Living Well with AI: Virtue, Education, and Artificial Intelligence

Nicholas Smith
Darby Vickers

Follow this and additional works at: https://digital.sandiego.edu/philosophy_facpub

Part of the Philosophy Commons
Abstract
Artificial intelligence technologies have become a ubiquitous part of human life. This prompts us to ask, ‘how should we live well with artificial intelligence?’ Currently, the most prominent candidate answers to this question are principlist. According to these approaches, if you teach people some finite set of principles or convince them to adopt the right rules, people will be able to live and act well with artificial intelligence, even in an evolving and opaque moral world. We find the dominant principlist approaches to be ill-suited to providing forward-looking moral guidance regarding living well with artificial intelligence. We analyze some of the proposed principles to show that they oscillate between being too vague and too specific. We also argue that such rules are unlikely to be flexible enough to adapt to rapidly changing circumstances. By contrast, we argue for an Aristotelian virtue ethics approach to artificial intelligence ethics. Aristotelian virtue ethics provides a concrete and actionable guidance that is also flexible; thus, it is uniquely well placed to deal with the forward-looking and rapidly changing landscape of life with artificial intelligence. However, virtue ethics is agent-based rather than action-based. Using virtue ethics as a basis for living well with artificial intelligence requires ensuring that at least some virtuous agents also possess the relevant scientific and technical expertise. Since virtue ethics does not prescribe a set of rules, it requires exemplars who can serve as a model for those learning to be virtuous. Cultivating virtue is challenging, especially in the absence of moral sages. Despite this difficulty, we think the best option is to attempt what virtue ethics requires, even though no system of training can guarantee the production of virtuous agents. We end with two alternative visions – one from each of the two authors – about the practicality of such an approach.

Keywords
Aristotle, Virtue ethics, Artificial Intelligence, Wellbeing, Higher education

Disciplines
Philosophy

Notes
Nicholas Smith is an assistant professor at Alabama Agricultural and Mechanical University. His primary research interests are the intersection of virtue ethics and artificial intelligence, and the epistemology of religion. Darby Vickers is an assistant professor of philosophy at University of San Diego. Her research focuses on questions at the intersection of ethics, epistemology, and philosophy of education. She draws much of the inspiration for her work on contemporary issues from ancient Greek philosophy, particularly Plato and Aristotle.
Living Well with AI: Virtue, Education, and Artificial Intelligence

Nicholas Smith
Darby Vickers

Abstract
Artificial intelligence technologies [AI] have become a ubiquitous part of human life. This prompts us to ask, “how should we live well with AI?” Currently, the most prominent candidate answers to this question are principlist. According to these approaches, if you teach people some finite set of principles or convince them to adopt the right rules, people will be able to live and act well with AI, even in an evolving and opaque moral world. We find the dominant principlist approaches to be ill-suited to providing forward-looking moral guidance regarding living well with AI. We analyze some of the proposed principles to show that they oscillate between being too vague and too specific. We also argue that such rules are unlikely to be flexible enough to adapt to rapidly changing circumstances. By contrast, we argue for an Aristotelian virtue ethics approach to AI ethics. Aristotelian virtue ethics provides a concrete and actionable guidance that is also flexible; thus, it is uniquely well-placed to deal with the forward-looking and rapidly changing landscape of life with AI. However, virtue ethics is agent-based rather than action-based. Using virtue ethics as a basis for living well with AI requires ensuring that at least some virtuous agents also possess the relevant scientific and technical expertise. Since virtue ethics does not prescribe a set of rules, it requires exemplars who can serve as a model for those learning to be virtuous. Cultivating virtue is challenging, especially in the absence of moral sages. Despite this difficulty, we think the best option is to attempt what virtue ethics requires, even though no system of training can guarantee the production of virtuous agents. We end with two alternative visions— one from each of the two authors— about the practicality of such an approach.

Keywords
Aristotle, Virtue ethics, Artificial Intelligence, Wellbeing, Higher education

Alabama Agricultural and Mechanical University, Huntsville, USA
University of San Diego, San Diego, USA

Corresponding author(s):
Nicholas Smith, Social Sciences, Alabama Agricultural and Mechanical University, 4900 Meridian Street N, Huntsville, 35811, USA. email: nicholas.smith1@aamu.edu 4900 Meridian Street N
Introduction

In 2014, Stanford ran a panel debate entitled “Does teaching ethics do any good?” In the last two decades, universities have increasingly dedicated themselves to ethics—particularly, applied ethics—curricula (Darwish, 2015). Public attention to and concern about the rapidity of scientific and technological progress has motivated a growing emphasis on applied ethics, and particularly a desire for applied ethics courses in higher education (Darwish, 2015). However, a 2013 study in ethics in engineering showed that moral education did not improve either moral judgment or knowledge of responsible conduct in research (May & Luth, 2013). The ineffectiveness of ethics teaching is concerning, especially if courses in applied ethics are one conduit through which we expect the next generation to learn how we might live well with the artificial intelligences that increasingly inhabit our world. In particular, we need effective and flexible ethical training that can prepare future generations for living in a world in which novel ethical situations crop up with novel technologies and their applications and for designing AI systems that are likely to benefit rather than to harm.

AI tools are increasingly present in our lives. In this paper, we consider the ways in which we live with the various autonomous tools, software, applications, etc., that are not directly programmed but rather modify themselves based on data during a training process, which we call “AI”. Discussions of how AI will transform the world abound, both in traditional media and online, particularly since the release of ChatGPT-3.5. At the heart of these discussions is the question of what it takes to live well with AI and how AI can be used to improve our lives without damaging the things we cherish. Given these questions, we must consider how to design, build, implement, govern, and regulate AI technologies. We must consider the normative and applied moral concepts needed to live well with AI, and how we might train future generations to ensure that they can engage with and deploy these concepts in their lives.

University ethics courses are one venue for training the next generation to act well. Increasingly, universities are requiring ethics courses—usually applied ethics—for those who might one day become the creators of novel AI. One dominant approach in applied ethics curricula is the principlist approach. An alternative approach is a consequentialist approach. A consequentialist approach weighs the consequences of an action to determine the best outcome. This is an approach advocated by some computer scientists and electrical engineers—especially in areas such as algorithmic fairness—because it is straightforwardly quantifiable (e.g. Card & Smith, 2020; Chen et al. 2023). Anecdotally, we have often heard from computer scientists and engineers that they prefer consequentialist methods because they are easy to implement. We think the issue is not so straightforward. Consequentialism is a system that can evaluate specific policies or systems of policies based on their consequences or aggregated consequences, but is centered on specific actions rather than a holistic approach to how to live. For that reason, it is an insufficient basis for training students to live well with AI. Moreover, it is difficult to predict the consequences of novel situations. Mathematical models of fairness demonstrate additional problems, e.g. an algorithm may produce fair results according to one mathematical model, but unfair results according to another. Developing a deep critique of consequentialist approaches would require an entire paper of its own, which is something we hope to do in future. While consequentialist ethics have some defenders, principlist approaches are
that we develop, follow, and instill in people norms and principles that will appropriately govern
the way we live with AI. Many argue that if we simply find the right rules, follow them
appropriately, and ensure they are passed on, we will live well with AI. This seems particularly
problematic in the case of AI; rules may become quickly obsolete and tend to be reactive rather
than anticipatory. In other words, in a rule-based system, humans are likely to be chasing after
the right way to live, rather than providing a dynamic, living rubric as to what it means to live
well with new technology. We think a more fruitful approach for training the next generation to
live and act well with AI is Aristotelian virtue ethics.

We argue that Aristotelian virtue ethics provides all of the theoretical tools needed to
guide our moral decision making in developing and implementing AI. Aristotelian virtue ethics
is a useful guide to thinking about ethical action when the future is opaque, because it is a highly
flexible system that is responsive to the ethically salient features of each situation. The reason
that it is flexible is that virtue ethics is agent-based rather than action-based. Virtue ethics does
not prescribe a set of rules, it requires exemplars who can serve as a model for those learning to
be virtuous. Virtue is similar to a skill, which is acquired through guidance from more
experienced moral actors, practice, and feedback. The virtuous person has a skillset which allows
them to determine the ethically salient features of a situation, identify the appropriate virtue
required, and be motivated to take appropriate action.

However, implementing virtue ethical ideals is easier said than done. It is unclear whether or
how soon virtue ethics can provide actionable moral guidance with respect to the development
and implementation of artificial intelligence technologies. Virtue ethics requires rigorous ethical
training for agents, and ethical exemplars who can provide illustrations of what it looks like to
live well with AI. It also requires moral exemplars who have expertise in the technologies
themselves, sufficient to help them correctly identify the ethically salient features of those
technologies. There are few, if any, people who undergo such training and it is unclear if there
are any relevant ethical exemplars for us to follow.2 In short, we argue the Aristotelian virtue
ethics approach would provide the best basis for living well with AI, despite being difficult to
implement.

We begin this paper by considering virtue ethics, briefly. We show how it provides a
better rubric for living well with AI than principlist or consequentialist approaches. Our view is
that, despite the importance of a virtue ethics approach to living well with AI, doing this is
incredibly difficult. In the final section of the paper, each author provides an answer to the
question of whether and how such a system might be implemented. Nicholas Smith argues that
there is, in theory, little hope for implementing this virtue ethics approach. Darby Vickers argues,
based on her experience in teaching virtue ethics, that there is some hope for educating students
to live well with AI, even if it is extremely challenging.

---

2 While Aristotle presupposes that such exemplars exist—including figures like Socrates, Plato’s Socrates argues
that there are no experts in virtue (e.g. Apology, esp.19e-20c, 21b-23b, 24d-26a; Meno, esp. 89e-100b).
Aristotle’s Virtue Ethics, Briefly

In this section, we offer an account of Aristotle’s virtue ethics. Aristotle argues that humans all ultimately aim at *eudaimonia,* the highest human good; “*eudaimonia*” is the exercise of the unique function or activity of humans (*NE* 1.7, 1098a12-20, esp. 16-17). He identifies this unique human function as rational activity in accord with virtues of character (*NE* 1.7, 1098a12-20, esp. 16-17). Virtues of character range from virtues still widely acknowledged, like “courage”, “generosity”, and “justice” to the more obscure “*megalopsychia*” (commonly rendered in English as “magnanimity”) and “*megaloprepia*” (commonly rendered in English as “magnificence”). A virtue, in the Aristotelian sense, is a deeply rooted character trait which is stable across a variety of contexts and over the course of one’s life.

We acquire virtues by a process Aristotle calls “habituation”. This primarily involves the imitation of others more morally advanced than oneself, responding to novel moral situations, being corrected by one’s betters, and eventually coming to “see” what a certain situation calls for (*NE* 2.1-2.4 1103a14-1105b18.). For example, the brave person is one who, when confronted with a grave danger, sees whether, how, and why the circumstance calls for the danger to be confronted or avoided. The generous person is one who sees whether, how, and why money or other resources ought to be given. *Mutatis mutandis* for the other virtues. This is often condensed by saying, for example, that the generous person gives money “to the right person, in the right amount, at the right time, for the right reason, and in the right way.” (*NE* 2.9 1109a28-29) Overall, what habituation trains someone to do is to “see” the salient ethical features of a situation and react accordingly. This not only involves a trained way of understanding typical situations, but also a way of seeing that allows one to do the same thing for novel situations.

Habituation is a process by which a person becomes virtuous through a combination of instruction, practice, and feedback. For example, a child becomes generous by learning to share resources with others appropriately. At first, children are told or forced to share in appropriate circumstances, such as sharing a toy with a sibling after playing with it for an extended period of time. Then, when children share spontaneously, adults around them praise them for sharing.

---

3 We hew closely to Aristotle’s model, as opposed to choosing among its many neo-Aristotelian children. This is primarily because we believe that Aristotle’s model of virtue ethics provides the most coherent model of ethics pedagogy. It is outside of the scope of this paper to defend this view among alternative virtue ethics views or to defend that this view is right while all others are wrong. Rather, we aim to show that this view provides a coherent theoretical framework for living with AI that has distinct advantages over principlist approaches. We thank an anonymous reviewer at *Theory & Research in Education* for encouraging us to clarify our position with respect to the various Aristotelian models of virtue ethics.

4 We use the Greek word, “*eudaimonia*” instead of a rough English equivalent, like “flourishing” or “wellbeing” to emphasize that “*eudaimonia*” usually connotes an objective state of the soul that indicates that the person lives well over a lifetime.

5 Citations from Aristotle will use the more intuitive “*NE*” to indicate the *Nicomachean Ethics,* rather than the more usual “*EN*”. All citations are from the Oxford Classical Text of the *Nicomachean Ethics.* All translations are Vickers’s.

6 Or, second highest, depending on what you make of Aristotle’s argument in book 10 that concludes that the best life is a life focused on exercising, primarily, of the intellectual virtues (*NE* 1177a12-1179a33). We feel addressing the relationship between *eudaimonia* and the theoretical life is outside of the scope of this paper; our paper will focus on the character virtues. We thank an anonymous reviewer from *Theory and Research in Education* for encouraging us to address this point.
Once children learn to associate sharing with praise, they begin to share of their own volition, motivated by the pleasure that the praise induces and begin to share more frequently. As they do so, they receive feedback from parents, mentors, and peers about the circumstances in which it is appropriate to share. For example, sharing a pencil with a friend who lacks one is good, while sharing a prescription medicine with someone is inadvisable. As children mature, they realize that sharing—under appropriate conditions—is important in and of itself because they learn the importance of having resources for doing what one wants and undertaking good actions. This realization motivates them not only to share more, but also to think carefully about the most appropriate ways to share. Once this realization occurs, then agents share more spontaneously. As their character transforms, sharing becomes easier and more automatic. The final stage in the process is when these agents get pleasure from simply the act of being generous and their emotions and desires are calibrated such that a combination of their good instincts and judgment lead them to act in accordance with virtue.

Of course, habituation is not so linear as the above narrative makes it seem. For example, children spontaneously share with others, even before they are told to share and before praise reinforces sharing. In addition, some children may be more inclined to share than others, either because of natural temperament or growing up in an environment that is either resource-rich or heavily interdependent. Mutatis mutandis for the other virtues. The purpose of the simplified illustration is to demonstrate that, in order to learn to be virtuous, one must undergo scaffolded, guided practice of virtuous actions.

During habituation, the actions an agent undertakes serve to shape the character of that agent. However, the agent undergoing habituation does not do virtuous actions in the same way as someone who is already virtuous. While anyone can undertake a virtuous action, only the virtuous person does so while fulfilling three additional conditions. The conditions are (NE 1105a26-32): (1) knowing the action is virtuous, (2) understanding why the action is virtuous and being motivated to do it for that reason, and (3) doing it from a firm and unchanging disposition. In other words, if we have virtues, we cannot lose them easily and, in general, we will be motivated appropriately and act appropriately. Aristotle describes character development somewhat like athletic training. The more one undertakes virtuous actions, not mindlessly, but with deliberate practice (doing them knowingly and for the right reason), the more one conditions oneself to act virtuously in each circumstance. One’s character cannot be shaped by a single sort of virtuous act alone, just as one will not become strong through bicep curls alone. Rather, virtue requires that one calibrate the desires and emotional responses that guide one’s actions to the situation in which one finds oneself (NE 2.6: 1106a15-1107a26, esp. 1106b17-29).

What separates genuine virtues from more generic positive character traits is the presence of phronesis or “practical wisdom”. This is a kind of good judgment about what is genuinely

---

7 For a literature review, see Rheingold et al., 1976: 1148
8 Aristotle considers that losing the virtues is something that can only occur with (1) someone who was virtuous and (2) when something profoundly bad happens to them that would significantly alter their emotional responses to the world. Aristotle thinks that one of the things that is important about someone who is virtuous—as opposed to someone who acts well but is not virtuous—is that the virtuous person is in tune with the world such that her desires and emotions align with doing the virtuous thing in the circumstances in which she finds herself. Someone who is suffering from extreme tragedy might lose that alignment with the world, but still act well or might act poorly because their judgment was temporarily unseated by grief. However, Aristotle thinks that virtue runs so deep in a person that, given time, that she would return to being virtuous (NE 1.10 1100b4-1101a13)
good and how to go about attaining it, based on a combination of understanding the circumstances and responding with the correct virtue based on those circumstances (NE 6.5: 1140b5-8, 20-21; Vallor 2016 18-19). Aristotle says that “for virtue makes the goal correct and *phronesis* makes the means to that goal correct” (NE 6.12 1144a8-9). The idea here is that Aristotle has already established the general end towards which all action is directed i.e. *eudaimonia* constituted by virtuous activity. *Eudaimonia* as a goal is almost always too vague to allow for successful practical reasoning about what one should do in any given situation. The virtues of character function to provide a finer-grained specification of one’s end as relevant to some particular situation in which one finds oneself.

For example, one part of a happy life is providing resources for others to enable them to act well and achieve in ways they might not be able to do otherwise. This is part of living a happy life because a happy human life involves living in a thriving community. So, generosity points individuals towards giving the appropriate amount, at the appropriate time, to the appropriate people. This is important so as to neither hoard resources for oneself nor to overshare and deplete one’s ability to continue to act well. All virtues require *phronesis*, because *phronesis* points to the virtue that needs to be deployed in a particular context. Thus, no one truly can exercise any virtue until they have also acquired *phronesis*.

A corollary to this is a view called the unity of the virtues, according to which one requires all the virtues (including *phronesis*) to be virtuous. *Phronesis* is only achievable if the person can use it to call into action the appropriate virtue for the circumstance in which one finds oneself. In other words, there is not a separate *phronesis* for each virtue; (e.g. there is no *phronesis* of bravery separate from a *phronesis* of generosity), one has every virtue if they have any virtues and if one is lacking any virtue one has no virtues. The virtues rise and fall together. The virtues are not only similar in kind (they are all states of character that generate a type of correct action), but that they are not individual states of character but aspects of good character. The virtues are unified because virtues are not individual states of character each virtue is deployed in a particular situation, but good character overall involves having one’s emotions and desires tuned to the world in such a way that one is inclined away from the more distant vice in all cases (NE 2.6: 1106a15-1107a26, esp. 1106b17-29).

The virtues are not separable. While someone may, through practice and good instruction, gain proper emotional calibration in a single area of their life such that their emotions predispose them to act well, this is extraordinarily challenging to pull off. Imagine an ER doctor who is incredibly good at her job and taps into her well-calibrated emotions in the ER to be able to couple her instincts with her medical training to great effect. However, her instincts are only well-calibrated to the ER; when she comes home, she is easily irritated by her family, distracted, and immoderate. Clearly, this doctor is not holistically thriving; rather she is partitioning her life in such a way that she acts well in one sphere, but not in another. She is not acting well from a firm and unchanging character; we might say she is a good doctor, but she is not a virtuous person. Moreover, it is possible to imagine a scenario in which her partitioning of the various spheres of her life breaks down and her emotions prevent her from making a good decision in the ER.

The unity thesis is often abandoned in contemporary formulations of virtue ethics for several reasons. Some hold that the unity thesis makes virtue too hard to achieve. Others find it intuitive that people can have some virtues but not others (e.g. a soldier could be courageous
Foot (1983), Walker (1993), Hursthouse (1999), Badhwar (1996), and Watson (1984), for example, all reject essential pieces of the unity thesis. Walker and Foot hold that certain virtues may be incompatible, while Badhwar (1996: 308), for example, argues that virtues need only be unified “in a particular domain” in an agent’s life, and that possessing a virtue in one domain does not imply the existence of that virtue in another domain. Russell (2009), and Annas (2011) defend weakened versions of it, according to which completely unified virtue is an ideal to which we aspire, or that the virtues are at least not completely separate.

While the unity thesis is often an unpopular or controversial thesis, we think it is essential for virtue ethics. It is true that the unity thesis suggests that it is extremely difficult to be a good person on the whole, i.e. to be the kind of person about whom others correctly say “X is good” full stop. However, being virtuous is difficult. It also strikes us as reflecting, correctly, that what any situation calls for will be determined by a complex mixture of all of one’s virtues. [see, e.g. Watson 1984] Someone who has achieved complete virtue and moral situational sensitivity is rightly called a “phronimos” or a “sage”. This kind of person is a model of, and a metric for, good action. (NE 1113a31-33; Hursthouse 1999, ch.1; Vallor 2016 37)

A common commitment of both ancient and contemporary Aristotelian virtue ethics is that morality cannot be codified in a finite list of rules. (Hursthouse 1999 32-34; Annas 2004 63-65; Vallor 2016 24-25) While there may be small handful of inviolable moral rules (e.g. do not murder, do not commit adultery) these are few and far between, and hardly offer comprehensive guidance. Morality as an area of study does not admit of the specificity of mathematics or the sciences, and we should not expect to find a finite list of rules that constitute a complete decision procedure for ethics. (NE 1.3 1094b11-1095a2) First, there are simply too many variables that must be accounted for when making any moral decision. You might think that one can get around this by proposing an indefinitely long list of rules that takes account of every possible situational variance. This is not a decision procedure that any actual human could construct or follow, and therefore fails to provide comprehensive moral guidance. Ethics is, if nothing else, something that humans do; attempting to fill a purported gap in a moral theory by means of something impossible be done is a nonstarter.

Further, the guidance provided by the virtues and phronesis are strongly agent and situation relative. The virtues demand different things from people with different backgrounds, means, knowledge, and skills. If two people are presented with a circumstance calling for generosity and both have the means to give but one is a multimillionaire while the other has a minimum wage job, generosity might demand a higher degree of giving from the former than the latter. If a situation calls for courage, situations may differ as well. Imagine someone is trapped in a burning building and the fire department cannot arrive quickly. If one person on the scene

---

9 Hursthouse (1999: 120) identifies her view, and Watson’s, as being similar to Badhwar’s view in that it amounts to a kind of “limited unity”. We accept this interpretation.

10 Aristotle claims that these rules concern actions that do not admit of a mean (NE 2.6 1107a10-17). The type of actions like this are rare enough that they cannot provide guidance in most cases, but only in exceptional cases of base actions.

11 If one proposes simply adding exceptions to general rules as needed for new cases, such a system is ad hoc (after all, the moral rules one would appeal to are exactly what is in need of an exception) and provides insufficient generality to deal with novel cases.
knows to breach a door and has an ax and the other does not, inaction would speak badly of the former’s character but not the latter’s character.

### Why Virtue Ethics

Virtue ethics has advantages over other normative systems regarding dealing with novel situations. An ethical system that can flexibly deal with novel situations is necessary in an age with rapidly changing technologies. The benefit of a virtue ethics approach is that there is at least some basis by which to judge what it would look like to live well with AI. Aristotle’s account of *eudaimonia* attempts to describe what it means for a human to live well in general—to live a life of rational activity in accordance with virtue. What it means to live well with AI is to consider how AI technologies might detract from or enhance our ability to engage in such activities and determine how to proceed accordingly. In other words, the virtues that we would need to live with AI are already well-known, the question is only how to apply them appropriately to current situations. Moreover, the way in which the proper use of such technologies would be determined depends on how the sage would use such technologies given the salient features of a situation. Virtue ethics, therefore, can guide designers and engineers of such technologies, and inform all members of a society who interact with these technologies how to appropriately interact with them.

The future, and much of the present, is opaque. Determining what is at stake in a moral decision is often difficult or impossible. We are unable to know the minds and wills of other agents. Our understanding of the future is murky and is certainly not clear enough to make specific plans to guarantee specific moral outcomes. We cannot know what new technologies will be developed or adopted into wide use; we cannot predict the consequences of technology we cannot envision.

We, like Vallor, hold that this opacity is particularly concerning when it comes to technological change. Vallor, describes our current situation as one of “acute technosocial opacity”. (Vallor 2016 6) For large chunks of our history, humans have had the luxury of assuming, reasonably confidently and reasonably correctly, that technology would continue to work much as it had in the past, or at least would change in broadly predictable ways. In the former case, the future would be sufficiently like the present such that whatever moral norms and rules we had would suffice to ensure that we could continue to live well in the face of technological change. Or, in the latter case, we would at least be able to foresee the impacts of certain technological changes and have sufficient time to prepare ourselves for new moral and social situations that updated technology would bring. However, current technological change is fast and unpredictable: familiar technology changes, new and unexpected tools are introduced,

---

12 This is distinctly different from an account like Vallor’s, which argues that a special set of virtues are needed to deal with current technological change.

13 Vallor (2016 8, 9) contends that this current technosocial opacity is “unprecedented”, and that the moral problems raised by current technology somehow require a unique response. While current technological advances and changes contribute to the opacity of the future, we are unsure this opacity is unprecedented. How, for example, are we to judge whether early adopters of fire, which greatly increased the bioavailable nutrients of most food, had any conception of the consequences of things like increased brain size. Did those who innovated the assembly line
and our moral and social life can quickly become something unfamiliar. In the face of this opacity, we cannot expect a moral system with fixed rules, especially ones that provide fairly detailed guidance or are tailored to present circumstances, to guide us in living well with changing technologies. Extant rules may have to be twisted and stretched to the breaking point to enable even a semblance of good moral decision making if we insist on a rule-based approach of some sort. (Vallor 2016: 6-7) Faced with this circumstance, we ought to consider whether virtue ethics, which never purported to reduce ethics to a finite list of rules to follow but nonetheless offers insight into living well, would enable us to cope with an opaque future.

Problems with Rule-based Approaches

Alternatives to virtue ethics tend to be either rule-based or based on calculi of ends. We confine ourselves in this paper into addressing principlist approaches. Principlist approaches, which create governing rules or guidelines, are the dominant approaches to technology ethics. Rules are an ineffective way to guide action in novel or unpredictable situations, because any such formulation would be unlikely to address questions that arise from the specifics of the situation. Unexpected, or unexpectedly complex, problems can only be dealt with post-hoc.

A good example of this comes from the case of gene-editing (and especially human gene-editing) in biomedical ethics. Despite the fact that there are general rules in the biomedical field (e.g. “do no harm”), specific rules must be created for each new innovation. Indeed, when Nobel Prize winner Jennifer Doudna discovered the way to undertake targeted gene editing with CRISPR Cas-9, she realized that she needed to convene a new set of councils to govern the use of the new biotechnology. In her book, A Crack in Creation, she details a number of cases where she met with scientists and business people who wanted to use the new biotechnology for questionable purposes (Doudna & Sternberg, 2017: 184-188, 213-216). This convinced her that she must convene a summit to create a set of specific rules to govern the experimentation and innovation using CRISPR Cas-9.

The CRISPR meeting convened in January 2015—along with subsequent summits—shows the drawbacks of the rule-based approach in the biomedical field as well as in governing technological advancements in AI. The original summits on which Doudna based her ideas—Asilomar I (1973) and II (1975)—were convened to determine rules on gene-editing during the initial recombinant DNA revolution. While CRISPR Cas-9 is much more powerful and accurate than recombinant DNA gene-editing that was discussed at the Asilomar summits, the technologies are similar insofar as they both alter genes. The rules that came out of the Asilomar conferences were highly specific to recombinant DNA and are an insufficient basis for regulating gene-editing more generally. Given that leaps forward in AI technologies differ much more widely than the two approaches to gene-editing mentioned above, it seems unlikely that any single rule-based approach could cover all cases, while providing sufficient specificity.

production process really have any way of understanding the subsequent changes in consumer behavior and natural resource consumption that would result from a massively enhanced industrial sector? Even the transistor or the microchip changed the world profoundly by, among other things, enabling spaceflight. Further, it isn’t clear that it matters whether current technological and social change are unprecedented. The fact is, we are aware both that we are unable to predict the future impacts of present technology, and that we nonetheless need some way of guiding our decision making.
The principlist approach has been used in the AI ethics literature, often by explicitly citing inspiration from, or directly importing, principles laid out by Beauchamp and Childress in their seminal text *Principles of Biomedical Ethics*. A good example of this comes in Floridi and Cowls (2021) which argues that AI ethics can use Beauchamp and Childress’s four principles of biomedical ethics (Beneficence, Non-Maleficence, Respect for Autonomy, Justice), and that AI ethics should add a fifth principle regarding transparency.

According to this view, the hard moral work, though maybe not the hard empirical or policy work, has either been done or isn’t particularly mysterious. We know what to do, even if we are still working out exactly how to do it or struggling to codify these otherwise clear principles and rules into official institutional rules or laws.

From an Aristotelian perspective, this is, at best, severely misguided. If these rules or principles are clear, concrete, and actionable, they are likely insensitive to the variety of novel situations that will inevitably arise. This is a central concern for anyone interested in the ethics of AI and machine learning. After all, living well with AI requires guidance on moral matters related to technology that changes rapidly and unpredictably, and often does things that are beyond our understanding. We should expect to encounter unexpected consequences and outcomes, and processes that operate in unfamiliar ways. Even if we can codify rules and pass them on or teach them effectively, we should worry about encountering situations in the future to which the rules and principles fail to apply. In short, principles frequently fail to provide concrete guidance, especially in novel circumstances.

Some principlist systems invoke virtues in addition to a set of guiding principles. In general, these principlist systems foreground (usually exceptionless or absolute) rules and use virtues to fill in gaps, adjudicate in cases of conflict, and guide application. The rules and principles often appealed to in AI ethics make use of contested moral concepts; thus, applying these rules amounts to exercising good judgment. This seems very much like saying that appropriately applying these rules is a matter of *phronesis*, in which case it is not good rules doing the work of helping us live well with AI, but practical wisdom. One might think that appealing to these rules and principles, as long as they are the correct ones or sufficiently detailed, amounts to backing into virtue ethics without trying. However, we are in danger of

---

14 We note the relationship between Beauchamp and Childress’ principlist ethics and virtue ethics in footnote 14.

15 Floridi et al. (2021) synthesized six AI research and development ethical codes and compiled a list of 47 individual principles. They found that these 47 principles could in fact be effectively clustered around the four principles of biomedical ethics plus a fifth “transparency” or “explicability” principle.

16 E.g. Kant’s ethics, as spelled out in the *Metaphysics of Morals* uses virtues in conjunction with the categorical imperative, Beauchamp & Childress also invoke virtues in addition to their principles. Beauchamp and Childress (2012) note that virtues have a place in their principlist ethics, though they hold that virtues are just one of many things that need consideration in a normative system (Beauchamp & Childress, 2012: viii-ix, 3-4, 14) and have a distinctly non-Aristotelian view of virtue (for example, they deny the unity of the virtues; Beauchamp & Childress, 2012: 32) and hold that virtues can lead people morally astray, a claim which makes no sense in an Aristotelian virtue ethics (Beauchamp & Childress, 2012: 34). W.D. Ross combines what he calls *prima facie* duties—which are not absolute or exceptionless—with a set of four intrinsic goods, including virtue. (Ross, 2002) While Ross’s theory is interesting, its unique contribution to moral philosophy places it outside the scope of this paper. We thank an anonymous reviewer from *Theory & Research in Education* for suggesting that we deal with these views explicitly.

17 We provide a case study in the following section that demonstrates the issues we discussed with regards to principlist approaches to AI.
specifying our principles badly without *phronesis* as a guide.\(^{18}\) Aristotle’s virtue ethics takes the opposite approach to principlist approaches that embrace virtues. For Aristotle, virtues, and specifically *phronesis*, serve as the ultimate guide to action. Rules, when they exist, are secondary; Aristotle mostly invokes rules when discussing the creation of laws for a state.\(^{19}\) Laws, however, are established to promote virtue and they are not exceptionless, but must be adjusted based on circumstance.\(^{20}\) For this reason, virtue ethics by its nature provides a system that is more flexible and responsive to novel circumstances.

Additionally, given the cultural and social norms in AI research, principlist systems may be a particularly poor fit for AI ethics. In order for principlist systems to constrain behavior, interested parties must agree to rules, have a culture that encourages obeying rules, and have some mechanism for enforcing those rules. Unless they do, rules are merely suggestions. A particular issue can be seen with the attempt to put in place some sort of ethical rules for the development of AI. In an open letter, a group of AI influencers and developers called for a 6-month moratorium on developing AI tools as or more powerful than ChatGPT-4.\(^{21}\) The attempt at the moratorium was in parallel the moratorium enacted prior to Asilomar I in the 1970s; the Asilomar conferences put in place a set of rules that allowed research to proceed (Berg 2008).

The biomedical field already relies on strong sets of rules that govern every action and experiment that is done. This culture of rule-following is lacking in computer technology fields. The call for a moratorium, so effective in DNA research, failed in AI research. This may be, in part, because many of the interested parties in AI position themselves as rule-breakers, innovators, and industry disruptors. Culturally, this sector of the technology industry has profited from a “move fast and break things” mentality that is antithetical to—or unlikely to encourage—rule-following. (Benjamin 2019: 11-17) Thus, principlist approaches face two barriers: changing the culture of the technology community and creating and implementing a set of rules.

Aristotelian virtue ethics has cultural change baked into it. It requires educating the next generation, especially those who will have an impact on the future of technology—those in

---

\(^{18}\) There are, of course, some approaches to virtue ethics that are less hostile to principlism, though many of these approaches involve some deviation from the Aristotelian position we take, or a very loose conception of what constitutes a rule. Bakhurst (2018, esp. 682) argues that principles can be incorporated into a general, virtue ethical moral outlook or ethos as a mode of expressing that outlook or an aid to moral instruction (though Bakhurst also concedes that adherence to principles must be guided by phronesis, which would seem to make his view very close to ours). Haydon (2009) explores R.S. Peters’s view that moral education, while ultimately aimed at developing an appropriately rationalist and deontically oriented moral cognition, should incorporate a Humean (rather than Aristotelian) sort of virtue that is conducive to behaving in line with rationalist morality. Hursthouse’s (1999) contemporary and avowedly Aristotelian account of virtue ethics includes what she calls “v-rules” (though these have the form of “Do what the virtuous person would do” and are thus rules in only the loosest sense). Aristotle himself notes that there seem to be at least some hard and fast rules, in the sense that some things are bad “in themselves” (e.g. murder and adultery) and there is no possible virtuous way to do them (*NE* 2.6 1107a10-17; 5.1 1129b19-25). These actions, and thus any rules prohibiting them, are too few and far between to comprise a complete ethics. The standard of moral goodness is acting as the virtuous agent would (*NE* 2.4 1105b7-9).

\(^{19}\) E.g. *NE* 2.1 1103b3-7, 5.1 1129b11-19, 6.13 1144a11-23

\(^{20}\) Aristotle makes it clear that some things cannot be guided by rules (*NE* 5.10 1137b27-30), and that even those things guided by rules have exceptions that require differential application (*NE* 5.10 1137b13-27). The virtue involved with this application of rules is “decency” (“ἐπιείκεια”). We discuss this below in the section “Exceptions to rules.”

\(^{21}\) [https://futureoflife.org/open-letter/pause-giant-ai-experiments/](https://futureoflife.org/open-letter/pause-giant-ai-experiments/)
design, governance, etc.—to shape the character of those who are making ethical decisions. This approach will likely provide a stronger foundation for envisioning a way to govern and live well with AI than a set of rules.22

A Case Study of Problems with current rule-based approaches to AI ethics

Analyzing every extant AI ethics framework would be a prohibitively large task. Floridi (2019) estimated that, circa 2019, there were over 70 such frameworks being advanced by various groups, each with their own areas of interest, aims, specializations, agendas, and ways of specifying their principles.23 Here, we explore one paradigmatic example, a version of a principle of beneficence. The IEEE’s Ethically Aligned Design framework states that, taken collectively, the principles they have developed “Embody the highest ideals of human beneficence within human rights” (17) Somewhat more specifically, one of their seven principles states that “A/IS creators shall adopt increased human well-being as a primary success criterion for development” (21) with the recommendation that “A/IS should prioritize human well-being as an outcome in all system designs, using the best available and widely accepted well-being metrics as their reference point.” (21)

The IEEE version of the principle of beneficence needs more detail in order to provide concrete guidance to those who develop or work with novel technologies. To fill in those details, the creators of the principles included specific guidelines for how their principles should be used. Below, we consider some paradigmatic examples.

The IEEE states that their principle of wellbeing should be interpreted with reference to the best available and widely accepted metrics of wellbeing. They hold that “There is now sufficient consensus among scientists that well-being can be reliably measured” (71) as well as the following positions:

- We encourage A/IS creators to consider the wide range of available indicators and select those most relevant and revealing for particular stages of the A/IS technology’s life cycle and the particular context for the technology’s use and evaluation. That is, measures of well-being that may be well-suited to wealthy, industrialized nations may be less applicable in low- and middle-income countries, and vice versa. (71)
- For A/IS to promote human well-being, the well-being metrics should be chosen in collaboration with the populations most affected by those systems—the A/IS stakeholders—including both the intended end-users or beneficiaries and those groups whose lives might be unintentionally transformed by them. This selection process should be iterative and through a learning and continually improving process. In addition, “metrics of well-being” should be treated as vehicles for learning and potential midcourse corrections. The effects of A/IS on human well-being should be monitored continuously.

22 We thank an anonymous reviewer at Theory & Research in Education for encouraging us to spell out our position more clearly.
23 Mittelstadt (2019) argues for a similar point to ours, using a meta-analysis of “84 public-private initiatives have produced statements describing high-level principles, values, and other tenets to guide the ethical development, deployment, and governance of AI” (Mittelstadt 2019: 501)
throughout their life cycles, by A/IS creators and stakeholders, and both A/IS creators and stakeholders should be prepared to significantly modify, or even roll back, technology that is shown to reduce well-being, as defined by affected populations.

- Therefore, even though a product modification may increase well-being according to one indicator or set of A/IS stakeholders, it does not mean that this modification should automatically be adopted. (79-80)

It is not clear that there is sufficient consensus among scientists that well-being can be reliably measured, and there is certainly no consensus among philosophers as to what wellbeing is, much less how to measure it. Further, this is not an exhaustive list of ways that IEEE has elaborated their principles, merely a selection of some ways they have chosen to do so regarding beneficence.

Just as it would be a prohibitively large task to analyze every extant ethical framework for AI research and development, comprehensively analyzing every philosophical implication of the IEEE’s position on wellbeing alone would be a prohibitively large task, given that their Ethically Aligned Design is unwieldy and near 300 pages, excluding references and links to supporting studies and documentation. While their thoroughness is to be commended, we trust that our reader will be content if we can demonstrate our point with just a few examples.

The IEEE also makes policy and best practice recommendations, but these are vague and based on the assumption that those following these recommendations have a thorough understanding of what wellbeing consists in. For example, two principles for best practice are:

- A/IS creators should prioritize learning about well-being concepts, scientific learnings, research findings, and well-being metrics as potential determinants for how they create, deploy, market, and monitor their technologies, and ensuring their stakeholders learn the same. (73)

- A/IS creators should adjust their existing development, marketing, and assessment cycles to incorporate well-being concerns throughout their processes. This includes identification of an A/IS lead ombudsperson or officer; identification of stakeholders and end users; determination of possible uses, harm and risk assessment; robust stakeholder engagement; selection of well-being indicators; development of a well-being indicator measurement plan; and ongoing improvement of A/IS products and services throughout the lifecycle (80)

Given our Aristotelian starting point, our view on these principles should be clear: even with the additional specifications and clarifications, principles cannot, on their own, ensure that we live well with AI. Even if these principles are useful for thinking through possible courses of action—

---

24 The most straightforward way of seeing this is simply to note that there are many competing accounts of what wellbeing is. Major camps include hedonists who hold that it is a matter of experiencing pleasure (Crisp 2006; Feldman 2004), desire satisfaction theorists who hold that it is a matter of fulfilling some (usually idealized) set of desires (Brandt 1979; Railton 1986), objective list theorists who that it is a matter of obtaining certain goods regardless of one’s subjective interest in them (Finnis 2011; Fletcher 2013), and perfectionists that it is a matter of developing certain aspects of one’s nature (Kraut 2007). A classic general introduction to the wellbeing literature is Appendix I of (Parfit 1984) If there is no agreement as to which of these (or some other) views of wellbeing is correct, then there is surely no consensus as to how wellbeing ought to be measured. At best, there might be agreement among adherents of a particular theory.
e.g. designating ombudspeople and determination of possible uses, harm and risk assessment—these will only promote wellbeing if the actions of those ombudspeople are virtuous and good determinations are made about how to habituate users to use technologies well, based upon those assessments. *Mutatis mutandis* for the majority of these principles about best practices.

Moreover, such principles fail to innovate much beyond traditional best practice principles that govern engineering more generally. The main innovation is putting “A/IS” in front of most of these ideas. If that is indeed the case, one might wonder why such a designation is necessary. Does it indicate that those innovating in AI fail to understand the principles apply to them or does it indicate that those innovating in AI fail to comply with standard engineering practices? In either case, it seems that such an elaborate set of principles are unlikely to solve these problems. Indeed, improving the character of those in the field is likely to have a far greater impact on our ability to live well with technology than a 300 page list of rules.

Exceptions to Rules

In addition, there is a general issue with rules considered from an Aristotelian perspective. In *NE* 5.10, Aristotle explains that there are inevitably exceptions to rules that govern behavior. This arises from two circumstances: (1) legislators being unable to imagine future circumstances in which applying such a rule would render an injustice (*NE* 5.10 1137b13-27) and (2) cases where exceptions arise not from any deficiency in the rule, but from the nature of the object (*NE* 5.10 1137b27-3). Three types of exceptions might be one-off cases, cases which require a deviation to compensate for a historical trend of unfairness, or cases which are completely novel and unanticipated. These are precisely the sort of cases that manifest as technology changes. For example, algorithms tend to learn from and amplify historical bias in the data upon which they are trained. While we might make a principle to test algorithms sufficiently to root out algorithmic bias before algorithms are employed to generate decisions in the real world, there is unlikely to be a specific rule about how to handle cases in which algorithmic bias affected individuals before the bias was discovered. Moreover, each of these cases is likely to be unique to the particular way in which the decision affected the individual.

There are no exceptions to virtue ethics, because virtues merely lead one to act in the best possible way given the situation at hand. When this sort of action departs from standard rules, Aristotle says that the virtue responsible for this adjustment is decency. There are, of course, cases in which the manifestations of two specific virtues might clash. However, *phronesis* helps the sage solve the issue as to which virtue is more important to manifest in those cases. We argue that IEEE erred by being too detailed to govern new situations or too vague to be informative and applicable, not that the IEEE has made uniquely bad or misguided recommendations.

---

25 Above, we describe how non-virtuous people can perform virtuous actions. Someone who often performs virtuous actions because they understand they are the correct way to act, but is not virtuous, is called *enkratic*. Vickers discusses *enkracia* in detail in her ending.

26 We revised this section, thanks to comments from an anonymous reviewer at *Theory & Research in Education*.

27 Aristotle, here, is considering the laws that would govern a *polis* (city-state). Aristotle suggests throughout the ethics that a state must have laws (written as rules), but those rules should be aimed at habituating virtuous citizens (*NE* 2.1 1103b3-7, 5.1 1129b11-19, 6.13 1144a11-23) and the enforcement of those laws require judges and statesmen who are capable of understanding when it is virtuous to bend, adjust, or not apply the rules as the ethically salient features in a situation dictate.
Asilomar’s recommendations show similar flaws. Indeed, everyone’s principles have these flaws because these flaws are inevitable in attempting to govern the development and use of this unpredictable technology in opaque circumstances.

A Case Study of Kantian Machine Ethics

The principlist approaches that we have discussed thus far are approaches that consider how humans should make rules to live well with AIs. There are alternative approaches, particularly in machine ethics, that concern how to regulate AI behavior and interaction with humans. Machine ethics is the branch of AI ethics that concerns how to program AIs to make ethical decisions. Some of the principlist work in machine ethics uses a Kantian framework, suggesting, for example, that the categorical imperative can serve as a basis for moral decision making frameworks in AIs (e.g. Ulgen, 2017; White, 2022). Ulgen (2017) considers how a Kantian framework might be applied both weak AIs controlled or monitored by humans (such as autonomous weapons) and to artificial general intelligences; White (2022) focuses on designing Kantian artificial moral agents.

Kantian principlist approaches to machine ethics suffer from the same flaws described above, insofar as they veer between being too detailed to govern new situations and too vague to be informative and applicable. The categorical imperative prescribes that one act on a maxim that one could will to become a universal law; in other words, universalizing that maxim would not undermine the action taken by the agent. For example, one should not make a lying promise because if everyone made lying promises, it would undermine the purpose and value of promising. While the categorical imperative provides concrete guidance about some actions, it fails to do so about others. Indeed, Kant’s categorical imperative, laid out in the *Groundwork for the Metaphysics of Morals*, is merely a beginning to Kant’s moral theory. In addition to the categorical imperative, Kant lays out a complicated, multi-level procedure for how to interpret and apply the categorical imperative to our daily lives, including a set of virtues with which we should act. Even with this procedure, Kant acknowledges that judgment is often required and that there will always be exceptions that cannot find their resolution in the rules laid out (Kant, 1991: 6:411).29 Given that no machine has the ability, currently, to exercise judgment, such an approach is likely both misguided and premature. Even in the human-machine rule-generating approach Ulgen (2017) advocates, AIs lack the judgment necessary to correctly apply the generated rules.

28 We thank an anonymous reviewer from *Theory & Research in Education* for suggesting that we tackle this viewpoint.
29 Ulgen (2017) acknowledges this: “a limited sense of rational thinking capacity can be programmed into the machine but it will not have the self-reflective and deliberative human capacities, as developed under the Kantian notion of rational beings, so that the machine will not be able to assess a given situation and exercise discretion in choosing a particular action or not.” (Ulgen, 2017: 82)
Practicality: Can Virtue Ethics Provide a Basis for How to Live Well with AI?

It is prima facie important to discuss how one should use or live with AI in a way that actually provides guidance in novel situations. Given our Aristotelian perspective, our answer to the questions of the sort “what should we do or how should we live with AI?” is “whatever the sage would do with AI”. While this is the answer that a virtue theorist in some sense should give, this is understandably unsatisfying to those who want an applied ethics of artificial intelligence to provide concrete or specific guidance. Once there are moral sages with sufficient understanding of AI technologies, concrete and specific guidance comes from determining what the sage would do in a given situation. However, the implementation is complicated by the lack of concrete rules or guidance provided by a virtue ethics framework and by the lack of easily-identifiable moral exemplars. Thus far, we have considered the theoretical justifications for an Aristotelian approach to virtue ethics; it is worth considering whether what we propose is practical.

We consider two different perspectives from which one can evaluate the practicality of a virtue ethics approach, each one written by one of the two authors of this paper. The combination of the two endings is particularly important when considering how higher education can respond and adapt to the challenges to educational institutions posed by powerful AI tools. Smith’s ending deals with the more theoretical reasons why it will be very difficult to implement Aristotelian ethics in higher education settings. Institutional change is difficult and massive change would be needed in order to make a real impact. Smith’s ending is a sobering reminder that overhauling undergraduate education is not easy and that interventions at the undergraduate level may be too late to shape the character of students.

Vickers’s ending deals with what she sees in her ethics classroom as she has been teaching this material. Vickers’s ending is a ray of hope that even some limited forms of training might be helpful, even if it would be better to overhaul education to help students deal with a world of changing technology. Vickers has noticed, both among faculty and among students, that there is a desire to figure out what it would take to live well with AI and that faculty and students are willing to take certain radical steps to figure out what that would look like. Both authors hold that an Aristotelian virtue framework must be a part of this conversation as it is emerging and that we should take it seriously even if it is challenging to implement.

Nicholas Smith’s Ending

We can answer our moral questions about the use of AI with old tools; we need no new principles or methods. However, an Aristotelian approach to AI is easier theorized than implemented. Finding virtuous people is a serious challenge, yet, exemplars are vital for moral learning. Compounding this problem is that, among a vanishingly small population of virtuous

---

30 Currently, there is a wide variety of experimentation. At [university name omitted for anonymity], where Author 2 works, there is a famous class in the Department of Communication that asks students to undergo a digital detox for a week and give up their phones. The New York Times recently showcased a variety of unusual classes that provide students sanctuary from the digital world.
people, we must ensure that at least some of them have the relevant technical knowledge. Under an Aristotelian framework, people would need to develop *phronesis* in order to make correct judgements about how AI research should proceed and to create and implement any principles that might aid in this process. Since we endorse the unity of the virtues, the acquisition of *phronesis* implies the acquisition of all virtues. In addition, we must ensure that those who have it are involved in developing, deploying, and making policy for AI.

A virtuous person who knows nothing at all about AI would, when presented with complex questions about the appropriate use of often inscrutable technology, be sufficiently self-aware to realize that they were the wrong person to ask. However, we cannot focus on exclusively training people to acquire technical knowledge, because even great technical knowledge of AI provides no guidance about how one ought to use it and towards what ends. A solution to this problem would require training virtue along with training the technical aspects of AI. Virtue ethics is highly sensitive to the capacities and knowledge of the agents involved in a given moral situation, so even completely virtuous agents might reasonably say, when presented with a moral question involving a complex new technology “Why are you asking me what to do? All I can say is that I don’t know enough. You need someone who knows how that works.”

Developing virtue is hard work, as is becoming an expert in a new technology. Finding someone who has already done both is unlikely, and the educational demands of creating someone who meets that standard are immense.

If we want a virtuous person with a high level of technical knowledge, we’ll need to both habituate and educate them. Each of these are time consuming and resource intensive processes. In order to start a program to nurture virtuous people with sufficient technical knowledge, at a minimum, we would need at least one person who possesses both virtue and the relevant technical knowledge. This person would serve as an exemplar and provide guidance about using novel technology in novel situations, someone who can help habituate and guide other people, and so on. If a program along these lines can be developed, we can likely perpetuate it. This program would be easier to maintain than to start because once started, there should be increasingly good role models for new generations to imitate.

We might pursue both technical education and character development simultaneously, by trying to ensure that high level technology and science educators are, if not fully virtuous, at least of sufficiently well developed character that they can point their students in the right direction and be worth imitating. We probably are not in a good position to identify morally upright technology educators than we are to successfully screen a pool of incoming technology students for those who are already properly habituated. It might be possible to develop course material that itself contributes to the development of virtue, and could be used by even non-virtuous educators. However, it’s not clear that such one-size-fits-all course material could be

---

31 Formal education alone is not responsible for ensuring one’s character development. While educators are uniquely well placed to shape character, others, such as near-peers, parents, mentors, and employers may be able to fill in some gaps.

32 For discussion of modestly sized examples of this approach, see (Orona and Pritchard 2022, Croce and Pritchard 2022). This approach, however, is limited. For example, this approach involves making pedagogical use of virtuous exemplars (i.e. paradigmatic or model virtuous agents the non-virtuous can look to and learn from) of the sort discussed in (Zagzebski 2017) to fill in for a lack of actual virtuous people in the classroom. However, we shouldn’t expect accounts of these exemplars to be self-evidently clear; exemplars must be explained in the right way for
sufficiently attentive to the nuances of virtue, especially in the cases in which it is administered by instructors who themselves may lack the requisite level of moral development. This would help ensure the proper overlap in habituation and high level scientific education.

However, the combined technology and character education approach has other issues. If we rely on character development taking place during the process of higher education the subjects of these efforts will overwhelmingly be people who have already developed their characters in a certain way. The standard Aristotelian view is that, after a certain point, character becomes fairly fixed. This should not be surprising; it is part of the definition of virtues that they are deeply rooted character traits that are neither gained nor lost easily. While the strictest forms of this view—that character just does not change after a certain point—seem false, it seems plausible that the amount of effort involved in changing the characters of those whose character is fixed is far greater than that involved in developing the character of a younger, more impressionable subject.

None of the above issues would be, on their own, decisive objections to educating people such that they are both virtuous and have the requisite technological knowledge and skills. However, when taken together, there are clearly substantial, perhaps collectively insurmountable hurdles. Still, embracing this approach is the best we can do. It is better to fail at this, having realized that we must aim at virtue if we are to live well with technology, than to accept defeat from the outset and settle for an approach to living with technology that is fundamentally insensitive to the problems with the currently dominant approaches.

Darby Vickers’s Ending

Like Smith, I believe that there are serious challenges to overcome in the virtue education that we propose. However, I am more optimistic than he is that we can make a significant positive impact at the college level, contingent upon institutional changes. My optimism comes from three main sources: Aristotle’s concept of *enkratia*, recent research on brain plasticity in adulthood, and anecdotal personal experience teaching ethics courses. I address each of these three pieces in turn.

Aristotle understood—and communicates in the *Nicomachean Ethics*—that virtue is extraordinarily difficult to achieve. Indeed, if a virtuous person must have all of the virtues, they must have the virtue of *megalapsychia*. Only a truly great person can have the virtue of *megalapsychia* (sometimes translated as “magnanimity”) (*NE* 1124a1-4). The ancient Greek adjective “*megalapsyche*” literally means “great souled” and Aristotle says that the *megalapsyche* person is someone who is particularly suited for great deeds and expends her energy only on great causes because she is worthy and capable of achieving great things (*NE* 1123b15-30). The *megalapsyche* person is the best and most worthy (*NE* 1123b25-30) Only a select few can ever be *megalapsyche* for this reason. However, Aristotle’s schema allows that there is a way to undertake virtuous actions without being virtuous, namely *enkratia*.

students to get the most out of them. The perfect example, explained badly, is useless to a student unable to make sense of it for themselves.

Aristotle expresses this view, among other places, at. NE 1114a15-22.
The enkratic person has enough training in virtue to often be able to choose the virtuous actions but falls short of being virtuous because of a lack of the correct emotional calibration (NE 1145b1-3, 9-15). For Aristotle, as we mentioned above, being virtuous involves more than doing the virtuous actions, but understanding why they are virtuous, undertaking them for that reason, and doing them from a firm and unchanging character. In addition, the virtuous person feels pleasure from doing the virtuous action and that pleasure stems not from the action itself but from the fact that the action was virtuous. For example, imagine I find $100 on the road. Both the virtuous person and the enkratic person would try to find the owner or turn the money over to authorities to seek the owner. However, the virtuous person would feel no qualms about turning in that $100 and feel pleasure from doing it, even if $100 in their pocket would be nice, while the enkratic person might feel a desire to keep the money and feel pain or regret at losing the possibility of another $100 in their pocket. In addition, the enkratic person might not experience pleasure merely from doing the right thing.

Enkratia serves as both an intermediary step towards virtue and as a possible endpoint that creates respectable members of a society. For example, Aristotle argues that virtue requires the correct calibration of one’s emotions to the situation that one is in. There are some mental disorders, for example, that prevent one from being able to correctly calibrate one’s emotions to the situation, e.g. chronic depression or psychothymia. Aristotle argues that individuals with these disorders can be enkratic, but not virtuous because they are unable to fully calibrate their emotions to the world (NE 7.5 1149a1-6).

Educating for enkratia remains challenging because it still requires developing the correct skills for determining what to do in various situations. The enkratic person is not a rule-follower, but is instead someone who is able to determine— at least on balance— how to act, even if her emotions are not appropriately calibrated to the situation. For this reason, no one-semester college class can develop enkratia; like with virtue enkratia needs to be something that students develop over time through the process of habituation. The rapid improvement of AI has increased the interest of departments and administrators, as well as the general populace, in ethics. I hope that we can generate enough momentum with this to provide ethics training in addition to incorporating this habituation and skill-building in ethical thinking throughout the curriculum, especially for those students who are training in technical fields.

An additional question is whether higher education is an appropriate place to start training in virtue. Smith is correct that Aristotle believes character is fixed after a certain point and that no amount of training will be able to change a fixed character (e.g. NE 3.5 1114a15-22, 1114b17-1115a3). In addition, Aristotle believes that a student without the proper upbringing can never be virtuous and that ethical training will not work on them (NE 1.3 1095a3-11,1095b5-12). However, Aristotle envisioned that students would begin the theoretical aspects of their ethical training at about college age, because students needed to be sufficiently mature to benefit from such training (NE 1.3 1095a3-11). In my experience, this is a fertile age for ethical training; college students are quite receptive to Aristotle and are particularly interested in, for example, using his ideas on friendship to evaluate their own lives. Yet, it is true that most of the college students I teach have lacked the sort of environment that habituates students toward virtue, and Aristotle would say they are therefore the wrong candidates for becoming virtuous. I think, given recent understanding of brain plasticity, it is clear that it is challenging, but possible, to develop
students for virtue— or at least enkratia— at a later time in their development than Aristotle envisioned.

Character is malleable at any age, but it becomes increasingly difficult to change as a person ages. Yet, recent research demonstrates that all sorts of traits, once thought to be fixed, including farsightedness caused by aging (presbyopia), are able to be altered through an intensive training process.\(^{34}\) Human brains are surprisingly plastic and retain the ability to make changes even after the preliminary bursts of neuronal growth and pruning that happen throughout childhood and periodically through adulthood, with a final major brain development in mid-to-late 20s.\(^{35}\) Even beyond this final period of brain growth, the brain continues to change and humans are able to harness that plasticity to learn new skills, form new habits, and innovate (Ericsson & Pool, 2016: 26-49.). Some skills are hard to form after childhood (such as language learning and perfect pitch\(^{36}\)). Character does not appear to have a hard and fast developmental cut off. While character development may become more difficult as adults age, I hypothesize that the reason character changes less is because the average person does not want to put in sufficient effort to make changes. I hypothesize that teaching brain plasticity and growth mindset\(^{37}\) to students would help encourage them toward character development at a later stage, even if that change is more difficult.

One possible objection to this claim is as follows: even if character change is possible, ethics courses are notoriously bad at changing student behavior. Studies of students in ethics courses show that students do not behave better in general (May & Luth, 2013; Schwitzgebel, 2013), and any successes are modest (e.g. Antes et al., 2009;\(^{38}\) Schwitzgebel et al., 2020). Schwitzgebel concludes that ethics education is ineffective and may, indeed, make people worse (e.g. Schwitzgebel, 2009; Schwitzgebel & Rust 2009, 2010). Indeed, Schwitzgebel thinks that ethics professors may act less well than those in other subdisciplines of philosophy (e.g. Schwitzgebel, 2009).

Proponents of Moral Foundations Theory (MFT) provide an explanation for the lack of effectiveness of moral education. Jonathan Haidt argues that students do not improve because we are trying to address the wrong part of them. (Haidt, 2012: 103-106) He argues that our moral decisions are primarily emotional and we only reason about them after the fact. (Haidt, 2012: 61-64) The model that moral foundations theorists use can be analogized to a rider on an elephant. The rider is the reasoning and the elephant is the emotions. Haidt explains this as the idea that the emotions lead our reasoning, at least for the most part, and much of what our reasoning can do for us is to provide ad hoc reasoning for our emotional decisions. As MFT proponents believe that emotions are the source of our moral reasoning, they think that ethics courses are worse than useless. Indeed, Haidt thinks they simply make people better at creating reasoning to justify their

\(^{34}\) Ericsson & Pool, 2016: 36-37.
\(^{35}\) E.g. see Pujol et al., 1993.
\(^{36}\) Concerning perfect pitch, see Ericsson & Pool, 2016: xii-xvi.
\(^{37}\) Dweck, 2016.
\(^{38}\) Antes et al.’s meta-analysis shows that ethics instruction was most effective when it was interactive and provided students with the skills to identify ethical problems and think through solving them. They also found that case-based exercises were effective, because they allowed students to practice ethical decision-making (2009: 397-398). These effective approaches are similar to the ethical pedagogy Aristotle recommends.
ethical decisions, particularly when those decisions put them at odds with other members of the community.

The account of MFT proponents fails to capture cases in which reasoning clearly influences ethical decisions. For example, lots of individuals change their views on ethical treatment of LGBTQ+ individuals not because of any emotional experience, but rather because they come to understand through reading or discussions with others (e.g. Sterelny 2012: 155-160). This undermines the contention of the most die-hard MFT proponents, who argue that reasoning has little power to overcome emotional responses and ethical behavior. Aristotle’s view that acting virtuously (or viciously) requires a combination of desires, emotions, and reasoning seems closer to what actually happens when we make ethical decisions than the almost exclusively emotional reaction that is proposed by MFT views.

Moreover, while Schwitzgebel’s studies—which Haidt (2012: 103-106) and other MFT theorists cite—are interesting, these studies employ a limited idea of what it means to be ethical. Studies need to trace some concrete and measurable behavior, so they tend to use proxies for ethical behavior such as donating to charity or refraining from spending money on meat products (e.g. Schwitzgebel, 2013; Schwitzgebel et al., 2020). These are crude principlist standards that provide little, if any, indication about a person’s character. Take donating to charity as an example. Donating to charity may be an instance of generosity, but only if the person in question donates the correct amount of appropriately-sourced resources to the correct cause, at the correct time. Studies which simply look at whether or not someone donated to charity (or even how much) cannot possibly capture the necessary ethical features of a situation; such a proxy is so crude as to be useless.

However, Schwitzgebel and the MFT theorists are correct that most ethics classes are indeed ineffective in training behavior. Changes in behavior cannot stem simply from learning principles or calculi or learning about various ethical theories. Rather, moral training can only happen through a combination of learning the skills to identify the salient ethical features of a situation, understanding what virtue is called upon, understanding how to act, correctly calibrating one's emotions to the world such that they motivate one to act correctly, and receiving guidance and feedback. In other words, moral learning is a sort of apprentice learning, where students must have access to a learning environment that is seeded for them such that they can practice moral decision-making (Sterelny 2012: 165-171). Most ethics classes are not structured in this way.

My hope is that we can create an ethics curriculum that will move students toward enkratia. While this is optimistic, I see potential for two reasons. First, I have seen first hand how students respond in the classroom to the sorts of teaching that we describe in this article. Second, Part of the reason is that the rapid improvement of AI has increased the interest of departments and administrators, as well as the general populace, in ethics. I have hope that this may lead toward popular support for the sort of curricular changes that we advocate. I engage each of these reasons in turn.

I begin with my own experience. I have taught a number of introductory ethics classes using Aristotle, upper division seminars focusing entirely on Aristotle, and AI ethics classes where we consider Aristotelian ethics as well as alternative pictures about how ethical skills are

---

quired. The responses from students are impressive and, in some cases, transformative. While Smith is certainly right that habituation cannot happen in a single semester in a college class, I think that regular engagement with this material may provide students with a starting point for becoming enkratic. Indeed, they acquire from these courses a set of skills which they can practice in the moral decisions that they make in their day-to-day lives. In the last part of my upper division ethics classes, I have students write a reflection paper about whether and how they think that the class has impacted them. While I teach them about Schwitzgebel and Haidt’s perspectives, the majority of the students state that the classes caused them to think differently about decisions they were making and many of them claim that these classes changed their behavior.

In technology ethics courses, the most prevalent topic students brought up in reflections were internships and jobs. The courses that I teach are upper division courses that are one way for computer science and electrical engineering students to fulfill their ethics credit. [Institution removed for anonymity], as a Catholic institution, requires that all students take a class designated as “Ethical Inquiry”. For this reason, the majority of students in these courses come from STEM disciplines and most of them have never taken an ethics class previously. One of the things that I have students do towards the end of the course is to write a reflection paper on what they gained from the course. In the reflection papers, I noticed something striking; several students were either considering or reconsidering where they wanted to work or intern based on the information from the class. A few students even told me that they regretted not taking the course earlier, because they would have liked to think more carefully before accepting post-graduation jobs where they weren’t sure how the organizations would consider ethics of AI. Some voiced their concerns about working for military contractors, weapons manufacturers, or big tech. Others were concerned more specifically about whether superiors would take them seriously if the students voiced ethical concerns. This seems like evidence not only against the MFT view of ethical change, but also in favor of teaching ethics at the college level to build ethical decision-making skills.

In addition, I want to consider the possibility of public support for infusing ethics into the college curriculum. The rapid improvement and projected ubiquity of AI has increased the interest of departments and administrators, as well as the general populace, in ethics. Some of this interest is reactionary– people are afraid of misuse and they are hoping that by forcing students to take an ethics class they might end up with better behavior. There are organizations, like the engineering grand challenges, that are trying to create people with technical know-how and some sort of ability to anticipate long term consequences. While I don’t think any of these solutions are good ones, I think that the inclination is right. Namely, we need to be training the next set of technically inclined students to be virtuous if possible, or at least to be something approaching enkratic. I have anecdotal evidence for such a position being persuasive to students and causing them to change how they think about their lives and sometimes alter their behavior (from a combination of student testimony and student evaluations).

In short, both authors propose that ethics training is the correct way to go about helping the next generation to live well with AI. We both acknowledge the difficulty in the proposed path. As universities regroup and reflect on their position in the face of ChatGPT and other AI tools, we hope that those in power take seriously both the possibilities and the challenges of an Aristotelian approach to living well with AI.
Acknowledgements
The authors would like to thank Kelly Clark, Earlene Fernandes, Kaite McKenna, Benjamin Pacini, as well as audiences at APPE 2023, SEAC 2023, NAAPE 2023 the BYU-I Tri-Ed Society, and the reviewers from Theory and Research in Education. We would also like to thank students from Computer Ethics (Fall 2021) and AI Ethics (Spring 2022, Spring 2023) as well as Dan Tigard and the students from ChatGPT and AI Ethics (Spring 2023).

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Nicholas Smith 0000-0002-7724-771X
Darby Vickers 0000-0003-3284-2520

Works Cited


Schwitzgebel, E. (2013 manuscript).”*Do Ethics Classes Influence Student Behavior?*”


**Author Biographies**

**Nicholas Smith** is an assistant professor at Alabama Agricultural and Mechanical University. His primary research interests are the intersection of virtue ethics and artificial intelligence, and the epistemology of religion.

**Darby Vickers** is an assistant professor of philosophy at University of San Diego. Her research focuses on questions at the intersection of ethics, epistemology, and philosophy of education. She draws much of the inspiration for her work on contemporary issues from ancient Greek philosophy, particularly Plato and Aristotle.