Prompt 4:
”AI is not an objective, universal or nuetral computational tequnique that makes determinations without human direction. Its systems are embedded in social, political, cultural and economic worlds, shaped by humans, instituions and imperitives that determine what they do and how they do it. They are designed to discriminate, to amplify heirarchies, and to encode narrow classifications. When applied in social contexts such as policing, the court system, healthcare, and education, they can reproduce, optimise and amplify existing structural inequalities, This is no accident: AI systems are built to see, intervene in the world in ways that primarily benefit the states, institutions and corporations that they serve.”

Who is John?

Absolute power corrupts absolutely

This sentiment is generally used to verbalise the larger idea that human nature is corrupt, irrationally so, when given absolute control. There have been theories to sidestep this ´problem´, with some recommending an agent born and crafted to govern the maginificent Kaalipolis, no more, no less, systems bent on avoiding the concentration of power to the individual, such as our beloved *democracy*, and some wish to avoid the folly of human nature all together!
Human nature is emotive, sentimental, and irrational in regards to certain facts and knowledge, and our ability to reason is heavily impacted by it. Systems that wish to about the issues with human nature, they do that by idealising an agent that can apply logic and reason perfectly, uninfluenced by its own sentiment, and hence govern in the best way possible. A man that can comprehend and consider all relevant facts in the process of reasoning, and be uninfluenced by that which is irrelevant. A man that does not engage in post-hoc rationalisation (starting with a desired conclusion and working to rationalise it) and instead accepts whatever conclusion the torch of reason leads him to.
Perhaps this perfect man cannot exist, and no absolutely powerful man can remain uncorrupted, but fianlly, with modern technology, we have Artificial Intellgience, an entity capable of reason, logic, untethered to emotion and the likes. AI is what we have been looking for all along. Whilst we still trust humans to be at the top of the social organisation heirarchy, AI’s implementation is currently a project being undertaken in many sectors of society, accompanied by new ethical dilmenas and proposed issues. Lets call this AI ’John’.

Crawford in this quote has quite *nihilistically* almost, shown her skepticism that AI can be the perfect man, instead classifying it as a tool of the perverse powerful, that is touted as unbiased and facetiously used to achieve the goals of the corrupt man, under the guise of *infallible reasoned output*. Her point is not wrong, the developement of AI certainly has a plethora of issues, least of which are implicit discriminatory outputs and model biase. However, not only do I think that this is a is a non persistent point of contention, but that the rational choice is to submit to an AI’s governance, with some added nuance.

In the essay, what I aim to prove is why AI bias is non persistent and hence Crawford’s objection is also not one that will be persistence throughout Ai discourse.

Why John doesn’t exist yet

To start with the decontruction of the aprehension and rightful suspision of ’ unbiased AI’. AI, in its current conception is, to be blunt, extremely advanced text prediction. There is no actual understanding of the matter of what is being said, it is simply outputting the most apporpriate response, based on its traning. Current AI is very similar to Searle’s Chinese room.

Searle's chinese room experiment states that there is a man inside a box, with a book with references to chinese phrases and their appropriate replys. The man does not know chinese or understand it in any regard. A person can write something in chinese and give it to the man in the room. The man in room then refers to the book he has and appripriatly hands out another card. Now does the man actually know chinese?

The difference is that with current AI, the book of translations has mulitple answers for the same inpute, and the man in the room rolls a dice to see which output he should accordingly give. The analogy works because AI works in a similar manner. The training allows it to seive through the data of the random map generated and accordingly turn it into an appropriate output. This can be appropriately called LLM (Large Language Models).

With the current state of AI then, Crawford is correct. Instutions in power create the AI, train it and tune it to serve their purposes, and then implement it. But this just goes to show that current AI is not even close to the conception of the AI that we wanted. Current AI is not John, not even close to it. To compare their workings would almost be to make a category error. A LLM is just a man in the chinese room where the book is written by the intitution in power. LLM does not apply 'reasoning', it simply gives complex predetermined outputs by mashing what is written in the book at random.

Why Crawford doesn't actually hate John

The reason Crawford's objections to current AI is not a persistent issue, or rather, will not be a persistent issue is because she is arguing against a rudimentary state of AI which is flowed in creation. The reason LLMs are biased is because its training data does not make it an entity that reasons, it makes an entity that can parrot the patterns of reasoning it has been provided. When presenting Modus Ponens reaasoning, it does not enegage in deduction, rather it gives an output based on the the examples of arguments that resemble a major premise-minor premise-conlusion form. The in which calculators perform logic and do arthmetic or even algebra and calculus is not at all how LLMs work, meaning it is not a reasoning entity. A future developemnt of AI will see a higly advanced computer capable of doing actual raw logic and engage in reason, rather than emulate what has been fed to it.

There are 2 posible objections I have concieved of that Crawford's camp could present here, one casting doubt on whether 'John' can actually exist, one talking about the issues with John's criteria of evalution, and similar that, one regarding whether 'John' is actually what society should want.

Defending John's Existence

Let's enage with the first one, the idea that John can actually exist. This objection, at its heart casts doubt on the idea that a machine that can engage in higher order reason can actually exist. This means that the objector here is saying that there is relevant differenec between 'general' reason and mathematical reason. A computer's ability to do classical arthimatic and algebera comes from the formalisation and implementation of logic as boolean algebrea, and since any Turing complete machine can do boolean algebra, a computer can engage is mathematical reasoning. Logic is much of the same. Any statement can be broken down into its contituent premises and its following conclusions, and be written in logical notation, presented as a truth table etc. From this, does it not follow that a computer can engage in general reasoning, as it is identical to mathematic reasoning?

John isn't to blame!

Now here the second objection will come into play. Crawford might further object that I have not

dealt with her argument properly, and that for what purpose John employs reason is what contributes to the biase, and whether John will sacrifice the common man for the instituion.
Lets giver John a problem. A beggar has come to a bank, asking for a 500 usd loan to start a tea stall. John has to decide whether to grant the beggar a loan or not.
Lets walk through what John will be doing. The basis on which we will recive a yes-no output is dependant on what is John's goal. If John's goal is to make the bank the most money, Crawford's objection will stand, as John clearly will give an output that may be beneficial to the bank at the cost of the beggar.

Here I can take two lines of objection. First is that Crawford would be incorrect to say that this is an issue of AI being "not an objective, universal or nuetral computational tequnique that makes determinations without human direction." Her objection would be to the objective of the bank, not John. John's determination, as we concluded earlier, absolutely can be devoid of human direction, but it can only work towards an objective given to it. Crawford's argument of AI working towards the biase of institution and enforcing heirarchy will apply equally to any entity that has to make descisions, not just an AI. Hence her objection is not inherent to AI, but inherent to the existence of institutions itself.

For the second line, lets accept that AI is biased for the bank by working for its objective. But this can be alleviated. We can implement systems that look, not for the profit of one stakeholder, but idea of 'purely additive benefit'.
If John assesses that the beggar will be able to return the money back to the bank(with interest ofcourse) then we can tell John that he must give an output based on benefit of both. The sum of the benefit to the beggar and bank both must be considered. As long as neither party loses, the loan should be given. Even if the bank makes no benefit, as long as loss will not occur, the loan should be granted, and we can clearly allow John to proceed as such.
A generalisation of this sentiment would be: For whatever instituition an AI must serve, the AI should chose the option that maximises the benefit of all stakeholder, but not at the cost of any stakeholder.

Possible descision	Benefit of A	Benefit of B	Take descision?
1	Non negative	Non negative	Yes
2	Non negative	negative	no
3	Negative	Non negative	no
4	Non negative	Non negative	Yes

A table representing the proposed principle

So both rebuttals work against the possible objection, as in both cases we have determined that the issue is not inherent to AI, and Crawford's objections lie within the institution's imperitive itself, and the quote in the prompt is refuted.

Now we can come to the the question of "should society want John"? I agree with Crawford that John, as implemented by an institution, in application is going to be biased. But from that Crawford extrapolated that hence AI is the issue. The error I believe is being made here is that Crawford does not actually object to AI, but to the sentiment around AI, that it is a "objective, universal ….. direction." This sentiment allows these institutions to delfect blame away from themselves and alleviate themselves from responsibility. But in a world where institutions used AI transparently, saying that they are responsible for the descision taken by the AI, Crawford's issue no longer stands,

as again, her objection would like within the institution, not the AI itself. In this world, there is no morally sginificant difference between a human working towards the imperitives of the institution and the Ai working towards the same, since responsibility in both cases is of the institution and the AI is just an entity to which the institution has offloaded the burden of reasoning to.

Defending John, step by step (consolidation of arguments with reference to the prompt)

A sentence by sentence decontruction and objection of the prompt:

AI is not an objective, universal or nuetral computational tequnique that makes determinations without human direction.

This is refuted by the fact that its computation technique and determniations can be devoid of human direction and influence, and it is logically possible for a John to exist. The issue presented is true for current forms of AI, but it is not a persistent issue whatsoever and will be alleviated with time.

Its systems are embedded in social, political, cultural and economic worlds, shaped by humans, instituions and imperitives that determine what they do and how they do it.

AI computational systems whilst being embedded in social, political, cultural and economics can absolutely work for institutions and their imperitives, however the issue here as stated previously, would lie in the nature of institutional imperitives, not AI itself. AI can embody imperitives not of the institutions, and with the pricniple of purely additive benefit, can employ the imperitive of the institution with no cost to the population.

They are designed to discriminate, to amplify heirarchies, and to encode narrow classifications. When applied in social contexts such as policing, the court system, healthcare, and education, they can reproduce, optimise and amplify existing structural inequalities

AI like LLMs can be designed to discriminate, but that is because it is not AI, no reasoning is actually happening, it is purely the output of the institution, given via a faux independant agent (the AI just parrots the institution's discrimiation, and hence is the instution itself). John can reason independent of the institution and IF it reasons for the imperitive of the instution, the point of contention would be the imperitive, not John. A great example can be the implementation of AI in education and correcting of essays. Students of different cultural backgrounds which generally have examples and cultural viewpoints that are discrimnated against in normal checking, see the same discrimination when their essay is checked by AI because LLMs parrot the checking tequniques of the discriminatory teacher. With a John correcting however, analysis of the essay and correcting is free of discrimination, as the ideal AI employs reason to evaluate the efficacy of an example, coherence of a demonstrated cultural viewpoint, withtout the implicit bias that a teacher or LLM may have by their nature. Hence Crawford's objection here agains, lies within the instituion, and is regarding the current state of AI, which is a non-persistent argument.

This is no accident: AI systems are built to see, intervene in the world in ways that primarily benefit the states, institutions and corporations that they serve."

Whilst is it true that current LLMs are purposefully built to benefit and work for the state and corporation that employs them, that isn't inherent to the implementation of AI. AI systems being built to see and intervene in specfiic ways is just instutions intervening in the world as they always have, now under the guise of 'pure reason'. If Crawford agrees that AI is not nuetral computation and application of reasoning, then as long as we can agree that current AI is not John, and that John is the true state of AI, Crawford's objection lies not within AI, but the social sentiment and misconception around AI.

Why we must submit to John.

With my view on all this made clear, this section is dedicated to seeing, to what extent should John be a part of society.

Often atheists, famously Christopher Hitchens, like to call the tri-omni god some variation of a

”despotic, dictatatorial ruler, a celestial North Korea!” This sentiment has never made much sense to me. If the tri-omni god did exist, then does it not follow that we should rationally submit to it? If a being, definitioally, knew what is perfect for you, and wants you to be the happiest you can be, isn’t the rational course of action to submit to it and follow its instruction? You, a non omnicient being, possibly do not(or could be under a misapprehension) know the perfect way to acheive what is best for you, but a omnicient and omnibenevolent being , by definiton, wants the best for you and knows exactly what is it. Hence the rational course of action would be to submit to the tri omni being.

John is an entity, that in its most perfect form, is, in the domain of all to know about humans and information directly relevant to their existence, omnicient. John can also be given an impertive to be ”omnibenevolent”(we recognise in the above arguments that John can only employ reason in so far as to fulfil teh imperitive given to it). John is also, in its perfect state, capable of perfect reasoning (the same manner in which an ideal turning computer is capable of oerfect mathematical output). So does it not accordingly follow, that we must submit to the output of a perfect John which has been given the imperitive of ultimate human wellbeing? It certainly seems to be the case?

Now realistically, a perfect John cannot exist, we can only get ever-close to one. But we can use the reasoning above, and build a system where John is given a say in administration, and accompanied by humans as a means of confirmation if John’s reasoning seems to be faulty. What this would allow us to do is get as close as possible to the most ’correct answer’, with a check on John’s output via verification of the output. If one agrees with the idea that even non perfect John has the capability to employ reason better than humans, then using John in adminstration, given the correct imperitive follows, assuming better administration is the goal.

A possible objection could be that, for John to hold such power, he must be absolutely omnipotent to be employed in such a manner.
This can be responded to with a simple calculator. When a calculator spits out the result for 123133*34242 , one takes that result at face value and uses it to command economies, put people in cars and tanks, and in general, rely on it, despite the disasterous consequences of the possibility that the answer was incorect due to a muon flipping a single bit. John works the same way. Since mathematical reason is the same as general reasoning, we can rely on John for the same reason we rely on the calculator.

Another objection that may be considered is that no such universal imperitive can be applied to John to parralel omni-benevolence, meaning it cannot be given administrative power.
This objection may work, but at a high intellectual pricetag. If one porports that there is no imperitive that can sufficiently represent ’that which is best for us’ then the idea of political adminstration makes no sense, because then voting for the best leader is also a non-rational conception, as there is no imperitive that a candidate could have that suffiently makes us want to vote for them. Voting would be relegated to an irrational act of simply chosing an arbitrarily prefered candidate. As long there is some sembelence of the perfect impertive an adminstrator should have, John’s implementation in adminstration to some extent is a rational consequence.

Final words

The actual existence of John is very far away, and perhaps humans are not smart enough to build John. But as long as the idea of John is graspable and possible beyond the horizon of the sea of possibilities, Crawford’s conception of AI and its consequence is not an objection to actual AI, or the implementation of AI, simply a true commentary on how institutions with perverse intent take advantage of people’s misconception of AI to further consolidate power. Luckily, the future of John does not look that bleak yet.