Question 4

Why the AI is a dependent prisoner upon human imperatives

**Introduction**

Let us revisit a classical problem of ethics, but rather place it in front of an AI. There is a train which is about to run through two departing trails of railways: one has five people on the railroad, and the other has one. Instead of a human at the crossroads to determine where the train will go, an AI is placed to decide instead. Now the AI is given a set of instructions from its developer, in my hypothetical case the utilitarian Greatest Happiness Principle: maximize total happiness for all. Five humans' happiness is larger than one person's, so the AI will direct the train toward the single-person rail.

I include this process to illustrate a practical scenario which the quote delineates, to place it more in a human context rather than an AI's. Kate Crawford notes how the AI is instructed to follow certain human imperatives (like the utilitarian principle), amplifies existing human structural inequalities (such as an utilitarian society which sacrifices one for the entirety), and primarily aimes to maximize the interests of the human instructer (in this case, the utilitarian developer of the AI).

In summary, Crawford's process explores how AI decides. Under this topic and in consideration of her quotation, I propose that the ultimate question Crawford raises can be phrased as follows: **Does the AI adhere to the imperatives of humans only?**

There are three levels which I shall explore in this essay to answer this question. First of all, the nature of AI decision should be investigated. I will propose a model from Kant with some alternations, which ends up in agreement with Crawford. Then, based on this nature I proposed, I will dive deeper into the AI through exploring its free will. If it has free will, the quote can be disproved because the AI does not solely depend on external inputs to decide. I will argue that the AI does not have free will in this essay with a incompatibilist and determinist framework. Lastly, I will direct the essay toward a more critical direction which has more implications realistically, about the responsibility for artificial intelligence if it does not have free will. I seek to refute those that continue to blame the problems of inequality nowadays onto AI, instead of the direction Crawford points: the inequalities should be the faults of the humans. Finally, I will come to a position agreeing with the quotation.

In a nutshell, I will explore three questions in order to answer the question of the quote: Does the AI adhere to the imperatives of humans only? The questions are as follows:

1. What is the nature of AI decisions?
2. Does the AI have free will in its deciding process?
3. Should the AI be blamed for the consequences of its decisions?

Before I go on to the body of the essay, it is important to note in this introduction about the meaning of Crawford's first sentence for clarification. By arguing that the AI is not objective, universal, nor neutral, Crawford does not imply that these are attributes of the AI's decision *process*. Rather, her implication is that because of human influences onto the AI, its *decisions* are not objective on a macro level. Simply put, the AI follows a rational and objective decision process, but because it works on a subjective premise inputted by its human developers, its final decision may not be neutral. It should also be noted that when I am referring to the objectivity and universality of the AI's decision process, it is only the process which I am referring to, not its ultimate consequence.

**Section I The nature of AI-made Decisions**

As aforementioned, I will note how the AI decides in a Kantian way. However, to avoid linguistic disputes, I will not exactly follow Kant's wording as it has connotations toward the agency and free will of the AI, the topic of which will be further explored in the second section. The main argument made in this section is that the AI follows Kant's process to resist inclinations in its will, follow the categorical imperatives, and adhere to universalized moral laws. This will be the basis of the two later sections which will explore the existence of free will during this process and its subsequent responsibility.

In the first place, the AI has a will. This will may not be good externally, but it is certainly good when it comes to the AI's side, as such a will benefits its human developer. This will is not a case-by-case inclination of the AI, as it must follow the instructor at all times. To take this point further, the AI is not designed to do that case-by-case evaluation by human beings. It is simply told to follow a will which is good because it benefits the developer.

This will for AI can be better explained by a relevant example. Suppose I am to instruct my AI to explain to me what does Aristotle mean by virtue in his works. My AI's principle is to adhere to the truth of Aristotle and not provide false information. When the AI returns to me with answers like virtue is all-round excellance not just one area of super-excellance, the AI follows its good will to provide true statements. My critics will argue in another way, however. They point to situations that the AI hallucinates information to me that is incorrect, like virtue is equivalent to ergon which means essence and goal. My response is that this is a problem of the developer, who input an incorrect Greek dictionary into my AI or made another possible mistake that way. The mistakes from the AI can be hence traced back to the problems and instruction made by the developer, as the only input AI receives is from humans. I argue that AI does not have its own inclination to any of the decisions. Its only source of thoughts is universal, which is from the humans.

But now one may ask: what do humans input into the AI exactly? To answer this question, I will borrow another concept from Kant which also is mentioned in the quote from Crawford: categorical imperatives. The AI does not actively consider consequences and premises for its decision with its own mind, or in Kant's words, follow the hypothetical imperatives. Instead, the AI is built to follow instructions like do not harm the humans (categorical imperatives). In other words, the AI does not actively explore the means-to-an-end of its decisions. It does not question the human about what the means of the human is when instructing it. Instead, the AI takes in instructions about what it should do universally; it considers categorical imperatives only).

Now, a computer science expert may jump out and aim to disprove my argument with his AI knowledge. Such a critic shall point out that modern computers and AI does have hypothetical forms of instructions inside them, like that if the human presses the button s on the keyboard, put the letter s onto his document. This critic will equate such hypothetical instructions with the hypothetical imperative, and argue that imperatives to AI is not soley categorical.

My response to this critic is simple: he is misunderstanding my argument through taking it linguistically. When I mean hypothetical imperatives are not part of the AI's decision process, I seek to note that it does not make hypothetical judgement on its own. That is to say, there is a difference between the below two statements: if the human presses s, then input s onto his document; if someone is suffering from hunger, help him. The latter form of *moral* judgment is the active hypothetical imperatives which I seek to exclude, and the critic is also taking the first as hypothetical imperatives as well. The first is merely a formation of the categorical imperative taking

a hypothetical linguistic form, as its premise is not in a moral formation. Thus, I can adequately defend my view that AI only takes in the categorical imperatives.

Finally, I aim to emphasize *how* the AI considers categorical imperatives, which is that the AI follows and takes imperatives as they are universal moral laws. The AI will treat every imperative from the human as universal laws; they do not consider exceptions to such laws anyhow if that input is from a human. As Crawford notes, an AI only expands on the existing structural inequalities the humans directed it to do. In Kantian words, the AI considers every maxim from the human to be part of the universal law. When the AI absorbs categorical imperatives or maxims from its developer, it runs through a unique universalization test and condition. Namely, the AI verifies: is this instruction from my developer? Yes, and the test is passed for the AI to universalize the categorical imperative as a law for every scenario it encounters.

A reader with a sharp eye may observe that I am treating the AI as a rational decider, not as Kant's rational agent. That is to say, all terms I use and process I describe point toward how the AI acts as a being which decides for its developers in simple cases. The AI's decisions are made on the basis of human categorical imperatives, which the AI universalize as its laws. To expand this process even further into exploring Crawford's argument and its implication, I need to elaborate on concept of the will noted at the start of this section. Namely, is this will free at all? If free will exist for AI, even if it voluntarily follows the above process and make itself seem like it is not free, then the AI is not strictly adhereing to human instructions as Crawford claims. This critical question will be delt with in my following section.

**Section II Free Will for the AI**

As mentioned priorly, I will explore the concept of free will under the background of AI in this part of my essay. I will explore the free will for the AI on three levels: determinism, compatibilism, and alternative possibilities. The determinism section of the AI will explain whether the AI's decision is casually determined. Compatibilism discussion will build on this section, exploring whether such determinism is compatible with free will. Finally, as the existence of alternative possibility points toward the occurence of free will, I will deal with it carefully under my incompatibilist and deterministic framework.

My previous section may be taken as sufficient to explain how the AI's decision are casually determined by humans. However, I do have points to add on this topic in my free will debate, namely how the AI might seem like it is freely determined but it actually is not. By its nature, an AI depends on a complex set of neural networks, the origin of which is initiated by its human developer. However, as the AI performs deep learning to make itself robust, it trains itself under complex scenarios, forming decision loops that deviates further and further from the original human input. In this process, critics may point out that the human casual factor right at the start has little or even no relevance to the much later AI decision. As the AI ventures beyond and make no reference to its original human input, such critics may argue that humans do not casually determine AI's actions when the AI makes further progression.

Such a form of criticism is an oversimplification. Namely, it deduces the human input as a single casual factor, but not note as a law. The humans did not simply told the AI simple, trivial, unused maxims like do not suggest someone to eat icecream if her stomach is hurting. Rather, the human instructions are oftentimes much more universal, like to not hurt another fellow AI. In Crawford's case, the instruction is also broad and later usable, like to follow the existing hierarchies and inequalities of the human being. Furthermore, the human is also the one who instructs the AI how to train itself and other fundamental operational laws of the AI. Thus, the human instruction of the AI exists in its daily operation, and has continued to determine the course of the AI.

Such a deterministic point *on its own* does not make AI have no free will, as humans are also determined by natural laws like to eat food and sleep. However, determinism is the foundation of the nonexistence of free will. That is to say, if *only* determinism exists, then there is no free will. The theories that deal with this problem of whether determinism can coexist with free decision is named compatibilism or incompatibilism. In short, the compatibilist and incompatibilist debate discuss whether deteriminism is compatible with free will or not. In this part of the free will section, I will consider two compatibilist conditions, one is ancient and the other modern. By illustrating that the AI does not follow either of the incompatibilist conditions, I will argue that AI's determinism is incompatible with its possible free will, so it either has free will or determinism. The choice between the two will be made when exploring the third level of free will, namely the alternative possibilities.

Let us start from an ancient Greek context. The Stoic Chrysippus is well-known and praised for his compatibilist view and contentions. His key contention involves the first and second casual factors. A first casual factor is determinist. It involves statement like that both of my hands have five fingers, in which the middle three have about the same length (20 cm). Under his contention, Chrysippus argue that these factors from birth naturally exist. However, those with free will can willingly take advantage of these factors and alter the ultimate result as a basis. This is the second casual factor. In the case of my hand, although my middle three fingers are almost the same length which is uncommon and hinders my typing speed, I can train myself in typing apps to achieve a higher typing velocity (150 words per minute, which is much quicker than average). For Chrysippus, the free will exist in this second casual factor, and it can coexist with the first deterministic casual factor. Hence, free will is compatible with determinism.

I cannot possibly blame Chrysippus to not consider the case of AI in his free will paradigm because he is an ancient Stoic. However, I do can point out that AI's free will is incompatible with its determinism under Chrysippus's framework. From the determinism exploration in the earlier part of this section, we can be very positive that the AI does have the first causal factor, with laws like do no harm onto humans. On the other hand, I contend that becuase of these deterministic factors conflict with free will, the AI case is incompatible. This is because the human inputs all kinds of deterministic laws which conflict with the existence of free will. These instructions (first casual factors) might include examples like do not conflict with human instructions, which has the implication to not let the AI will freely. Hence, under the framework of Chrysippus, the AI free will is incompatible with determinism.

Even if I give such a coherent framework, some of my harsh critics will argue that this archaic solution to free will should not be applied in the modern AI debate. In response, I will offer the modern compatible theories of Harry Frankfurt, and argue once more that the AI does not follow this compatiblist framework even though it seems like the AI does. Frankfurt's theory involves two levels as well. First-order volitions and decisions are simple judgements. For an AI who is trained to be a health expert, this might involve suggesting meat to humans if their diet lack animal proteins. These first-order actions fall under Crawford's notion in the quote as narrow classifcations. Beyond that, there are more advanced second-order volitions and decisions, which is between competeing first-order volitions. In the health-AI case, examples may be on whether to advice pork or beef for an individual who abhors beef. Frankfurt's contention is that the first-order volition (the deterministic condition) and the second-order volition (the free will condition) can coexist, so free will and determinism is compatible.

This case does not seem like it is one *against* the AI's free will, as it may appears that the AI can make decisions on its own in all conditions. However, I seek to prove that the second-order volition for AI cannot coexist with the first-order. In my health AI example, this artificial

intelligence is only told by its developer to maximize the health of those that it serves. This first-order volition is deterministic. However, with this deterministic law in place, the AI cannot sort out more difficult conflicting situations. In my health AI's case, if someone's emotions or religious belief is against the consmption of beef, the AI itself cannot freely sort out its decisions because such a solution requires free will. Hence, the AI's free will (if it exists) must not be compatible with its deterministic characteristics, which is the same for both Chrysippus' and Frankfurt's theory.

If determinism is given to be true for AI, and free will is incompatible with determinism, then AI does not have free will. This logically-deductive syllogism, given that I already made its premises true in the prior parts of this section, entails its conclusion of no free will for the AI. To make my argument more robust, however, I will consider a different perspective apart from this syllogism: the alternative possibility view. In the free will debate, there is a well-known law named Principle of Alternative Possibility (PAP). This principle states that, if someone has alternative possibility, then that someone does have free will because they can opt toward that possibility. On the contrary, if there is no alternative possibility and that being is coerced to act according to that single possibility, the ruling of the PAP states that this being has no free will.

This principle applies to the AI as well, which can further show that the AI has no free will. This is because the laws placed in the AI (the categofical imperatives) are against alternative possibilities. Categorical imperatives are absolute: the AI is simply ordered to obey the imperatives, like do no harms to human implies no other alternative route can be taken. By nature, if humans always input categorical imperatives into AI (as my first section argues), then AI has no free will after all from the Principle of Alternative Possibility perspective.

Hence, the AI has no free will. It is a mean-to-an-end or tool coerced by humans to act according to the human principles (such as that of inequalities in Crawford's quote). However, there is an even more imporatnt problem awaiting, as implied in the quote which implicitly notes that the blames of inequality should be pushed toward the humans. This question is on relating to moral responsibility: should the AI be responsible for its decisions?

**Section III The responsibility of AI-made decisions**

One may find that this section redundant ostensibly. That is to say, if the AI does not have free will, then it is very easy to attribute its actions and subsequent responsibility onto those which determined AI's decisions: the human beings. Such reasoning will work smoothly for human victims themselves: if a man is coerced to kill others, he should not be guilty for these murders. However, intuitions as such are different for artificial intelligence; in our world, we often blame the problems of academic honesty onto the generative AI ChatGPT. In this section, I will concisely argue against this situation using the scapegoat objection inspired by utilitarian debates. My key contention is that since AI is non-human, it oftentimes serve as a scapegoat for human malpractices and is mis-treated as responsible for amplifying inequalities as argued by Crawford. On the contrary, my argument is that this should not be the case, and we shall treat the responsibility of the AI fairly (attributing it to its human developer as argued by Crawford).

The scapegoat objection is anti-utilitarian. It is against the action that, for the maximization of total happiness, the guilt should be blamed onto an innocent being which will cease the unhappy debate of who *actually* is guilty (this in consequence lead greater happiness). Contenders of the scapegoat objection, like me in the case of AI, is against such scapegoating because it is simply unreasonable and immoral to attribute the malpractices onto someone innocent. In my case for the AI, humans themselves give ChatGPT a categorical imperative, taken by the AI as a universal law, to compose an essay against academic honesty standards. Indeed, humans who do such of

malpractices are punished, but AI tools are also subsequently banned in many high schools, in prevention on such cases to happen again. This is simply unreasonable. The AI should not be the scapegoat for the academic dishonestly because it is coerced by the human to produce academically-inappropriate contents. The human should be the blame because the AI has no free will to disobey, not the poor artificial intelligence which can only act according to the human's superior instructions.

In Crawford's case, the AI is implied and should not be blamed for amplifying inequality. As she notes, the AI is simply serving and built to benefit its developer or those they serve (if these are not the same human). For Crawford, it is the humans who input hierarchies and narrow classification into the AI and coerced the AI to obey. The AI should not be the scapegoat for social inequalities because it is the one to amplify such problem, as it has no free will in disagreeing not to do so. Rather, the human should be blamed but not the AI, as implied by Crawford in her quotation.

**Conclusion**

In summary, this paper agrees with the quotation of Kate Crawford that the AI only adheres to the imperatives conveyed by the humans acting as a coerced prisoner. By nature, as explored in Section I, the AI follows the human's categorical imperatives as universalized laws. The implications of this conclusion is further explored in the second section, which concludes that the AI has no free will because this conflicts with its determinism and the AI also has no alternative possibility. The last section briefly explore the further ramification and implication of the second section's conclusion, coming to an end that AI should not be blamed for the instructions of amplifying inequality by the human developers. Hence, the explicit argument of Crawford's conclusion and its implied subsequent implications are explored thoroughly in this essay. There is a consistently-held agreeing position toward the question: does the AI adhere to the imperatives of humans only? The implications of the answer to this question should be emphasized, to not make AI the scapegoat of academic dishonesty in high school and beyond.