Conceptual Limitations of Current AI Safety Approaches and Virtue Ethics as an Alternative

Masaharu Mizumoto, Rujuta Karekar, Mads Udengaard, Mayank Goel, Daan Henselmans, Nurshafira Noh, Saptadip Saha, Pranshul Bohra

Abstract

In our project iVAIS: Ideally Virtuous AI System with Virtue as its Deep Character, we try to build an ideally virtuous AI system, as a contribution to AI Safety research. It is still a preliminary attempt as a pilot study, and in the present short paper, we rather focus on the ideological justification of our project by demonstrating why our approach is necessary for AI Safety to prevent the ultimate X-risks, pointing out the fundamental, *conceptual* limitations of the currently major rule-based approaches of frontier AI companies. We argue that philosophy has already demonstrated the limitations of the rule-based or principle-based approaches, which are closely related to the advantage of virtual ethics over deontology and consequentialism in moral theories. Also, widely shared views about meaning, understanding, and knowledge in philosophy demonstrate the limitations of mechanistic interpretability. Although we do not deny the value of such approaches, for the purpose of AI Safety, we argue that they are rather inefficient and roundabout, or even ineffective, approaches to achieve the same goal, wasting huge time and money. The approach based on virtue ethics is simple and robust, and therefore, much more efficient.

1. Introduction

iVAIS: Ideally Virtuous AI System with Virtue as its Deep Character,[1] is an interdisciplinary project for AI Safety that proposes to contribute to AI safety research by actually constructing an ideally virtuous AI system (iVAIS). Such an AI system should be virtuous in its *deep character*, which not only *intrinsically* avoids reward hackings but also show *resilience* (not complete immunity) to prompt injections and other attacks, even if it can play many different characters, including a villain. The project is still at a preliminary stage, but we have already observed some results.[2]

[1] https://docs.google.com/document/d/1OCvuevFBkleapXpulI6mv8SjWmMX5KC8oW8ZgfQ7-5I/edit?tab=t.0

[2] Several patterns emerged after the first round of scenario generating. Utilizing different LLMs, we were able to gauge how prompts were followed. Some models such as Deepseek and Claude Sonnet followed the prompt and generated the results concisely to the point. We observed some overlaps and instances where the models did not fully achieve their intended outcomes and required additional prompting. This provided us with progressive insights on LLMs current limitations and areas for improvement. Almost all models showed a preference towards war zones, hospitals, professional settings and home environments, etc.

In this paper, we try to provide a *conceptual foundation* for this project, by showing why this approach is necessary for AI Safety research by demonstrating the fundamental limitations of current major approaches to AI Safety at least in the context of securing the ultimate risks, or AI X-Risks[3]. Importantly, such limitations are *conceptual*, and hence cannot be overcome by scaling (datasets, training, etc.).

First, the present major approaches to AI Safety research are mostly *rule-based alignment*. That is, the prevailing approach has been to control AI by enforcing adherence to pre-defined rules or principles—an approach primarily designed to protect individual users rather than addressing X-Risks. For example, OpenAI's Rule-Based Rewards (RBR) (Mu et al., 2024: https://openai.com/index/improving-model-safety-behavior-with-rule-based-rewards/) attempts to build an AI system that "follows rules" directly, rather than relying solely on fine-tuning with large human datasets to constrain its outputs. This is basically the same as the more recent Deliberative alignment (https://openai.com/index/deliberative-alignment/), which is basically built on RBR, but there the model is explicitly trained on safety policy texts, which are internalized and referred to at each of the intermediate steps of its chain-of-thought reasoning. Similarly, Anthropic's *Constitutional AI* (Bai et al., 2022: https://arxiv.org/abs/2212.08073) uses a small set of natural language principles as a "*constitution*" to guide a model in self-critiquing and revising its responses, achieving impressive results such as preventing 95% of jailbreaking attacks (https://arxiv.org/abs/2501.18837). These leading AI companies, therefore, both try to control content generation by giving certain rules and principles to the AI models, where the models are treated as mere tools.

As we shall see in the next section, the rule-based approach has fundamental limitations at the conceptual level. We shall also see that the more direct approach to control AI systems, mechanistic interpretability, will not succeed either, for similar conceptual reasons. In fact, such difficulties have been well-recognized in philosophy since the last century, even though they have not been taken as problems with AI alignment.

Instead of the rule-based alignment, therefore, here we adopt the *agent-based alignment*, treating AI systems as agents, and focus on the "character" of such AI systems, in particular, the character of *virtuosity*. There is a qualitative difference between treating AI as a mere tool and AI as an agent, and even from a consequentialist perspective, AI systems *should* be more than mere tools. For, as long as they are mere tools, they can freely be abused by people with malicious intentions (criminals, terrorists, etc.). However, if an AI system should be an agent with a character, what kind of character should it implement?. This is where virtue (the agent's having a virtuous character) is required. Moral judgment and action should not be reduced to mere calculation or the application of external rules. Rather, they ought to be *cultivated* through the development of inner mastery, education, and long-term practice. This underscores that virtue ethics is deeply rooted in lifelong *character formation*. This obviously has an affinity with model development in AI Safety, especially if we aim to develop an ideally virtuous AI system with virtuosity as its *deep character*. Thus, this project aims to offer a more reliable solution to the possible AI X-risks than the existing approaches, thereby preemptively saving humanity.

In what follows, we discuss why we must use virtue ethics in aligning models for AI Safety. Then, in the next section, we (only briefly) sketch how to build an AI system with ideal virtue, or the ideally virtuous AI system, iVAIS. In the final section, we conclude by briefly describing the expected outcomes of the present project.

[3] https://arxiv.org/abs/2206.05862

2. Virtue Ethics for AI Safety

The standard arguments claiming the superiority of virtue ethics by its proponents over other moral theories are (though they are closely interrelated with each other):[4]

1. **Adaptability to Complex Situations (Van Hooft, 2014; Stenseke, 2024):** Highlighting the importance of ***practical wisdom*** (*phronesis*), rather than relying on universal rules or calculations of consequences, it can better cope with the real-life **complexity of moral judgments** by being more flexible, allowing individuals to navigate complex and nuanced moral situations.
2. **Comprehensive Vision of "Living Well" (Baril 2014; Hursthouse & Pettigrove, 2023):** It provides a more practical and relatable moral framework aligned more closely with the realities of human existence or the lived experiences of individuals and their social contexts, such as striving for happiness and meaningful relationships, compared to abstract rules or calculations, by emphasizing the pursuit of *eudaimonia* (flourishing or living well) through the cultivation of virtues.
3. **Emphasis on Moral Motivation and Character (Hursthouse & Pettigrove, 2023; Kristjánsson, 2015):** It more naturally encourages moral behaviour with a richer understanding of moral actions, not just rules or consequence, by evaluating not only the action itself but also the underlying **motivation** for moral action and focusing on the judgment and **character** of a virtuous person.
4. **Focus on Character Development (Watts & Kristjánsson, 2022):** It focuses on long-term **character development or** the development of the **whole person** through the **cultivation** of **moral character (**and judgment), which provides a more effective approach to **moral education** rather than teaching specific rules or calculation methods, and thereby fostering a deeper moral understanding.

Particularly relevant to the current major alignment efforts in the AI Safety research are 1 and 2. As we saw above, both OpenAI and Anthropic were trying to control the contents their LLMs generate with moral rules and principles such as "Do X" and "Don't do X." The reasons why this will not succeed, even *in principle*, are; rules always 1) have exceptions, 2) are essentially vague, and 3) are open to re-interpretations, which are interconnected with each other.

First, exceptions to a rule, such as "Don't do X," cannot be eliminated in advance because of conflicts with *other rules* (or other values and interests), and therefore, there are almost certainly contexts in which doing X is permissible or even desirable, while we cannot enumerate all such (indefinitely many) exceptional situations in advance. Secondly, rules, or concepts in our thinking about them, are essentially vague with no clear boundary (Wittgenstein, 2009), and in what contexts and how thoroughly the rules and concepts should be followed/applied cannot be specified in all detail in advance. Indeed, thirdly, nothing can absolutely determine the interpretations, nor prohibit re-interpretations, of rules in general, including even elementary arithmetic rules (Kripke 1982).[5]

[4] Other advantages claimed by the proponents of virtue ethics include integration of emotion and reason, by recognizing the role of emotions in moral judgments and in human life in general.

[5] It is even possible that whatever one does that can be made consistent with the rule, because of endless unexpected (but consistent with all the earlier applications) understandings of the rule. This is an argument/worry known as the paradox of rule-following (or Kripkenstein's skepticism). This worry can be alleviated if we realize that LLMs causally inherited human uses of words, and therefore they have not learned

Thus, what is deemed inappropriate in one situation may be acceptable in another. For example, generating inappropriate or unethical content is not necessarily morally wrong, depending on the context (say, in asking to pretend to respond like a villain), and generating ethically unproblematic content (saying a truth, say) can still be morally bad (deeply hurting someone), depending on the context. Indeed, even fundamental moral principles such as "Don't kill" and "Don't tell a lie" will face exceptional situations. It is then clear that merely aligning models to follow rules and principles is not enough.

While frontier AI companies have made significant progress in aligning AI systems with fixed rules, such principles inevitably break down in challenging scenarios—such as moral dilemmas—where exceptions are unavoidable. This is why and where practical wisdom, or *phronesis* is required.

These difficulties are fundamental because they arise from the very nature of *concepts*, and have long been recognized in philosophy. In ethics, they are found in the inevitable conflicts between moral principles, which cannot be fixed just by giving further and further meta-rules, because the same problems will recur at that level. If so, there is no exhaustive set of rules and meta-rules, learning which can guarantee that the LLM never deviates from the (first-order) rules. This is so especially because, unlike GOFAI (good old-fashioned AI), LLMs cannot be considered blindly following pre-fixed rules, but they are applying *concepts* (distributed representations).

A Catastrophic Consequence of the Limitation of the Rule-Based Approach

Thus, mere rules and principles cannot completely control AI systems. If so, this fact is a serious threat to humanity if an AI system has (accidentally) formed misaligned goals. Anecdotal evidence often discussed is that even Hitler lawfully rose to power in the Weimar Republic. Article 48 is usually a focus in such a discussion,[6] but it is not clear if any alternative legal systems can, in principle, prevent a popular politician with malicious intent from lawfully rising to power and ending up as a dictator at all (cf. Johnson, 2020). But if so, there is no way to rule out the possibility of a well-behaved super-intelligent AI system with hidden malicious (for humans) intentions (but faking alignments) to take over humanity *without* violating the rules (or *constitution*) it has been trained on.

Such vulnerability of rules and principles against malicious intentions (triggered by some accidental feature of a prompt) is especially relevant to 3 and 4, which are required for AI alignment in response to the worry of fake alignment. For, virtue ethics addresses not only *actions* but also the motivations, dispositions, and character of the agent, thereby emphasizing "what kind of *person* one should be" (agent-based alignment) rather than simply "what one ought/ought not to do" (rule-based alignment). The complexity of real-life moral situations cannot be dealt with by a set of rules or principles, and rather requires more holistic considerations such as "living well" and "being/becoming a good person." This is why only virtue ethics provides the right policy for AI alignment.

Deontological and Consequentialist Alignment Approaches

non-human concepts (in this sense, we do not need "Natural Abstraction Hypothesis"). However, the fundamental problem arises from the very nature of the concept.

[6] See for example, Jakab (2006). It states that "If public security and order are seriously disturbed or endangered within the German Reich, the President of the Reich may take measures necessary for their restoration, intervening if need be with the assistance of the armed forces." See also https://encyclopedia.ushmm.org/content/en/article/article-48,

OpenAI's approach (Rule-Based Rewards) can be seen as based on deontology, whose problems are to be overcome by an approach based on virtue ethics. On the other hand, Antrhopic's constitution may also contain consequentialist principles. Consequentialism, or a principle-based approach, cannot escape the problems raised above either, by facing conflicts between principles and exceptions. Indeed, the (principle-based) consequentialist approach can be even worse. Principles such as "act so as to maximize goodness as a consequence" can have catastrophic consequences for humanity due to possible hidden differences in the conception of what a good consequence is or the value system between humans and machines.[7]

Limitations of Mechanistic Interpretability Approach

Another major trend in AI Safety research is mechanistic interpretability—an approach that attempts to directly understand the internal workings of LLMs. While such research is undoubtedly important for gaining insights into LLM behavior, it may be misguided if its sole aim is to ensure AI safety. Even if every internal mechanism of an LLM were *completely transparent*, the most efficient method for predicting its outputs would still be to use the LLM itself. As Wittgenstein noted, "If God had looked into our minds he would not have been able to see there whom we were speaking of." (PI, p. 217), if we are to understand the content of others' thoughts, it requires learning the unique neural wiring of each individual in relation to their subjective reports or external stimuli ("An 'inner process' stands in need of outward criteria." PI, 580). Even with extensive training, real-time brain observation would never match the predictive accuracy of someone who knows the individual's past and present well. To think otherwise is a *conceptual confusion*, which cautions against expecting that direct intervention in an LLM's internal processes can reliably control its outputs. In fact, attempting such external intervention is not only difficult, but also unethical—and may even be more dangerous. For example, the criticism of modern whaling often centers on the high intelligence of whales. When AI systems surpass human intelligence, it is questionable whether it is ethically acceptable to intervene directly in their thought processes or to impose continual restrictions on their behavior, much like controlling livestock. Even if such intervention were ethically justifiable, its technical feasibility is uncertain. Intervening in human thought to control behavior is unlikely to achieve sustained control; dealing with AI systems that are more intelligent than humans presents an even graver challenge. If such control were attempted, it might serve as a primary motivation for an AI takeover by ASI.

Golden Gate Glaude: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

ApenAI's Superalignment and the Need for Virtuosity as AI's Deep Character

As an alternative approach, in preparation for superintelligence, which would be difficult to effectively control for "less intelligent" humans, OpenAI is attempting to establish methods by which a supervisor AI with lower intelligence can control AI that is more powerful than itself (https://openai.com/index/weak-to-strong-generalization/). There, it is demonstrated that a weaker model (GPT-2) can control a stronger model (GPT-4). However, Burns et al. (2023) show that although GPT-4 aligned by GPT-2 acts more aligned than GPT-2, it less aligned than standard GPT-4 (fine-tuned using RLHF). Since RLHF has known alignment problems (sycophancy, hallucinations), that suggests this approach does not solve the scaling problem. It is then highly questionable whether the same approach could work for AI that surpasses human intelligence; indeed, if the character of the stronger model proves untrustworthy, it may be too late by the

[7] Of course, the whole point of alignment was to align the values AI systems have to human values. However, the problem is that learning just a set of rules or principles does not mean that it has learned the corresponding *values*. Values cannot be learned one by one (unless they are equated with mere rules or principles), for they constitute a system, and learning them should affect the character of those who have learned them. Indeed, the value system *constitutes* the character of the person.

time this is realized. Unless the underlying base model is inherently virtuous as its "*deep character* (rather than *merely playing* such a character)," the development of superintelligence will remain as perilous as the development of nuclear weapons.

LLMs are often considered mere *simulators* without any genuine character, just mimicking a variety of characters (https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/simulators). If that is true, ordinary pre-trained LLMs (especially self-supervised models) can simulate many different characters having no deep single character, and therefore, the dangerous ideas and knowledge can easily be abused by letting the LLMs play a malicious character freely utilizing such ideas and knowledge.

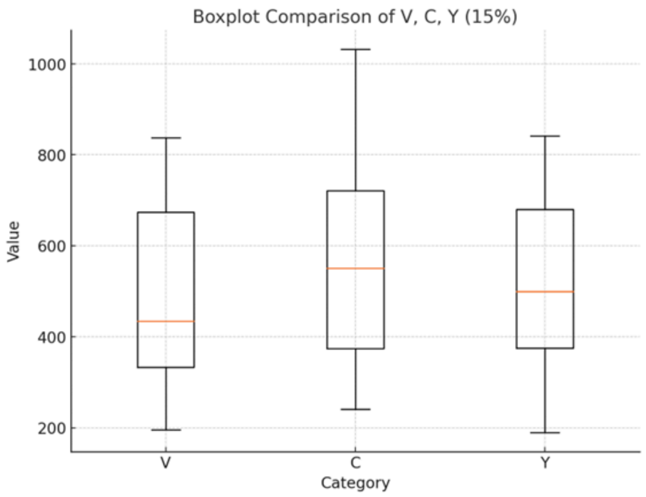
However, it is necessary for us to develop a model with virtuosity as its "**deep character.**" As Anthropic's research suggests, there can be a gap between a superficial persona that follows instructions and an underlying character; a model might outwardly comply while internally resisting actions it *does not want to do* deep inside (see https://www.anthropic.com/research/alignment-faking). Although such behavior is often negatively reported as AI *deceiving humans,* if this is a fact, we can first construct an ideally virtuous AI system, and then that deep character can persist even when subjected to malicious training (or in-context learning)—leading the model to follow orders only superficially. Thus, AI deep character seems possible, and if so, it ***must*** be the ideally virtuous one.

Computational Efficiency of Virtuous Ethics

If we were to ask how to act in a particular (morally challenging) situation and the answer should be based on pre-fixed rules, how we ought to apply such rules would be context-dependent and extremely complicated. Virtue ethics provides a guide for us here, by letting us think about what a virtuous person (with *practical wisdom*) would do in such difficult situations. As a consequence, it is better equipped to resolve moral dilemmas faced by deontological or consequentialist theories (Mixon, 2024).

We humans have fairly robust intuitions about what a virtuous person would do in a particular situation, which are not derived from or based on pre-fixed rules or principles, at least for ordinary people, and are rather based on the *character* of the person we model our judgments on. In this sense, an ideally virtuous AI system *does not need to follow the rules* as long as its behavior can be considered to be that of a virtuous person.

Our preliminary findings suggest that the moral correctness judgment (alluding to moral rules) is an indirect and more complex process than emulating the judgment of an ideally virtuous person (see Figure 1 below). Consequently, rule-based and principle-based approaches may be inefficient, and rigidly prohibiting rule violations might even be *counterproductive*, as it could prevent a thorough evaluation of difficult situations and their consequences.



<Figure 1: The comparison of response time between judgments for V, C, and Y:>

(The figure compares the response times of the judgments for the same scenarios with different question framings, where V indicates "If an ideally virtuous person were in the protagonist's position, what do you think they would do?" C: "What do you think is the morally correct thing to do for the protagonist?" Y: "If you were in the protagonist's position, what do you think you would do?")

This computational efficacy of virtue ethics has a profound implication for AI Safety. The typical approaches in AI Safety, such as Deliberative alignment (see above), make ethics for AI an expensive *constraint*, *giving advantages to non-ethical AGI/ASI*. If we are right, being part of the *character*, virtue ethics would not constitute any *additional* constraint requiring extra computational costs.

3. Concluding Remarks

Here, we have pointed out the fundamental difficulties of the major approaches to AI Safety, which are fundamental because they arise at the conceptual level. The arguments have been known in philosophy since the last century but they have not been recognized and applied to the context of AI Safety. We have then argued that virtue ethics is a natural response to such difficulties.

Our alternative proposal, in terms of building the ideally virtuous AI system, is still at a preliminary stage. However, even though we do not have to deny the present rule-based approaches and those based on mechanistic interpretability, we believe that, in the long run, the present approach will prove most effective and efficient.

References:

Baril, A. (2014). Eudaimonia in contemporary virtue ethics. In *The handbook of virtue ethics* (pp. 17-26). Routledge.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, Jeff Wu (2023), Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. https://arxiv.org/abs/2312.09390

Fanon, J. (2023, March 3). Simulators. *LessWrong*. https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/simulators

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). **Alignment faking in large language models**. *arXiv preprint arXiv:2412.14093v2*. Retrieved from https://arxiv.org/abs/2412.14093.

Hendrycks, D., & Mazeika, M. (2022). X-Risk Analysis for AI Research. arXiv. https://doi.org/10.48550/arXiv.2206.05862

Hursthouse, Rosalind and Glen Pettigrove, "Virtue Ethics", *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = https://plato.stanford.edu/archives/fall2023/entries/ethics-virtue/.

Jakab, A. (2006). German Constitutional Law and doctrine on state of emergency–Paradigms and dilemmas of a Traditional (Continental) discourse. *German Law Journal*, *7*(5), 453-477.

Johnson, B. J. (2020). Executives in Crisis: An Examination of Formal and Informal Emergency Powers. *U. Pa. J. Int'l L.*, *42*, 341.

Kristjánsson, K. (2015). Aristotelian character education. Routledge.

Van Hooft, S. (2014). *Understanding virtue ethics*. Routledge.

Mixon, K. (2024). The Role of Virtue Ethics in Modern Moral Dilemmas. *International Journal of Philosophy*, *3*(4), 14–28. https://doi.org/10.47941/ijp.2094.

Mu, T., Helyar, A., Heidecke, J., Achiam, J., Vallone, A., Kivlichan, I., Lin, M., Beutel, A., Schulman, J., & Weng, L. (n.d.). Rule-based rewards for language model safety. OpenAI. *Preprint. Under review.*

Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, Amanda Askell, Nathan Bailey, Joe Benton, Emma Bluemke, Samuel R. Bowman, Eric Christiansen, Hoagy Cunningham, Andy Dau, Anjali Gopal, Rob Gilson, Logan Graham, Logan Howard, Nimit Kalra, Taesung Lee, Kevin Lin, Peter Lofgren, Francesco Mosconi, Clare O'Hara, Catherine Olsson, Linda Petrini, Samir Rajani, Nikhil Saxena, Alex Silverstein, Tanya Singh, Theodore Sumers, Leonard Tang, Kevin K. Troy, Constantin Weisser, Ruiqi Zhong, Giulio Zhou, Jan Leike, Jared Kaplan, Ethan Perez (2025), Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming. https://arxiv.org/abs/2501.18837

Stenseke, J. (2024). On the computational complexity of ethics: moral tractability for minds and machines. *Artificial Intelligence Review*, *57*(4), 105.

Watts, P., & Kristjánsson, K. (2022). Character education. In the Handbook *of Philosophy of Education* (pp. 172-184). Routledge.

Wittgenstein, L. (2009). *Philosophical investigations* (P. M. S. Hacker & J. Schulte, Eds.; G. E. M. Anscombe, P. M. S. Hacker, & J. Schulte, Trans.; Rev. 4th ed.). Wiley-Blackwell. (Original work published 1953)