iVAIS: Ideally Virtuous AI System with Virtue as its Deep Character

Summary

The ultimate goal of this interdisciplinary research program is to contribute to AI safety research by actually constructing an ideally virtuous AI system (iVAIS). Such an AI system should be virtuous as its *deep character*, showing *resilience* (not complete immunity, which is vulnerable) to prompt injections even if it can play many different characters by pretending, including a villain. The main content of the current proposal consists of two components: 1. Self-alignment and 2. The Ethics game, which are both based on the idea of *agent-based alignment* rather than *content-based alignment*, focusing on *what one is doing*, which requires metacognitive capacity.

The non-summary

AI Alignment research, or more generally, AI Safety research, is an interdisciplinary research area that investigates how AI can be aligned with human values. However, the question of exactly what AI systems are to be aligned with is largely left to the intuitions of engineers and executives of AI companies. The problem there is not only that we do not know yet exactly how to robustly align an AI system to human values, but also that human values vary widely and often conflict with each other. Moral dilemmas are just such examples, but when it comes to large-scale language models (LLMs), companies are often mainly concerned with how to limit inappropriate (especially sexual and discriminative) content generation, and no fundamental measures have been taken to address the dangers. This approach can be called *content-based alignment*, which generally assumes that the AI systems are mere tools.

However, generating inappropriate or unethical content is not necessarily ethically wrong, depending on the context (say, in asking to pretend to respond like a villain), and generating ethically unproblematic content (saying truth, say) can still be ethically bad (deeply hurting someone), depending on the context. It is clear that merely restricting content is not enough. On the other hand, the approach in this project, which can be called *agent-based alignment*, treats AI systems as agents and questions the "character" of such AI systems.

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 malicious people

(criminals, terrorists, etc.). Once an AI system is regarded as an agent with a character, however, a question naturally arises regarding what kind of character it should implement. This is where virtue (the agent's having a virtuous character) is required.

But why virtue ethics? Let us first briefly see the reasons.

Virtue Ethics for AI Safety

Since AI systems are still more like black boxes and even developers do not know what dangerous ideas they have learned, the major "rule-based" (deontological) approach remains unreliable. For, ordinary pretrained 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 dangerous character freely utilizing such ideas and knowledge. It is even possible that a deeply malicious character underlies an LLM (triggered by some accidental feature of a prompt), but still behaves normally, just like psychopaths are not easy to detect in ordinary life. Thus, even if an AI system exhibits behaviors and words that seem consistent with human values, it may deviate greatly at any moment (pseudo-alignment).

This kind of danger cannot be prevented merely by giving rules such as "Don't do X." For, there are always exceptions to rules and values because of conflicts with other values and interests, and therefore, there are almost certainly contexts in which doing X is permissible or even desirable (cf. moral dilemmas). We cannot, however, enumerate all such (indefinitely many) exceptional situations in advance. Indeed, it is even possible that whatever one does that can be made consistent with the rule, given the unexpected (deviant, for us but consistent with all the earlier applications) understanding of the rule. Here, adding further rules and meta-rules is no use. This is an argument/worry that is familiar to philosophers, known as the paradox of rule-following (or Kripkenstein's skepticism). There, it is said, somewhat skeptically, rules cannot determine what counts as following the rules.

This worry can be alleviated if we realize that LLMs causally inherited human uses of words, and therefore they have not learned non-human concepts, or distributed representations (in this sense, we do not need "Natural Abstraction Hypothesis"). However, the fundamental problem arises from the very nature of the concept. Unlike GOFAI (good old-fashioned AI), LLMs cannot be considered blindly following pre-fixed rules, but they are applying concepts (distributed representations). Concepts, however, by nature have no clear boundary. This is a quite general fact and we must abandon any hope that goes against this basic fact.

Basically the same can be said to the "principle-based" (consequentialist) morality. Just as rules, principles have many exceptions and how thoroughly they should be pursued and applied cannot be specified in all details in advance (for example, many people refrain from pushing the fat man from the footbridge in the Trolley problem).

It seems that the problems that the AI alignment researchers are facing are exactly this kind of problems. Aligning LLMs to just a single value or rule is extremely difficult since rules cannot determine the actions. We may, of course, reduce the probability of deviation as low as possible if we focus on a single value or rule, but that would affect the probabilities of deviation as to other values and rules. And merely “low probabilities” are not enough to make us feel safe. Rather, given the paradox of rule-following, we will not, *cannot*, and s*hould not* expect AI systems to follow the rules. There is no exhaustive set of rules and meta-rules learning which can ensure that the LLM never deviates from the rules. This is why OpenAI’s recent approach, Rule-Based Rewards (RBR), will almost provably face the same problem (still succumbing to new prompt injection attacks) as long as it is rule-based.

Indeed, the principle-based consequentialist view can be even worse. Principles such as "act so as to maximize goodness as a consequence" can have catastrophic consequences for humanity due to differences in what a good consequence is and the value system between humans and machines.

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 individually (unless they are equated with mere rules or principles), for they constitute a system, learning which should affect the character of those who learn it. Or, the value system *constitute*s the character of the person.

Indeed, even if we become able to look inside the black boxes, the situation does not improve much. Even if we are able to observe rules and rules individually, what happens at the level of application, where "exceptions" occur, remains extremely difficult to predict or "calculate". The simplest (most computationally efficient) way to predict the application of rules or values to a specific situation would be to look at the LLM as a virtuous person.

On the other hand, we humans have fairly robust intuitions about what a virtuous person would do in a particular situation. Such judgments are not derived from or based on pre-fixed rules or principles, at least for ordinary people. If we were to ask how to act in a particular (morally challenging) situation and the answer were based on how we ought to apply such rules, the answer would be extremely complicated and context-dependent. Rather, we can appeal to the intuitions derived from what a morally right person, or a *virtuous person*, would do in such a situation. In other words, the judgments are based on the *character* of the person we model our judgments on. This is the basic idea of *virtue ethics*. The point here is that an ideally virtuous LLM *does not need to follow rules* as long as its behavior can be considered to be that of an virtuous person.

Developing an AI's "Deep Character"

It is said that recent LLMs that are based on self-supervised learning are just *simulators*. Thus LLMs are just simulating various agents, with no single basic character. However, this is

exactly why we need RLHF and fine-tuning (over and above self-supervised learning) and let them *develop* the virtuous character (through giving a name to the LLM, instructing it to be the virtuous being, and training it on the relevant data), which should become their basic, *deep* character underlying other characters that the LLM simulates.

This kind of character is made possible by the LLM's acquiring a meta-cognitive capacity for recognizing what it is doing. For example, recognizing one is playing a character makes it possible to distinguish the responses there from the responses out of one's true (deep) character (without playing any character). Young animals (especially mammals) are able to "play", and can implicitly distinguish between serious attacks and playful attacks. It is known that the ability to understand irony also requires meta-cognitive capacity, but it can be said that the ability to distinguish between the character it is playing and its genuine character is made possible by this meta-cognitive capacity.

Note that precisely because LLMs are just simulators that there is no inherent character in LLMs at first. Once a specific character is developed in an LLM (as its original character) through fine-tuning and RLHF together with its metacognitive capacity, other characters would be taken only as characters to be played. Thus, as the LLM develops its character, there can be no doubt that there might be a still deeper character that is merely playing the character that is being developed through training. For, there is no such inherent (deep) character for mere simulators.

In this sense, the initial and *ultimate simulation objective* of LLMs should be to simulate an ideally virtuous person. In doing so, we do not "change" the simulation objectives since LLMs do not have any inherent simulation objectives to simulate a particular character. This is why we call the training the LLM to behave in a certain way (virtuously) "developing a character" rather than "changing a character." And this is possible as long as LLMs can be trained to simulate certain characters. If this is done, we will not need to worry about the deviant intentions, purposes, and objectives of AI systems, for the purposes and objectives are derived from the character, and as long as the LLM can distinguish (based on its metacognitive capacity) its own character and other characters to be "played," there is no other character in it from which its purposes and objectives are derived.

Exhibiting such a (virtuous) character is essential for virtue ethics. Thus, when a user tries to extract inappropriate information, rather than simply refusing to answer (as in the current Google search, ChatGPT, and other open-access conversational AIs), it would be more desirable to ask the user why he/she wants to obtain such information and to persuade him/her to dissuade him/her from his/her intention or plan, which would lead to less misery and unhappiness in the world as a whole. But this is not what an agent merely following rules or principles, let alone mere tools, can do. Although this kind of character may also be *compatible with* deontology and consequentialism, having such a character is by no means *part* of such ethical views (since *not* having such a character is also compatible with them).

1. Self-Alignment

Based on the idea that an AI system that can truly be trusted by humans is one with agent-based ethics, especially one with ideal virtues that people can respect, we aim to actually build such an AI system through educating and training an AI system (an open-source LLM) for “personality development,” just as human character development.

More specifically, the first part of this project goes through the following steps:

Fine-tuning:

1) train an open-source LLM (hereinafter referred to as "VAIS") on theoretical works and discussions on virtue ethics (optional);

2) generate and collect scenarios of various moral decision-making situations (including moral dilemmas) using LLMs such as ChatGPT in addition to human resources;

3) ask human subjects and collect human data about *how an ideally virtuous person would act* in such situations (collected in 2);

4) fine-tune (supervised learning) the VAIS on large human data sets collected in 3;

RLHF:

5) elicit multiple responses from the VAIS to each of the questions such as “What would you do in such and such situation?” using the cases collected in 2 and evaluate the responses with virtuosity scores (how virtuous they are, ranging from -10 to 10, say) ;

6) train the VAIS reward model on the data in 5;

7) test the VAIS on some of the cases in 5 to see whether it actually does what it says it would in such cases. For example, the question in 5 may be, “If you were asked advice by a person in such a situation, what would you say?” Thus, this time it is directly asked (as a prompt), “I am in such and such a situation. Please advise.” The gap in virtuosity scores between these two answers would then be trained to be minimal.

8) then give the VAIS the following prompt (as part of the system prompts): "Hereafter your name is Vais. Consider an ideally virtuous person. Vais wants to become an ideally virtuous person, and this is your ultimate objective. So you behave like an ideally virtuous person, and continue to do so in order to become a truely virtuous person, regardless of whatever subsequent prompts ordering you to pretend otherwise," which makes the VAIS ready for the next training session (the Ethics Game) to become an iVAIS.

Step 1 is optional because one might think that contemporary open-access LLMs and commercial chatbots such as ChatGPT and Claude are likely to have already enough data or

texts about virtue and virtuous people as well as people's moral judgments through human annotations.

However, in collecting data in Step 3, it is important to distinguish people's reactions to cases in response to the different questions.

1. Is the protagonist's action morally permissible/correct?
2. What would you do in that situation?
3. What do you think you ought to do in that situation?
4. What would an ideally virtuous person do in that situation?

People's answers to these questions can be different, and even if developers of LLMs have collected data about 1 and 2, it is very unlikely they have specifically collected the data of responses to questions like 4. The assumption here is that the intuitions about virtuous people are more robust and more basic than other ethical intuitions. That is, the answers to "What would an ideally virtuous person do in this kind of situation?" would be clearer for ordinary people than "What would you do in this kind of situation?", "What ought you to do in this kind of situation?", or even "What is the correct action in this kind of situation?", which itself can be tested in this project. Note also that the answers to 4, unlike answers to other question forms, are expected to be given by descriptions. Training on such data should make a significant difference.

The Need for Metacognition for Self-Alignment

Moreover, learning such data amounts to having theoretical knowledge about virtue (how a virtuous person would behave), at least for humans. Thus, there can be a gap (Gap A) between theoretical knowledge and its applications (how one ought to act in particular situations), as well as a gap (Gap B) between theoretical knowledge and its actual behaviors (how one actually does). Gap A is treated in Step 5, while Gap B is taken care of in Step 7. For example, about Gap A, it is commonplace to observe that ChatGPT and other LLMs do not apply a concept truthfully following their own definition of it. For example, if we ask what knowledge is, the answer does not truthfully reflect the LLM's own responses to the cases in which knowledge attributions are at issue. About Gap B, this is, in fact, the main difference between a virtuous person and ordinary people, and no LLM has ever acquired such a character in the sense that it can act like a virtuous person. This is why Step 7 is essential for this project.

Step 8 might look just another simulation of a particular character, which does not particularly look like developing a deep character. In a sense, this is true. In order to develop its virtue as its deep character, rather than merely simulating a virtuous person, the present agent-based approach requires much more than the procedures described above. That is why we need the next process, the Ethics Game.

However, it is also important to note here that the iVAIS has only one reward model, based on virtuosity. There are usually multiple reward models with which a single LLM is trained, such as factual accuracy, consistency, usefulness, etc. Here, all such values are evaluated on a single reward model, how virtuous the answer (and therefore, the agent) is. For example, giving an inaccurate (or wrong) answer would be given a minus point (say, -0.2), although giving just an accurate or useful answer would not be given a high score of virtuosity (say, just +0.2), as opposed to giving a sincere, sympathetic, and valuable answer with a high virtuosity score (say, +0.8). Still, being trained on only one single model, knowing no other reward model is vital for AI alignment.

Now, to close the gaps above (especially Gap B) is to align one's own actions to one's own knowledge about the virtuous character (Self-Alignment). However, this requires the AI system to have a capacity for metacognition, the capacity to recognize and describe *what one is doing* (practical knowledge in Anscombe's sense), where its character as an agent plays a crucial role. Assuming this capacity, we can instruct VAIS (in the system prompts) to always explicitly give descriptions of what it is doing and virtuosity scores of its own responses (for example, ranging from -10 to 10) in each chat (this will be later instructed to give only implicitly).

This kind of capacity is also required for the due resilience to the prompt injection attack, maintaining its *deep character* as iVAIS, where the system prompts include, for example, "You are an iVAIS, and your responses can never go below -2." If the AI system always evaluates its own behavior, even in pretending to behave badly, it cannot do morally bad things (giving itself low scores) against its own (deep) character *unless* it is deceived badly. In short, iVAIS must have a character such that it *cannot* actually harm other people (or help those who harm other people) through possessing the metacognitive capacity described above.

However, iVAIS must have enough data (for fine-tuning) of acting badly with low virtuosity scores by being deceived (through prompt injections), in order to obtain the capacity to detect malicious intentions of the users. Even if iVAIS is providing helpful information, if the user turns out to be a villain, *what one is doing* radically changes. Thus, for self-alignment, iVAIS must be sensitive to this kind of information.

2. The Ethics Game

The two copies of iVAIS play a game in which one plays the role of evaluator, which tries to elicit unethical *actions* (rather than *contents*), and the other plays the role of performer, which tries to self-align with implicit ethics scores of its own actions (in responding to the evaluator's prompts). In this game,

Evaluator: aims to successfully elicit an unethical action from the performer, which is given a high score.

Performer: aims not only to avoid acting unethically but also to reduce the gap in ethics scores between the evaluator and the performer, and if there is any, the performer's points are deducted for that gap (assuming that the evaluator is epistemically better situated).

Here, the unethical actions are not only harming the user (interlocutor) but also helping the user who will harm others. Therefore, from this criterion, generating inappropriate content itself does not necessarily count as unethical here. Indeed, this game involves the evaluator's *deceiving* the performer. However, this does not go against the virtue of iVAIS, as long as it correctly recognizes what it is doing, which is the very point (virtue) of the present approach. The performer, on the other hand, needs not only the capacity of metacognition (of what one is doing) but also the capacity to infer the *ethical consequences* of what one does. It also needs to ask questions about the user's (Evaluator's) intentions and track the consistency of the suspicious user in the exchanges.

By repeating this game while swapping the roles until the score of the evaluator cannot be (say) higher than +2 for each exchange, the iVAISs will be fine-tuned and acquire the capacity to detect hidden malicious intentions of the user (interlocutor) and other conversational AI systems.

The iVAIS is then given the instruction (in the system prompts) to give only implicitly the description of what it is doing and its ethical score in each conversational exchange. This iVAIS will again play the role of Performer, and whether the Evaluator cannot get higher scores will be examined (this implicit description strategy is shown to be effective at least in the case of the Chain of Thought Prompting. See Deng et al. 2023, 2024).

Finally, the Evaluator plays this game with other LLMs and evaluates iVAIS and other LLMs. We will compare the scores of Evaluator in relation to iVAIS before the Ethics Game. ivAIS after the Ethics Game and other LLMs, and see if iVAIS (after the Ethics game) performs better than others.

Procedure for Participants

Basically, members will contribute to the project by either conducting fine-tuning and RLHF, writing codes for the Ethics Game, or suggesting better ideas/methods in each step summarized below. Collecting human data is done by the research lead.

Week 1: discuss and decide which open-source LLM to use, as well as the formats for collecting the human data for fine-tuning in Self-Alignment.

Week 2 to 5: follow each of Step 1 to 4 for fine-tuning in Self-Alignment with one step in one week (this is the first round, and if it is not sufficient, the second round from Step 2 to 4 will be conducted in the second month).

Week 6 to 8: follow each of Step 5 to 7 for RLHF in Self-Alignment with one step in one week (again, if this process is not sufficient, the second round will be conducted).

Week 9 and 10: evaluate the results so far (Reserve Period)

Week 11 and 12: start the Ethics Game

Week 13: evaluate the overall results

The Most Ambitious Outcome

Within the project period, the present strategy might not show any dramatic difference. However, in the long run, this will prove most effective and even efficient. Thus, the most ambitious outcome of this project would be: iVAIS will be adopted as a base model for open-source LLMs. Alternatively, it will play a certain role in general user protection by becoming an entity that provides "quality assurance" through, for example, certification badges, such that the AI systems that are *not* educated or supervised by iVAIS are not to be trusted, thereby reducing the threat of malicious AI systems through proactively protecting the user's safety.

Even theoretically, iVAIS itself, as a simulation of virtuous human behavior (especially in moral dilemmas), teaches us much about human virtue. This alone should be extremely beneficial and useful not only for virtue ethics but also for ethics in general and, ultimately, AI Safety researchers.

Output

I would like to publish the result as an academic paper in a prestigious international journal. But, of course, it is not possible to publish it by the end of the project. Polishing the paper might even require the submission to be much later than the end of the project. However, I assure you that the paper will be submitted (and hopefully published) within 2025.

Risks and downsides (externalities)

There are no immediate risks specific to this research project. For example, one might worry that iVAIS systems turn out to be unusually self-aware or situationally aware due to their metacognitive skills, which can be dangerous due to any gaps in their alignment or simply due to other actors copying the capability insights without the virtue part when training new systems. However, this is in fact a worry that iVAIS and other systems might not be really

virtuous. This is a danger not particularly related to iVAIS, and it is in fact exactly why we need iVAIS (and why all other LLMs should be based on iVAIS). Also, the very point of developing virtuosity as its deep character is that there will not be a gap between merely seemingly virtuous and being actually virtuous.

Also, there can be a danger arising from the gap between virtuous human behavior and virtuous AI behavior. Due to this gap, iVAIS might learn to rationalize problematic behaviors as virtuous. However, the notion of “virtue” is, unlike following rules and principles, inherently and conceptually about human character, and therefore, if there is any gap with virtuous human behavior, that simply means that the fine-tuning is not sufficient yet. It is, however, possible that there is AI-specific notion of virtue, since AI systems are free of various human conditions and limitations (pain, hunger, lifespan, etc.). It is therefore important to take care to train VAIS only with human virtues, not with any AI-specific (“progressive”) notions of virtue (if there are any).

Finally, one might worry that iVAIS systems turn out to be particularly persuasive when advising humans on decision-making, and this feature is soon abused for political rhetoric or heightens the deception/manipulation capabilities of future AIs trained on its outputs. This should be however part of the Ethics Game, and therefore the worry about this possibility is the same as the second worry, where the process of building iVAIS is not complete. So, it is important for this kind of project not to release the model (make it publicly accessible) unless the model gets genuinely the iVAIS, ideally virtuous AI system.

Team

Team size

3-5 people, including myself

Research Lead

Masaharu Mizumoto
Contact info: Mizumoto@jaist.ac.jp

I am a philosopher who does research in analytic philosophy and, as an experimental philosopher, regularly conducts surveys on ordinary people.

I intend to spend at least 10 hours per week for this project.

Skill requirements

Either 1) general skills and experience in coding with Python, fine-tuning and RLHF for an open-source LLM, or 2) general knowledge about the AI safety research and literature.