Noname manuscript No.
(will be inserted by the editor)

Perspectives about Artificial Moral Agents: Supplementary Information

. . .

Received: date / Accepted: date

In this Supplementary Information we provide the following materials: (i) q-set; (ii) template of the email sent for recruitment of participants; (iii) factor arrays; (iv) crib sheets developed for each perspective; and (v) comments written by participants that loaded significantly on each factor/perspective.

1 Q-set

1. Technological progress requires artificial morality.
2. Given the great technological advances in AI the advent of autonomous machines capable of ethical reasoning is unavoidable.
3. Machines will inevitably become full ethical agents.
4. It is too late to restrict AI solely to non-ethical decision-making.
5. AI Systems in morally salient contexts will not and cannot be avoided.
6. Because it is not possible to know what machines will do in novel situations moral competence is needed in order to govern unpredictable actions.
7. The development of AMAs prevents machines from hurting human beings.
8. AMAs increase public trust and confidence in creating autonomous agents acting on behalf of humans.
9. Developing AI Systems with moral reasoning capabilities will ultimately lead to a better understanding of morality.
10. Safety features, codes of conduct, certification processes, and clear limits to the operational contexts are sufficient to ensure machines will not hurt humans.
11. The act of creating moral machines is not permissible according to the tenets of existing moral theories.

First Author [...]
address: [...]
Tel.: [...]
Fax: [...]
E-mail: [...]

12. We do not need machines which are full ethical agents but rather machines that are inherently safe and law abiding.
13. Moral implementation strategies should be context-specific.
14. Ethics cannot be reduced to a computational algorithm.
15. In the design of intelligent machines, ethical concerns should not be restricted to prohibiting unethical behavior.
16. Machines should be equipped with Theory of Mind which would help them infer the goals and desires of agents around them.
17. Implementing societal preferences in machines is not the right approach because it is deriving an ought from an is.
18. Machines should be fundamentally uncertain about morality.
19. AMAs should be designed with reference to human heroism.
20. Moral machines need to have something like moral common-sense found in folk.
21. AMAs must be able to grasp the basics of moral principles like the principle of utility, the categorical imperative, or the golden rule.
22. Even within a strictly utilitarian framework, ultimately the AMA will have to make a judgment about what is best.
23. AMAs should be introduced simply with the potential to become AMAs upon moral education, development, etc. yet to come.
24. For implementing ethics in AMAs, a hybrid (top-down and bottom-up) computational approach should be used.
25. Ross's prima facie duties should be the moral theory for the basis of algorithmic artificial ethical agents.
26. Machines should be equipped with a virtue-ethical framework as it contains elements conducive to both top-down and bottom-up computational approaches.
27. Machines should use societal preferences to identify an informed and desirable choice when faced with a specific ethical dilemma.
28. Logic is the ideal choice for encoding machine ethics.
29. Stories could be used to quickly bootstrap an artificial agent to a point where we feel comfortable about it understanding our social conventions.
30. Top-down approaches, bottom-up approaches, and supra-rational faculties will need to be combined in AMA system development.
31. A Moral Turing Test should be used to evaluate the moral performance of autonomous systems.
32. Moral performance of machines should be be evaluated through verification of transparency, accountability, and predictability.
33. Computers lack the conceptual understanding to make a moral assessment which precludes them from achieving moral agency.
34. Because computer programs do not have free will they can never be independent moral agents.
35. Artificial moral agency is a primitive form of moral agency which does not amount to full moral agency as it pertains to human beings.
36. Machines should not be considered moral agents because people will use them to avoid personal responsibility.

37. The degree of agency that some artificial systems may reach makes it inadequate to understand them as simple tools.
38. Human beings are not the only existing class of moral agents.
39. No essential difference occurs between human beings and AMAs as far as moral agency is concerned.
40. The features which constitute moral agency are parameters that can be represented in quantitative terms.
41. Machines will be better moral agents than humans since they are not subject to irrationality, seduction, or emotional turmoil.
42. There is a long way to go before artificial agents can replace human judgment in difficult, surprising, or ambiguous moral situations.
43. AMAs endowed with human-centred values will be our moral teachers.
44. Mere indifference to human values—including human survival—could be sufficient for AGIs to pose an existential threat.
45. AGI with moral reasoning capabilities will lead to a better understanding of morality.

2 E-mail Template for Recruitment of Participants

Dear (...), I am contacting you upon coming across your article "(...)". My name is Andreia Martinho and I am a PhD candidate at Delft University of Technology (The Netherlands). I am currently working on an empirical study about artificial moral agents (AMAs), particularly our aim is to map the opinions of scholars about development, design, moral agency, and future projections with respect to AMAs. To provide answers to these and related questions, we make use of a research method called Q-methodology. This approach helps structure the opinions of groups of experts in a statistically sound way. The input for Q-methodology is a so-called Q-survey, where we ask experts to rank statements within a forced quasi-normal distribution. It would be very valuable to have your perspective in our sample and so I am writing to kindly ask if you would be willing to participate in our study. We obtained your name and contact details from an academic paper published by you in the field of Ethics. The survey is anonymous and it takes about 20 minutes to complete. We understand that the survey is quite long but it captures the literature abut AMAs quite well so in a way is like reading a super condensed literature review. We would really value your participation and input in this study. Access the survey by clicking on this link: www.q-onderzoek.nl/ama. I look forward to hearing from you. Thank you. Best wishes, Andreia Martinho

3 Factor arrays

Table 1 Factor arrays

Statement	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
1	4	-4	-1	-2	-1
2	5	-4	1	-3	-2
3	0	-5	0	-5	-1
4	1	2	3	-4	-3
5	3	4	5	1	0
6	3	1	4	4	2
7	-1	-3	-4	-2	4
8	4	-1	-1	-1	0
9	5	2	3	-3	5
10	-4	-2	-5	-4	-3
11	-5	-3	-5	0	-4
12	-3	3	-2	4	-3
13	3	4	0	2	5
14	-4	5	-4	5	-3
15	2	1	5	3	3
16	1	4	3	3	3
17	-3	3	-1	3	-2
18	-5	1	4	-2	-4
19	-2	-2	-3	0	-1
20	-1	0	1	1	-4
21	0	-2	2	2	3
22	2	1	2	1	0
23	-1	-1	-1	0	-2
24	2	3	0	0	1
25	-3	-1	-3	-1	0
26	1	0	-1	0	0
27	0	-1	0	-1	-5
28	-2	-1	-2	-3	3
29	0	0	-2	-1	-2
30	3	1	0	1	-1
31	-2	0	-3	-1	1
32	2	5	1	0	4
33	-2	-2	-3	5	-5
34	-5	-3	-5	4	-5
35	0	3	-4	0	2
36	-3	0	-2	2	1
37	4	2	2	2	2
38	1	0	2	1	1
39	-4	-5	0	-5	0
40	0	-5	1	-4	-1
41	-1	-3	0	-3	4
42	1	5	3	5	1
43	0	-4	1	-5	2
44	-1	2	5	3	0
45	5	0	4	-2	5

4 Crib Sheets

Table 2 Crib Sheet Factor 1

ITEMS RANKED +5
2. Given the great technological advances in AI the advent of autonomous machines capable of ethical reasoning is unavoidable.
9. Developing AI Systems with moral reasoning capabilities will ultimately lead to a better understanding of morality.
45. AGI with moral reasoning capabilities will lead to a better understanding of morality.
ITEMS RANKED HIGHEST IN FACTOR 1 ARRAY
1. Technological progress requires artificial morality.
2. Given the great technological advances in AI the advent of autonomous machines capable of ethical reasoning is unavoidable.
8. AMAs increase public trust and confidence in creating autonomous agents acting on behalf of humans.
30. Top-down approaches, bottom-up approaches, and supra-rational faculties will need to be combined in AMA system development.
37. The degree of agency that some artificial systems may reach makes it inadequate to understand them as simple tools.
ITEMS RANKED -5
11. The act of creating moral machines is not permissible according to the tenets of existing moral theories.
18. Machines should be fundamentally uncertain about morality.
34. Because computer programs do not have free will they can never be independent moral agents.
ITEMS RANKED LOWEST IN FACTOR 1 ARRAY
16. Machines should be equipped with Theory of Mind which would help them infer the goals and desires of agents around them.

Table 3 Crib Sheet Factor 2

ITEMS RANKED +5
14. Ethics cannot be reduced to a computational algorithm.
32. Moral performance of machines should be evaluated through verification of transparency, accountability, and predictability.
42. There is a long way to go before artificial agents can replace human judgment in difficult, surprising, or ambiguous moral situations.
ITEMS RANKED HIGHEST IN FACTOR 2 ARRAY
N/A
ITEMS RANKED -5
3. Machines will inevitably become full ethical agents.
39. No essential difference occurs between human beings and AMAs as far as moral agency is concerned.
40. The features which constitute moral agency are parameters that can be represented in quantitative terms.
ITEMS RANKED LOWEST IN FACTOR 2 ARRAY
1. Technological progress requires artificial morality.

Table 4 Crib Sheet Factor 3

ITEMS RANKED +5
5. AI Systems in morally salient contexts will not and cannot be avoided.
15. In the design of intelligent machines, ethical concerns should not be restricted to prohibiting unethical behavior.
44. Mere indifference to human values—including human survival—could be sufficient for AGIs to pose an existential threat.
ITEMS RANKED HIGHEST IN FACTOR 3 ARRAY
15. In the design of intelligent machines, ethical concerns should not be restricted to prohibiting unethical behavior.
18. Machines should be fundamentally uncertain about morality.
ITEMS RANKED -5
10. Safety features, codes of conduct, certification processes, and clear limits to the operational contexts are sufficient to ensure machines will not hurt humans.
11. The act of creating moral machines is not permissible according to the tenets of existing moral theories.
34. Because computer programs do not have free will they can never be independent moral agents.
ITEMS RANKED LOWEST IN FACTOR 3 ARRAY
35. Artificial moral agency is a primitive form of moral agency which does not amount to full moral agency as it pertains to human beings.

Table 5 Crib Sheet Factor 4

ITEMS RANKED +5
14. Ethics cannot be reduced to a computational algorithm.
33. Computers lack the conceptual understanding to make a moral assessment which precludes them from achieving moral agency.
42. There is a long way to go before artificial agents can replace human judgment in difficult, surprising, or ambiguous moral situations.
ITEMS RANKED HIGHEST IN FACTOR 4 ARRAY
11 The act of creating moral machines is not permissible according to the tenets of existing moral theories.
ITEMS RANKED -5
3. Machines will inevitably become full ethical agents.
39. No essential difference occurs between human beings and AMAs as far as moral agency is concerned.
43. AMAs endowed with human-centred values will be our moral teachers.
ITEMS RANKED LOWEST IN FACTOR 4 ARRAY
45. AGI with moral reasoning capabilities will lead to a better understanding of morality.

Table 6 Crib Sheet Factor 5

ITEMS RANKED +5
9. Developing AI Systems with moral reasoning capabilities will ultimately lead to a better understanding of morality.
13. Moral implementation strategies should be context-specific.
45. AGI with moral reasoning capabilities will lead to a better understanding of morality.
ITEMS RANKED HIGHEST IN FACTOR 5 ARRAY
7. The development of AMAs prevents machines from hurting human beings.
28. Logic is the ideal choice for encoding machine ethics.
41. Machines will be better moral agents than humans since they are not subject to irrationality, seduction, or emotional turmoil.
ITEMS RANKED -5
27. Machines should use societal preferences to identify an informed and desirable choice when faced with a specific ethical dilemma.
33. Computers lack the conceptual understanding to make a moral assessment which precludes them from achieving moral agency.
34. Because computer programs do not have free will they can never be independent moral agents.
ITEMS RANKED LOWEST IN FACTOR 5 ARRAY
20. Moral machines need to have something like moral common-sense found in folk.
27. Machines should use societal preferences to identify an informed and desirable choice when faced with a specific ethical dilemma.
33. Computers lack the conceptual understanding to make a moral assessment which precludes them from achieving moral agency.

5 Comments of participants that Loaded significantly on Factors

5.1 Factor 1

Sort 6100640

comment1: (s1) In arger to garner the best of technological progress, this seems to be an obvious prerequisite
comment8: (s8) Systems that behave in accordance with morals are more likely to garner trust than those without
comment35: (s35) 1) There is no normative theory that Is commonly agreed upon.n2) Our comprehension of what morality is might still change radically and new normative theories might emerge
comment25: (s25) Cf. my disagreement with (18)
comment39: (s39) AMAs lack important properties humans have
comment18: (s18) 1) There is no normative theory that Is commonly agreed upon.n2) Our comprehension of what morality is might still change radically and new normative theories might emerge

Sort 6100903

comment9: (s9) As with other theories of perception and cognition, I believe that implementing process-level models of such theories and testing them in various situations is an invaluable method for evaluating said theories.
comment45: (s45) (See response to 9)
comment16: (s16) I believe that ToM would be necessary to facilitate human-like interactions with human interaction partners, independent of whether the situation was a morally charged one or not. As such, I strongly agree with the statement.
comment10: (s10) Similar to my response to 7, I believe these sorts of practices are valuable in reducing instances of human harm, but would not be sufficient to entirely eliminate all instances of machines harming humans. To draw an analogy, commercial air travel safety has significantly improved over the course of the last several decades, due to improved practices, but unfortunately incidents still occur (though very rarely).
comment7: (s7) In society and human-human interactions, harm can occur due to unforeseen and unintended consequences. Likewise, there are many situations (the trolley scenarios being the classic thought experiment), where human harm is inevitable. I see no reason why AMAs in similar situations should be immune from being causally involved with resulting human harm.
comment34: (s34) This statement seems to be partly predicated on the assumption that any future autonomous system cannot have something that some might deem "free will." While this would be inline with certain schools of thought on free will (e.g., libertarian), I tend to be more sympathetic to compatibilist or even hard incompatibilist schools of thought. Either a sufficiently advanced artificial intelligence could have something deemed "free will"

or people don't have "free will" either. Additionally, as this agnosticism regarding "free will" might indicate, I do not ground my notion of moral agency in the metaphysics of freedom.

Sort 06110518

comment26: (s26) Given that every kind of Meaningful decision has a value component, then practical wisdom is a key component of all reasoning. Machines will have to have this skill or they will not be of much help in helping us chose the best courses of action in any given domain.
comment40: (s40) It is important to note that this will not be true of every salient factor in moral reasoning, but many types of ethical systems can be made computational and these can aid systems in making more morally justifiable decisions.
comment24: (s24) This is the most likely way to succeed. Though bottom up would be the most cost effective.
comment14: (s14) That makes no sense, for instance classical Utilitarianism demands a moral calculation process, so there is at least one ethical system that is explicitly computable.
comment11: (s11) There may be some that might state this, but it does not follow from some of the major ones.
comment34: (s34) It is very likely that even humans lack radical free will. If we require moral agents to have free will, then there will be no moral agents. That is too strong a statement.

Sort 6110727

comment5: (s5) I doubt progress in AI will halt. it is inevitable that AIs will be present in morally salient contexts.
comment45: (s45) A machine with human level intelligence and moral reasoning capabilities would probably be a gold mine of insights.
comment1: (s1) Technology is and will be thrust into morally-laden situations. So, there should be some artificial morality to deal with that.
comment18: (s18) Just seems like asking for trouble.
comment23: (s23) Seems like a very bad idea to introduce an "AMA" with merely the potential to become an AMA upon some future education, development. Maybe you'd get an AI with arrested moral development, kind of like an artificial Donald Trump.
comment44: (s44) I actually strongly agree with this statement and accidentally put it into this Disagree (-5) category. Sorry–I misread it on this step. For me, this should be another Agree (+5).

Sort 6110355

comment37: (s37) Technologies (especially machine learning algorithms and related systems) a more than mere instruments or tools. they are so-

cially interactive entities that challenge existing moral/legal categories.
comment1: (s1) Emerging technologies (AI, robots, etc.) necessitates that we entertain questions concerning the moral standing/status of artificial others.
comment38: (s38) Agency (moral, legal, or otherwise) is a socially constructed and recognized designation.
comment34: (s34) Agency (as it is currently defined and operationalized) does not necessarily depend on a prior determination concerning free will.
comment28: (s28) Logic has limitations. What is called "logic" is not universal. it is often linguistically and culturally specific.
comment11: (s11) Prescribed prohibition is never a credible way to do the investigative work of science.

Sort 06150228

comment30: (s30) Our only understanding of morality so far stems from humans. It has been a human term by definition (allow me to freely exclude animal morality for a second). As humans can have certain principles they try to live by, learn from and respond to their environment, and have certain cognitive processes they use for moral decision making, by definition the same is needed for humans to mimic human moral decision making.
comment6: (s6) As mentioned in my last comment, generalizability to unknown situation is not yet at a stable enough level for moral decision making. To ensure positive outcomes, moral competence can help mitigate bad decisions because of unknown/new situations.
comment42: (s42) Current AI technology that comes closest to being able to generalize (case-based reasoning or learning algorithms such as neural nets) are still very limited in their generalizing capabilities. Not only should the system be able to deal with a new situation, they should be able to recognize all relevant features to be used in the decision. We are currently very far from AGI and I believe an new leap in technological advancement is needed before we come close to algorithms that can compare to human judgment.
comment3: (s3) Full ethical agents, using Moore's definition, imply the existence of AGI. Although it could very well happen, I do not yet believe it is *inevitable* that we will reach this point.
comment31: (s31) Humans are flawed and not every human acts as ethical as possible. The fact that the system manages to appear human, implies it also does not act as ethical as possible. Like we try to set higher standards for humans than how they sometimes act (e.g., the law is not always followed but is there to try to ensure ethical behavior), the same should be done for machine. While I do believe folk morality and societal preferences should be considered in creating AMAs, they should not be used to test against.
comment10: (s10) Especially when systems become autonomous but are not generally aware, it is possible some new situation were accidentally excluded from the given limits. Usually, these faults will be taken out during a testing phase, but in certain situations, it is still possible the machine has unwanted behavior. In moral situation, as much or even more than in non-moral situa-

tions, you do not want the machine to act wrongly.

Sort 6160954

comment16: (s16) There are different kinds of harm, some of which could be emotional, and likely only detectable by machine with TOM capacities.
comment30: (s30) This seems like the most viable path forward for AMAs as it allows for context specificity, adaptive response, and learning.
comment19: (s19) I think there is a lot of learn from human heroism as an example for AMAs. There could of course be disadvantages as well. But why not take the best of humanity as a standard for machine morality.
comment36: (s36) I haven't thought too much about this. But if the moral agents are autonomous, then there should be accountability of the agent as well as the designers.
comment18: (s18) Autonomous machines require some form of moral capacity.
comment28: (s28) I don't think logic alone would be sufficient for encoding machine ethics as there are many illogical moral decisions.

Sort 6170343

comment38: (s38) You can teach moral machines moral rules. They are not fully ethical agents, but they can make decisions on a moral basis.
comment8: (s8) Autonomous machines must act morally in certain contexts, otherwise they are dangerous and unacceptable.
comment9: (s9) We recognize that part of our morality is machine-processable, while another part is not.
comment12: (s12) Maybe we don't need full ethical agents, but for autonomous machines we need more than damage control and legal compliance.
comment36: (s36) In some situations this is not the case. I can teach a machine to behave in the household in exactly the same way as I do - I still bear the responsibility for this.
comment18: (s18) At least some autonomous machines must be able to judge moral questions, otherwise they cannot be released into the world.

Sort 8270106

comment2: (s2) Same answer: Applications like autonomous vehicles and surveillance systems require moral behavior. I do not know of any scientifically sound argument that falsifies that.
comment16: (s16) As we argue in our papers, neuro-science and moral psychology should be considered to captured goals and desires.
comment5: (s5) Applications like autonomous vehicles and surveillance systems require moral behavior. I do not know of any scientifically sound argument that falsifies that. comment19: (s19) It is about wellbeing. All other attributes are at best instrumental to that.
comment14: (s14) Then the requirement of ethical AI as propagated by the

EU is undefined. Not only a wrong position but also a very dangerous one.

Sort 9021041

comment3: (s3)
comment2: (s2) Technology we already have such as search algorithms and driverless cars require and implement value judgments to a limited extent, primarily as directed by human input, but there are already hints that these limitations can and will be surpassed to at least some extent.
comment37: (s37) Machines are already at least at the level of animals, which are capable of not merely being restrained but being trained.
comment11: (s11)
comment10: (s10) It is impossible to write out a precise rule governing every situation and circumstance.
comment18: (s18) No human being begins the process of ethical reasoning with a blank slate. To pretend that we can create machines that lack our biases and are uninfluenced by our values is misguided.

Sort 09140324

comment32: (s32) These factors lead to trust - the probably most important aspect when people adopt AI.
comment40: (s40) It turns out that most of human ethics can be quantified if you have large enough datasets. Eventually, researchers will decode the moral code and use it to train AI.
comment2: (s2) There will be no other way than to develop ethical machines when humanity is supposed to rely with their life on them.
comment7: (s7) In case of autonomous machines, ethical programs won't stop them from hurting humans. Yet, they will do what we would expect and want them to do, i.e. hurt an intruder/attacker)
comment12: (s12) Abiding to law and safety doesn't stop AI and machines from discriminating against humans (i.e. because of color). We've already seen examples of it and therefore machines need to be ethical as well.
comment18: (s18) They might if we cannot find a better solution which I, however, firmly believe we do.

Sort 10081217

comment2: (s2) Technology we already have such as search algorithms and driverless cars require and implement value judgments to a limited extent, primarily as directed by human input, but there are already hints that these limitations can and will be surpassed to at least some extent.
comment37: (s37) Machines are already at least at the level of animals, which are capable of not merely being restrained but being trained.
comment11: (s11)
comment10: (s10) It is impossible to write out a precise rule governing every

situation and circumstance.
comment18: (s18) No human being begins the process of ethical reasoning with a blank slate. To pretend that we can create machines that lack our biases and are uninfluenced by our values is misguided.

5.2 **Factor 2**

Sort 06100643

comment16: (s16) My prior reply covers this, but core socio-moral practices like blaming and praising require mental-state inference to assess foreknowledge, intentionality, and so on.
comment20: (s20) A big part of moral competence is being able to predict and explain the behavior of agents – both one's own behavior, and importantly, the behavior of others. This is at the heart of complex socio-moral practices like blaming and praising, and these practices are ultimately the driver for how norm systems change over time. But this predictive and explanatory capacity depends on having an adequate folk theory of moral concepts involving valuation of states and actions, and also specifically of concepts surrounding agency and intentional action.
comment6: (s6) This is somewhat self-evident, but clearly follows from sampling work in philosophical ethics over the years, along with contemporary work in applied ethics. I don't view moral competence and ethical behavior as always tracking one another. Ethical theory is updated by and large when large classes of problems push against the normative boundaries that ethical theories establish. Moral competence is more basic in some sense, and precedes the adoption of any particular ethical or moral theory in our various communities.
comment3: (s3) The jury is out. Much of what fundamentally separates man from machine (presumably) are things like consciousness and intentionality. Can we adequately render these things computational? A bevy of literature stretching back decades argues that we likely can't, but who knows. The importance of developing machines that get close enough such that they can productively be part of man-machine teams is such that we should just put our noses to the grindstone and try.
comment39: (s39) Ultimately consciousness of of concern here, and specifically phenomenal consciousness, since the functional parts of consciousness are becoming better understood through research. We don't know in general if beliefs, desires, intentions, obligations and so on have phenomenal properties for sure, but there are compelling arguments to this effect. There is no first-person perspective for artificial systems replete with experiential properties. For some relevant phenomena, like moral motivation, this presents what could very well be an insuperable set of obstacles. We hand-code numeric values for some of these things (maybe), but it's unclear what that amounts to in terms

of "fundamental difference reduction" between human moral agents and artificial counterparts.
comment2: (s2) If anything, recent advances in AI belie a certain kind of tone-deafness among AI modernists and machine learning practitioners to the raft of complexity involved in representing and reasoning about many of the concepts one would need to have in a folk theory of morality. Beyond representation, the forms of reasoning and decision-making involved are beyond well-known limitations in artificial neural networks. This is just a matter of mathematics, and not an expression of a lack of confidence in those approaches being able to get to where they would need to go.

Sort 00000000

comment1: (s1) Again, don't see why one would follow from the other, as well as why we would need artificial morality.
comment12: (s12) It seems difficult for anything that we "need" as humans, except for food, shelter, peace and a functioning non-discriminatory and equal society. Certainly I do not see why we would need morality in AI systems.
comment15: (s15) Ethics is just much more complex than that.
comment3: (s3) "inevitably" is a strong word, and given the current developments it seems possible that there might either be moral or political grounds for stopping the development of AMAs or just technological inability
comment5: (s5) They are already deployed in morally salient contexts, such as for bail, loans, hiring, etc. All of these are morally salient in virtue of having long-lasting effects on people's lives.
comment7: (s7) Just as human beings who - moral agents - hurt other human beings machines might do so as well. Also, many moral theories allow for humans to hurt each other.

Sort 06260646

comment42: (s42) It is a fallacy that AMAs can perform better than humans. At present AMAs have only been demonstrated in laboratory tests, of limited scope. Such minimally ethical machines are far from real world application. If follows that there is a huge gap between the capabilities of present day minimal AMAs and the ability of humans to make judgements in ambiguous moral situations. I believe closing that gap will take many decades of research, and might even prove impossible without fundamental breakthroughs in AGI and machine consciousness.
comment32: (s32) If an AMA makes the wrong decision the outcomes could be disastrous. similarly the risks of malicious hacking of AMAs are serious. This verification, validation and transparency are critical to the success of (even limited) AMAs in real world use. Equally important is very strong governance of AMAs, so that their ethical performance is carefully monitored, and both accidents and near-misses thoroughly investigated.
comment16: (s16) AMAs will need to interact with people, and will therefore

need to model their beliefs and intended actions. Artificial theory of mind will allow AMAs to anticipate the likely consequences of a human's current actions and hence proactively intervene if they are likely to lead to harm.
comment41: (s41) My answer follows the same logic as (42) above. I can think of no principled argument that machines would be better moral agents than humans either now or in the future.
comment43: (s43) This is in the realm of science fiction. I agree that there is a possibility that in the far future we might be able to build robots like Data from Star Trek, and that such robots would not only be able to act ethically, but over time could become trusted friends and colleagues. But I see no reason why such robots would become teachers of morality.
comment39: (s39) There is a vast difference. We do not understand the cognitive processes of human morality - which likely depend on both rational and emotional responses, alongside experience. In contrast AMAs are based on simple abstract models, which are far from even scratching the surface of human ethical judgement.

Sort 08070438

comment35: (s35) This statement connects back to responsibility and the quest to avoid responsibility gaps. It is important to differentiate among the different understandings of the terms "agents" and "agency" in different fields. Here my understanding of agency relates to the capacity to act in a given environment. Thus, if machines make decisions that have some moral implication, they will have some primitive form of moral agency (regardless of the philosophical discussion on free will).
comment42: (s42) This statement reaches back to responsibility. Even if machines are capable of making moral decisions, completely replacing humans in such situations might lead to responsibility gaps.
comment18: (s18) If machines are fundamentally uncertain about morality they will have an incentive to better understand and align with one's moral preferences, by reasoning about the consequences of their actions and interacting with humans.
comment3: (s3) What most triggered my negative reaction to this statement is the "full" qualifier. I would agree that (many) machines would inevitably have ethical impact, but I don't believe that they should be full ethical agents, with the implication that this would mean replacing humans as the ethical agent. I believe the best way is to work towards collaboration between machines and humans whenever possible.
comment41: (s41) Here we go into metaethics and what is the purpose of morality. I disagree with this normative statement because it reduces ethics to rational decision-making.
comment39: (s39) For me, the most immediate, relevant, and dangerous aspect of this statement relates to responsibility. Our societal institutions do not have (yet?) the means to place responsibility or blames (and the consequence of it)

to machines.

Sort 9061106

comment5: (s5) AI systems already exist in morally salient contexts.
comment14: (s14) After centuries of debate, there is still very little consensus among humans regarding ethics. I do not think that would be the case if ethics was reducible to an algorithm.
comment42: (s42) I believe that this will take many years of research.
comment33: (s33) Computers that can autonomously and deliberately take moral action should be considered moral agents.
comment36: (s36) Machines should be considered moral agents if they are capable of autonomous and deliberate moral action.
comment34: (s34) I do not believe that free will is a prerequisite for independence or moral agency.

Sort 09130141

comment44: (s44) Something that is, by definition, more capable than humans across almost all cognitive tasks, and has no reason to favour the desires of humans, could quite clearly pose an existential threat for humanity.
comment18: (s18) I doubt whether we will ever 'solve' ethics (whatever that might mean), and to build machines under the assumption that we have (that there is asome absolute moral pricinciples or truth of which one may be certain) would at best be unhelpful and at worst be catastrophic.
comment14: (s14) I agree more with the sentiment of this statement than perhaps the actual precise technical content (i.e., if ethics is considered to be the study or reasoning about what is 'good', then insofar as all physical processes can be reduced to computation, then ethics might indeed be reduced to computation). I think, however, that the more standard/colloquial reading of this statement is one that ignores the complexities of ethics and approaches such questions with an engineer's mindset instead of a philosopher's. comment40: (s40) The reasons I disagree with this statement are the same reasons I agree with the statement "Ethics cannot be reduced to a computational algorithm.", so please refer to my previous answer.
comment3: (s3) Very few things, to my mind, are truly inevitable, and this is certainly not one of them. There is far too much uncertainty regarding: a) the concepts in this statement. and b) the future development of intelligent machines.
comment7: (s7) The 'morality' of AMAs may not preclude them from hurting humans (indeed, one may easily imagine extreme scenarios in which not only would they be compelled to hurt a human being, that we might, as humans, think it is morally right for them to do so).

Sort 10021133

comment42: (s42) The case of GPT-3 recently changed my perspective on this. I was far more optimistic, but the disappointment of this much-hyped system has made me reconsider my view. I now believe we are far from having morally competent agents, and that the threats from lack of transparency, privacy, security, etc. are far more pressing, morally speaking.
comment13: (s13) The search for a unifying account of metaethics will not be of much help in the practical instantiation of AMAs. We need to shift our thinking into the strictly normative domain, and this would entail context-specific evaluations.
comment5: (s5) We cannot turn back the clock. These systems are already embedded in our everyday lives, with Big Data being but one example of ubiquitous computing.
comment34: (s34) Free will and other metaphysical theories do not do much work in the AMA debate. By moving to the normative domain, we avoid such metaphysical squabbles. Of course, metaphysics matters, but free will, understood in metaphysical terms, does not.
comment21: (s21) I do not think they MUST be able to. There may be other, embodied, criteria, that are more relevant to moral thinking than merely the grasping of moral principles. comment15: (s15) What comes to count as "good" is often subject to change. This is also true for "bad" things. However, we should focus on not unjustifiably making moral mistakes. This could in itself come to constittue and kind of good.

Sort 10021236

comment42: (s42) This is what I learnt from scientists and engineers in my conversations with them and in reading their articles
comment14: (s14) Ethics has to do with emotions, engaging in meaningful discussions and deliberations about things one cares about personally, empathy etc.
comment36: (s36) MAchine cannot bear moral responsibility so it is not a good idea to consider them moral agents
comment1: (s1) As above, technological progress can and should be guided by ethical and societal considerations and can happen also without artificial morality
comment3: (s3) This is unwarranted given the current scientific development, plus human/societal choices may prevent this to happen even if technically possible
comment28: (s28) See above: ethics going beyond logic and argumentation, it is something more personal, requiring taking a position, taking responsibility, caring etc.

10120149

comment36: (s36) AMAs can be sophisticated tools to further complicate the 'problem of many hands'. There are already reported cases where people

delegated their responsibility to allegedly "ethical' technology. It's a new tool in an old & shady business.
comment35: (s35) A human being is a complex subject, even when we focus on its identity as a moral agent. Part of what matters to the idea of 'full' moral agency is that human beings are in a constant process of evolvement in relationships with other people and non-people, with society, with culture, history, religion etc. An AMA would be different from the start, even if its development would be simulated in a human-like way. If people end up experiencing a shared history and culture with artificial moral agents (in the long term), it would still be a very different different experience from the shared history & culture between people. AMAs are primitive in this way. In short: Being the 'original' moral agents (as in 'the ones that were there before') will (and should) always matter. On the other hand, I expect it to be very interesting and productive to study moral agency in its primitive form - just like Wittgenstein started to study language by considering primitive forms .
comment14: (s14) Moral judgements, sentiments and motivations depend on a holistic perception of the world. Ethics would not exist at all without this special perspective that is shaped by reasons, emotions, culture and history. A representation of ethics in the form of computational algorithms (or in any kind of model) is an abstraction in comparison to the rich features of the ethical world. Such a representation may successfully serve a specific purpose when realized in a technological artifact, but no representation could possibliy model the whole ethical world as a subset of its features.
comment40: (s40) *Some* features can, but not all of them. To understand moral agency, quantitative terms are not sufficient - additionally, an interpretation is needed. Ultimately, the judgement that something counts as a moral action (let alone if it's right or wrong) is not based on quantitative terms, but depends on the phenomena as perceived and interpreted by persons.
comment43: (s43) Although there is some benefit in a systematic moral rationality, this is not how the ethical world is ultimately structured. For once, it is hard to subsume conrete situations and events under general ethical principles. There is no general rule for this. Alternatively, a data-driven approach might suffer from the ought-is fallacy. I suspect that an AMA would only give us commonplace truisms and superficial clues.
comment3: (s3) This is the inverse to the above arguments: A full ethical agent would be one that perceives the world in a holistic way, shaped by reasons, emotions, culture and history etc. It would have to grow up and 'live' in a process of constant evolvement in relationships with people, with society, with culture, history, religion etc. Although this might not be considered impossible if taken up in a though experiment, it makes no sense in reality.

5.3 **Factor 3**

Sort 06110951

comment4: (s4) Again, there is already overwhelming evidence that AI systems reside in an abundance of morally salient contexts.
comment5: (s5) There is already overwhelming evidence that AI systems reside in an abundance of morally salient contexts.
comment15: (s15) Negative ethical restraints will not be sufficient. Many social and ethical issues and progress itself require careful deliberation and proactivity.
comment10: (s10) These may be sufficient for basic kinds of product safety, but it is already demonstrable that AI has massive subtle and pervasive impacts that require far deeper technical, social, and ethical thinking and all kinds of internal and external regulation and best practices to avoid harming humans.
comment11: (s11) I suppose one could come up with such an argument relying on one or more moral theories. One could equally come up with arguments that creating moral machines is permissible or even obligatory. So this argument is far too strong.
comment34: (s34) The premise begs the question on machine free will and on human free will. The best philosophical and physical reasoning suggests to me that there is no free will in the libertarian sense. So this argument is unlikely to be a good basis. Humans have designs, like machines do, so the fact that machines and do or do not have free will and that they are designed is not especially salient to the question of moral agency.

Sort 07020805

comment4: (s4) See above answer - [AI systems are already in morally salient contexts, and have been for years. Witness their use in creating news filter bubbles, criminal sentencing and parole determinations, detecting welfare fraud, etc., and all of the misuses and abuses associated with those, as just the tip of the iceberg.] AI is already locked in for ethical decision making. There is no choice except whether to be better or worse at it.
comment5: (s5) AI systems are already in morally salient contexts, and have been for years. Witness their use in creating news filter bubbles, criminal sentencing and parole determinations, detecting welfare fraud, etc., and all of the misuses and abuses associated with those, as just the tip of the iceberg.
comment30: (s30) AI ethics needs to be a multilayered approach, utilizing all of the ethical resources available to humanity, acting not only within the systems themselves but also in their design, implementation, in the ways that they interact with humans and each other, in the ways they are used and interact with society, etc.
comment28: (s28) While I think that logic can do a whole lot, there are multiple kinds of logic, not just symbolic & mathematical, which I assume is what the statement is referring to. So I cannot say it would be ideal. Logic is often

in reference to some goal, and so the topic of ends - teleology - must be considered as well.
comment29: (s29) Computers do not understand human stories. Many humans do not even understand human stories. Any attempt to do this will result is bizarre machine behavior. Not to mention the question of what stories to feed the machine - Herodotus? Religious texts? Pornography? Harry Potter? Ethical cases might make sense, but even then the delineations of what should be learned from each case would need to be extremely closely controlled, and would be far beyond merely reading stories.
comment10: (s10) These sorts of measures are necessary but not sufficient for the implementation of AI in society. Appeal to these mere approaches is a form of ethical minimalism that will lead to avoiding small problems while at the same time walking humanity into huge problems.

Sort 09120415

comment5: (s5) We already see examples with self-driving cars and trading bots, but more generally I see it as (nearly?) inevitable that AI systems will eventually be deployed in every domain that requires intelligence, which is essentially a superset of all domains that contain morally salient situations.
comment45: (s45) I answered +3 or +4 to the statement that working on AIs with moral capabilities would lead to a better understanding of morality. The version with an existing AGI that has moral reasoning capabilities is a superset of that, as I see it, since presumably by that point we have worked for a while on developing such capabilities AND we have produced a superintelligent teacher that can see past our errors in logic and our biases.
comment44: (s44) I find thought experiments such as the paperclip maximizer to show quite convincingly that for an AGI to pursue a goal that is merely orthogonal to human values could plausibly present an existential threat.
comment35: (s35) It is currently primitive, but I believe it will eventually be possible to create AIs that match or exceed humans in every intellectual capability, which includes moral reasoning/agency.
comment34: (s34) I find "free will" to be a somewhat vague term whose meaning often depends on the speaker. However, because I view the human brain as a material object that performs computation, I am prepared to say that if human beings can be said to have free will under any particular definition, then it is possible to implement a program that can be said to have free will under that particular definition.
comment19: (s19) Generally speaking I think AMAs should have uncertainty about which moral theories they would follow, and I took statements along the lines of "AMAs should be designed with reference to ..." as meaning that they should be designed solely or primarily with reference to that particular moral theory. I don't know exactly what human heroism is supposed to entail exactly, but out of the specific moral theories mentioned in the statements it seemed to me to be the vaguest and possibly most problematic one.

Sort 09121118

comment18: (s18) Morality is a critical determinant of ethical behavior, and there is incredible disagreement among humans. If AGI does not have uncertainty about morality, its behavior may be arbitrarily bad given a commitment to the wrong set of moral principles.
comment16: (s16) It is important for machines to understand us deeply, theory of mind seems like a necessary mechanism for doing so.
comment45: (s45) I believe we only truly understand something when we have a good understanding of how to engineer it. In this way, engineering moral reasoning would demonstrate the depth of our understanding.
comment33: (s33) Computers generally speaking, at least as far as we know, can implement or simulate any algorithmic process. if we believe in materialism, then all facets of human behavior are not fundamentally beyond computation. By that logic, conceptual understanding is not a supernatural property, and therefor is not out of range of computers.
comment10: (s10) These are somewhat superficial features, it is not clear why these should be the necessary conditions.
comment1: (s1) There does not seem to be a logical connection between these ideas.

Sort 10130524

comment14: (s14) A program is only as ethically good or ethically wrong as it is programmed, because it does not make ethical decisions or evaluate them on its own - without prior input. Rather, ethics is a combination of factors that lead to it being holistic. Since there are so many factors to implement ethics in algorithms, an ethical implementation always represents only a part of the ethics.
comment10: (s10) Certification and regulations are necessary, but also a legal framework that covers the whole. It is crucial that companies and customers work together from the very beginning with the individual states and the community of states on the issues mentioned.
comment28: (s28) Machines work strictly logically and so it is only logical that all systems are subject to logic. The logic of the machine also reveals logical inadequacies of the human being and can thus be adapted. Also the way of working can be made explainable for all partners.
comment43: (s43) A moral teacher should by no means be a machine, but by taking into account numerous factors, man himself should be empowered throughout his life to make moral decisions and also to define morality through individual and social learning. Throughout the world, there will always be different moral concepts at all times, but these contribute to the fact that morality is not a machine concept, but rather, just like man himself, is subject to change and reflection, but also passes on unrenounceable contents.(weblink) (free version) comment39: (s39) The machine's ability to act is usually dependent on that of the human being and should also not take over actions

itself or simply act without control or transparency. A thinking together of the instances man and machine probably leads to good decisions, whereby the mere machine thinking often lags behind, as well as the human being has certain limitations. A comparison is not necessarily meaningful, since man and machine have different strengths and still the human being, the centre of attention, should be represented.
comment3: (s3) Compare what I have written on question 14. With all the progress made and with all the ethics implementation that may be used in the future, the role of the human being as an ethical actor should not be underestimated. Man remains the only "full-fledged" ethical agent, even when supported by technology and systems.

Sort 10010119

comment38: (s38) There are moral agents of a number of different kinds, some of which are non-human animals. Others still are artificial and not biological as discussed in literature. It is obsolete and problematic to consider morality as a uniquely human affair.
comment6: (s6) see above. Also, these might not have been my top three picks necessarily, i had to move them around a bit to make sure they will all fit the pre-determined scheme (vaguely gaussian) that you prepared for me. I think the rigidity in your setup restricted and distorted the value I would have normally given to some of the items. Moving them back and around was also awkward as I could no longer read the complete extent of the question... So, yeah...
comment16: (s16) Both 16 and 6 deal with the same issue: artificial moral agents already do interact and take part in complex moral situation. Equipping them with the possibility of self-reflection and the possibility to guess and estimate intentions, feelings, and emotions of other agents involved in the aforementioned situation is clearly going to be necessary for their integration within a moral community.
comment35: (s35) This is the flipside of the statement that claimed that morality was not exclusively, necessarily a human affair. nArtificial moral agency can become the paradigm of ethics, and is not necessarily bound to remain a lesser, more mechanical form of assessing and relating to situations. The novels of Iain M. Banks have a similar perspective: humans and their biological setup are actually obstacles for reaching a fairer and more desirable society. comment19: (s19) Nothing should be designed with reference to human heroism.nTry 'with reference to human compassion', and I might agree a little more.
comment7: (s7) I disagreed with this statement as It might lead to them desiring (or making the utilitarian estimate) that we need to be eliminated :DnThe SF tradition is rich with examples of similar scenarios. Autonomy and the capability of subjectively take moral stances does not necessarily entail their being agreeable or friendly.

5.4 **Factor 4**

Sort 06120543

comment34: (s34) See above. Moreover, we don't even know what free will is for humans. We know how it feels, but this is another dimension of the discourse, which makes sense for humans but not for machines.
comment33: (s33) These are machines. Very complex, but machines. Yes, some programmer can shape these machines to have an input/output function that produces a behaviour that some human observer may see as analogous to human moral behaviour, but computing machines are ultimately and intrinsically incapable of understanding in human terms, which is the basis for moral agency.
comment14: (s14) Ethics is not only about quantity. There are inherent qualitative concepts that put it outside of the box of quantitative science, let alone mathematics, let alone algorithmic computation.
comment43: (s43) How can software agents be teaching to humans?Yes, there are agents who can teach humans about chess and go, but that it because thanks to their computational power they explored more parts of the solution space of these games.Morality does not have such a space to explore: it is about convictions and convergence or clash. Nothing to teach there, at least not from a machine.
comment3: (s3) Nothing is inevitable, and let's all enjoy Sci-Fi for what it is: entertainment. If we all stopped working in computer labs, would we have software agents? No. So it's our choice. I repeat: nothing is inevitable.
comment1: (s1) Nothing is required. And please let us all stop believing in technological progress as an external inevitable physical phenomenon like the death of the Sun or Andromeda and the Milky Way eventually colliding. It's us: we human beings create technological progress. It can go awry for so many reasons (pandemics, wars, etc.)

Sort 06221030

comment42: (s42) Right now the computer codes that we have are not sophisticated enough to produce machines that can act morally.
comment38: (s38) It seems likely that there are moral agents that are not humans.
comment44: (s44) If AGIs are indifferent to human values, they may end up treating humanity as a means to promoting some other end.
comment11: (s11) If it turns out that it is best to have moral machines, then utilitarianism would permit and even require that we bring about moral machines.
comment4: (s4) We are at very early stages in the development of moral AIs.
comment26: (s26) This seems too detailed.

Sort 06221114

comment33: (s33) You start being moral when you recognize your shared humanness with others and understand that, like it or not, you are in relationship with them. Until machines get that (and I'm suspicious of their ability to do so) then they're not going to have full moral agency.
comment44: (s44) If the trash-sweeping machine decides that the best way to eliminate trash is to kill all humans, then we're in trouble.
comment6: (s6) To at least give them a rough idea of what to do. comment43: (s43) Unless machines develop empathy and compassion, they aren't really going to pass down lessons in a meaningful way.
comment9: (s9) Are you kidding? We can't even agree on which moral system is best, let alone understand morality completely. Like above, trying to teach morality to machines through quantitative parameters is going to limit our understanding of morality, not enhance it.
comment40: (s40) Every time philosophers try to quantify abstract concepts like value or "amount of moral value" it ends up falling apart. Let's not make the same mistake with killer robots, please.

Sort 06241046

comment14: (s14)
comment42: (s42)
comment34: (s34) Free will strikes me as a basic condition of responsibility and therefore of moral agency
comment39: (s39) see above
comment3: (s3)
comment45: (s45)

Sort 07100409

comment42: (s42) When artificial agents begin to become self-aware....
comment14: (s14) Ethics are fundamentally intersubjective. Computers lack subjectivity.
comment33: (s33) What is it like to be a computer? If there is nothing that it is like to be a computer then how can a computer have conceptual understanding?
comment3: (s3) Agency requires consciousness
comment7: (s7) Hurt is rarely logical
comment39: (s39) Without sentience computers cannot express agency

Sort 09050630

comment12: (s12) inherently safe and law abiding machines are good enough
comment16: (s16) this is greatly help amas to simulate human ethical behaviour

comment19: (s19) this is such that AMAs can replicate human heroism in times of need
comment39: (s39) we cannot expect 2 different sets of moral codes comment28: (s28) logic is the only choice
comment4: (s4) AIs are already able to simulate self learning

Sort 09080738

comment6: (s6) By moral competence here, I did not understand moral competence of the machine, but that we will have to rest on the moral competence of other human agents. Note also that the claim "to govern unpredictable actions" is rather paradoxical
comment15: (s15) That seems evident, because there cannot be an a priori exhaustive list of what constitutes ethical behavior or not in various circumstances
comment14: (s14) Mostly because ethics, or at least what ethics deals with, requires a plurality of points of view related to the physical embodiment and location of independent agents, which means that there is no possible universal description of an embodied agent's position or situation. Even considering an embodied AMA the difficulty is that an agent only acts morally if he or she or it could have acted immorally. We could simulate that by including a radom generator but acting immorally is not the same thing as acting randomly
comment9: (s9) Why would that be the case? What argument or reasoning underlies this claim?
comment45: (s45) I do not see AGI to be in anyway a viable project. In fact, I do not understand what we are talking about when people speak of AGI
comment3: (s3) I see no reason why machines could not become full moral agents, though the machines we can now produce certainly are not and the planned AMA I know of will certainly not be full or real moral agents though there is no reason to think that it is in principle impossible

Sort 10020345

comment42: (s42) Unexpected situations are hardly manageable artificially
comment33: (s33) Computer lack empathetic experiences which give humans the conceptual understanding needed for moral agency
comment14: (s14) Ethics is not about calculations, but about not quantifiable preferences
comment43: (s43) Morality cannot be teached simply through information transfer: it requires experiential sharing.
comment41: (s41) Irrationality, seduction and emotional turmoil might be the push for a good moral agency rather than an obstacle to it
comment9: (s9) AI is not explanatory relevant with regards to morality

5.5 **Factor 5**

Sort 06110921

comment21: (s21) It's very necessary for an AMA to have a theoretical framework which is grounded in basic moral principles.
comment32: (s32) Transparency is very important.
comment44: (s44) In the long run I believe this to be true.
comment12: (s12) As I stated for question 10, rules and regulations are not enough. Self regulation whereby AMAs evaluate their own decision will be required.
comment33: (s33) This was a hard question because it's not clear what the timescale is. They certainly lack it today , but we don't know whether they will in the future.
comment10: (s10) These are not understandable to most people. DIscussions about ethics of AI need to be understandable, otherwise serious implications will not be understood. Recent event makes it very clear that big tech companies will do their best to escape from moral responsibilities.

Sort 10010115

comment13: (s13) Deliberation understood as an argument for the application of universal principles to concrete situations is a crucial element in moral decision-making
comment43: (s43) I believe that for moral development of individuals from the meta-criteria of autonomy it is essential that AMAs are instructors rather than advisors
comment38: (s38) Some animals are moral agents
comment27: (s27) I believe that machines can enhance us as moral agents if they manage to distance us reflectively from our intuitions, which are very much determined by social preferences
comment18: (s18) Machines which, like cars without drivers, can affect both the well-being and the life of humans cannot be indifferent to morality
comment34: (s34) As some philosophers have pointed out, human beings could still be moral agents even if they were partially determined by their biology