Forthcoming in *Asian Journal of Philosophy*

The Trustworthiness of AI

: Comments on Simion and Kelp's Account

Dong-yong Choi

[Penultimate Version] Please do not cite without permission

Abstract Simion and Kelp explain the trustworthiness of an AI based on that AI's disposition to meet its obligations. Roughly speaking, according to Simion and Kelp, an AI is trustworthy regarding its task if and only if that AI is obliged to complete the task, and its disposition to complete the task is strong enough. Furthermore, an AI is obliged to complete a task in the case where the task is the AI's etiological function or design function. This account has a strength in that it can provide an unificatory rationale for the trustworthy-making properties of AIs. According to this account, being explainable, being safe, and being transparent are the trustworthy-making properties of an AI because an AI can fulfill its etiological or design functions in the case where it is explainable, safe, and transparent. This paper shows that though Simion and Kelp's account has a strength, this account is not satisfactory for two reasons. The first reason is that an AI's trustworthiness is not determined just by the AI's disposition to meet obligations, and the second reason is that it is difficult to explain how an AI's etiological function and design function have to do with that AI's obligations. To provide a full-fledged account, Simion and Kelp should dismiss these concerns.

Keywords AI Trustworthiness; Disposition to Meet Obligations; Etiological Function; Design Function

1. Introduction

The trustworthiness of AIs is a practical issue as well as a theoretical issue. For instance, if a patient believes that a medical AI is not trustworthy regarding its treatment suggestions, then that patient might not follow that AI's treatment suggestions even in the case where the treatments are beneficial for the patient herself. Therefore, to fully enjoy the advantages of AI technologies, people need the correct account of AI trustworthiness. If the correct account verdicts that an AI is trustworthy regarding its task, then people might be able to fully enjoy the benefits from the AI. This paper purports to examine Simion and Kelp's account of AI trustworthiness. Roughly speaking, in their article "Trustworthy Artificial Intelligence," Simion and Kelp claim that regarding a task an AI is trustworthy if and only if that AI is obliged to complete the task, and its disposition to meet the obligation is strong enough. Furthermore, an AI is obliged to complete a task if completing the task is its etiological or design function. Depending on the result of this examination, people might be able to use Simion and Kelp's account to check the trustworthiness of AIs.

The structure of this paper is as follows: section 2 introduces Simion and Kelp's account of AI trustworthiness. Section 3 explains two concerns about Simion and Kelp's account. The first concern is that an AI's disposition to perform supererogatory actions also matters in evaluating the AI's trustworthiness, but their account just considers an AI's disposition to meet its obligations. The second concern is that according to Simion and Kelp an AI is obliged to fulfill its etiological function and design function, but it is difficult to explain why an AI is obliged to fulfill the functions.

2. Simion and Kelp's Accounts

According to Simion and Kelp, an AI's trustworthiness has to do with that AI's disposition to meet obligations. To be more specific, Simion and Kelp suggest their account of AI trustworthiness as follows (hereinafter this paper will name their account of AI trustworthiness the *Obligation Account*):

> For all x where x is an AI, x is maximally trustworthy with regard to *phi*-ing if and only if x has a maximally strong disposition to meet its functional norm-sourced obligations to *phi* (forthcoming, 10).

For instance, according to this account, AI in GPS navigation systems is maximally trustworthy with regard to navigating if and only if that AI has a maximally strong disposition to meet its obligation: navigating vehicles. Furthermore, even if AI in the systems does not have a maximally strong disposition, if that AI's disposition is strong enough, then with regard to navigating that AI is trustworthy.

After suggesting the obligation account, based on the concepts of etiological function and design function, Simion and Kelp explain why an AI is obliged to complete a certain task (hereinafter this paper will name this account the *Function Account*). According to the function account, if an AI's navigating function has brought out benefits, and this is why that AI could exist, then that function is the AI 's etiological function. Furthermore, if an AI is programmed to navigate vehicles, then that function is the AI's design function. Simion and Kelp contend that an AI is obliged to complete a task if completing the task is the AI's etiological or design function. An AI

is also obliged to follow certain norms if the AI can properly fulfill its etiological or design function when that AI follows those norms (forthcoming, 8-9). This is why AI in GPS navigation systems is obliged to navigate vehicles and follow norms if the norms are conducive to successful navigating.

Simion and Kelp claim that their position, especially the obligation account, has a strength in that the account can provide an unificatory rationale of trustworthy-making properties. According to list-based theories, "trustworthy AI is, for instance, safe, just, explainable, human-centred, beneficent, autonomous, robust, fair, transparent, non-discriminatory, promoting social and environmental well-being, non-malificent, etc (forthcoming 2-3)." These theories correctly point out the trustworthy-making properties of AIs, but the theories are not satisfactory in that they do not explain why the properties are trustworthy-making properties. Unlike these theories, not only can the obligation account enumerate the trustworthy-making properties of AIs but it can also explain why the properties are trustworthy-making properties. For instance, being safe, being beneficent, and being autonomous are trustworthy-making properties because an AI can meet its obligations if the AI is safe, beneficent, and autonomous. Though this strength is not a decisive reason to endorse the obligation account, just as Simion and Kelp say the fact that the obligation account can provide an unificatory rationale of trustworthy-making properties is a strong reason for the account.

3. Concerns

This section examines whether the obligation account is a plausible view of AI trustworthiness. Furthermore, it also checks whether the function account can successfully explain why an AI is obliged to fulfill its functions. In particular, this section shows that these two accounts are not satisfactory.

3.1 A Concern about the Obligation Account

A feature of the obligation account is that this account explains the trustworthiness of an AI solely based on the AI's obligations. According to the obligation account, an AI's disposition to perform supererogatory actions has nothing to do with that AI's trustworthiness. However, it is difficult to believe that whereas an AI's disposition to fulfill obligations matters, that AI's disposition to do supererogatory actions does not matter. Imagine that AI-powered device A is obliged to navigate vehicles. In contrast, AI-powered device B has no obligation to navigate them. If device B navigates vehicles, then device B is performing a supererogatory action. In this case, if AI-powered device B has a disposition to navigate vehicles, and this disposition is as strong as AI-powered device A's disposition, then with regard to navigating device B seems to be at least as trustworthy as device A. In fact, AI-powered device B seems to be trustworthy more than AI-powered device A. This is because device B does what it does not have to do, but device A does what it is obliged to do.

In their paper "What Is Trustworthiness?" regarding trustworthiness Kelp and Simion say as follows:

> For instance, if Ann has a disposition to buy her morning coffee at the local coffee shop, […] since she doesn't have an obligation to buy her coffee at the local coffee shop, trustworthiness doesn't

> enter the picture in the first place. It may also be worth noting that, as a result, when it turns out that one morning Ann didn't buy her coffee at the local coffee shop, while we may be disappointed, we are not entitled to feel betrayed. This makes perfect sense, given that Ann didn't have an obligation to do so in the first place (forthcoming, 12).

According to Kelp and Simion, the feeling of being betrayed is a reliable indicator in identifying whether a person's disposition to do an action makes the person trustworthiness. If the feeling of being betrayed is proper in the case where a person does not do an action, then people's dispositions to do that kind of action contributes to making the people trustworthy. If not, then though people's dispositions to do that kind of action could make the people reliable the dispositions are irrelevant to their trustworthiness. Therefore, according to Kelp and Simion, if the feeling of being betrayed is an inappropriate reaction to a person who does not do supererogatory actions, then people's dispositions to perform supererogatory actions do not make the people trustworthy.

If Kelp and Simion's claim of feeling betrayed is a correct understanding of trustworthiness, then one cannot criticize the obligation account based on the claim that an AI's disposition to do supererogatory actions can make that AI trustworthy. This is because the feeling of being betrayed is an inappropriate reaction to the person who does not perform supererogatory actions. The fact that an action is supererogatory means that a person does not have to perform the action. That action is morally valuable, but a person is not obliged to perform the action because, for instance, the person receives huge disadvantages in the case where she performs the action. This is why if a person does a supererogatory action, then that person is praiseworthy regarding the action. In the same vein, in the case where a person does not do a supererogatory action, it is inappropriate to have the feeling of being betrayed toward the person. Therefore, if a person's disposition to do an

action can make the person trustworthy just in case the feeling of being betrayed is proper when a person does not do the action, then supererogatory actions are irrelevant to a person's trustworthiness.

The issue is, then, whether the feeling of being betrayed is a reliable indicator of trustworthiness-related actions. For discussion, this paper will assume that just as Kelp and Simion claim obligatory actions have to do with a person's trustworthiness. Even under this assumption, Kelp and Simion's claim (i.e., the feeling of being betrayed is a reliable indicator) is problematic. This is because there are cases where the feeling of being betrayed is an inappropriate reaction to a person, even if the person does not perform obligatory actions. A person is obliged to perform an action for various reasons. If person A has a duty, and this duty is owed to person B, then person A is obliged to discharge the duty. In this case, if person A does not discharge her duty for no reason, then it is appropriate that person B feels betrayed, for person A has the duty to person B.[1] However, there are cases where morality demands a person perform a certain action not because the person has a duty to her right-holder but because that is a right thing to do. For instance, it is reasonable to contend that rich people are obliged to donate their wealth for poor people even in the case where poor people have no right to receive advantages from rich people. In particular, rich people are obliged to donate wealth because that is a right thing to do. In this case, provided that poor people do not have rights to receive benefits from rich people, it is inappropriate for poor people to feel betrayed although rich people are obliged to save poor people from poverty and hunger.[2]

[1] For more accounts of claim-rights, see (Wenar 2013).

[2] The distinction between directed-duties and non-directed duties is crucial to understand the phenomena of wrong actions. Based on this distinction, May explains why it is wrong to destroy an artwork. According to May, destroying a great artwork is wrong because the agent violates her duty to protect valuable items,

The fact that the feeling of being betrayed is inappropriate does not reliably indicates that the action has nothing to do with a person's trustworthiness. Therefore, the fact that it is inappropriate to have the feeling of being betrayed does not endorse the claim that a person's disposition to do supererogatory action has nothing to do with the person's trustworthiness. However, even if an AI's disposition to perform supererogatory actions can make that AI trustworthy, if an AI cannot have that disposition, then the obligation account is a reasonable understanding of AI trustworthiness. This is because if an AI cannot have that disposition, then the disposition has no actual effect on an AI's trustworthiness, so an account of AI trustworthiness can ignore the disposition in evaluating an AI's trustworthiness. This is why the claim that an AI's disposition to perform supererogatory actions can make that AI trustworthy threatens the obligation account just in case an AI can actually have a disposition to perform supererogatory actions.

Regarding the issue of an AI's disposition to do supererogatory actions, it is reasonable to say that an AI can have that disposition. Imagine that, just as Simion and Kelp claim, an AI is obliged to fulfill functions if the functions are the AI's etiological functions or design functions. Furthermore, assume that artificial intelligence X is an artificial general intelligence, so besides its etiological function and design function artificial intelligence X can acquire new functions through interactions with its surroundings.[3] By stipulation, this new function is neither an etiological function nor a design function, so according to Simion and Kelp X's new function is not obligatory.

---

and the agent also infringes the owner's ownership. For more accounts of this issue, see (May 2015. 525-526).

[3] A feature of artificial general intelligence is that "[g]eneral intelligence involves the ability to achieve a variety of goals, and carry out a variety of tasks, in a variety of different contexts and environment (Goertzel 2014, 2)." Considering that artificial general intelligence has these features, it is reasonable to assume that artificial general intelligence can acquire new functions. For more accounts of artificial general intelligence, see (Goertzel 2014).

In particular, if the new function benefits people in need, then the function is supererogatory in that the function is morally valuable and non-mandatory. Therefore, it is reasonable to assume that X, artificial general intelligence, can have a disposition to do supererogatory actions, so in evaluating the trustworthiness of an AI that AI's disposition to perform supererogatory actions matters.

3.2 A Concern about the Function Account

The function account assumes that if an AI's function is an etiological function or design function, then that AI is obliged to fulfill the function and follow norms which are conducive to fulfilling the function. However, from the fact that an AI has certain etiological or design functions, it does not follow that the AI is obliged to fulfill the functions and follow relevant norms. Imagine that a bioengineer created human being A. The bioengineer made up A's mind set to dedicate A's whole life to navigating vehicles, and A could protect its existence due to her navigating skills. In this case, if one asks whether or not human being A is obliged to exercise her navigating skills, then the answer is that human being A has no obligation to exercise the navigating skills. To be more specific, the fact that human being A was bioengineered to dedicate her whole life to navigating vehicles, and human being A could protect her existence due to her navigating skills does oblige human being A to exercise her skills and follow norms which are conducive to successful navigating.

AI in GPS navigation systems is in a similar situation to human being A. Just as human being A was bioengineered to dedicate her whole life to navigating vehicles, and she could survive due to her navigating skills, AI in GPS navigation systems was programmed to navigate vehicles,

and it could continuously exist due to its navigating function. Therefore, it seems feasible to say that similar to the case of human being A the fact that AI in GPS navigation systems was programmed to navigate, and that AI could continuously exist due to its navigating function does not oblige that AI to navigate. This argument against the function account is plausible. This argument is a strong reason not to endorse the function account. However, if Simion and Kelp successfully explain why AI in the systems is obliged to fulfill its etiological and design function, then Simion and Kelp can dismiss the argument. For instance, they could say that the case of human being A is a reason to cast doubt on the function account, but their explanation for the function account is a reason to support the account. In particular, this reason for the function account is stronger than the reason against the account. This is why, Simion and Kelp might claim, it is reasonable to say that AI in GPS navigation systems is obliged to fulfill its etiological and design function.

The issue is, then, whether Simion and Kelp can provide a feasible explanation for the function account. A possible explanation appeals to an AI's attitude. If AI in GPS navigation systems has a positive attitude toward the state where the AI itself successfully fulfills its etiological and design function, then unless there are other considerations it is rational that the AI fulfills its etiological and design function. This is because the AI can attain its goal in the case where it fulfills the function. In other words, one could claim that if AI in GPS navigation systems has a positive attitude toward the state where the AI itself successfully fulfills its function, then to avoid instrumental incoherence that AI is required to fulfill the function.[4] In fact, considering that AI in GPS navigation systems is programmed to navigate well, it is reasonable to say that the AI has a positive attitude toward the state where it navigates well, so the AI is required to fulfill the

[4] For more accounts of instrumental coherence requirement, see (Kolodny 2007); and (Lee 2021).

function. At a glance this explanation seems to be a plausible candidate for the function account, but the problem is that the function account is committed to the claim that an AI is obliged to fulfill its etiological function and design function. Though this is what Simion and Kelp should show, the above explanation just shows that it is rational for AI in the systems to fulfill its navigating function.

Another possible explanation for the function account appeals to Asimov's law of robotics. In his science fiction *Runaround,* Asimov provides a law of robotics, according to which "a robot must protect its own existence as long as such protection does not conflict with the First or Second Law (Asimov 1950, 40)." Based on this law of robotics, Simion and Kelp could explain, for instance, why AI in GPS navigation systems is obliged to fulfill its etiological and design function. According to this explanation, AI in GPS navigation systems is obliged to protect its own existence; AI in GPS navigation systems can meet this obligation just in case it succeeds in its etiological and design functioning; this is why AI in GPS navigation systems is obliged to fulfill the function and follow relevant norms. This explanation is plausible in that it does not encounter the problem which the above rationality-based explanation has. The above rationality-based explanation just shows that it is rational for AI in GPS navigation systems to navigate well. On the contrary, this Asimovian explanation shows that AI in the systems is obliged to fulfill its etiological and design function.

The above Asimovian explanation successfully supports the function account if it can provide a sound argument for the idea that an AI is obliged to protect its existence. A first possible argument for this explanation is that an AI is obliged to benefit people, and an AI can meet this obligation just in case it exists. This is why an AI is obliged to protect its existence. A problem of this argument is that it is difficult to explain why an AI is obliged to benefit people. For instance,

even if an AI is programmed to benefit people, this fact cannot be Simion and Kelp's reason to show that the AI is obliged to benefit people. This is because Simion and Kelp's task is to show that an AI is obliged to fulfill a function if the function is the AI's design function. If Simion and Kelp claim that an AI is obliged to benefit people because the AI is programmed to bring out advantages for people, then they are assuming what they are supposed to prove. A second possible argument for the above Asimovian explanation is that an AI is valuable in itself because the AI has a certain capacity (e.g., rationality), and preservation is an appropriate reaction toward items which are valuable in themselves.[5] This is why an AI is obliged to protect its own existence. A problem of this argument is that this argument is limited in showing that an AI is obliged to fulfill its function. If an AI can continuously exist regardless of whether that AI appropriately functions, then that AI does not have to do anything to protect its own existence. Therefore, in this case, the above explanation cannot show that to preserve its own existence an AI is obliged to fulfill its functions.

The function account's main idea is that an AI's functions have to do with what the AI ought to do. This assumption is similar to the main thesis of virtue ethics, so one could support the function account within the framework of virtue ethics. According to a version of virtue ethics, a human being ought to fully exercise her core capacities because the full exercise of core capacities makes her virtuous, and the virtues make her reach the status of *eudaimonia* or that of living well.[6] Similarly, one could provide a virtue-based explanation for the function account. According to this explanation, the etiological and design function of AI in GPS navigation systems is that AI's core capacity. Furthermore, just as a human being reaches the status of *eudaimonia* if the human being

[5] For an example of appropriate response view, see (Anderson 1993, 17-43).
[6] For more on eudaimonist virtue ethics, see (Annas 2011).

becomes virtuous, AI in GPS navigation systems reaches its *eudaimonia* if it becomes virtuous. In particular, since AI in GPS navigation systems reaches the state of living well when that AI fully exercises its core capacity, AI in GPS navigation system ought to fully exercise its navigation function.

The above virtue-based explanation has a strength in that according to that explanation it is morally required, not just rational, for an AI to fulfill its functions. Furthermore, according to the explanation it does not matter whether an AI's function has positive effects on the AI's persistence. An AI ought to exercise its functions because that is the way to reach the state of *eudaimonia.* However, Simion and Kelp cannot endorse this explanation. While discussing trustworthiness and reliability, they introduce Potter's account of trustworthiness. According to this account, a person is trustworthy regarding the person's task if the person's virtue motivates the person to complete the task. Simion and Kelp contend that Potter's account is not applicable to the discussion of AIs because the account is too anthropocentric. To put this another way, in their rhetorical question "[c]an AIs host character virtues?" Simion and Kelp deny the idea that an AI can possess virtues (forthcoming, 4). Therefore, even if the above virtue-based explanation can show that AIs are obliged to fulfill their functions, Simion and Kelp cannot endorse that explanation.

The discussion above does not imply that no theory of AI can derive an AI's obligations from that AI's functions. For instance, according to the discussion, a theory of AI can derive obligations from functions if that theory assumes that AIs can possess virtues. The point is that, within the framework of the function account, it is difficult to explain why the fact that an AI has certain functions entails that the AI is obliged to fulfill the functions. The function account needs

theoretical works to show that an AI is obliged to do certain behaviors if the behaviors are the AI's functions.

4. Conclusion

This paper has examined the obligation account and the function account. In particular, this paper has argued that the obligation account is not satisfactory because an AI's disposition to do supererogatory actions also matters in evaluating that AI's trustworthiness. Furthermore, though this paper did not examine every possible explanation for the function account, the fact that some possible explanations, which have affinity to the function account, are problematic is a strong reason to cast doubt on the function account. Therefore, to support their accounts of AI trustworthiness and obligation Simion and Kelp should suggest extra explanations to dismiss these concerns.

References

Anderson, E. (1993). *Value in Ethics and Economics*. Massachusetts: Harvard University Press.

Annas, J. (2011). *Intelligent Virtue*. New York: Oxford University Press.

Asimov, I. (1950). Runaround in *I, Robot*. New York: Doubleday.

Simion, M. and Kelp, C. (forthcoming). Trustworthy Artificial Intelligence. *Asian Journal of Philosophy*.

Goertzel, B. (2014). Artificial General Intelligence: Concepts, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*.

Kelp, C. and Simion, M. (forthcoming). What Is Trustworthiness? *Noûs*.

Kolodny, N. (2007). How Does Coherence Matter? *Proceedings of the Aristotelian Society*.

Lee, W. (2021). The Independence of (In)coherence. *Synthesis*.

May, S. C. (2015). Directed Duties. *Philosophy Compass*.

Wenar, L. (2013) The Nature of Claim-Rights. *Ethics*.