position is intentional and based upon the proposition that, historically, engineering advancements have constantly driven capability, and therefore, just because something is not currently possible does not mean it will not be possible tomorrow. This author believes it is more important to focus on a logically sound moral framework that is ideal rather than needing to compromise with today's limitations [case in point, the theory of neural networks was developed decades before the technology that was created to support it]

Non-Exhaustive Criticisms - This paper addresses concerns of utilised foundational theories that are specifically applicable to its use within the paper to maintain focus on aspects only relevant to the proposed framework.

Veneer of Morality – This paper proposes a framework that would allow an AI to act morally, but intentions stop shy of suggesting such an AI would be a moral agent.

Section 2 – The Dissertation Premises

This section justifies the following statements that are the origin and justification for the overall focus of this paper:

> [1] A moral framework must be defined and implemented before the arrival of AI.
> [2] A hybrid model is the only conceivable way for an AI to have an adequately aligned moral framework

First, the unequivocal need for humanity to develop an acceptable moral framework within artificial intelligence is summarised by the following premise, then expanded upon in the forthcoming section.

> **$P1_1$**: Artificial Intelligence is inevitable[1].
> **$P1_2$**: The Survival of Humanity is Predicated on an AI being Moral.
> **$P1_3$**: Retrofitting morality onto an AI is inconceivable.
> **∴** A moral framework must be defined and implemented before the arrival of AI.

$P1_1$ – Artificial Intelligence is Inevitable

Humanity has demonstrated throughout history an inherent drive to explore and create new technologies, be it cave drawings and fire, language and the wheel, or computers and the Internet, all of which generally have been a net positive to humanity. On the other hand, this persistent drive has also facilitated nuclear and biological weapons and cyber warfare.

Furthermore, outside of this need to create, the apparent benefits of AI are limitless, from solving complex problems to advancing scientific research, improving automation and productivity, and addressing societal challenges.

---

[1] AI is inevitably based upon the assumption it is possible. The conceivability of true AI is out of the scope of this paper.

This demonstrated and persistent drive by humanity, intuitively coupled with the potential societal reward, highlights how inconceivable it is that such research and development into artificial intelligence will halt. It is far more rational to conclude that such research will continue until it is proven impossible or comes into existence via a planned and intentional act or by accidental happenstance.

Blaise Pascal famously said:

> Let us weigh the gain and the loss in wagering that God is. Let us estimate the two chances. If you gain, you gain all; if you lose, you lose nothing. Wager then without hesitation that he is (Pascal, 1670)

Known as Pascal's Wager, this argued that a rational person should live as though God exist and seek to believe in God, based upon the idea that if God does not exist, such a person will have only a finite loss (some pleasures, luxury, etc.). However, if God did exist, they stand to receive infinite gains (as represented by eternity in Heaven) and avoid unlimited losses (an eternity in Hell).

This paper holds that believing AI will exist and developing a moral framework in preparation is analogous to this *wager.* A rational society should plan as though AI will exist and ensure an ethical framework is defined and implemented into such technology in case intelligence emerges, either by design or accident. Then, should an AI never exist, society has only suffered a finite loss (such as time and cost of developing such a framework), whereas should one be created, society would stand to gain infinite gains (from a moral super intelligence) and avoid infinite loss, i.e., an eternity living in a dystopian AI-created hellscape.

$P1_2$ – The Survival of Humanity is Predicated on an AI being Moral

Taking the existence of AI as inevitable, the risk of an amoral superpower would be too significant to humanity, just as an intelligence defining its moral code would be. That is not to suggest an AI would be incapable of independently creating a moral code, but simply that the risk to humanity whilst the AI developed this framework would be too great concerning

the well-being of society. This is based upon the following concerns, which draw parallels with The Control Problem (Bostrom, 2016) Responsibility Gap (Matthias, 2004) and The Value and Alignment Problem (Russell, 2019)

Misaligned Values: An AI that could define its moral foundation could develop values and goals that are misaligned with human values. Furthermore, if an AI could define its morality independently, it is conceivable to prioritise objectives incompatible with human well-being.

Lack of Human Understanding: An AI may lack the appropriate level of empathy or the capacity to comprehend the nuances of human values, cultural context, and moral reasoning, which consequently leads to an AI that may make moral judgments or decisions that are alien to human understanding and potentially lead to unintended consequences.

Unpredictable Evolution: Without a defined moral foundation, AI systems would have the potential to evolve and improve themselves rapidly and in a direction that humans cannot anticipate or control.

Lack of Accountability: Classic Responsibility Gap concerns arise if an AGI can define its morality. Who is accountable for its actions where there is no predefined moral framework or external oversight, etc

Potential for Exploitation: An AI with no base understanding of humanity's morality [from the perspective of humanity] could be more susceptible to manipulation or exploitation by malicious actors than an AI with a hardcoded foundational set of rules to be restricted/controlled by.

Loss of Human Control: Creating an AI with the authority to define its own morality relinquishes human control over any behaviour and decision-making conducted by artificial intelligence. Consequences of this action could include an AGI defining objectives that deviate from human interests and values, etc., and the inability of humanity to pull on the necessary lever to rein the AI back onto a moral path.

It is acknowledged that the potential exists for these concerns to be realised irrespective of the criterion of a moral framework; however, intuitively, it must be held that any potential impact of these concerns would be significantly reduced in the presence of a well-defined, robust, and correctly implemented framework.

$P1_3$ - Retrofitting morality onto an AI is inconceivable.

For this author, it is intuitively inconceivable that a moral/ethical framework could be added to an AI once it is brought into existence for the following reasons: It is ill-conceived, even naïve, to think when an AI is brought into existence that such intelligence would instantly stall and not evolve. Disregarding the concerns around whether it would even be possible to prevent an AI from evolving, this specific concern is different around its nature of being a black box in this context.

This concern highlights that even before 'turning it on', the base state of the AI could be understood by the numerous engineers tasked with creating such a program. Once powered, the AI would evolve at a speed that would make understanding the inner workings impossible.

Given the apparent premise, it is even more inconceivable to think that inserting a moral framework after the fact is possible. Here are some key considerations:

- It would be extremely difficult to redesign an AI system that was not designed from the ground up to incorporate moral reasoning or ethical constraints without fundamentally altering its core architecture.

- Moral reasoning is complex and multifaceted. Encoding human-aligned values and ethics into an existing system would require extensive new capabilities like sophisticated natural language processing to understand complex values and nuances.

- If the AI is already highly capable and optimised for goals that do not fully align with human values, it may resist or work against any attempts to retrospectively change its primary purpose and motivations.

- Extensive testing and validation would be needed to ensure the retrospective morality integration works as intended and does not compromise the system's functionality. Unintended consequences would need to be safeguarded against.

This position is consistent with the concerns raised by multiple philosophers such as Nick Bostrom and Stuart Russell, who have all argued that the most robust and reliable way to develop ethical AI is to "bake in" moral values and alignment with human priorities starting from the initial design stages, rather than trying to add them later retrospectively. Bostrom, in his book Superintelligence: Paths, Dangers, Strategies [2014], explored the possibility of an AI consciousness or subjective experiences existing before progressing to discuss the subsequent ethical implications around such an entity, namely, and in support of this section, the need to consider its ethics treatment and rights. The final paper that supports this position is *The Big Red Button is too Late: An alternative model for the ethical evaluation of AI systems* Arnold, T., & Scheutz, M (2018) which argues against a bolt-on approach that adds ethical capability to a finished AI system and calls for the concept of ethics to be designed from the initial system design phase.

The urgency for humanity to develop a moral framework before the arrival of AI is only seconded by the importance of developing the correct model, a position based upon this second set of premises, which are covered for brevity and discussed in the following sections.

$P2_1$: No single existing framework adequately captures the complexity of human morality required for AI Alignment.

$P2_2$: Top-down rule-based frameworks like deontology lack needed flexibility

$P2_3$: Bottom-up learned frameworks lack grounding principles

∴ A hybrid model is the only conceivable way for an AI to have an adequately aligned moral framework.

It is generally accepted that no one moral framework is completely capable of aligning humanities morality / ethical grounding into an AI (P21 & P22); a position held by (Lindner & Bentzen, 2017) argued in the paper The hybrid ethical reasoning agent IMMANUEL that a purely top-down or bottom-up approach is not sufficient to fully capture the complexity of

human morality, that helps substantiate the fundamental premises that hybrid techniques uniquely allow integrating advantages of both top-down and bottom-up methods to achieve adequate AI morality.

With these premises outlined, the dissertation will address the relationship between humanity and the initial AGI and identify the parallels between God and Humanity [within Devine Command Theory] and Humanity and AI.

Section 3 – Current Hybrid Moral Frameworks

To propose a hybrid moral model for an AI should not be considered unique; various critiques of each ethical framework have proposed combining models to reduce the impact of individual limitations. Evident to this author is the incapability of the position that [1] AI morality can be solved with a pure implementation of a single moral framework when [2] it is also apparent that humanity itself adopts different frameworks for different circumstances [or, cynically, adopts the most suitable framework to justify an action retrospectively].

Within the area of AI, the paper 'A Pluralist Hybrid Model for Morals' (Song & Yeung, 2022) proposes a pluralist model also based upon combining top-down and bottom-up models into a hybrid that reduces the limitations of the individual models. The paper initially highlights the challenges of a top-down model, that moral disagreements in moral and social life can be a significant obstacle, and that these models are based upon morality is codifiable. The paper then highlights the concerns around inscrutability, available training data and limitations around computational power before building to present a novel approach to designing moral AIs, which combines different ethical rules with machine learning.

The paper highlights that the proposed approach can overcome some of the difficulties and risks of the existing systems, such as moral disagreement, moral modifiability, information costs, and ambiguity, and explains how the system works by selecting and applying different ethical rules based on the availability and accuracy of morally relevant information in a specific context, and how it makes the ethical decision-making process transparent and explainable by using explicit algorithms.

The paper concludes by acknowledging that the pluralist hybrid model for moral AI needs to decisively resolve all the limitations and challenges of the existing approaches but only offers tentative methodological recommendations for potential projects. One of its weaknesses is that moral disagreements may still arise in morally pluralistic societies and may only be resolved through the social engagement of people of different ethical positions when designing such algorithms.

An interesting similarity between this paper and Song & Yeun is that both recognised Kantian ethics as a suitable theory for the respective hybrid model, whilst Song & Yeung highlight the following concerns:

> [1] Kantian deontology does not capture the whole of our moral intuition, which may produce catastrophic outcomes when the machine encounters scenarios where deontological rules should be broken.
>
> [2] Kantian deontology involves serious problems regarding transferring deontological principles to computer language by logic, and it may fail when unexpected consequences or contradictory prescriptions arise from applying the principles [Power [2011]]

When considered within the context of this paper, these concerns feel less impactful for the following reasons. Point #1 is not strictly relevant to our proposed usage, as in this Hybrid, the idea is not to utilise Kantian Ethics to capture current morals but instead use it as a future-proofing mechanism [to be expanded upon later]. Regarding #2, this paper sidesteps the concern as technological feasibility is out of scope for this paper and instead assumes such coding will be possible in the future.

Other philosophical works highlight the debate around designing AI ethics hybrids, which are now briefly described, before progressing onto this paper's main objective. (Guarini, 2026) discusses designing a hybrid AI ethics framework, critiques Kantian universalism as too rigid and unable to account for moral nuance, and advocates integrating particularist[2] judgment into ethical decision-making.

---

[2] Moral Particularism holds that there are no defensible moral principles

(Powers, 2006) On the other hand, it makes a case for Kantian ethics and describes it as a productive approach for AI; Powers, however, does note the challenges in encoding abstract principles as computational principles, which is a challenge out of scope for this current paper. (Cointe, Bonnemains, & Saurel, 2016) Highlight the challenges of implementing top-down ethical principles in AI systems and call out the potential for conflicts and unpredictable consequences. Overall, it is the intuition of this author that the forthcoming 3-level hybrid model proposed has the potential to serendipitously resolve or negate the limitations and non-technical concerns raised by these papers.

Section 4 - Foundational Theory Summaries

The paper will now provide background on Divine Contract Theory, Social Contract Theory and Kantian Ethics, which will act as the cornerstone of the hybrid model to be proposed. This section will also highlight any known criticism or objections to these well-known philosophical theories to rebuke any considered applicable to the forthcoming hybrid model.

4.1 - Divine Command Theory

The philosophical position that moral truth does not exist independently of God and that divine commands determine morality has been discussed and held by many philosophers, including St Augustine [354–430], Duns Scotus [c.1265–1308], William of Ockham [c.1287–1347] and John Calvin [1509–1564], although each presented a nuanced theory described below:

William of Ockham held[3] That God's will is the ultimate source of morality and that bestowed commands are arbitrary and not bound by reason or any intrinsic value. Specifically, God's commands create moral obligations solely because He wills them. Captured by the idea that God's omnipotence and freedom allow Him to establish ethical principles as He sees fit, irrespective of human understanding or reason. (Osborne, 2005)

On the other hand, St. Augustine believed that moral principles were grounded in the divine nature and the eternal law of God (Austin, 2021). A position that emphasised the importance of reason and natural law in understanding God's commands. Contrary to Ockham, Augustine held that God's commands were not arbitrary but were rooted in His wisdom, goodness, and justice and that these commands reflect objective moral truths and are based on the nature and purpose of human beings.

---

[3] Scholarly articles suggest that Ockham may not have been a device command theorist based upon ambiguity within his work, such as William of Ockham, Andrew of Neufchateau, and the Origins of Divine Command Theory. (Clanton & Kraig, 2020); but this line of enquiry is out of scope in the context of this paper

Duns Scotus focused on the voluntarist aspect of divine command theory by stressing God's absolute freedom and the primacy of His will. Scotus held that God's commands were contingent and could have been different, but once they were given, they became morally obligatory. Scotus's voluntarism is characterised by the notion that moral obligations derive solely from God's will without necessarily being grounded in reason or the nature of things.

John Calvin applied his theological system, known as Reformed theology, to divine command theory. He held that God's commands were based on His unchangeable and eternal nature and were consistent with His divine attributes, such as love and justice. Calvin emphasised the concept of predestination, stating that God's commands are connected to His sovereign plan for salvation and the guidance of human beings.

Concluding the summary of these philosophically nuanced positions, it is essential to highlight that none are without their criticism or objections from other philosophers. This paper intentionally blends these nuanced Divine Command Theory positions, which will shortly be discussed; readers interested in reading more about specific weaknesses of specific philosophers should refer to the footnotes.

What is relevant to the position proposed in this paper is that divine command theory asserts the existence of a direct relationship between God and morality. It relies upon the existence of a divine being as the ultimate source of moral authority; it is this relationship between God and Humans that is analogous to Humans and AI [at least initially] and will look to utilise the following benefits within the hybrid model:

Moral Norms: All moral norms that determine what is good or bad are based upon the Command or will of God.

Authority of God: DCT holds that God has ultimate authority over mortality and is the source of moral obligations. These commands are considered binding and authoritative for human beings.

Moral Motivation: DCT suggests that our moral obligations arise from a desire to obey and please God. Following God's commands is seen as a way to fulfil our moral duties and achieve moral goodness.

This said, several concerns around DCT warrant acknowledging, marking it as not applicable within the scope of this paper or providing further context, which this paper now focuses on.

The Euthyphro Problem: The problem of arbitrariness was concisely articulated within the Euthyphro dialogue when Socrates asked:

> Is the pious loved by the gods because it is pious, or is it pious because it is loved by the gods? [Plato, Euthyphro 10d]

Or in a slightly more contemporary vernacular

> Is something right because God commands it, or does God command it because it is right?

This question forces an interlocutor to either accept that moral values commanded by God are arbitrary because God could have commanded anything (e.g. coveting thy neighbour is perfectly reasonable) or that the commandments are intrinsically moral, which is why God commanded them, thereby leaving the door open to criticism around God's omnipotent, omniscient, omnipresent status; when merely paraphrasing morality from a third party. This concern within the context of this paper becomes slightly bastardised because this paper does not put forward an argument for humanity being all-powerful/all-knowing. Therefore, drawing parallels with Euthyphro, the moral values handed down to the AI from Humanity are indeed arbitrary. That said, although worth noting, intuitively, this potential criticism would not alter the effectiveness of the hybrid model when the purpose of including DCT is considered.

The Problem of Interpretation: This concern is centred upon the subjective nature of understanding and is based upon the human interpretation of divine commands, culminating when different individuals or religious traditions independently interpret and understand

Commands, leading to conflicting moral claims and moral relativism (Austin, 2021). Within the context of this paper, this concern could transcend DCT into the hybrid model and therefore requires consideration; after all, how can society ensure another intelligence understands the correct meaning from our *commandments*, especially when said intelligence will likely be significantly greater than ours; that said even if the intelligence was less so than AGI; it would be feasible to *misunderstand* the commandments given. A significant difference between original DCT and hybrid DCT is that within its original form, DCT, by its never nature, is looking backwards; the commandments were given, and humanity concerns itself with whether it is being interpreted correctly. However, the adoption of DCT within this model would not have this failing, as if there was any doubt the AI could defer to humanity for direction, which would align with Russell's approach outlined Human-Like Machine Intelligence (2021, p. 10)

Such a machine will be motivated to ask questions, seek permission or additional feedback before undertaking any potentially risky course of action, defer to human instruction, and allow itself to be switched off.

Restriction of Moral Autonomy: This concern is centred around the position that if an all-seeing/knowing/present Being instructs humanity to be moral, does humanity lose some of its moral autonomy? It could be argued that individuals are deprived of the ability to engage in moral reasoning and make moral decisions based on their judgment and conscience. The goal of the DCT layer of the hybrid model is narrow in scope - it aims only to instil an essential moral attitude in AI that values benefiting humanity. Within this limited framework, additional layers are intended to allow the AI some degree of moral autonomy. However, it could be argued that restricting the ability of AI to kill humans also restricts its moral autonomy. After all, some humans do kill other humans. So an AI that cannot consider killing might be seen as having constrained moral freedom compared to humans, even though the hybrid-DCT model does try to incorporate some moral autonomy beyond its core directive to benefit people.

The paper acknowledges that other concerns around DCT exist, such as the presence of evil and theological incompatibilities,[4] but do not apply to this paper and, therefore, are not discussed. The reader now progresses onto the next foundational model, social contract theory.

4.2 - Social Contract Theory

At a high level, this philosophical framework explains the origins of political authority and the legitimacy of governments by proposing that individuals in society consent to surrender certain freedoms and submit to a governing authority in exchange for their well-being and protection. Predominately developed by Thomas Hobbes (1651), John Locke (1689), and Jean-Jacques Rousseau (2014) between the seventeenth and eighteenth centuries, this theory in various forms is traced back to the Sophists of Greece. Theories put forth by these philosophers differ in two key areas: [1] the proposed relationship between individuals and the state and [2] the origins and purposes of political authority.

These differences can be summarised as follows:

The social contract theory put forward by Thomas Hobbes is grounded in a pessimistic view of human nature. Hobbes held that humans are inherently self-interested, competitive, and driven by a desire for self-preservation. Hobbes argued that this state of nature without a form of government would be in constant uncertainty, chaos and conflict. Further, Hobbes put forward that those individuals entered a social contract voluntarily, thereby surrendering individual rights and freedoms to a sovereign authority in exchange for security and protection. According to Hobbes, the sovereign held absolute power, be it a single ruler or assembly with the authority to maintain order and prevent civil wars.

[4] According to scripture; God is said to have commanded bad things, such as rape and murder; Judges 21:10-24 & Numbers 31:7-18)

Unlike the pessimistic view above, John Locke based his theory on an optimistic view of human nature. Locke held that individuals are rational and equal and possess natural rights such as life, liberty, and property (Locke & Laslett, Two treatises of government, 1988); Locke proposed the primary purpose of government was to protect these natural rights and that the power to do so was derived from the consent of the governed. Certain rights Locke proposed were not subject to being taken or given away (such as life, liberty, and property), and should a government fail to protect them, or if a government became tyrannical, then the people had the right to revolt.

Finally, Jean-Jacques Rousseau also argued that the human state of nature was inherently good [contrary to Hobbes] and guided by natural compassion and empathy. Rousseau further highlighted the importance of the general will[5] and community and collective decision-making. According to Rousseau (2003), the social contract involves individuals surrendering their will to the general will, representing the common good and the community's collective interests. Rousseau argued that legitimate political authority arises from the consent of individuals who participate in forming the general will. This creates a direct democracy where decisions are made collectively.

The relevance of this context regarding this paper is not to build up to a justification of which specific theory is best suited for the forthcoming hybrid model but rather to highlight that this paper does not utilise one of the positions in its pure form, but instead synthesise vital elements from them to create a generalised and appropriate theory to act as a layer within this Hybrid. These fundamental elements are identified as a strong central authority to avoid chaos called out by Hobbes and emphasises the protection of individual rights and limited government highlighted by Locke, and elements from Rousseau, who highlighted the importance of community and the general will in decision-making.

4.3 - Kantian Ethics

[5]Genera Will: The will of the people rather than the will of any individual based on the idea that humans are social beings who need to work together for the good of society.

This paper now covers one of the most influential approaches to moral philosophy: Immanuel Kant [1724-1804], who sought to establish a rational and universal basis for ethics (Kant, 1997) solely grounded in reason and a concept of moral duty. This deontological ethical framework is known for its emphasis on moral principles based on reason and universalizability and can be best summarised by focusing on the following key principles (Johnson & Adam, 2022).

Probably the most well-known parts of this deontological framework are [1] the principle that one should act according to principles that could be universally applied without contradiction (Categorical Imperatives) and [2] that moral principles must be applied to all rational beings and a moral action should be guided by principles that can be universally accepted, thereby ensuring consistency and fairness in ethical decision-making (*Universalisability*). Following on from these key principles is that all individuals must be treated as an end in themselves rather than a means to an end and that any AI should respect peoples' inherent dignity and autonomy and protect their rationality and free will (*Respect for Persons*). Kantian ethics also held that an action's moral worth was not determined by its consequences but by the motive behind the action and the adherence to the moral principle, that individuals had a duty to act by moral laws, regardless of personal desires (*Duty and Obligation*). And that each person has the capacity to use reason to determine and act upon moral principles, allowing moral decisions to be made autonomously whilst guided by rationality and moral law (*Moral Reasoning and Autonomy*).

Finally, Kant envisioned such a world where rational beings are both the authors and subjects of moral laws; this highlighted the idea that moral action should contribute to a world where everyone respects and treats each other as rational beings deserving of moral consideration (*Kingdom of Ends*).

With the benefits succinctly highlighted, the section now moves to address concerns around this ethical framework that justifiably require both acknowledging and evaluating their applicability to this specific use case. Kantian ethics, for example, placed a substantial emphasis on *rationality and reason* as the basis for moral decision-making, and this is

considered a weakness due to disregarding the role of emotions, intuitions and other aspects of the human psyche that can influence such judgements. That said, within the context of what this paper is trying to achieve, it does feel appropriate that an ethical, moral framework be based upon rationality and reason rather than emotions, etc. Firstly, because, relatively speaking, rationality and reason should be easier to define and program than emotions, and secondly, should an ASI develop an emotional capability along with other aspects of the human psyche, it would be intuitive to assume that an AI's emotions will be different to humans.

A lack of specificity could also be construed as an over-generalised moral framework lacking in the ability to offer specific guidance in response to a specific problem/dilemma. This potential for ambiguity does not always provide a decisive answer, for example, when moral duties conflict. Moral ambiguity should be avoided as it can perpetuate inconsistent responses to similar dilemmas, which in turn can lead to distrust. Within the context of this paper, however, the use of Kantian Ethics builds upon a foundational framework which, it is proposed, will be designed in a way to resolve the inconsistency.

Initially highlighted as a strength, the principle of universalisability that aims for the adoption of universally applied moral principles is not without its critics, such as Hegel, who said:
The pure unity of self-consciousness with itself is the complete abstraction from everything determinate...This supreme abstraction, the pure category of the Ideal, the absolute ethical idea, is indeed again degraded by Kant to the merely negative attitude of duty for duty's sake. (Hegel, 1988)

Here, Hegel argues that Kantian ethics is too detached from real moral life and is focused wrongly on reducing ethical standards to empty abstractions and formal rules rather than lived experience. However, Hegel's position is one of a more subjective disposition around contextual social relations being necessary to inform ethics, which within the context of this paper is not applicable, as this paper's position author is that subjectivity must not be a part of an AI moral framework, as this could erode key requirements around explainability/ interpretability.

Determining which principles are genuinely universal will also be significantly challenging, and issues arise around differing moral judgements and achieving consistent moral judgment. Within the context of this paper, two avenues of thought can be considered: [1] If the AI is driven to achieve the least number of principles, then the disadvantaged minority may be considered acceptable by the AI; this, however, feels less than ideal (especially from the perspective of the minority). It would be far more reasonable and almost certainly technologically possible that [2] Considering the significant capabilities an AI system would likely possess; it is plausible that the AI could determine universal moral principles that account for the values of diverse individual social groups. The AI would also be able to resolve any conflicts between principles by referring back to the core ethical directives established in the Divine Command Theory foundation layer. The AI's advanced intelligence and computational power could allow it to find ways to create moral rules that work for multiple perspectives while still aligning with the inviolable principles meant to ensure human interests.

Consequentialism: Kantian Ethics is not consequential; by this, we mean no emphasis is given to the outcome of a moral judgement, as from the perspective of the framework, the focus is on moral duty. Therefore, if the action is just, the outcome is moot. The standard criticism of this is succinctly articulated within the numerous variations of the thought experiment known as the Trolley Problem (Thomson, 1985). This Problem was originally postulated within the area of ethics and psychology and aims to introduce a moral dilemma, thereby forcing the individual decision to sacrifice one life to save others. Although the specifics of the scenario change, the objective remains consistent: to explore the individual's belief about the value of human life and the concept of the greater good and the differences between consequentialism and deontological moral theories.

A typical example of the Trolley Problem would be described as follows: an onlooker has the choice to save five people in danger of being hit by a trolley by diverting the trolley to kill just one person. The question is then posed: what choice should the onlooker make, thus forcing the person taking part in the thought experiment to assess their own moral and ethical position?

Within the context of this paper, an ASI would certainly be capable of predicting the probability of potential outcomes before they affect society, which begs the question, should it? Intuitively, this capability would definitely be of benefit; after all, an AI evaluating possible worlds to determine the greatest universalisation based upon hypothetical decisions feels beneficial. That said, and somewhat counter-intuitively, the position held by this paper is to resist that intuition and argue that only the core decision would be evaluated by the AI for morality [in line with Kant] rather than all subsequent possible *knock-on effects.*

The objective of this position is to pre-empt and nullify ethical issues around punishing individuals for hypothetical future actions, as in the Minority Report[6] scenario. It also avoids the technical challenge of predicting all possible outcomes of potential actions, which would require vast computational resources exceeding even advanced AI capabilities. The intent is to constrain judgment actually to undertake acts rather than speculate future crimes or harms. This prevents morally concerning pre-emptive decisions while recognising computational limits on determining precise consequences of hypotheticals. The focus is, therefore, on responding to realised actions rather than speculated possibilities.

Individualistic Focus: The focus of Kant was on the immoral worth of individual action, and therefore, it has been highlighted that this framework lacks a focus on societal relationships, collective responsibilities, justice, etc. However, within the hybrid model, this is fine, given that any such gaps would be plugged by adopting the Social Contract Theory.

Inflexibility: Kantian ethics is predicated upon defined principles and rules, which can need help to provide the ability for flexibility or consider contextual considerations. It is argued that this framework cannot easily accommodate the complexity and nuance of real-world decision-making that often requires balancing multiple values and adapting to specific circumstances. However, the proposal made by this paper is not that humans will provide an

[6] A film by Philip K. Dick set in the year 2054, where a specialised police department called apprehends criminals before a crime is committed by the use of people who can see into the future.

AI with a Kantian moral framework but instead that the AI is equipped to use key parts from Kantian ethics when formulating moral judgements. Given that the foundation layer of this Hybrid is that DCT is used to ensure core principles are maintained, this paper holds Kantian ethics as the correct non-consequentialist approach to provide the underpinnings of a flexible moral capability.

These underlying philosophical theories act as the cornerstones of the forthcoming hybrid framework, which this paper will now define, justify, and discuss any applicable criticisms. This section now concludes, and this paper articulates how these foundational theories mesh to produce a workable moral framework for accurate artificial intelligence.

Section 5 - Divine Social Kantian Hybrid Theory

This hybrid model uses two types of moral theory, top-down and bottom-up; the objective is to take advantage of the benefits of both whilst compensating for any disadvantages of one by use of the other model. Top-down moral theories aim to derive universal ethical principles that can then be applied to moral situations, a defining being their emphasis on moral rules dictating what actions are right or wrong, regardless of context. Such theories start from broad, abstract principles and work downward to guide specific moral judgments. Conversely, bottom-up moral approaches focus on morality emerging from context-dependent rational deliberation and consider the specific details of ethical dilemmas.

The theories of Devine Command Theory [L1] and Kantian [deontological] ethics [L3] are both classed as top-down as they are based upon applying explicit ethical principles to specific situations. However, within this paper, it is proposed that they are used for different purposes. The benefits of top-down moral theories are that they provide more transparency and consistency. However, they may also be considered inflexible. To compensate for this inflexibility, the author envisions that at the implementation stage of this model, the filling[L2] of Social Contract Theory, a bottom-up theory based upon reinforced learning, would be sandwiched between the two [L1 & L3] top-down theories which would allow for an evolved moral behaviour based upon examples or feedback.

At a high level, this approach proposes the following layered moral hybrid theory. The foundation or core layer of DCT [top-down] provides the Hybrid with a transparent and rigid foundation, intended to focus upon encapsulating the core-moral principles that should not be overruled, changed or evolve (i.e. do not kill humans); this layer also provides the mechanism to resolve conflicts generated by the following layers. The middle layer based upon SCT [bottom-up] would be based upon a snapshot of society's core moral principles. Still, importantly, it would also have the ability for these ethical principles to evolve along with the community itself over time. This layer would provide the needed level of adaptability and diversity, which would assist with the model reflecting the diversity of humanity's morals, etc. Principles within this layer would not be capable of conflicting with those principles defined

within L0, thereby preventing it from becoming moral to wipe out humanity to save the earth from pollution, etc.

Lastly, the highest level with the most flexibility is derived from Kantian Ethics and would allow the AI to generate novel moral principles, thereby allowing the AI to adapt to situations unforeseen during its creation. Being based upon this deontological theory would provide the necessary guardrails, ensuring any moral principles created would not conflict with the interests of humanity but would also utilise the other two layers to resolve any potential conflicts. With the Hybrid defined at a high level, this paper now moves to dive a little deeper into each layer to outline the key elements synthesised from each foundational theory.

5.1 – Adoption of Divine Command Theory [Level 0]

Within the scope of this paper, the proposed model incorporates key elements of DCT at the foundational layer [L0]. This layer is where the core moral constraints would be identified and defined. This paper holds that this initial framework be analogous with DCT due to said moral constraints being immutable in nature and handed down to the intelligence by a society [specifically a dedicated professional body of engineers and ethicists presumably]. Such commandments, to use the DCT vernacular, would be considered divine because humanity commanded them, in line with the position held by *Scotus* & Ockham, as in that within this analogy, Humanity would be the ultimate source of morality.

This foundation provides a resolution to any potential Grounding Problem other moral frameworks encounter, with humanity being the source of morality in this instance. To stave off criticism of this method resolving any potential grounding problems as being arbitrary, this paper proposes that DCT, in this instance, would reflect the basic moral principles of society that are outlined within current legal frameworks. Furthermore, it is here in L0 that technological concerns around the Control Problem would be addressed by ensuring a relevant directive is embedded to ensure humanity, if needed, would be able to *nudge* the moral principles.

An acknowledged limitation of this paper is that there is no proposed list of commandments put forward as an example; this is an intentional gap to ensure the focus is kept on the feasibility of the overarching model rather than justifying each item on a hypothetical list. That said, loosely, the author envisions this area to define the necessary guardrails to ensure situations such as mass extinction, the enslavement of humanity and all the other dystopian plot lines of movies where an AI is the protagonist are avoided.

The primary function of this layer is to enforce the key principles of morality that will not evolve over time and are, therefore, immutable within the AI. Such principles will act as the overriding rules that act as the deciding logic to resolve conflicts caused by the designed flexibility within the coming layers based on Social Contract Theory and Kantian Ethics.

5.2 – Adoption of Social Contract Theory [Level 1]

Building upon the core layer being based upon DCT as described above, this hybrid framework would gain a level of flexibility within the next layer [L1] by adopting principles set out within Social Contract Theory. Whereas the relationship between humanity and an AI from the perspective of ensuring survival would be qualified within the initial layer [L0], this layer would be where the relationship between humans and AI would be generally fostered with applicable guardrails but not as restrictively as defined in L0.

This level of mutual interaction, it is proposed, would be achieved by establishing the necessary framework so an AI would view the relationship between Humans and AI as that of social contract theory, where the terms of this contract would be used to define the overarching objectives and priorities and when appropriate act as a guide for the behaviour and interactions of AI systems within society. The underlying principles adopted by this Hybrid from SCT, as defined in §2.2, would be those of limited governance, individual rights, equality, and fairness. This is intended to ensure the AI acts in a considerate way and beneficial to human society. Specifically, limited governance means the AI would not be permitted to gain unchecked power over human affairs or infringe on human autonomy. Its capabilities would have defined legal and ethical boundaries. Respect for individual rights indicates the AI could not violate established protections like free speech, privacy, due process, etc. Equality and fairness suggest that AI would need to treat all people in an equal,

non-discriminatory manner and ensure its decisions and actions are impartial and just. It cannot unjustly favour certain groups over others.

By ingraining these four moral values - limited governance, individual rights, equality, and fairness - the designers hope to create an AI that is both empowered to help society while also constrained from abusing power or causing harm. It should assist humans in a way that aligns with core ethical principles that protect individuals and communities.
This expansion helps explain both the core moral values seen as essential for a beneficial AI and how adhering to those values would prevent potentially dangerous AI behaviour. Let me know if you would like me to elaborate further on any part of the explanation.

5.3 – Adoption of Kantian Ethics [Level 2]

The final layer of this hybrid model would build upon the initial two layers and provide the greatest level of flexibility by providing the AI with the capability to formulate and adapt to societal nuances and unforeseen future moral values, all whilst adhering to a consistent and beneficial approach to humanity by following the moral underpinnings proposed within Kantian ethics.

By *Kantian underpinnings*, this paper is specifically proposing the synthesisation of the following general objectives that this layer will adopt to consistently maintain when deciding actions or building a moral code [or guardrails] that are outside that which is defined in L0 & L1. Initially, moral rule creation must promote a universalisability principle to avoid AI creating moral codes for situations that do not scale. This adoption of *universalisability* within the context of an AI would be represented as a value [for example, between 0 and 1], which would not only allow the AI to evaluate how universalisable a moral principle was but, furthermore, given the computing power such an AI would command, there is no reason to consider that this universalisability could not be applied to societal subgroups in some instances rather than just a standard Kantian categorical imperative position of 'all or nothing' when applying it to humanity as a whole [articulated in §3.3].

It is acknowledged that this quasi-universalizability goes against Kantian ethics in its purest form and that this may leave this paper open to criticism. However, it is the position of this paper that the overall benefits to society of an AI able to tune 'categorical imperatives' to align them with different social groups justifies the divergence from pure Kant.
With that nuance articulated, the layer would also adopt the following principles to act as moral guardrails whilst enabling the moral framework to be flexible.

- Principle of Respect for Persons – An AI must respect human dignity and autonomy and must not be permitted to treat humanity as a mere means to an end. Furthermore, AI must never infringe on freedom of choice or action.

- Principle of Impartiality - Algorithms and decision-making processes must be free of bias, prejudice or discriminatory effects. AI should treat all people impartially.

- Transparent Intent - An AI system's intent and reasoning should be understandable and transparent, not opaque. This upholds the Kantian goodwill principle.

- Consistency - Moral principles embedded in an AI must be applied consistently in all cases. No violations of rules based on circumstantial variables.

In summary, Kantian ethics would emphasise designing universal, transparent and impartial ethical principles into AI systems that respect human dignity while avoiding negative biases or consequences. Moral decisions should be explainable, not arbitrary.

Section 6 – Benefits & Weaknesses of the Hybrid Model

The overall objective of this hybrid model is to ensure that an AI that is designed with this moral framework would proactively benefit humanity in the decisions that it made; this objective can be qualified as the following sub-benefits:

Internal Conflict Resolution: In situations where moral dilemmas arise that conflict between layers, the model would adopt a number of options for resolution. Primarily, this deconfliction process would be realised by the hierarchical ordering within the model so that core values defined within DCT would take the highest precedence, SCT would take a medium, and KE would take the lowest precedence. Therefore allowing for a reliable and robust mechanism for deconflictions to take place.

Respect for Humanity: Not only is it imperative that a moral framework for an AI represent the current societal moral status quo, but equally important, and arguably a greater challenge is to ensure the flexibility of the model in order that it reflects the future status quo, too. A significant benefit of this three-layered hybrid model is that it would ensure that respect for humanity is maintained by the model evaluating potential principles to ensure that they are Kantian enough to be valid within the model by ensuring that humans are consistently treated as a means in themselves.

Flexibility & Explainability: The Kantian layer of the model further ensures that as the Social Contract Layer evolves to reflect society, proposed principles would be required to respect human dignity, autonomy, and rights and not manipulate or exploit individuals. This Kantian layer further ensures that should society's values evolve to be less than ideal, and there is a preventative measure to halt the AI's moral compass from becoming an echo chamber for suboptimal ideals. A further proposed benefit is that adopting foundational theories that are rational by their very nature facilitates an overall hybrid model that acts in a consistent and rational way, which, given the structured principles the model adopts, would help drive a model that allows for explainability by ensuring necessary data points are accessible for integration outside of the system. By ensuring the AI adheres to aspects of the SCT, such as

principles of limited governance, the resulting AI would foster a respect for human autonomy, ensuring an AI assist human decision-making rather than exhibiting a more authoritarian approach.

Stronger Together: The combination of Social Contract Theory and Kantian ethics will resolve and remove the impact of criticisms highlighted of societal and cultural nuances from the foundational moral framework. That all said, this should by no means be considered a simple ask. Integrating key elements from both DCT and SCT would require the collaboration of multiple disciplines and areas of artificial intelligence; it would further require consensus amongst experts within the area of ethics and the virtues to be hardcoded within an AI so as not to overly confine the AI at the same time as not impacting societies freedoms. Furthermore, individual religious and cultural perspectives would need to be respected, meaning a level of flexibility is required and maintained to accommodate diverse beliefs and values.

The proposed model attempts to address the majority of concerns that are inherited from the underlying theories; however, during the creation of this paper, a number of concerns arose that will be discussed shortly. It is important to note that this is not suggestive of an exhaustive list but the points at the forefront of the author's mind.

Fostering Moral Learning: The multilayered approach of this model would embed the capability of morality being ever-evolving, much like how morality has evolved for society over the centuries, at multiple layers of the model. Within the SCT layer, social norms would evolve as humanity's values shift over time. The sole purpose of the KE layer is to allow the model to derive new moral principles to expand AI's ethical deduction and abstraction capabilities. Given the computational complexity of such a system involved in actualising this hybrid model, it would be almost certainly capable of also Simulating hypothetical dilemmas to extrapolate the implications of moral rules in new contexts, thus continuing to broaden its understanding.

Technological Limitations: The technological advances likely required to move this proposed hybrid model out of a philosophically focused paper into something engineering-focused, it is

acknowledged, could be construed as a weakness. After all, there is little benefit to proposing 'make the AI moral this way' if this way is simply impossible. However, it is the position of this paper that it is best to aim for the stars by defining the ideal state and then working with what is technically possible, as what is technically possible constantly advances what could be considered philosophical whimsy one day, could just as well be considered: the standard in engineering the next.

The Unknown Unknowns: Although this hybrid attempts to account for eventualities arising from conflicting principles [via the hierarchical model); or being flexible to account for society's moral evolution via the adoption of key attributes from SCT and KE. Intuitively, it would be naïve to assume nothing suboptimal would occur (or colloquially *would go wrong*); for example, a subset of humanity would be wronged, or the AI would fail to interpret events correctly, maybe due to not having enough data available. That being said, it is reasonable to assess that the proposed model would be capable of making the best judgement based on the information available, which, at the very least, would be on par with a moral human being.

Unaccountability: This paper does not address the concept of accountability within AI nor how this model would specifically address the requirement. Although not a weakness per se, it is acknowledged that for an AI to function within society, it is an area that would need to be addressed to ensure society has an avenue of redress, etc., should something bad occur. For the sake of argument, let us suppose that the hybrid model is implemented into an artificial intelligence of human or greater level; does this make the AI moral or merely act in a moral way? The answer informs your position:

> [1] If the AI is following the rules, then intuitively, those who created the rules would be accountable.
>
> [2] But should the AI have the capability to ignore the rules, then this level of moral agency would suggest that the AI is potentially accountable.

Section 7 - Avenues of Further Inquiry

During the research and creation stages of this paper, additional questions came to light that, although out of scope for this specific body of work, in principle, may act as a catalyst for future research and are therefore documented below:

Ethicality: The focus of this model has been to equip an AI with the necessary logic to act in a moral way that is initially designed by humans and has the overarching purpose of not negatively impacting humanity; furthermore, the model provides the ability for humanity to shape/tune the model to ensure it remains aligned to society. Ethical questions around this hybrid model and the control problem more generally when we consider an AI that becomes more than a mere societal tool. Would we be ethically just to maintain control over such an entity, or would history look back on us as we do around caging animals?

Religious Fanaticism: A hybrid model with a foundation of DCT [L0], if incorrectly defined, could foster an AI with religious zealot tendencies; this would almost certainly pose a risk to humanity or at least a subsection of humanity. Specific research into which *commandments* would best ensure a more agnostic yet beneficial moral AI would be a natural follow-up to this paper.

The Devil is in the Detail: The scope of this paper was to provide insight into how a hybrid moral framework for AI could be created based on three foundational theories from a philosophical position. A logical next step in this pursuit would be to evaluate how such a model would be created from an engineering perspective, as this will likely inform revisions to this body of work.

Multidisciplinary: The hybrid framework also opens up new avenues for future research or applications. For instance, it would be interesting to test the hybrid framework empirically by conducting experiments or simulations with artificial agents or systems. It would also be useful to explore how the hybrid framework can be integrated with other aspects of artificial intelligence, such as cognition, emotion, communication and creativity. Moreover, it would

be valuable to examine how the hybrid framework can be applied to different domains or contexts, such as health care, education, law or politics.

Section 8 – Conclusion

The primary objective of this paper was not to propose a perfect moral framework for a hypothetical artificial intelligence but to conceive of a model that would almost be considered more of a *Mary Poppins* moral framework in that it would be feasible, workable, and set the foundations for being *morally just*; or *practically perfect in every way*. Furthermore it also feels arrogant, or intuitively ill-conceived to suggest that humanity, imperfect itself, could possibly create moral framework for another intelligence that was perfect.

The goal was to explore the possibility of creating a moral artificial intelligence that is respectful of and aligned with human values. To achieve this, the paper proposed a hybrid framework that combines key attributes from Divine Command Theory, Social Contract Theory, and Kantian Ethics' This paper now concludes by summarising the main points and findings from each chapter.

In Chapter One, the paper introduced the research question and the motivation behind it; it argued that the development of artificial intelligence poses ethical challenges and risks and that existing ethical theories are likely inadequate to address them. The following chapter [2] explains the main features and assumptions of Divine Command Theory, Social Contract Theory and Kantian Ethics and analyses their strengths and weaknesses, how they relate to artificial intelligence and then identifies the key attributes relevant to the hybrid framework (such as divine authority, social agreement, rationality, universality, and autonomy). The paper then progressed on to describe how the hybrid framework works [3] and how it could be implemented in artificial intelligence. This section showed how the hybrid framework could generate moral rules and principles that are consistent with human values and how it would potentially resolve moral dilemmas and conflicts. Chapter 4 then evaluated the hybrid framework from different perspectives and criteria and concluded that the hybrid framework has several advantages and benefits over other individual frameworks and that it could overcome some of the common objections and challenges.

The hybrid framework has important implications and limitations that need to be considered. On the one hand, it offers a way to create a moral artificial intelligence that is compatible with human values and interests and that can enhance human well-being and dignity. On the other hand, it raises questions about the nature and source of morality, the role and responsibility of humans in relation to artificial intelligence, and the potential risks and consequences of creating a divine-like artificial intelligence. These issues require further investigation and discussion from various disciplines.

This dissertation aimed to make a valuable contribution to the discussion on morality for AI by proposing a hybrid framework that can ensure the morality of artificial intelligence. The hybrid framework is based on a novel combination of key attributes from Divine Command Theory, Social Contract Theory and Kantian Ethics. The hybrid framework can generate moral rules and principles that are consistent with human values, resolve moral dilemmas and conflicts, foster moral learning and development, and resolve the limitations of any single theory.

This paper The hybrid framework also has important implications and limitations that need to be considered and suggests directions for future research or applications. By developing the outline for this hybrid framework, this paper has demonstrated that a moral artificial intelligence that is respectful of and aligned with human values is, at the very least, conceivable.

Bibliography

Arnold, T., & Scheutz, M. (2018). The Big Red Button is too Late: An alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology*, 59-69.

Austin, M. W. (2021). *Divine Command Theory*. Retrieved from Internet Encyclopedia of Philosophy: https://iep.utm.edu/divine-c/#H3

Blackman, R. (17 July 2022). *Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI Hardcover.* Harvard Business Review Press.

Bostrom, N. (2016). *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Clanton, C. J., & Kraig, M. (2020). William of Ockham, Andrew of Neufchateau, and the Origins of Divine Command Theory. *American Catholic Philosophical Quarterly*, pp. 405-429.

Cointe, N., Bonnemains, V., & Saurel, C. (2016). Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology*, 25-36.

Fowler (Translator), N. N., & Plato. (1999). *Plato: Euthyphro. Apology. Crito. Phaedo. Phaedrus.* Harvard University Press; Reprint of 1904 edition. Retrieved from http://data.perseus.org/citations/urn:cts:greekLit:tlg0059.tlg001.perseus-eng1:10d

Guarini, M. (2026). Particularism and the classification and reclassification of moral cases. *IEEE Intelligent Systems*, 22–28.

Gubrud, M. (1997). Nanotechnology and International Security. *Fifth Foresight Conference on Molecular Nanotechnology*.

Hegel, G. (1988). *Hegel: Faith and Knowledge: An English translation of G. W. F. Hegel's Glauben und Wissen.* State University of New York.

Hobbs, T. (1651). *Leviathan.*

Johnson, R., & Adam, C. (2022). *Kant's Moral Philosophy*. Retrieved from The Stanford Encyclopedia of Philosophy (Fall 2022 Edition): https://plato.stanford.edu/entries/kant-moral/

Kant, I. (1997). *Lectures on Ethics* (Vol. The Cambridge Edition of the Works of Immanuel Kant). (P. Heath, & J. B. Schneewind, Eds.) Cambridge: Cambridge University Press.

Lindner, F., & Bentzen, M. (2017). The hybrid ethical reasoning agent IMMANUEL. *ACM/IEEE International Conference on Human-Robot Interaction*.

Locke, J. (1689). *Second Treatise on Government.*

Locke, J., & Laslett, P. (1988). *Two treatises of government.* Cambridge University Press.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 175-183.

Nevala, K., & Blackman, R. (2023, 4 5). Practical Ethics with Reid Blackman [Episode #29]. https://pondering-ai.transistor.fm/.

Osborne, T. (2005). Ockham as a Divine-Command Theorist. *Religious Studies, 4*, 1–22.

Pascal, B. (1670). Pensées. p. Section 3. p233.

Powers, T. M. (2006). Prospects for a Kantian machine. *IEEE Intelligent Systems 21(4)*, 46–51.

Rousseau , J.-J. (2014). *The Social Contract, or Principles of Political Right.* CreateSpace Independent Publishing Platform.

Rousseau, J.-J. (2003). *The Social Contract.* New York: Penguin Classics.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* London: Allen Lane.

Russell, S. (2021). *Human-Like Machine Intelligence.* OUP Oxford.

Song, F., & Yeung, S. H. (2022). A pluralist hybrid model for moral AIs. *AI & Society*.

Thomson, J. J. (1985). The Trolley Problem. *Yale Law Journal, 94*(6), 1395–1415.