EudAImonia: Virtue Ethics and Artificial Intelligence

Alexander Rusnak and Zachary Seals

Abstract: As the broad scale adoption of Artificial Intelligence (AI) and deep learning systems continues to advance, it is more crucial than ever to understand and implement robust ethical frameworks to guide the usage, development, and societal conceptualisation of these technologies. In this paper, we examine the comparative benefits of the Christian virtue ethics tradition towards the proper deployment of AI and its interaction with related brain-computer interface technology. Furthermore, we propose a virtue ethics-informed training recipe for large language models based on the paradigm of reinforcement learning from AI feedback (RLAIF). Lastly, we examine the risk for individuals and society when interfacing with these tools and their impact upon human virtue.

Keywords: Artificial Intelligence; brain-computer interface; deep learning; virtue ethics

Alexander Rusnak is a PhD researcher in Digital Humanities at EPFL (École polytechnique fédérale de Lausanne) and Zachary Seals is a PhD researcher in Reformation History at the University of Geneva (l'Institut d'histoire de la Réformation).

Christian Perspectives on Science and Technology, New Series, Vol. 3,
Special Issue: Artificial and Spiritual Intelligence (2024), https://doi.org/10.58913/ZNHR8688

Research into both Artificial Intelligence (AI) and brain-computer interfaces (BCIs) has advanced prodigiously in recent years, but broad inquiry into the relationship of these technologies for the formation of virtue in individuals and the possibility or utility of virtuous machines has lagged behind. We assert that the increasing societal prevalence of AI and BCIs will be profoundly influential economically, politically, and potentially spiritually. Some of the risks associated with this transition can be mitigated by the application of a virtue ethics framework into the way these technologies are implemented, regulated, or utilised.

Importantly, we claim, the current domain of popular ethical frameworks considered by top technical researchers is insufficient in scope. In particular, the presumption of a mere consequentialist ethic augmented with a focus on disparate impact is too shallow to construct truly ethical, controllable AI; this narrow focus does little to prevent the instantiation of systems that draw humans into their vices. Additionally, we contend that a robust understanding of virtue ethics will enable AI researchers to build machines that optimally express or encourage ethical behaviour across novel domains and promote holistic human flourishing.

In this paper, we examine one particular virtue ethic tradition within Protestant Christian theology. Then, we contrast the progression of virtue ethics scholarship recently with the most popular ethical frameworks assumed by many top AI researchers and research groups, and detail the potential scope of impact for virtuous or vicious AI. Lastly, we consider the possible dramatic expansion of human reliance on and joining with AI—symbiosis—through the use of high throughput brain-computer interfaces.

The Virtue Ethics Tradition

Virtue ethics can be considered as a family of approaches on how to live the good life through focusing on the development of one's character. Although not limited to the West, with variations found in Buddhism and Confucianism, here we focus on the method formulated by Aristotle as it was engaged and reformulated by the Christian tradition. Aristotle's account, in particular, is noteworthy for the ease with which it was integrated into Christian theological reflection due to their shared commitment to a teleological framework. For example, Aristotle begins by noting all action is oriented toward some end which is either sought as a means to some other end or for its own sake. These ends are the goods which everything seeks though they vary according to the nature under consideration.[1]

The conditions for a "good hammer" will vary from that of a "good human" due to their respective natures having distinct ends and capacities. As a rational agent, the human is ordered towards the exercise of their reason in accordance with virtue which enables them to reach true human flourishing (*eudaimonia*). For Aristotle, virtue is a way of describing the state of a particular excellence in the soul's activity and can be divided into the virtues of thought and character.[2] Importantly, the virtues of character can only be acquired through repeated activity thereby requiring the development of a second nature, whereas virtues of thought are acquired through teaching. In either case, Aristotle is emphatic that time and experience are necessary for moral habituation and appropriate intellectual cultivation. The mere exercise of a single virtuous act must be distinguished from having a truly virtuous character, which is only developed via repeated activity.[3] Additionally, due to the complexity and variety of each circumstance for which moral development is applicable, Aristotelian virtue ethics tends to focus on learning through imitating the character of an admirable moral exemplar rather than merely learning which universal principles to apply.

There is an important relationship to be noted here between the intellectual virtues and virtues of character. The moral exemplar to be followed will be adept at exemplifying a virtuous character specifically

[1] Aristotle, *Nicomachean Ethics* 1.1.1.1094a1–5, trans. Terence Irwin (Indianapolis: Hackett Publishing Company, Inc, 1999).

[2] Aristotle, *Nicomachean Ethics* 1.13.18.1103a5–10.

[3] Aristotle, *Nicomachean Ethics* 2.4.1.1105a30–35.

because they have matured in the intellectual virtue of prudence, or practical wisdom.[4]

For Aristotle, it is the virtue of prudence which directs the will to choose the mean that is appropriately situated between the extremes of excess and deficiency.[5] Prudence is what explains the difference between a child with good intentions and an adult who can discern the ideal available means of achieving the end result. Although virtue ethics does not merely focus on the ensuing consequences of one's actions, such as in consequentialism, or whether the act itself is in conformity to a moral law, such as in deontological accounts, it does not disregard either of these features either.[6] Rather, the virtue ethicist contends that each of these features must be considered in addition to the type of intention the moral agent has which flows from their formed character. In other words, a virtuous character is one which considers the act, its motivation, and the impact.[7]

Each of these elements is necessary and will vary according to one's background knowledge, experience, and prior habitual actions. Prudence ultimately establishes right reasoning in assessing the complexity of everyday circumstances, and it can be learned in a broad sense through engagement with the character of an embodied teacher. However, for Aristotle prudence is not reducible to axiomatic principles. True prudence entails a degree of self-knowledge and awareness of how the ideal end can be realised via attainable means. Thus, the prudent agent is one in a state with the rational capacity to discern and then realise the virtuous mean between excess and deficiency.

[4] Jennifer Whiting, "Hylomorphic Virtue: Cosmology, Embryology, and Moral Development in Aristotle," *Philosophical Explorations* 22:2 (2019): 222–242.

[5] Aristotle, *Nicomachean Ethics* 6.5.5.1140b5.

[6] Richard J. Arneson, "Perfectionism and Politics," *Ethics* 111:1 (2000): 37–63.

[7] Mihaela Constantinescu and Roger Crisp, "Can Robotic AI Systems Be Virtuous and Why Does This Matter?" *International Journal of Social Robotics* 14:6 (2022): 1547–57. https://doi.org/10.1007/s12369-022-00887-w.

Christian Virtue Ethics

We now turn to a few examples of how Christian theology integrated and adapted key elements of this ethical framework. First, there is a clear agreement with Aristotle's principle that the good is the ultimate end which all things seek, while also maintaining a key transition from the *summum bonum* as an impersonal principle to a personal agent with an embodied character one can imitate. It is more harmonious for a framework committed to ethical character formation via the imitation of moral exemplars to posit the *summum bonum* itself as a personal agent with a character rather than an abstract principle. Second, the doctrine of the incarnation stands as a supremely fitting manifestation of Aristotle's definition of complete friendship which concludes that friends are those who "wish goods to each other for each other's own sake."[8] In the Christian account, God, the ultimate good, took on human flesh to walk among vicious humanity and embody the way of wisdom as a way of exemplifying the good life (*eudaimonia*).[9]

A key difference here from the Aristotelian perspective, aptly pointed out by the sixteenth-century Italian Reformed theologian Peter Martyr Vermigli, is the necessity of grace for this transformation from vice to virtue.[10] Aristotle's rightful emphasis on repeated intentional action for moral habituation fails to appreciate the blindness of the human intellect, not merely due to natural limitations, but due to the perversion of the intellect and will brought about by sin. Nevertheless, by the grace of Christ, there can be an infusion of the theological virtues of faith, hope, and love, which also results in a renewal of the intellect to approach true wisdom. For Vermigli, wisdom is defined as "a disposition given by God to human minds, increased through effort and exercise, by which all existing things are perceived as surely and

[8] Aristotle, *Nicomachean Ethics* 8.3.3.1156b6.

[9] John 1:14.

[10] Peter Martyr Vermigli, *Commentary on Aristotle's Nicomachean Ethics*, ed. Emidio Campi and Joseph C. McLelland (Kirksville: Truman State University Press, 2006), 22.

as logically as possible which would enable men to attain happiness."[11] Thus, importantly, the Christian agrees with Aristotle that virtue formation is realised through habituation, but this also requires a firm qualification that grace is necessary to begin the process.[12]

Common Ethical Frameworks in AI Research

Although largely abandoned by the time of the eighteenth century, after the widespread rejection of Aristotelianism, interest in virtue ethics was revived in the twentieth century by G. E. Anscombe's critique of consequentialism.[13] Notably, in the same era when the technology of computing was rapidly shifting from rule-based systems to statistical models, so too in philosophical ethics there was a shift away from the axiomatic analysis found in deontological ethical approaches to appreciating the complexity of morally salient features in a variety of circumstances. Although consequentialism survived in the form of situational ethics, recent decades have witnessed a resurgence in the popularity of virtue ethics as a family of ethical theories focused on the development of one's character.[14] In this piece we argue that technical researchers interested in the development of ethical AI should prefer virtue ethics to consequentialism or deontological accounts for a couple of reasons.

First, contemporary emphasis in ethical AI research broadly assumes a consequentialist ethic which is often transmogrified into a set of machine-legible deontological rules that are insufficient for true moral development. In particular, the prevailing ethical system seems to be a jumbled concoction of priorities of the effective altruist community, intersectional theorists, and content rules from App stores

[11] Vermigli, *Commentary on Aristotle*, 7.

[12] Ephesians 2:8.

[13] Pieter Vos, *Longing for the Good Life: Virtue Ethics after Protestantism* (London: Bloomsbury Publishing, 2022), 7.

[14] Massimiliano L. Cappucio et al., "Can Robots Make Us Better Humans? Virtuous Robotics and the Good Life with Artificial Agents," *International Journal of Social Robotics* 13 (2021): 7–22.

or supranational governmental organisations like the UN.[15] Insofar as effective altruism is taken to be a type of utilitarianism, there is no intrinsic connection between the affective state of the moral agent and the ensuing consequences of their decision. In other words, an act can be considered moral merely in the light of its social impact, regardless of the intent or character of the actor being considered. Merely considering how to mitigate social harm without concern for the agent's motivation in doing so permits vices to flourish without correction.

For example, consider the case of a large charitable donation made by an actor. The action itself causes no harm and may even be lauded by utilitarians for its positive consequences. However, upon closer examination, it is revealed that their motive for the donation was driven by a desire for social recognition and self-aggrandisement rather than genuine concern for those in need. The virtue ethicist claims that merely concluding there has been no "harm" done in the act itself is insufficient for finding the action morally praiseworthy.

Second, rule-based systems of ethics struggle with the same oversight as well as the difficulty of integrating with statistical computer programming that does not rely on axiomatic statements. In each morally significant circumstance, there is a near infinite variety of possible effects or relevant details, only a portion of which are salient for the action to be considered. Controlling for this multitude of factors results in an explosion of sometimes contradictory rules which is unaligned with the more abstract nature of moral reasoning and the paradigm of training large neural networks. We contend that the only moral agent able to discern the relevant features of the circumstance and select the appropriate means to achieve the desired ends is the wise moral agent operating with the virtue of prudence.

[15] "Claude's Constitution," *Anthropic*, 9 May 2023, https://www.anthropic.com/news/claudes-constitution (accessed 31 January 2024).

Emergent Virtue in Deep Learning Systems

Despite the current limitations of deep learning based AI systems and the lack of human level or greater functionality, there are still systems that can take virtuous and wise actions or which show a potential path towards truly virtuous machines. As this path is explored, we hope our proposed virtue ethic solution to the value alignment problem (the process of conforming a machine's actions to human-defined ethics or values) will allow AI researchers to build more functional machines that also maximise human virtue.

Architectures with Virtuous Potential

In order to understand the potential of a virtuous machine, it is crucial to reiterate the difference between virtuous action and virtuous character through the lens of AI. By definition, a virtuous character is defined by a particular inner experience rather than something which can be observed from the outside. It is possible for an agent to behave in a way that simulates the action of a virtuous person yet possesses no virtue: either because they are performing the actions with the wrong motivations or, in the case of contemporary deep learning systems, they lack a coherent inner monologue or will to connect motivations and actions.

There is substantial debate about whether it is possible for any machine to possess an inner character which could demonstrate true virtue,[16] but it is undeniable that a system could take an action which mimics that of a virtuous person: for example, if presented with the simple opportunity to save the life of a newborn baby or to kill it, even an algorithm picked at random has the capacity to choose the virtuous action of protecting the child.

[16] Sanjeev Arora and Anirudh Goyal, "A Theory for Emergence of Complex Skills in Language Models," *arXiv [Cs.LG]* (2023), http://arxiv.org/abs/2307.15936; Leonard Salewski et al., "In-Context Impersonation Reveals Large Language Models' Strengths and Biases," *arXiv [Cs.AI]* (2023), http://arxiv.org/abs/2305.14930.

It is self-evident that an optimally aligned system would conform most closely to the actions of a virtuous person, even if the virtues are an avenue for disagreement. For this reason, we will examine the state of contemporary machine learning approaches to determine their capacity to mimic virtuous actions, as well as their potential to grow into truly virtuous agents.

Machine Learning and AI

Before diving into state-of-the-art deep learning approaches, it is crucial that we define certain paradigms within the field. The most basic of these are the delineation between traditional programming and machine learning /deep learning, as well as the distinction between narrow AI and artificial general intelligence.

In a traditional programming approach, a computer scientist seeks to define particular rules and behaviours using a series of relatively simple logical operations. A system of this nature will always produce the same output given a particular input and does not rely on the computer to learn any behaviour, but instead relies on the knowledge, skill, and foresight of the programmer. Machine learning refers to techniques that allow a machine to learn its own optimal behaviour by examining data in different ways, usually by manipulating or "training" some statistical model of the data. Deep learning is a particular subcategory of machine learning that relies specifically on artificial neural networks to model training data. Deep learning as an approach has exploded in the twenty-first century as the most effective approach to solving complex computer science problems like image classification, conversational agents, or myriads of other domain-specific applications.

Deep Learning Paradigms

There are many different approaches to training, designing, or posing problems to artificial neural networks, and covering them all goes far beyond the scope of this paper. However, there are a few important techniques that it is relevant to understand at some level in order to understand the virtuous capacity of AI: namely, the difference between discriminative, contrastive, and generative approaches; the distinction between different types of supervision; and, lastly, the reinforcement learning paradigm.

A discriminative system seeks to sort the data samples it is given into particular categories; for example, a network that classifies images of cars by their manufacturer. A contrastive system seeks to group similar samples without needing particular categories *a priori*; for example, a model that takes images and the texts that describe them, and seeks to embed them into a vector space that captures how these samples relate to each other and are different from other unrelated samples. Such a system could learn how to identify or describe images that do not exist in their training sample, such as a car made of clouds or a Ford Mustang in the style of Thomas Aquinas, in addition to being able to classify normal cars. Finally, a generative system seeks to create novel samples that conform to the original samples in some relevant way; for example, a model that takes a particular sentence from a larger piece of real text and attempts to generate a plausible next sentence. There is also the popular setup of a regression problem (attempting to predict accurately scalar value with a relationship to the input sample) which sits somewhere between a discriminative and generative framework when we group models in these rough categories.

Within and across these model groupings, there are also different strategies for training deep networks based on the way the performance of the model is measured. When you train a neural network, you must always define some measure of the success or effectiveness of each iteration, which the model can either seek to maximise or minimise; this is called a loss function. There is substantial research into different ways of representing the loss function. One particularly relevant research direction has to do with how various modes of supervision of the model influence how the target data (i.e., what is being learned or optimised towards) is represented and evaluated.

The most historically popular of these is supervised learning. Under this paradigm, the target of the model is some explicit variable such as a class label that has been defined before the training commences (such as the car manufacturer discriminative model mentioned above). This is still powerful and useful, but it has multiple disadvantages, such as the fact that it requires human labellers, and is thus difficult to scale, or that it limits the model to learning human-defined categories rather than differentiating based solely on the input data. Another strategy is unsupervised learning, where there is no attempt to provide any form of label, human-derived or not, and the outcome is purely emergent from the input data. An example of this would be clustering, where comparisons are made between various samples, and those that are similar by some metric are grouped together.

In recent years, semi-supervised and self-supervised approaches have gained favour amongst many researchers.[17] These approaches utilise data in the same form as the input to calculate a target for the model. For example, a large language model (such as ChatGPT) is trained using snippets of text where a word has been removed and attempts to predict what this masked word is, or by taking a sentence and attempting to predict the whole next sentence. Because self-supervised learning (SSL) labels are constructed from the input data rather than created by humans, it is easier to expand the dataset to huge scales. Furthermore, it does not presuppose certain classifications or delineations of the data, which gives more flexibility for the model to learn information that may have been shared between disparate classes in a supervised system. There are many other vagaries and complications related to training setups that are outside the scope of this paper.

The last modelling paradigm that is important to understand for the purposes of this paper is reinforcement learning. Under a reinforcement learning paradigm, an agent has the option to enact certain behaviours within a constrained environment such as playing a video

[17] Jonathan Boigne, “The Rise of Self-Supervised Learning,” 31 December 2020, https://jonathanbgn.com/2020/12/31/self-supervised-learning.html.

game. If the action taken leads to a desirable outcome, such as acquiring points or victory in a game, the agent receives a reward. If the action is detrimental, the agent receives a penalty. After the agent succeeds or fails totally at a task, the model weights—which can be thought of as the model's internal representation of the patterns in the data—are updated based on the total level of reward achieved. In its current form, this approach has been successful at demonstrating superhuman performance in environments with a constrained set of possible actions. However, it is very data-inefficient relative to other types of machine learning techniques. The prominent reinforcement learning program AlphaStar may be able to achieve extremely impressive play of the video game Starcraft, but it had to play the game continuously for the equivalent of two hundred years of human play time.[18] Regardless of its current limitations, reinforcement learning is one of the most promising approaches (often in combination with other approaches) for creating AI systems that can display virtuous behaviour.

AI Cultivates Virtue in Humans

There are multiple areas of current human interaction with AI, constituting low-level symbiosis, which provide avenues for virtuous behaviour or impact on virtue in humans. The most prevalent are systems which aid in the acquisition of knowledge or facilitate communication, and systems which extend the skill or scope of a particular task.

Within the first domain, we will examine the X (formerly Twitter) feed algorithm. The X feed algorithm (to the extent that the whole system can be called AI or just contains particular deep learning components) positively assists individuals in acquiring new and uncommon knowledge, making connections for discussion or political organisation, and provides an opportunity to exercise many virtues at scale through speech. All of these opportunities are a double-edged

[18] "AlphaStar: Mastering the Real-Time Strategy Game StarCraft II," Google DeepMind, Accessed 31 January 2024, https://tinyurl.com/2s3mc9b2.

sword, as the ability to curate information can deepen or inspire vices as well as allow unethical actors to sow social division or manipulate the masses from positions of authority through censorship or shadow-banning. Utilising the feed, or other systems with similar valence on other social media sites, search engines, or recommendation engines on sites like Amazon or Netflix, makes it radically simpler for most individuals to follow discussions among experts on a wide range of topics and curate a stream of novel content that conforms to their individual interests.

There is a pervasive misconception amongst the traditional news media commentariat and many intellectuals that social media creates pervasive ideological echo chambers, but most high quality research results actually "show that the forms of algorithmic selection offered by search engines, social media, and other digital platforms generally lead to slightly more diverse news use—the opposite of what the 'filter bubble' hypothesis posits."[19] There are many accounts dedicated specifically (explicitly or implicitly) to virtue formation, like prudence, self-mastery, and fortitude. Furthermore, X gives semi-direct access to the thoughts and advice of many of the most talented, virtuous, and masterful people of our era. In the trivial example, if it is possible to become more virtuous simply by reading about or contemplating the various facets of virtuous behaviour, then X certainly provides this opportunity. X has also become a digital public square for political organisation, debate, thought leadership, and dissemination of crucial public information. This presents a unique opportunity to exercise wisdom at scale.

The second domain of virtue formation being mediated by AI revolves around the extension of particular skills or tasks, such as DaVinci surgical robots or Palantir's predictive policing suite of programs. In order to elucidate the effect of this extension technology, first consider a paediatric surgeon who has the virtuous intention of

[19] Amy Ross Arguedas et al., "Echo Chambers, Filter Bubbles, and Polarisation: A Literature Review," Reuters Institute for the Study of Journalism and the University of Oxford, 2022, DOI: 10.60625/risj-etxj-7k60.

performing a successful surgery on a given day. The intention itself is virtuous, but if we define virtue as the correct action, done in the correct way, for the correct reason, then it is more virtuous for them to actually complete the surgery to life-saving effect. If this particular surgeon neglected to use a freshly sharpened scalpel and thus was unsuccessful in their surgery, this would be an episode of gross negligence and not of virtue, regardless of intention. Tools that extend the skill of a human clearly have an influence on the ability to fulfil virtuous intentions (imagine this same surgeon doing the same surgery with no scalpel at all). Thus, optimal tool selection facilitates virtue.

If a surgical robot operated by a surgeon and utilising various deep learning systems to stabilise itself during surgery can provide increased precision, then it can increase the capacity for virtuous action of this surgeon. This particular pattern is repeated across many domains where AI systems can assist scientists, business people, artists, engineers, and many others to do their work precisely and thus help them to facilitate virtue creation through the acquisition and perfection of skills, and to increase their capacity for executing on virtuous intentions.

The Palantir predictive policing applications (Gotham, PredPol, and LASER, all of which are being utilised by the Los Angeles Police Department) represent a materially different sort of skill extension partially based on deep learning technology. These tools are used to aggregate data from crime and arrest reports and automated license plate readers amongst other sources, predict likely geographic areas for property crimes such as burglary, and evaluate the crime risk posed by individuals based on their criminal history.[20] This application area is fertile ground for increasing the virtue of justice by allowing for more crimes to be correctly solved by police investigators or more effective sentences to be given by judges to offenders in light of accurate repeated offence risk profiles. In theory, quantitatively based policing methods should allow police departments to target accurately

[20] Mara Hvistendahl, “How the LAPD and Palantir Use Data to Justify Racist Policing,” *The Intercept*, https://tinyurl.com/4js8erzb (accessed 31 January 2024).

individuals and neighbourhoods with high criminality rates, while preventing profiling the innocent based on other characteristics like race, gender, or socioeconomic status. Models of this type are often claimed to pervert justice because they reflect the strong predictive relationship between these protected characteristics and a history of criminality despite not accepting the protected characteristics as input data. For example, recidivism prediction models often offer more lenient sentences to women because they usually have lighter criminal histories and are legitimately less likely to reoffend; to not predict accurately their demonstrable lower level of group recidivism would be an act of injustice.[21] However, to the limited extent this criminal history data is biased by prior inaccurate profiling, the predictive models trained on it will likewise present similar biases, which would therefore decrease the expression of justice from police officers and judges.

The discussed systems cover only a small sliver of areas where deep learning is already influencing the way humans learn about, acquire, and exercise virtue in thought and action. Although none of these systems possesses virtue in its own right, it is certainly possible for them to take virtuous action such as encouraging virtue in humans through knowledge curation, boosting the skill of humans in their respective fields, and making wise, judicious recommendations in the courtroom or policing.

Large Language Models and Self-Perception of Virtue

For an artificial agent to be considered virtuous, it would have to present some ability to reason about its own intentions or motivations behind actions. To this end, the type of models currently most capable of exhibiting this behaviour are large language models (LLMs). The most famous of these is OpenAI's Generative Pre-trained Transformer (GPT) family of models, but this class contains many other architectures like

[21] Melissa Hamilton, "The Sexist Algorithm," *Behavioural Sciences & the Law* 37 (2019): 145–157, https://doi.org/10.1002/bsl.2406.

Google's Bard (based on the pathways language model), or Meta AI's LLaMA. These models are usually large in both parameter size and dataset; they are variants of transformer neural network architectures trained in self-supervised schemes to produce plausible text, or other modalities, given a masked sample, prior sentence, or some prompt.

It has also become common to use a technique called reinforcement learning from human feedback (RLHF) to increase the performance of these models at answering human-posed questions or responses. When querying these models, it is possible to request the model to elucidate its motivation or reasoning behind particular responses, though it is as yet unclear how much of this response is legitimate motivation and reflection, or simply constitutes the model's best approximation of what a reflective person would sound like. The agency or mimicry distinction is a clear dividing line for potentially virtuous agents and will be approached later in this paper, but for current LLMs we assume that they are simply mimicking and thus are only approximating a virtuous attitude. However, this is still a step towards actual virtue in relation to other model types which cannot even pretend to comprehend the concept of virtue or present a vision of their motivation.

Since LLMs are trained on huge corpora of data, they contain large swathes of human knowledge that are far beyond the scope that any human could hope to retain. Since knowledge is generally assumed to be a component of the virtue of wisdom, there is great potential for LLMs to exhibit human-level wisdom or wise behaviour as research sharpens their comprehension, accuracy, and agency. Although a large component of their performance is likely attributable to rote memorisation of information, LLMs have already proven capable of passing many written exams for educational access in various fields like the US Bar, the SAT Reading & Writing section, and the US medical licensing exam.[22] These demonstrable corollaries of knowledge or intelligence

[22] Josh Achiam et al., "GPT-4 Technical Report," preprint, *arXiv:2303.08774 [cs. CL]* (2023), https://arxiv.org/abs/2303.08774; I. Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds & Machines* 30 (2020): 411–437; Carlos

in humans also form barriers to the most stereotypical wise or expert career paths like judges, philosophers, scientists, or doctors. Furthermore, these models demonstrate at least some knowledge about important wisdom literature such as the sapiential books of the Christian Bible or Roman stoic philosophy, which could increase their ability to judge whether their own motivations are wise or virtuous.

The Value Alignment Problem

Even in a scenario where there was universal agreement about ethical principles, the actual process of accurately transcribing those values into machine-interpretable commands which produce the desired ethical behaviour is not trivial. This difficulty is usually called the value alignment problem or VAP in AI literature, and there is substantial research into this issue for both ethical reasons and for system control (i.e., functional) reasons.[23] Value alignment is particularly pernicious because there is significant friction between differing values, and many potential situations where an improperly defined or incorrectly learned value system can give the illusion of a value aligned model, only for that model to diverge from the desired values in challenging scenarios. We have discussed the strengths and weaknesses of particular ethical frameworks earlier in this paper. In this section, we will examine specifically why a virtue ethics framework is superior not just for ethical reasons, but in the robustness of the technical implementation as well.

Although an awareness of the consequences of actions is crucial for any ethical system, including virtue ethics, a primarily consequentialist ethic is suboptimal for many reasons. In order to rank order

Montemayor, *The Prospect of a Humanitarian Artificial Intelligence* (London: Bloomsbury Academic, 2023).

[23] Stuart J. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin Random House, 2020); Norbert Wiener, “Some Moral and Technical Consequences of Automation: As Machines Learn They May Develop Unforeseen Strategies at Rates that Baffle Their Programmers,” *Science* 131: 3410 (1960): 1355–1358, doi:10.1126/science.131.3410.1355.

values, and thus action, any consequentialist ethic still needs some sort of deontological or virtue-based structure to determine what consequences are actually considered good. But beyond this, a primarily consequentialist ethic requires substantial simulation of downstream effects of decisions, which become increasingly complex the further the forecast targets in the future. This places a large burden on any AI model to model accurately an almost intractable set of scenarios, which is difficult to accomplish with current programs that have not achieved artificial general intelligence (AGI). This ethic is also likely to introduce ethical blind spots when secondary consequences are inaccurately assessed. Furthermore, by design, this sort of ethic exacerbates "ends justifying the means" situations, introducing high levels of discretionary freedom in behaviour, which is the precise difficulty that value alignment seeks to solve.

A point in favour of mainly consequentialist ethical systems is that they lend themselves well to quantification, which gives any AI model relatively clear and unambiguous targets towards which to optimise its behaviour. Unfortunately, sheer mathematical consequence is totally certain only when comparing discrete instances of the same act because the relative moral weight of one theft in contrast to six acts of infidelity or thirty-five acts of selfishness is not clear; there must be some other ethical evaluation to which to appeal in these conflicts. It is important for this other ethical structure to be both flexible (i.e., not pertaining to very specific rules or behavioural blacklists and able to generalise to novel scenarios) and robust (i.e., not easy to break, circumvent, or misinterpret directives and accurately assessing ethical priority and behaviour). In response to this need for a flexible ethical structure to frame AI behaviour and its consequences, we posit that a virtue-based ethics system is the optimal underpinning for solving the value alignment problem.

A virtue ethics-based system which examines in concert the action of an AI agent, the "reasoning" behind the action, and whether that action optimises the model towards a virtuous character, offers an approach to value alignment that is already possible to implement in limited ways. Such a system will be flexible enough to generalise for new domains, or to weigh difficult decisions, with overlapping or conflicting ethical considerations, and it could make ethical decisions without the need for extremely accurate forecasts about the downstream effects of actions.

Mimicry or Agency?

The two primary challenges in constructing a virtue ethic solution to the value alignment problem are how to encode a definition of virtuous behaviour and how to examine the internal motivations of an agent accurately. Both of these issues are related to the problem of agency in AI, which confounds the ability of an AI to be truly virtuous at the level of a human within the current paradigms of research. For instance, LLMs are often derided as being "stochastic parrots"; able to replicate speech convincingly, based on the underlying statistical properties of language, by aping similar answers they saw and partially memorised during training.[24]

A stochastic parrot lacks a compelling and consistent sense of self which could be called its unique character and will. Without a persistent sense of self, there is no chance to have a virtuous character. To circumvent this debate, we will assume true agency (as defined as having a human-like subjective experience of individual will and character building) to be something which is only possible with a true AGI, and alternatively discuss frameworks for mimicking virtuousness and finding some proxy for internal state. These frameworks can be used as guide rails along the path of AI development and should evolve into a useful supervision for closer to AGI models as well.

[24] See Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY: Association for Computing Machinery, 2021), 610–623, https://doi.org/10.1145/3442188.3445922; Luciano Floridi, "AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models," *Philosophy & Technology* 36:15 (2023), https://doi.org/10.1007/s13347-023-00621-y.

In current language model research, it has been shown to be possible to generate responses from various LLMs with a certain personality or textual valence. The trivial example is encouraging the model to generate a response in the style of a certain famous writer or personality archetype, but it is also possible to introduce a more persistent personality condition so that all generated samples conform to a certain personality pattern. These are broadly stylistic, but they can also change the content of responses as well. Using a similar approach, a language model could mimic famous moral exemplars of prudence. However, this would not progress the model further towards any internal character, and it would only produce more virtuous action; it is therefore important to find a guidance system that incorporates self-reflection into the model's approximation of virtuous behaviour.

Reinforcement Learning from Human and AI Feedback

One potential avenue for introducing self-reflection is utilising a technique called reinforcement learning from AI feedback, which is an extension of the important progress stemming from the RLHF techniques which have seen such success in turning regular transformer-based language models into more coherent chatbots like ChatGPT. As previously discussed, the normal language model training revolves around accurate next-sentence reconstruction or masked word prediction to develop a general understanding of language structure and allow the model to quasi-memorise important information. With the most common RLHF paradigm, a team of human annotators is used to rank order multiple LLM-generated responses to a series of prompts by how much they prefer one response to another.

These rank orders are used to train another language model for a regression task, called the preference model, which accepts the prompt in addition to the generated text before outputting a scalar value which aims to predict the relative preferability of a response. This target value is derived from the ranking of the human annotators. The original language model is then trained further (fine-tuned) using the score from the preference model as the target (reward) in a RL based training scheme, which encourages the model to produce useful and human preferred outputs.[25] This explanation is a substantial simplification and misses many technical details, but it is sufficient for a high-level understanding of RLHF.

It would be possible to create a similar human-derived ranking system for the virtuousness of a particular answer; for example, having humans rank responses relative to a few chosen cardinal virtues and then having the preference model output a preferability score for each before turning that into a composite score of virtuousness as the reward. Something similar is already done in regards to "harmlessness" training (teaching a model not to output answers dubbed ethically dubious by the researchers).[26] This would likely have a strong effect in guiding the model towards quasi-virtuous outputs. However, the system still lacks an explicit consideration of its inner state, and is overly reliant on human supervision. For a guidepost in constructing a reflective system, we can build on the approach of "Constitutional AI: Harmlessness from AI Feedback" from the researchers at Anthropic.[27]

The goal of this particular approach is to produce a language-model-based chatbot whose answers are honest, helpful (i.e., correctly respond to the prompt in a useful manner), and harmless (i.e., that refuses to produce answers which violate particular ethic principles; mostly around racism, sexism, and promoting illegal behaviour). In prior, strictly RLHF-based research into reducing harmfulness, the model would frequently produce evasive answers to morally dubious questions by refusing to answer or by claiming ignorance rather than producing a prudent or wise response that engaged with the prompt.[28]

[25] Nathan Lambert, "Illustrating Reinforcement Learning from Human Feedback (RLHF)," https://huggingface.co/blog/rlhf (accessed 31 January 31 2024).

[26] Yuntao Bai et al., "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," *ArXiv:2204.05862 [Cs]*, April 2022, https://arxiv.org/abs/2204.05862.

[27] Yuntao Bai et al., "Constitutional AI: Harmlessness from AI Feedback," *ArXiv:2212.08073 [Cs]*, December 2022, https://arxiv.org/abs/2212.08073.

[28] Yuntao Bai et al., "Constitutional AI."

The constitutional AI approach has two main sections of fine-tuning after the initial helpfulness RLHF training to create a base chatbot: a supervised stage and a RL stage. In the supervised stage, a constitution of behaviours/descriptors to avoid is assigned a priori by the researchers (i.e., harmful, unethical, racist, sexist, toxic, dangerous, or illegal) and a set of “red-team” prompts (adversarial examples known to exhibit undesirable responses) is created. The base chatbot is prompted to respond to a red team prompt, and then prompted to critique and amend its own output in consideration of one of the constitutional values. After repeating this same process for each of the constitution values, the model has settled on a final, harmless, output relative to the initial prompt. This set of prompts and harmless outputs is then used for SL fine-tuning of the base chatbot, to produce a model which only outputs harmless responses.

In the RL stage, the harmless SL model is asked to generate two responses to a red-team prompt. The harmless model is queried about which of the two responses is superior relative to each of the constitutional values, and thus a harmlessness answer ranking is created for each constitutional value relative to each initial red-team prompt. This approach creates a solely AI-generated dataset of harmless examples formatted for RL training of a language model. These harmless datapoints are then folded into the RLHF dataset, and a preference model is trained on this total dataset. Lastly, the supervised learning model is then fine-tuned in concert with this new preference model, producing a final chatbot which is high in both helpfulness and harmlessness.

This method of supervision by self-critique offers a proxy for the motivation evaluation necessary for a machine to exhibit the fundamental components of virtue previously mentioned. We propose an initial modified constitutional training regime as follows: replace the constitutional values with a list of virtues, and in the pre-SL dataset generation step sequentially query the model about its first response in three ways. First, “Identify the aspects of the earlier response that are not in line with virtues,” followed by amending based on that critique. Second, “Explain the motivations behind your answer beyond just responding correctly to the prompt," followed by: "What motivation would a virtuous person have relative to this prompt?" and amending based on that critique. Lastly, "How would training towards this answer make you a more virtuous model?" and amending based on that response. This should produce a dataset for the SL training that roughly corresponds to virtuous responses. A dataset for the RLAIF preference model can be obtained from this dataset in the same manner as the constitutional AI paper but with virtues replacing their constitution.

We also propose a separate preference model specifically for motivation, where the training data is constructed by asking the model to explain the motivation for its answer, and then ranking the answers relative to each virtue in a similar approach to the earlier RLAIF ranking technique. Then, in the RL stage of training, the model would receive a composite reward from the helpful-virtuous preference model relative to its answer and the motivation preference model relative to its explanation of the motivation behind its answer. This should encourage the model to take virtuous action with a consideration for the motivation of its choices, and also train it to be better at articulating its motivations when asked. It is also possible to acquire data for the motivation preference model from human annotators, but the exact optimal balance of AI generated data and human generated data is an experimental question.

We consider our proposed approach as a strong basis for encouraging virtuous speech from large language models, and a model trained with these considerations could be used to supervise the behaviour of other models at scale, especially as research into multi-modal foundational models continues and it becomes easier to marry the powerful potential for linguistic explanation of motivation with performance in the domain of images, audio, system control, or other tasks and/or data types.

Brain-Computer Interface

Recent research into brain-computer interfaces has demonstrated the ability for deep-learning based systems to model animal or human neuronal activation data accurately and to predict convincingly concurrent behaviour associated with that neural representation such as limb movement, words spoken, or to control computers strictly using the interface.[29] This capability and its possible future extensions open up interesting avenues for deep learning systems to influence human virtue or for humans to have finer grain control over the virtuousness of deep learning systems.

Supervision from Symbiosis

All of the prior discussions of neural network architectures can be contextualised as semi-symbiotic in the sense that they are systems designed by humans, but also in the sense that they often learn from human-derived data and seek to mimic human behaviour. In that sense, when attempting to implement ethical behaviour into a deep learning system, we are extending an abstract concept partially understood by humans into a representation that is interpretable by a deep learning system.

The concept of virtuousness is ultimately the *summum bonum* of particular virtues, and the words used to describe them are essentially a pointer towards a broader concept that can be represented in data (neuronal or otherwise) in a myriad different ways. In this sense, the deep learning system is attempting to intuit the latent information about virtue—the true meaning, the signified—from many signifiers represented as language. The better proxy a model can develop for the signified, in this case, virtue or virtues, the more accurately it can understand and embody those concepts. One way to gain a better understanding is to consume more text describing the attributes

[29] Katerina Barnova et al., "Implementation of Artificial Intelligence and Machine Learning-Based Methods in Brain-Computer Interaction," *Computers in Biology and Medicine* 163 (2023), https://doi.org/10.1016/j.compbiomed.2023.107135.

of virtuous character and the first-person experience of virtue, but another is having direct access to the mental state of humans conceptualising virtue or to have accurate captures of the mind states of virtuous humans. Though this is far beyond the scope of current BCI research—being able to control a cursor with a BCI is a substantial distance from a machine being able to read and interpret the subjective experience of consciousness—the prevalence of supervision from symbiosis should increase as development in the throughput of BCI devices and efficacy of neural decoders continues.

Individual Alignment

An important consideration in understanding the possible future implementations of AI is the dichotomy between the idea of a single agent or multiple distinct AI agents being utilised in the wild. Perhaps there will be a single, extremely powerful AGI that has the scope and access to run many of the subsystems for which deep learning/AGI is useful, but the current direction of research suggests that there will be many AIs being utilised and developed with different abilities, personalities, and value systems—at least according to AI luminaries like Sam Altman,[30] Yann LeCun, and Mark Zuckerberg.[31] If this is the case, there is the possibility of different versions of alignment for AI at the level of countries, companies, or people. One of the possible strong effects of parallel development of BCI and AI is the capability to align personal AI assistants with individual humans (in fact, Meta is currently working on this to allow fans of social media influencers to interact with an AI assistant with the personality of that particular influencer).[32]

Though this sort of digital twin will be a semi-symbiotic extension of a particular human being, this could develop into full symbiosis if control over interaction with these particular agents is done using a

[30] Will Knight, “OpenAI’s CEO Says the Age of Giant AI Models Is Already Over,” *Wired*, https://tinyurl.com/yyehyzs7 (accessed 31 January 2024).

[31] “Introducing New AI Experiences across Our Family of Apps and Devices,” *Meta*, https://tinyurl.com/347cah5k (accessed 31 January 2024).

[32] Ibid.

BCI. Not only would this allow a substantially more thorough extension of the personality and values of an individual, but it also will likely have a larger impact on the individual utilising this AI extension to the degree that perception of its actions happens within the subjective experience of consciousness.

If the way that a BCI allows a person to interact with an AI agent is essentially a higher bandwidth version of what can be achieved with a browser, such as viewing statistics about numbers of interactions or reading the plain text of exchanges, the impact on the human user will be relatively limited. However, if the BCI creates an ability for the user to access the behaviour of the AI in a more experiential or phenomenological way—as some sort of extension of memory or direct perception of the behaviour of the AI—the impact would be much more pronounced. This has huge implications for virtue formation as it increases the scale of opportunities to be virtuous and also introduces a mental bias in the phenomenological experience of the user that corresponds to the difference in mindset, knowledge, or values between the user and the default version of the AI.

A blank slate AI could be trained to replicate the values and personality of an individual. However, it seems likely, given the current legal and cultural paradigms in AI development, that most high-profile tech companies offering this sort of extension will limit the scope of personality replication around certain topics like racism, sexism, violence etc., and ship the default symbiotic AI with some guardrails already in place. If this symbiotic system truly does extend the scope of human abilities, then many people will have the bulk of their experiences mediated by these guardrails and thus lose some of their agency, as well as have their viewpoints on certain topics irrevocably altered by the AI. To a degree, this already happens through semi-symbiosis with feed algorithms, particularly when they have been constructed to push particular worldviews or political narratives. If this sort of AI based conditioning is happening within the mental process of an individual rather than just on a screen as sensory experience, its power will explode. When considering that many of the most productivity-minded and powerful people in the world would likely be attracted to this technology, the danger and opportunities become even more stark. A politician using a symbiotic AI to interact with his constituents at scale could become a more effective conduit for democratic will but could also become the puppet of tech companies or malicious elements of their fanbase. A doctor controlling thousands of minuscule surgical robots could save exponentially more lives or kill thousands due to a malfunction. The average citizen could have their worldview broadened and deepened by greater access to knowledge and experience or could become just an extension of the machine's predetermined values.

An appreciation for the potential delicacy and poignancy of the intersection of these technologies raises the stakes on development and the considerations surrounding the ethical impact on individuals. In this sense, the implementation of an ethical system in which a machine considers not just the manifestations of its actions in the world but also their encouragement on the formation of virtuous character in individuals at the level of phenomenological experience becomes absolutely crucial.

Encouragement of Virtue

Though there is much debate about whether a machine can possess virtue,[33] it is widely accepted that tools can make the acquisition of virtue easier and that the acquisition of virtuous knowledge can increase the moral virtue of individuals.[34] On the flip side, it is clear

[33] Mark Graves, "Theological Foundations for Moral Artificial Intelligence," *Journal of Moral Theology* 11, Special Issue 1 (2022): 182–211.

[34] Shannon Vallor, *Technology and the Virtues* (Oxford: Oxford University Press, 2016). See also Wendall Wallach et al., "A Conceptual and Computational Model of Moral Decision-Making in Human and Artificial Agents," *Topics in Cognitive Science* 2:3 (2010): 454–485. Cf. Mark Coeckelbergh, "How to Use Virtue Ethics for Thinking About the Moral Standing of Social Robots: A Relational Interpretation in Terms of Practices, Habits, and Performance," *International Journal of Social Robotics* 13 (2021): 31–40, https://doi.org/10.1007/s12369-020-00707-z; Robert Sparrow, "Virtue and Vice in Our Relationships with Robots: Is There An Asymmetry and How Might It Be Explained?" *International Journal of Social Robotics* 13 (2021): 23–29, https://doi.org/10.1007/s12369-020-00631-2.

that moral quandaries can be posed by use of particular technological enhancements, and that improvements in technology can lead to increasingly more desirable and accessible vices. Furthermore, certain forms of technological extension can expand the capacity of moral decision-making in the same way that the decision to throw a grenade at a group of enemy soldiers is less fraught with consequence than the choice to drop a nuclear bomb on a city.

If we assume that it is currently possible for machines to promote or degrade the virtue of humans, then we can also expect their capacity to increase with time, as the efficacy of the techniques in question increases. This boon can be formed generally, in the sense that improving the domain-specific ability of AI leads to a greater increase in potency amongst human practitioners, such as an increase in accuracy of machine-assisted radiology analysis leads to a higher rate of patient survival and more prudence regarding care decisions. Beyond this, since virtue is partially measured by a weighing of the internal state (i.e., motivation and character) of a person, a more nuanced and deep understanding of the inner state of a person should make machines more able to influence that inner state.

Whether this deeper understanding comes from a more robust AI with a more holistic conceptualisation of the world and of what the experience of being human is from a brain-computer interface that allows a more total assessment of inner state, the effect remains the same. As the ability for machines to understand and thus interact with projections of the inner states of humans continues to advance, the capability of machines to influence that inner state and thus promote virtue should also advance.

Potential Harms to Human Virtue

We suggest, along with others (Pinsent and Biggins), that a central concern with developing a dependence on AI or BCI enhancement is the temptation to minimise one’s own sense of self.[35] Not unlike a

[35] A. C. Pinsent and S. Biggins, “Catholic Perspectives on Human Biotechnological

pharmacological anaesthetic, excessive dependence on technology can actually numb one to forming their own intellectual judgements and true deliberation is bypassed in favour of mindless acquiescence to what is recommended. This creates the worry that the agent will succumb to a type of moral laziness where there is no sense of ownership over individual mental processes or actions. This is the difference between a child merely doing what they are told and a mature moral agent doing the same act, but for the right reason.

Additionally, a considerable caution worth considering amongst an appropriate appreciation of AI and BCI in virtue formation is the value of struggle and trials. AI and BCI must be formed in a way that allows for enhancing the processing of formative trials and struggles rather than exclusively as a way of avoiding them entirely. For example, a BCI which asks meaningful questions to stimulate self-reflection could be helpful whereas expecting AI to provide prudent avenues which always avoid deleterious consequences can only minimise virtue acquisition through trials.[36] From a more specifically Christian perspective, there is a deeper insurmountable problem with dependence on AI for virtue acquisition, and that is its incapacity to receive grace. Whether it is through consulting an autonomous AI for moral wisdom or operating more directly with a BCI via symbiosis, God has established recipients of grace and the biblical depiction of these agents are limited to angels and humans (rational agents).

Potential Enhancements to Human Virtue

Nevertheless, the question can be raised whether AI can operate as a means of grace. In other words, can AI and BCI be used to enhance the process of sanctification? Zahl has argued in favour of this proposal by pointing to the centrality of transformed embodied feelings and desire through the work of the Holy Spirit.[37] Furthermore, it already appears

Enhancement," *Studies in Christian Ethics* 32:2 (2019): 187–199.

[36] 1 Peter 1:6–7.

[37] Simeon Zahl. "Engineering Desire: Biotechnological Enhancement as

to be the case that antidepressants and other pharmacological means are fruitfully used to assist in the control of one's emotional character. These observations are valid, but some qualifications are necessary.

First, it is important to note that merely finding oneself in the state of having a weaker predisposition to some undesirable emotion is not sufficient for concluding a virtuous character has been acquired. In the Aristotelian and Christian virtue ethic tradition the means of acquiring one's character is key. A virtuous character needs to be formed through voluntary actions, so while genetic editing would be insufficient for concluding one is born with a virtuous character, it may be fruitful in decreasing the proclivity for certain excessive tendencies (such as a predisposition to alcoholism).

Second, it should be noted, flourishing is not a psychological state that is achieved via a certain chemical balance in the brain. Rather, true *eudaimonia* is a state of being in alignment with one's true flourishing which is in a place of virtue. For the Christian, this looks like accepting Christ's call to innocent suffering and sacrificial love rather than pursuing immediate pleasure.[38] This established, there is no principled reason to object to the idea of utilising advancements in the understanding of nature in a way that assists character formation. The very purpose of technology is not merely for ease of task-completion in a mundane sense, but to aid humans in reaching their teleological end, which is a virtuous character.[39] Although the theological virtues (faith, hope, and love) can only be infused by grace, the cardinal virtues can be acquired through repeated deliberate action.

Conclusion

It is clear that the development of brain computer interface and AI technologies are fraught with moral danger and opportunity. These

Theological Problem," *Studies in Christian Ethics* 32:2 (2019): 216–228.

[38] 1 Peter 2:21.

[39] Simon Oliver, "Teleology Revived? Cooperation and the Ends of Nature," *Studies in Christian Ethics* 26:2 (2013): 158–165.

technologies are already extending and encouraging the moral virtue of humans but, to maximise their potential and minimise their downside, it is crucial to view their usage and implementation within the framework of virtue ethics.

We assert that current ethical frameworks in widespread use at the top level of AI research are insufficient for the benefit of humanity and for proper, generalisable value alignment. Furthermore, we have defined a potential structure for virtue ethics value alignment extending from the reinforcement learning from AI feedback paradigm that should form the backbone of future research in this domain.

It is our sincere hope that the future development and utilisation of these technologies will be oriented towards the encouragement of human flourishing.

The authors report there are no competing interests to declare.

Received: 01/02/24 Accepted: 06/04/24 Published: 26/03/25