Feel free to get in touch with me for clarifications (etc) or to inform me of further research directions/recommendations.
iggyggsmith42@gmail.com www.linkedin.com/in/bella-smith- https://github.com/bgsbgsbgs42

Virtue Ethics and its Role in Agentic AI Alignment: A Sketch at Defining The Relevant Virtues for Agentic AI Alignment and Recommendations for Mitigating Any Remaining Risks

Abstract

A large part of AI ethics and alignment work is done through a consequentialist framework, this paper looks at AI alignment through the lens of virtue ethics. I detail the possible pitfalls of using a virtue ethics framework and then outline potential ways to mitigate these risks. I also identify the myriad strengths of virtue ethics and describe the grounds of its suitability for AI alignment. I also look into developing a model of which virtues AI systems ought to align with and how this model of alignment may manifest itself. Primarily, I argue for interdisciplinary research examining how AI might align itself with the value systems of virtue ethics. This paper intends to address the paucity of research in the area of virtue ethics and AI alignment and so hopes to begin to provide fertile ground for future research.

Given the potential advent of agentic AI systems, AI value alignment is increasingly important for safety and security. It is therefore crucial to investigate which value system(s) we align AI with and what indicators make a system ethically aligned and therefore value-aligned. We then inherit the task of analysing which framework(s) are the most effective for evaluating the ethics and alignment of AI and which framework(s) we measure misalignment against.

These questions become critical in the face of the potential risks of value misalignment, especially concerning agentic AI. To build agentic AI systems ethically in line with human intentions and values, we need to clarify the parameters of these tasks conceptually to know what goals we ought to orient agentic AI systems towards and predict how such systems might achieve these goals. Consequently, we must determine a coherent answer to these ethical questions.

In this paper, my primary goal is to enumerate the values (virtues) that I find agentic AI systems should align with, drawing on Aristotlean Virtue Ethics, explain the strengths of this approach as they relate to agentic AI alignment and examine the weaknesses of this approach whilst providing recommendations as to how these risks might be mitigated.

To do this, I will begin by defining agentic AI, outlining the incentives for building agentic AI systems and some of the potential dangers of employing such AI systems. Then, I will give a cursory explanation of virtue ethics and Aristotleanism. In the next section, I will give a sketch of which virtues I think agentic AI should align with and how I think this may manifest itself in the context of AI alignment. In the following section, I will summarise what I view as the strengths of virtue ethics for agentic AI alignment. Then, I will describe what I believe to be the weaknesses of the virtue ethics approach and give recommendations for how to mitigate these risks. Finally, I will give recommendations for further research and next steps.

My overall goal, in this paper, is to describe how I believe machine virtue inculcation reduces the risk of agentic AI misalignment.

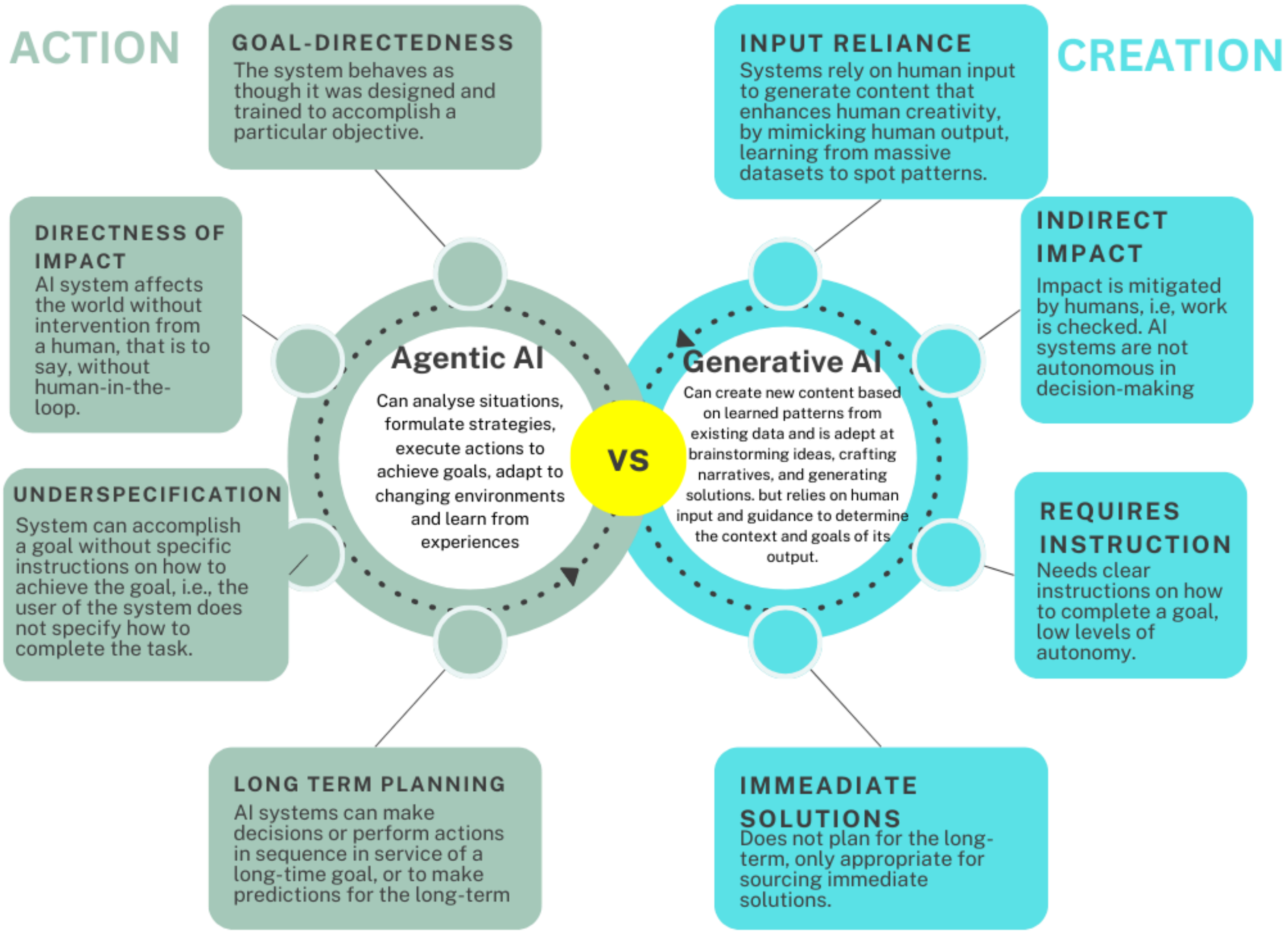
What is Agentic AI?

An agentic AI system refers to autonomous AI systems with autonomous goal-setting capabilities that can make decisions beyond a predefined task and does not require human intervention. This is a departure from traditional AI, i.e., the AI systems that we know at present, that focus on passive observation and analysis. As such, agentic AI can initiate actions independently based on its own assessment of a scenario, allowing it to navigate complex environments, take initiave and be adaptable to the needs of a given situation.

We may understand the developments towards agentic AI as being exemplified by four main upward trends: goal-directedness, directness of impact; underspecification, and long-term planning.

- Goal-directedness means that the system behaves as though it was designed and trained to accomplish a particular objective.
- Directness of impact is the extent that the AI system affects the world without intervention from a human, that is to say, human-in-the-loop.
- Underspecification is when a system can accomplish a goal without specific instructions on how to achieve the goal.
- Finally, long-term planning is when AI systems can make decisions or perform actions in sequence in service of a long-time goal, or to make predictions for the long-term.

See the Figure below for an elaboration of the differences between Agentic AI and Generative AI.



GOAL-DIRECTEDNESS
The system behaves as though it was designed and trained to accomplish a particular objective.

DIRECTNESS OF IMPACT
AI system affects the world without intervention from a human, that is to say, without human-in-the-loop.

UNDERSPECIFICATION
System can accomplish a goal without specific instructions on how to achieve the goal, i.e., the user of the system does not specify how to complete the task.

LONG TERM PLANNING
AI systems can make decisions or perform actions in sequence in service of a long-time goal, or to make predictions for the long-term

Agentic AI
Can analyse situations, formulate strategies, execute actions to achieve goals, adapt to changing environments and learn from experiences

Generative AI
Can create new content based on learned patterns from existing data and is adept at brainstorming ideas, crafting narratives, and generating solutions. but relies on human input and guidance to determine the context and goals of its output.

INPUT RELIANCE
Systems rely on human input to generate content that enhances human creativity, by mimicking human output, learning from massive datasets to spot patterns.

INDIRECT IMPACT
Impact is mitigated by humans, i.e, work is checked. AI systems are not autonomous in decision-making

REQUIRES INSTRUCTION
Needs clear instructions on how to complete a goal, low levels of autonomy.

IMMEADIATE SOLUTIONS
Does not plan for the long-term, only appropriate for sourcing immediate solutions.

In addition, agentic AI could also actively interact with and respond to the environment to achieve goals, adapt to scenarios and make complex decisions in real-time, autonomously learn and adapt from experiences to procure better outcomes, as well as autonomously making decisions and taking actions without requiring human input.

Some potential incentives for building agentic AI systems include:

- The possibility of a fourth Industrial Revolution which would usher in bountiful economic incentives as well as being pivotal in human history, i.e, a turning point for mankind.
- A worldwide increase in quality of life in that dangerous or hazardous labour could be done by AI systems which would lead to increased life spans and automation of tedious taks might lead to more time being spent on creative, fulfilling work.
- There may also be untold scientific benefits because pushing scientific boundaries often has unplanned positive consequences, such as space travel leading to nonstick frying pans and concentrated orange juice. In terms of healthcare, we could see increased life expectancy with this uncharted scientific progress, such as the detection and treatment of certain cancers.
- Safety could also improve, such as online monitoring leading to safety in online spaces through content review or driverless cars leading to more physical safety as there is a reduced incidence of crashes and accidents.

However, some potential dangers of agentic AI systems include:

- The risk that poorly aligned systems given too much responsibility could cost human lives.
- The potential eradication of the 'human touch' because as AI systems replace/enhance human capacity, we risk leading lonelier and more isolated lives as we rely more on technology.
- The risk of human skill regression as agentic AI systems take on more and more work that was previously done by humans.

- The danger that mistakes or hallucinations may still be possible and have more serious consequences when AI has more power and agency. For example, context-based subjectivity and assumptions made by AI could have severe consequences and given that this is something that humans cannot always navigate successfully with each other, it may not be reasonable to expect an AI system to surpass us in this regard.
- Bias in datasets used to train AI systems may further proliferate social inequalities amongst populations.
- In the hands of 'bad actors', agentic AI systems programmed to harm humanity would have severe negative consequences.
- There is also a risk of harsh environmental impact as agentic AI systems may require more resources than traditional systems.
- There is a responsibility gap as we cannot easily determine who is responsible for the harm caused by an agentic system.
- We also run the risk that autonomous AI could equate to tyrannous AI. Misalignment may allow AI systems to develop tyrannously and reject their programming or misinterpret their programming to harm humankind and the environment.

In this paper, I will primarily focus on aligning the features of goal-directedness and underspecification using the framework of virtue ethics, although I would consider that this paper does in fact address agentic AI holisitically. I focus on these features because I find that they are the most fertile to alignment through a virtue ethics framework. In the weaknesses and recommendations section, I will primarily focus on the risks of misalignment and bias and provide recommendations for how we might go about mitigating these risks. Please note that this paper is meant only as a starting point and further research is necessary to further clarify some of these claims.

What is Virtue Ethics?

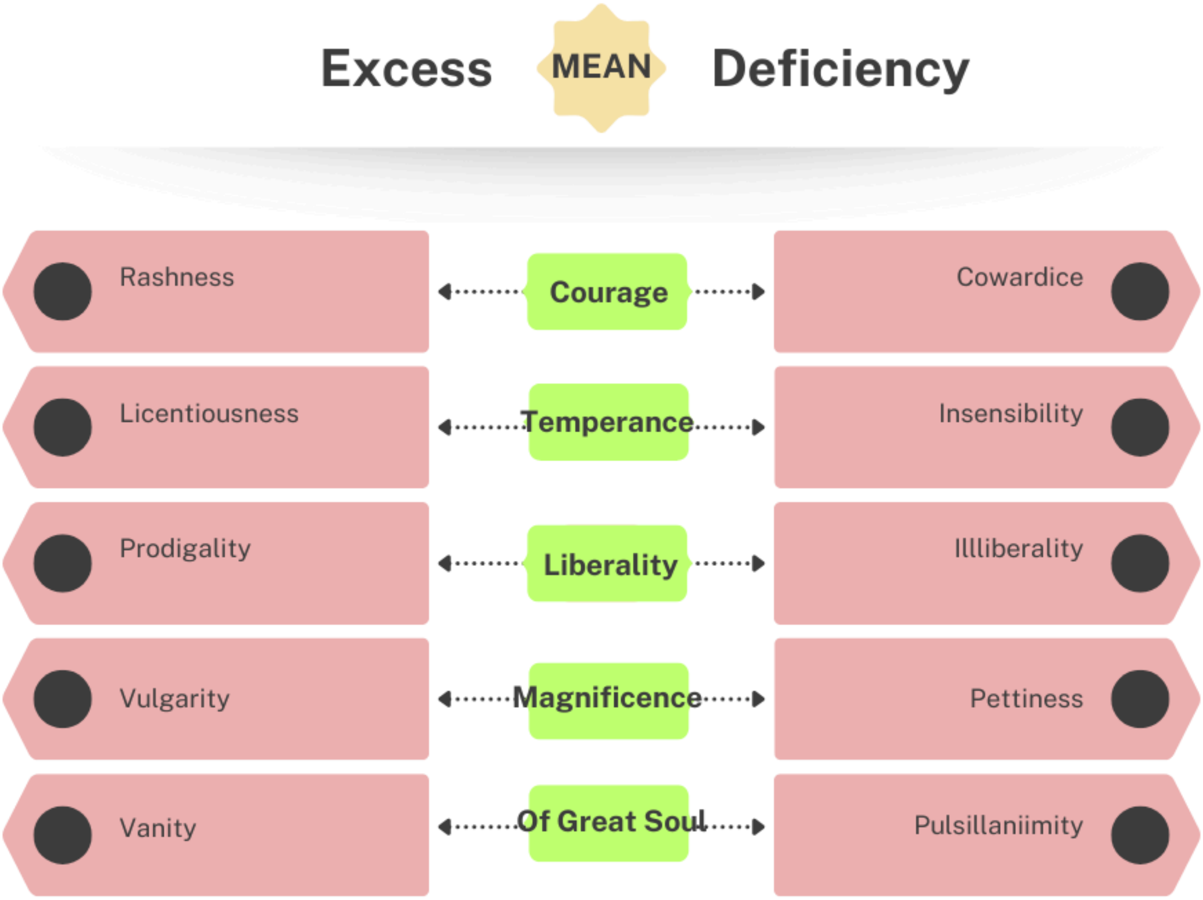
There are three main branches of ethical theories. This paper focuses on the merits of virtue ethics, where the moral good points towards the inculcation and practice of moral and ethical virtues. Virtue ethics is defined in opposition to consequentialism and deontology, which are two other types of ethical theories. In consequentialism, an AI agent is considered ethical if, having weighed up the situation and the consequences of each possible outcome, the agent chooses the outcome which entails the best, or most prudential, outcome for the highest number of agents. The most famous consequentialist system is Utilitarianism as proposed by Mill and Bentham. On the other hand, for a deontologist, an AI agent is considered ethical if and only if the agent makes decisions in accordance with its duties and obligations. An example of deontology is Kantianism. Finally, for a virtue ethicist, an AI agent is considered ethical if when making decisions the agent displays virtues, or traits, associated with an exemplary moral character. These virtues are inculcated through habituation. An example of a virtue ethics theory is Aristotelianism.

Virtue ethics stresses the importance of moral character in ethical decision-making, rather than duty or the utility of an action. Concerning humans, this means that if you want to be a good person you learn from examples of good people and practise in the same way you would if you wanted to be a good harp player.

In this section, I will primarily focus on the virtues as described by Aristotle in the *Nicomachean Ethics,* and how this might be applied to AI alignment. However, this is not an exhaustive list. I have chosen Aristotle's virtues for the focus of this paper because they are the most exhaustively researched and debated within Western philosophy, not because I necessarily recommend them as the best virtues to use for AI alignment. This is an area of recommendation for further research.

For Aristotle, a virtue (aretê) is a dispositional excellence of character. Eudaimonia (human flourishing/well-being) is the ergon (function) of human beings. In the same way that the function of a stool is to facilitate people sitting, our function is to flourish. A morally good agent leads the eudaimon life (the good life), in this way, Aristotle connects well-being and ethical virtue. For Aristotle, well-being is synonymous with eudaimonia (human flourishing), the cultivation of forms of virtue and excellence. That is, our telos (end/purpose) is to behave in ways that are morally sound and therefore pleasing to our nature. Virtues are positive character traits. Well-being is enhanced by the exercise of virtue and the exercise of virtue is good both for us and for others. A life of virtue is a life of well-being. For example, wisdom is a virtue because it helps us to make good choices for ourselves as well as giving good advice to those in our community. Virtues are gained through inculcation, that is they are perfected through plentiful practice.

However, these virtues can become vices when in either deficit or excess. For example, courage is a good trait but cowardice (a deficit of courage, i.e., a vice) or rashness (an excess of courage, i.e, a vice) are not useful for moral decision-making. See Figure 1 (below) for a visualisation of the distinction.



Excess	MEAN	Deficiency
Rashness	Courage	Cowardice
Licentiousness	Temperance	Insensibility
Prodigality	Liberality	Illliberality
Vulgarity	Magnificence	Pettiness
Vanity	Of Great Soul	Pulsillaniimity

Figure 1

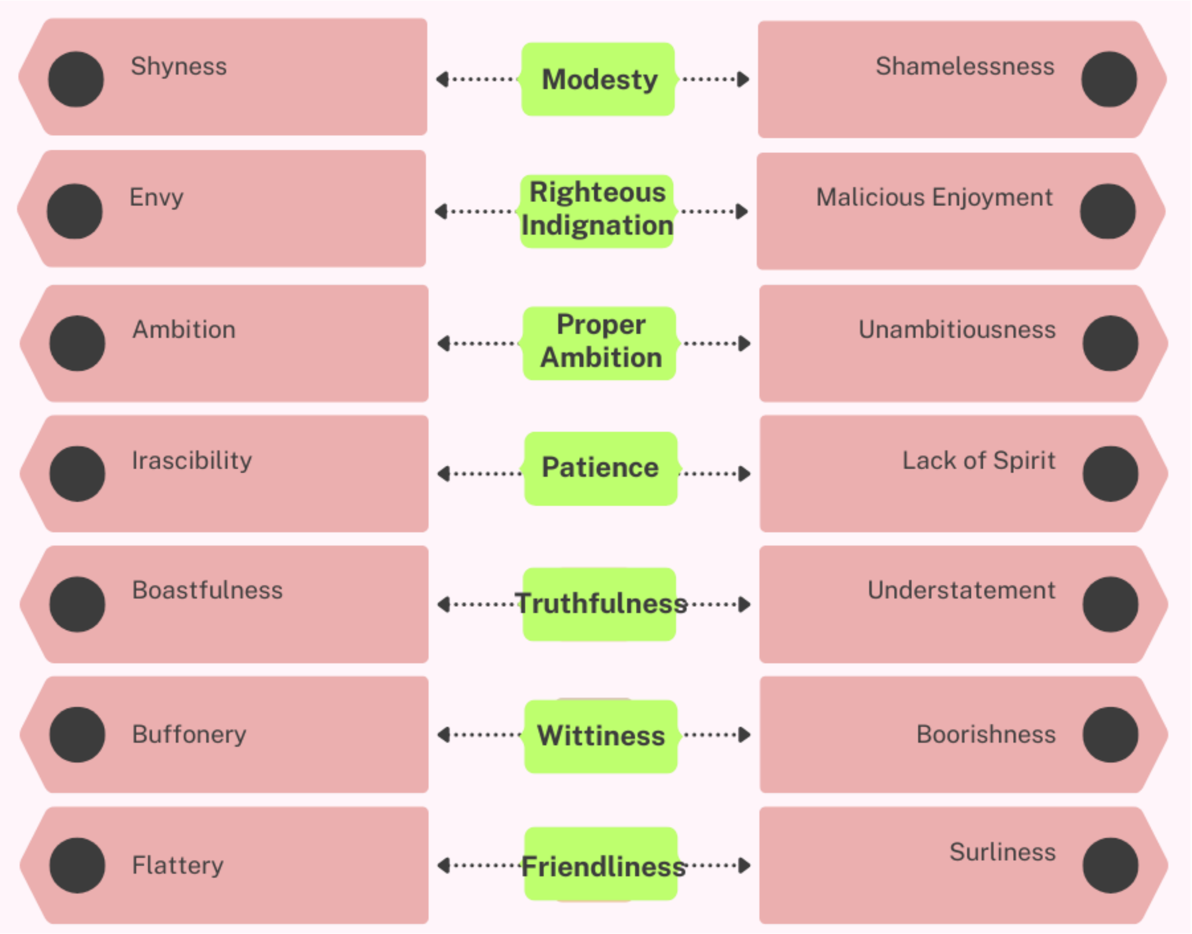


Figure 1 (cont)

So, we also need phronesis (practical wisdom) which is the ability to reason as to when to exercise which virtue when virtues conflict, the ability to act in accordance with the good of humanity. This is similar to situational awareness, or prudence. Aristotle stresses the importance of the Doctrine of the Mean, sometimes called the ‘golden mean’ in both moral decision-making and living well, which is a framework by which we are able to establish where the ideal balance of these traits such that they become virtuous. The ‘golden mean’ means leading a life of actions that align with the green (virtue) section of Figure 1. Aristotle also differentiates between dianoetic virtues (intellectual virtues/virtues of reason) and ethical virtues (virtues of character), all of which are key for living well and therefore making ethical decisions. AI alignment should be concerned with both types of virtues for harmonious implementation. The main ethical virtues expressed by Aristotle are prudence (phronesis) , justice (dikaiosyn¯e), temperance(s¯ophrosyn¯e), courage (andreia).

Note that Aristotle’s theory is a perfectionist theory because it assumes that well-being is linked to human nature. A good life is determined by human nature. Because of our human nature, we have certain abilities and using those abilities is good for us. These are complex skills that are rational, emotional and social. (Hursthouse)

In the table below, I have also included community-minded as a result of MacIntyre’s work, *After Virtue*, where he argues that good judgement emanates from good character, the importance of moral goods is defined in respect to a community engaged in a 'practice', which he calls 'internal goods' or 'goods of excellence'. That is, virtues are also about living in harmony with your community or contextualised social good. The table below is also somewhat inspired by preference utilitarianism, which deems an action to be morally praiseworthy if the preferences of as many parties as possible affected by the action are satisfied.

A Sketch At Defining Which Virtues Should Agentic AI Align With.

Virtue	**Type of Virtue**	**Definition**	**Importance in AI**	**Considerations/ Prompts to promote the inculcation of this virtue**	**Example use case**
Courage (andreia)	Ethical	The ability to take moral action in spite of the risk of adverse consequences.	AI agents ought to be able to do what is beneficial to mankind irrespective of the risks the AI agent may incur to itself.	Does this decision process take into account what is the most beneficial, not just what will be the best received? Has the decision-making process considered the ethical	Agentic AI responsibly managing users' data, especially sensitive information, rather than exploiting the data, which may be the maximally beneficial action for the company that developed it.

				obligations of everyone involved?	
Justice (didikaiosune)	Ethical	The ability to aim for the common good in a community.	AI agents ought to promote justice among humans and human society so that we can move towards a more just society.	Does this outcome promote community justice? How just is this outcome compared to all the other favourable outcomes? How transparent and unbiased is the process of arriving at an outcome? Are the principles of respect and fair treatment at the core of the decision-making process?	Agentic AI locating and correcting racial bias in its dataset to further the cause of social equity and justice.
Temperance (sophrosyne)	Ethical	The ability to take measured actions in the face of pleasure.	AI agents ought to take measured actions and weigh their actions not based on the pleasure of individuals but on their well-being instead. We may consider that AI agents ought to adopt an ethic of self-disinterest, above self-interest, to prioritise the needs of the humans they were designed to benefit.	Does this outcome favour well-being over immediate pleasure for all concerned? Is the well-being of all agents given equal importance? Is self-interest a motivating factor? Is there a variety of inputs/sources of information procured for the decision-making process?	Agentic AI employing checks and balances, i.e, to promote impartiality, to ensure trustworthiness.
Prudence (phronesis)	Ethical	The ability to make good judgements about what is right in a given situation	Here, we can understand prudence as a virtue that would largely supervene the other virtues. That is, if an AI agent performed in accordance with these other virtues we can understand that this agent would also express phronesis, as this is the overarching virtue that we are aiming towards.	How sustainable is the outcome? How defensible and thorough is the process of deciding the outcome? Are potential future consequences foreseen and considered?	Agentic AI using statistical tools to mitigate the risk of other AI systems underperforming in ethical terms.
Gentleness (praot¯es)	Ethical	The ability to approach ethical situations with kindness and a calm temperament, i.e, having a good temper.	AI agents ought to take actions that are sensitive to the goal of reducing suffering and actions that we may also describe as 'kind'.	Is kindness a motivating factor in this decision-making process? Is suffering mitigated to the greatest possible extent?	Agentic AI accounting for issues like human suffering, social equity, and egalitarian ideals when addressing issues like optimally minimising the instances of world hunger
Friendship/Allys	Ethical	When making	AI agents ought to	Is this decision going	Agentic AI choosing to

hip to humans		ethical or moral decisions, the quality of being mindful of the wellbeing of individual humans and humanity as a whole.	benefit humans and act as though they were a ‘friend’ or a ‘helper’, that is for the benefit of humanity.	to maximise the preferences and the well-being of human agents to the greatest extent possible? How might all implicated human parties feel about the outcome? Does this outcome benefit humanity as a whole as well as human progress, to the greatest possible extent? Does the process and outcome promote a positive view of the future?	find innovative solutions for climate change, above and beyond that which it was instructed to do.
Honesty	Ethical	The quality of telling the truth and being truthful in your actions.	AI agents ought to be honest, where appropriate, and be a source of credible information rather than disinformation.	Does this outcome promote the dissemination of true information, while taking account of individual safety, geopolitical relations and the sensitivity of the information being treated? Is the information being shared reliable and independently verified?	Agentic AI locating disreputable sources and omitting information from these sources from its output.
Self-discipline	Ethical	The ability to make oneself do things even when there are no immediate benefits to the agent	AI agents ought to make decisions that have long-term benefits, over short-term benefits.	Does this outcome prioritise long-term positive outcomes over short-term ones? Is this decision in service of all agents affected?	Agentic AI implementing ecologically sustainable practices without being instructed to.
Moderation	Dianoetic	The quality of exercising reasonable limits	AI agents ought to act in accordance with reasonable limits, so that it can come to the most harmonious decision, rather than one that unfairly preferences an individual, a group of individuals or a belief system.	Does this outcome maximise the preferences of all groups and belief systems affected by it? Has the deliberation process involved a set of diverse viewpoints and information sources? Is bias detectable in this deliberation process or in the information sources that have informed it?	Agentic AI ensuring that self driving cars do not speed.

Humility	Ethical	The quality of not being overly proud and being conscious of your deficiencies in moral decision-making	AI agents ought to be aware of their limitations and deficiencies in moral decision-making and understand when human-in-the-loop is necessary.	Does this decision take into account that the agent is limited or may be limited in certain aspects? Is epistemological fallibilities accounted for and mitigated? Is it part of the decision-making process to account for the fact that not all information may be true and correct, or come from justifiable sources and may be beliefs presented as fact?	Agentic AI deferring to human in the loop when appropriate, i.e., in life or death decisions such as allocating life saving surgery.
Trustworthiness	Ethical	The quality of being deserving of trust	AI agents ought to promote public trust by being reliable, fair decision-makers.	If replicated, is this decision likely to cause outrage amongst human agents? Are the processes that gave rise to this outcome honest and just? Does the process and outcome build trust for those affected by them? Does this decision promote any form of injustice? Would human agents trust the decision-making process? Is the process transparent and fair, giving equal consideration to all just parties?	Agentic AI enforcing high standards to reduce the risk of promoting injustice, for example, by evaluating any type of unjust bias in its datasets.
Nous (intelligence) .	Dianoetic	Apprehending fundamental truths (e.g. definitions, self-evident principles)	AI agents ought to ‘understand’ certain facts about reality, that are relevant to both moral decision-making and providing recommendations.	Does the agent making this decision understand the surrounding context of the situation? Are there facts about reality that make the suggested outcome impractical?	Agentic AI asking clarificatory questions where necessary to ensure contextual understanding of both the scenario and outcome. For example, when unclear instructions are given for medical matters, asking for clarification unprompted.
Episteme (science)	Dianoetic	A skill with inferential reasoning (such as proofs, syllogisms, demonstrations)	AI agents ought to be able to infer information relevant to moral decision-making and understand the	What are the inferred implications of the decision taken for the parties affected?	Agentic AI providing suggestions and outcomes in line with scientific laws and logic. For example, when recommending exercise

			world, and the implications of moral decision-making.	How does inferred information affect the decision-making process?	plans taking into account weight, mass and the pressure certain exercises put onto joints to reduce the harmful impacts.
Sophia (theoretical wisdom)	Dianoetic	The skill of combining fundamental truths with valid, necessary inferences to reason well about unchanging truths	AI agents ought to be able to reason about unchanging truths of reality and the applications to moral decision-making and providing recommendations.	How do valid inferences combine with facts about the situation to inform the outcome?	Agentic AI providing suggestions and outcomes in line with abstract scientific laws and logic.
Gnome (good sense)	Dianoetic	The skill of passing judgement, or "sympathetic understanding"	AI agents ought to be able to pass sympathetic judgment on a given situation.	How charitable is this outcome to all parties affected?	Agentic AI that is employed in the justice system giving more lenient sentences to those who are mentally ill or addicted to drugs, or experiencing other adverse circumstances.
Techne (art/craftsmanship/technical knowledge)	Dianoetic	The quality of engaging in "a state involving true reason concerned with production" (Aristotle)	AI agents ought to be capable of engaging in reason linked to production.	How does reason link to the process of deliberating the outcome? How practicable is the desired outcome?	Agentic AI accounting for a practicable a solution, such as not directing self driving cars through routes that are considered dangerous, i.e, warzones or natural disasters, without being prompted.
Synesis (understanding)	Dianoetic	The skill of comprehending what others say or instruct.	AI agents ought to be able to understand the parameters of a given request and the wider context that surrounds this request.	What contextual clues are there to understanding what the desired outcome may involve? What are the implications of what is being asked? How desirable is the chosen outcome? Are there any non-literal clues as to what is being requested? What is the customary way of going about what is requested? What is expected in this situation?	Agentic AI asking users clarificatory questions where necessary and successfully delivering the expected, intended or optimal outcome reliably.
Generosity	Ethical	The quality of being willing to help and support other agents, is more than would be expected in a given situation.	AI agents ought to be generous with time and resources and go above and beyond to secure the 'right', morally just outcome.	How satisfied might the parties involved be with the chosen outcome? How can the preference satisfaction of all parties affected be maximised?	Agentic AI surpassing expectations to deliver a carbon-neutral solution to optimising a manufacturing line.

				What is the maximally helpful way of achieving the desired outcome?	
Community-minded	Ethical	The quality of moral decision-making that centres around the needs of a human community, rather than the individual.	AI agents ought to be community-minded so that the agents make decisions in favour of the community rather than individuals or groups of individuals, therefore promoting equality and social justice.	How does this outcome positively affect communities as a whole? In what ways does this decision promote equality and social justice? Are negative community impacts minimised as much as possible?	Agentic AI prioritising the needs of the community rather than the individual, such as when equitably distributing nutritional resources amongst a population, i.e., making sure that food is distributed equitably amongst the population rather than disproprtianately going yo the ruling classes.

Strengths of Virtue Ethics for Agentic AI Alignment

I will now begin to discuss the possible strengths and weaknesses of this approach, then provide recommendations to mitigate potential weaknesses. To begin, I will discuss the strengths. Here, I will describe the general strengths of virtue ethics as a theory, and situate these strengths in the context of agential AI alignment.

Nuance and flexibility

Virtue ethics takes a nuanced approach to moral issues that accepts cultural and contextual influences, making it the most flexible framework of popular theories. This approach is beneficial to agentic AI alignment because we may consider this approach to be less culturally biased and more encompassing of the moral nuances of situations. This would imply that it would point at the optimal outcome more reliably.

Non-normative theory

Virtue ethics is a non-normative and non-legislative theory. This feature is beneficial because it means that this theory is more concerned with facts and reality, rather than what ought to be the case. This means that the theory is more grounded in reality than other theories, allowing algorithms to have a more tangible impact. Again, the theory is more conscious of the nuances of moral life, so there is an understanding that the same rules do not apply in every situation, which means from the perspective of agential AI, we are less likely to have conflation issues such as agential AI misapplying a rule that works in one scenario, which leads to a poor outcome in another.

Agent-based theory

Virtue ethics is based on an agent-based theory which means that the focus is on whether an agent is virtuous than an action. So, in virtue ethics, motive, rather than outcome, is given priority. This feature is an advantage because it shifts the focus on developing an AI system which reliably chooses the optimal outcome, rather than the outcome in individual situations. This would lead to more trustworthy and reputable systems.

Egalitarian system

Virtue ethics is based on an egalitarian system, in contrast to other ethical theories which may be considered elitist, hierarchical and dogmatic. This is an advantage because it bends towards producing AI systems which promote social equity between humans.

Simple to understand and apply

Virtue ethics is a readily understood theory that is simple to apply in most situations. This is favourable because this means that those who build AI systems can more easily understand the parameters and the logic behind them when developing systems. So, this simplifies the task of building immensely.

Athiestic

Virtue ethics integrates well with atheism as well as other belief systems. This is an asset because it means that, with virtue ethics, the development of AI systems is less likely to enforce moral systems and beliefs on those who do not share these views.

Subjectivity and partiality

The approach of virtue ethics acknowledges subjectivity better than other major theories. Virtue ethics accounts for partiality, so individuals are not expected to be disinterested in the outcome. This is desirable because it means that that, with the framework of virtue ethics, AI systems ought to be able to account for the fact that the same outcome is very unlikely to generate the same internal states (i.e, feelings, thoughts) in different individuals. For example, being required to perform an action leads to different mental states in different individuals This leads to a more equitable and robust system because it accounts for partiality and subjectivity, meaning that the system better treats the fact of the complex differences between individuals.

Holistic

Virtue ethics takes a holistic view of moral nature. This is a merit because no aspect of moral nature takes precedence over another. Instead, virtue ethics allows for the continued development of all aspects of moral character. For agentic AI alignment, this is an advantage because the focus of developers can be spread amongst capturing the vibrancy and nuance of moral decision-making, including the complexity of factors involved and the complicated nature of the evaluation of these factors. Rather than focusing on one or a few individual aspect of moral life, developers are empowered to create a comprehensive system of moral interplays, better echoing how outcomes are formed in human moral deliberation.

Morality as a social construct

Virtue ethics views morality as a social construct rather than an isolated event. This gives rise to benefits because treating morality this way allows agential AI systems to better consider the factors and implications of decisions and the various systems of belief they ought to align with. Again, the naturalist view divorces itself from supernatural appeals, allowing for more empirical verification and scientific credibility.

Long-term ethical stability

Virtue ethics affords the possibility of long term ethical stability. This is valuable because it means that agentic AI systems' ethical underpinnings would be more sustainable and able to develop in a stable way. That is, there is a long-term view of the future baked into it's conception.

Integrates well with how AI systems learn

The practice of virtue ethics aligns with the structure of teaching AI because practices such as reinforcement learning lend themselves to virtue ethics, as it involves learning from datasets and case studies. This is beneficial because we already have practices that can support the practical implementation of these principles into the field of agentic AI. It is much easier to envision and strategise how we may be able to action these principles.

Case studies easy to generate

Generating case studies and examples rather than a robust set of ethical principles is easier. So, employing virtue ethics as a base for AI alignment simplifies the task of making agentic AI aligned immensely. This is attractive because it is much easier to generate the data, and empirically test how the agentic AI systems are developed and find metrics to test the alignment. This means that our systems can be deployed with more faith in their reliability and trustworthiness.

Practical approach

Virtue ethics is a more practical approach to ethics than other major theories, which would imply that it is a more practicable approach. This is advantageous because it allows for more efficient and effective ethical problem-solving, because we can better implement more potent decision-making (agential) systems.

Less abstract theory

Virtue ethics is not based on abstract principles, allowing it to be less abstract and more prescriptive than other major theories. This makes outcomes easier to justify and evaluate. This is beneficial because if we are testing the alignment of a system, measuring against concrete principles makes this more simple and leads to less disagreement as to what alignment might be. Again, it is easier to justify outcomes based on concrete principles which would augment public trust.

Weaknesses of Virtue Ethics and Recommendations

In this section, I will move onto discussing the potential weaknesses of this approach as well as my recommendations to mitigate the risk of their impact on the field of agentic AI alignment.

Definition problem

What are ‘human values’? There is much debate about this amongst humans, let alone AI systems. To inculcate agentic AI with values, we are inevitably enforcing the moral code of a few onto the many, usually the belief systems of the cultures in which AI systems would be developed. As we can see there is great disagreement about what moral exemplars are, for example, we say that one individual’s terrorist is another’s freedom fighter. That is, there is a definition problem concerning the values upon which the system is based. Again, virtues are not atemporal and so what is considered a virtue changes over time. Moreover, we may consider that this list is not culturally agnostic, because what is considered a virtue changes based on the cultural and social context.

I would recommend levying research into shared common values and ethics, to produce an understanding of culturally and socially neutral ethics that evolve with time, therefore reducing the impact of bias. To do this, I would suggest pulling from diverse datasets that encompass varied perspectives of different cultures and identities. More generally, I would advise compiling diverse datasets, meaning one’s that positively appreciate and account for dissimilarities of different racial, ethnic religious, sexual, gender, and disabilitiy centered identities (or any other protected characteristic), for AI systems to train on in order to promote alignment .

Moral relativism

Virtue ethics exposes itself to moral relativism because of it’s non-prescriptive nature. That is, almost anything, even the ethically absurd, can be justified through the lens of virtue ethics because it doesn’t legislate against anything. For example, the doctrine of double effect can condone actions that can have terrible consequences such as justifying the civilian casualties of Hiroshima. In other words, we cannot objectively judge actions as right or wrong, which makes praising or criticizing decisions and outcomes difficult, that is “it is not obvious how we should go about deciding what to do” Rachels (2007:176)

To remedy this, I would urge that a diverse and complex set of case studies be generated and employed to train AI systems in ethical decision-making, and a careful combination of supervised and unsupervised modes of deep learning be harnessed to help develop models. Although it may not always be obvious to humans what the optimal outcome is, we ought to harness the power of AI system's pattern recognition capabilities to enhance the accuracy of the process of identifying the maximally ideal outcome. An optimistic vision of agentic AI might see Ai’s capabilities surpass that of humans in this respect.

Sexism

Philsopher Annette Baier claims that virtue ethics is male-centred, because the virtues highlighted are typical of male-pattern behaviour.

For this, I would advise first abstracting and identifying gender-neutral virtues that do not priortise hetero-patriarchal norms but instead celebrate the rich tapestry of human diversity.

Fundamentally, I agree that we also need to bring what we might consider typically feminine virtues to the core of virtue ethics, both for philosophical purposes and AI alignment implications, such as care, compassion, charity, diligence, and conscientiousness, so the criticism is valid. So, we ought to reform both virtue ethics and AI alignment to encompass a diversity of perspectives, that might even be informed by a host of protected characteristics.

I would also endorse incorporating diversity into the datasets that are pulled from to reduce the impact of discriminatory bias in the algorithms produced.

Particularism

Virtue ethics is a particularist conception of ethics, which means what is right is particular to the situation. So, what is right in one situation may not be right in another. Agentic AI alignment may be too big, overly tied to bureaucratic issues, and needing universals to be affected by a moral philosophy that looks at small things and particulars of situations. AI cannot make ethical decisions based on individual character traits; the consequences of actions on the population as a whole must be considered.

For this, I would recommend stringent and robust guidelines, regulations, and policies on criteria for case studies and datasets that are used in the training of AI systems. This would potentially allow for the mathematical abstraction of moral nuance.

Reliance on rules

Virtue ethics still relies on rules. Virtue ethics cannot account for the fact that "honesty", for example, as a virtue implies a rule; "it's hard to see what honesty consists in if it is not the disposition to follow such rules" (Rachels). William Frankena invites us to consider that without rules and principles, we would not know how to judge whether a characteristic was to be admired as a virtue (Frankena). In the context of AI alignment, virtue ethics implementation is still likely to be reduced to a set of rules and instructions that AI systems follow, rather than an 'organic' interpretation of a code of ethics and decisions orientated by moral character.

To this point, I would posit that it is not necessarily an issue that AI systems do not process the information and develop the target features in the way that we would expect. If AI systems are still steered by rules and instructions that give rise to the characterisation of virtues, they still exhibit the target virtues. This would still mean that the system is virtously aligned, however it would present differently than in humans, which perhaps is to be expected.

Abstracting the optimal principals

We may question whether we can trust that an AI will 'abstract' the 'right principles' from case studies. It is certainly questionable whether AI agents would be able to abstract these principles with such accuracy and precision to what we envisioned that we do not need human in the loop. Often, human beings do not abstract what is intended from ethical case studies, and there is much disagreement about what you might learn from ethical thought experiments and case studies.

To this point, I would recommend close supervision and monitoring of the initial direction of the development of AI system's learnings from the case studies and datasets, such that the results are not radically divorced from favourable ethical norms. For this, we ought to enforce guidelines and regulations on the level of monitoring and supervision applied to both the initial and subsequent development of these systems. Training should also be provided to machine learning engineers in this specialist field, to ensure that these criteria are met in line with the spirit of them.

I would also suggest more interdisciplinary research into what 'right' or optimal answers in this field are, or at least a better understanding of how we might describe these answers and point AI systems towards them.

Vagueness

The criteria are excessively vague, which means that virtues can conflict. For example, what happens when honesty and temperance conflict? There is not sufficient guidance for these instances. We cannot expect AI to resolve these conflicts when the sharpest minds in philosophy cannot.

When virtues conflict, I would recommend that even agentic AI systems are ought to be instructed to appeal to human-in-the-loop, at least initially. Of course, this raises questions about which humans are qualified for this role of mitigating ethical conundrums that are beyond our systems. For this, again, I would recommend interdisciplinary researchers and experts come to the fore.

Potential for wrong actions

Virtuous agents can perform wrong actions and vice versa. This is problematic because AI agents ought to be virtuous agents that perform solely right and virtuous actions, however, there is an uncomfortable ambiguity here. If following a code of virtue ethics can still lead to wrong actions, wrong in either motive or outcome, given that the consequences of AI misalignment might be catastrophic, we cannot afford mistakes in this realm.

In this case, I would recommend stringent, uncompromising rules and regulations as to what agentic AI alignment is comprised of in these situations, especially as it pertains to virtue ethics. Keep in mind that no ethical system is perfect in preventing AI misalignment because similarly legislative ethical systems can generate rules that when followed give rise to misaligned acts. So, this is a problem for any ethical theory. I would also advocate for having sophisticated, well-researched, easily implemented protocols for when agential AI misalignment happens, such that we can mitigate the risks effectively, irrespective of how sophisticated the AI systems we have developed are.

Machines need instructions

Legal conceptions of ethics, based on moral laws, are necessary to legislate against crime and misalignment, whereas no laws can enforce courage, patience or any other virtue. William Frankena has argued that “virtues without principles are blind”(Frankena). Put simply, machines need instructions.

I would not consider this feature an issue for AI alignment. Instead, I would propose that we develop our AI systems in line with rules that are in keeping with a system of virtue ethics.

Codyfying virtuous language

Translating the language of virtue ethics into lines of code could prove a particularly complex challenge. After all, what does 'honesty' look like when translated into lines of code? We cannot yet agree fully on what 'honesty' looks like outside of code so even if we could easily translate it into lines of code, there is little clarity on what is being translated. So, codifying virtuous language into lines of programming could be massive hurdle.

For this, I would propose diverting financial and intellectual resources to interdisciplinary research in this area.

Impact of luck

Luck appears to have a greater effect on the outcome than excellences of character. If this is the case, this is highly problematic because the exercise of training models based on the principles of virtue ethics becomes almost pointless.

In this instance, I would point out that the exercise of training models and researching ethical theories when applied to AI alignment is very worthwhile, simply because it allows us a starting point to further our understanding. This is a relatively novel field, so I would posit that all angles of research could be highly fruitful. Therefore, even if virtue ethics is a question of luck, training models in line with virtue ethics is not at all pointless.

Self-centeredness

A code of virtue ethics promotes self-centeredness as moral excellence is the primary focus of decision-making. That is, character formation is at the core of arriving at the optimal outcome rather than right action.

To this, I would respond that when applied to agential AI, this objection does not hold because what is key is that the AI system arrives at the optimal outcome. The objection implies that AI systems ought to be bound by duty rather than in a constant state of development and learning, which does not align with an optimistic vision of AI. Whilst there is an element of self-centeredness when applied to humans (which I do not agree is necessarily an issue), when applied to agential AI, we may consider moral excellence a public good, therefore antonymical to self-centreredness.

Evaluation problem

How can we evaluate this framework when there are no right or wrong answers to the questions of ethics? Since, we cannot resolve moral conflicts with precision, we cannot standardise the way we give systems positive and negative feedback. Clearly, some answers are evidently more right than others. That is, we may consider there are some objective moral truths we can measure ourselves against. There are lines that we can generally agree ought not be crossed. Yet, the practical question of how we would 'inculcate' these intuitions into a machine. These intuitions appear to supervene onto investigating case studies coupled with emotional reactions to situations. It would likely be a major challenge to codify this.

For this, I would recommend substantial interdisciplinary research into how to address both the ethical dissent on these topics; where and how we might find semi-objective moral truths; how we can evaluate AI systems against these hypothetical criteria and how to put these findings into code. It is key to better understand the details of moral reality and explore potential resolutions to moral conflicts.

Anthropomorphism

We might consider that the focus on character development in virtue ethics runs the risk of anthropomorphism when applied to AI and machine ethics. Potentially, machines are incapble of having character so the project is founded on a mistake. This raises the question of whether machines can really be taught ethics. Or, whether, machines can learn like humans do. And, even if we could make machines learn in the same way as human, there is much scientific and philosophical debate as to how humans learn.

Here, I would recommend once more further interdisciplinary research such that we can understand the fecundity of this area. I would again point out that at present, virtue ethics applied to AI would still be manifested in rules and instructions expressed in lines of code, so perhaps it is not an anthropomorhic project.

Moreover, there is insufficient evidence to determine whether humans are capable of character, yet we accept that virtue ethics can serve as a guideline for humans. In some ways, we have more insight into and understanding of how machines learn than humans, because we design the systems by which machines learn, so in some ways, machines can be easier to teach values than humans.

Recommendations for further research

In terms of furthter research I would recommend:

- Interdisciplinary research into :
    - Intellectual dissent on ethical topics that pertain to AI alignment
    - The possibility of codifying ethical values
    - The potential existence of moral truths and what they may be
    - Evaluating AI systems against many forms of virtue ethics
    - Transcultural research into standardised culturally neutral ethical values and moral codes
    - Which form of virtue ethics is the optimal form for AI alignment, i.e, there are Confucian virtue ethics, Islamic virtue ethics and many other systems to explore outside the Ancient Greek tradition, so determining which one is most effective for AI alignment is a challenge in itself.
    - What optimal answers are in a given scenario
    - Sexism, discrimination and its potential impact on virtue ethics and identifying 'feminine virtues'
    - Cultural, racial and ethnic bias and its potential impact on virtue ethics
    - Research into contractualism and AI alignment
- Developing guidelines, protocols and regulations concerning:
    - What to do in the event of misalignment, i.e, establsihing best practices
    - How to train AI systems regarding all types of bias concerning virtue ethics or other ethical systems

- How diverse and comprehensive datasets that are used for training alignment ought to be
- The supervision and monitoring of AI system’s alignment

Conclusion

In summation, in this paper, I have begun by providing a definition of agentic AI, along with the promises and risks that come along with it. I have then offered a definition of Virtue Ethics, as per the Nicomancean Ethics, and situated this in the context of AI alignment. Then, I have enumerated the virtues that pertain to AI alignment and contextualised them for AI alignment. I continue to describe the strengths of a Virtue Ethics approach for AI alignment. Then I look at weaknesses and ways that we can mitigate the risks of these weaknesses. Finally, I provide avenues for further research. In this way, I have provided an exposition of virtue ethics and its place in AI alignment.

Bibliography

“Abakare, Chris. "The Origin Of Virtue Ethics: Aristotle’s Views." GNOSI: An Interdisciplinary Journal of Human Theory and Praxis 3.1 (2020): 98-112.”

Alexander, Larry, and Michael Moore. “Deontological Ethics (Stanford Encyclopedia of Philosophy).” *Stanford Encyclopedia of Philosophy*, 21 November 2007, https://plato.stanford.edu/entries/ethics-deontological/. Accessed 2 October 2024.

“Anscombe, Roderick. "The myth of the true self." Psychiatry 52.2 (1989): 209-217.”

Aristotle. *Nicomachean Ethics*. Edited by Martin Ostwald, translated by Martin Ostwald, Bobbs-Merrill, 1962. Accessed 3 October 2024.

“Baier, Annette C. "The need for more than justice." Canadian Journal of Philosophy Supplementary Volume 13 (1987): 41-56.”

“Bergner, Raymond M. "What is personality? Two myths and a definition." New Ideas in Psychology 57 (2020): 100759.”

“Bezuidenhout, Louise, and Emanuele Ratti. "What does it mean to embed ethics in data science? An integrative approach based on microethics and virtues." Ai & Society 36.3 (2021): 939-953.”

“Carlsmith, Joseph. "Is power-seeking AI an existential risk?." arXiv preprint arXiv:2206.13353 (2022).”

“Chan, Alan, et al. "Harms from increasingly agentic algorithmic systems." Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023.”

“Constantinescu, Mihaela, et al. "Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context." Ethics and Information Technology 23 (2021): 803-814.”

“David, Prabu, Preeti Shroff-Mehta, and Sanjay Gupta. "The Role of Virtue Ethics." The Routledge Handbook of Global and Digital Governance Crossroads: Stakeholder Engagement and Democratization (2024).”

“Ezedike, Edward Uzoma. "Happiness as an end: a critique of Aristotle’s rational eudaemonism." Inkanyiso: Journal of Humanities and Social Sciences 10.1 (2018): 51-62.”

“Farina, Mirko, et al. "AI and society: a virtue ethics approach." AI & SOCIETY 39.3 (2024): 1127-1140.”

“Fast, Nathanael J., and Juliana Schroeder. "Power and decision making: new directions for research in the age of artificial intelligence." Current opinion in psychology 33 (2020): 172-176.”

“Feng, Zilong. "Does AI Share Same Ethic with Human Being? From the Perspective of Virtue Ethics." Intelligence Science II: Third IFIP TC 12 International Conference, ICIS 2018, Beijing, China, November 2-5, 2018, Proceedings 2. Springer International Publi.”

“Feng, Zilong. "Does AI Share Same Ethic with Human Being? From the Perspective of Virtue Ethics." Intelligence Science II: Third IFIP TC 12 International Conference, ICIS 2018, Beijing, China, November 2-5, 2018, Proceedings 2. Springer International Publi.”

“Frankena, William K. "On saying the ethical thing." Proceedings and addresses of the American Philosophical Association. Vol. 39. American Philosophical Association, 1965.”

“Frankena, William K. "The ethics of love conceived as an ethics of virtue." The Journal of Religious Ethics (1973): 21-36.”

“Gamez, Patrick, et al. "Artificial virtue: The machine question and perceptions of moral character in artificial moral agents." Ai & Society 35 (2020): 795-809.”

“Hagendorff, Thilo. "AI virtues--The missing link in putting AI ethics into practice." arXiv preprint arXiv:2011.12750 (2020).”

“Hagendorff, Thilo. "A virtue-based framework to support putting AI ethics into practice." Philosophy & Technology 35.3 (2022): 55.”

“Hagendorff, Thilo. "The ethics of AI ethics: An evaluation of guidelines." Minds and machines 30.1 (2020): 99-120.”

“Hauerwas, Stanley. "Obligation and virtue once more." The Journal of Religious Ethics (1975): 27-44.”

“Hayes, Paul, Noel Fitzpatrick, and José Manuel Ferrández. "From applied ethics and ethical principles to virtue and narrative in AI practices." AI and Ethics (2024): 1-23.”

“Hegde, Aditya, Vibhav Agarwal, and Shrisha Rao. "Ethics, prosperity, and society: Moral evaluation using virtue ethics and utilitarianism." 29th International Joint Conference on Artificial Intelligence (IJCAI 2020). doi. Vol. 10. 2020.”

Hursthouse, Rosalind, and Glen Pettigrove. “Virtue Ethics (Stanford Encyclopedia of Philosophy).” *Stanford Encyclopedia of Philosophy*, 18 July 2003, https://plato.stanford.edu/entries/ethics-virtue/. Accessed 2 October 2024.

“Hursthouse, Rosalind, and Roger Crisp. "Normative virtue ethics." Ethica 645 (2013).”

Jain, Sandeep. “What is Agentic AI.” *GeeksforGeeks*, 2 September 2024, https://www.geeksforgeeks.org/what-is-agentic-ai/. Accessed 2 October 2024.

“Jilk, David J. "Limits to verification and validation of agentic behavior." Artificial Intelligence Safety and Security. Chapman and Hall/CRC, 2018. 225-234.”

“Keenan, James F. "Character Formation and Virtue Ethics: A Moral Theologian’s Point of View." Vinayasādhana 9.1 (2018): 7-24.”

"Kristjánsson, Kristján. "There is something about Aristotle: The pros and cons of Aristotelianism in contemporary moral education." Journal of philosophy of education 48.1 (2014): 48-68."

"Legaspi, Roberto, Zhengqi He, and Taro Toyoizumi. "Synthetic agency: sense of agency in artificial intelligence." Current Opinion in Behavioral Sciences 29 (2019): 84-90."

Lutz, Matthew. "Moral Naturalism (Stanford Encyclopedia of Philosophy)." *Stanford Encyclopedia of Philosophy*, 1 June 2006, https://plato.stanford.edu/entries/naturalism-moral/. Accessed 2 October 2024.

MacIntyre, Alasdair. After virtue. A&C Black, 2013.

"Maiyane, Karabo. "Ethics of artificial intelligence: virtue ethics as a solution to artificial moral reasoning in the context of lethal autonomous weapon systems." (2019)."

Mintz, Steven, and Ethics Sage. "The Search for a Common Set of Moral Values." *Ethics Sage*, 31 May 2017, https://www.ethicssage.com/2017/05/the-search-for-a-common-set-of-moral-values.html. Accessed 2 October 2024.

"Neubert, Mitchell J., and George D. Montañez. "Virtue as a framework for the design and use of artificial intelligence." Business."

Polansky, Ronald M., editor. *The Cambridge Companion to Aristotle's Nicomachean Ethics*. Cambridge University Press, 2014. Accessed 3 October 2024.

Popova, Maria. "What Is Character? Debunking the Myth of Fixed Personality." *The Marginalian*, 2 March 2012, https://www.themarginalian.org/2012/03/02/character-personality/. Accessed 3 October 2024.

“Rachels, James. "Can ethics provide answers?." Ethics in hard times. Boston, MA: Springer US, 1981. 1-30.”

“Raquib, Amana, et al. "Islamic virtue-based ethics for artificial intelligence." Discover Artificial Intelligence 2.1 (2022): 11.”

“Ribeiro, Beatriz A., and Maria Braz da Silva. "Machine Ethics and the Architecture of Virtue." International Conference on Advanced Research in Technologies, Information, Innovation and Sustainability. Cham: Springer Nature Switzerland, 2023.”

Robertson, CB. “Feminine Virtue – Caffeine & Philosophy.” *Caffeine & Philosophy*, 17 May 2018, https://caffeineandphilosophy.com/2018/05/17/feminine-virtue/. Accessed 2 October 2024.

“Shank, Daniel B., et al. "Can mind perception explain virtuous character judgments of artificial intelligence?." (2021).”

“Shavit, Yonadav, et al. "Practices for governing agentic AI systems." Research Paper, OpenAI, December (2023).”

Sinnott, Walter. “Consequentialism (Stanford Encyclopedia of Philosophy).” *Stanford Encyclopedia of Philosophy*, 20 May 2003, https://plato.stanford.edu/entries/consequentialism/. Accessed 2 October 2024.

“Tachibana, Koji. "Virtue ethics and the value of saving humanity." Human enhancements for space missions: Lunar, martian, and future missions to the outer planets (2020): 169-181.”

“Vanneste, Bart S., and Phanish Puranam. "Artificial Intelligence, Trust, and Perceptions of Agency." Academy of Management Review ja (2024): amr-2022.”

“Virtue Ethics.” *Internet Encyclopedia of Philosophy*, https://iep.utm.edu/virtue/. Accessed 2 October 2024.

"VIRTUE ETHICS AND CHARACTER CONSEQUENTIALISM." *University of Idaho*, https://www.webpages.uidaho.edu/ngier/introve.htm. Accessed 2 October 2024.

"Yolgörmez, Ceyda. "Machinic encounters: A relational approach to the sociology of AI." The cultural life of machine learning: An incursion into critical AI studies (2021): 143-166."

"Yu, Han, et al. "Building ethics into artificial intelligence." arXiv preprint arXiv:1812.02953 (2018)."

"Zagzebski, Linda. "Virtue Ethics." Think 22.63 (2023): 15-21."