The hidden complexities of "moral circles" + a theory of change for longtermist MCE

Purpose

Some effective altruists aim to expand humanity's moral circle to encompass more or all of sentient life. However, the term "moral circle" is very ambiguous. Different definitions yield different conceptions of moral circle expansion, some of which could be much higher priority than others. In this post I discuss different ways we can define a moral circle, and offer some thoughts on which are most relevant to those interested in improving the long-term future of sentient life.

Overall, I'd actually rather use the term "moral circle" less, and instead focus on finer-grained consequences of work to benefit groups like nonhuman animals or artificial sentience. So why did I write this? A few reasons:

- Comparing definitions of the moral circle necessitates asking important strategic questions about what our advocacy should prioritise that I want to discuss. (This is my primary interest in the post.)
- If people are going to keep talking about "moral circles", I'd like us to be able to communicate more clearly about what we have in mind by that term.
- I hope seeing the ambiguity of the current approach will help create space for alternative models of what we aim to accomplish in advocacy for nonhumans—stay tuned for one of those from me in the coming weeks/months.

This post is peppered with my own opinions about strategic questions, which might be more useful to people who share components of my own view of how the future might unfold. Hopefully I did a decent job at indicating how my worldview influences my conclusions. I think a lot of discussion could be of fairly broad interest regardless.

The style of this doc is less "propose resolutions to some well-defined questions about moral circles" and more "explore a number of important issues using disambiguating moral circles as a unifying frame/motivation".

Feel free to jump around to parts that interest you most. A few sections that might be particularly important:

- Choice 1 Analysis: Are attitudes or behaviours a more important target for MCE work?
  Summary: Very tentatively, attitudes are a more important target with respect to their upside, but promoting behaviours may have slightly less backfire risk.
- How might AI developers and deployers have a disproportionate impact on moral patients?
  Summary: There's a few plausible mechanisms for this, especially if we expect alignment to succeed. I would be worried about specifically targeting developers in animal

advocacy messaging due to backfire risks, but this may be a consideration in favour of continued marginal investment in positive forms of outreach for groups like nonhuman animals in regions where AI developers disproportionately live.[1]

- Should we be interested in preventing moral circle *exclusion* instead?
  Summary: It might be easier to harm than to help moral patients. If that holds, tracking negative attitudes or behaviours (like hatred or disgust) might be more important than tracking positive ones (like moral consideration or affection).

Acknowledgements: this is based on a lightning talk I gave at EA NYC—though as you can see, it's expanded in length a little since then! I think I would have been substantially less likely to have written this without any of the following: the original lightning talks event, a grant from CRS, a stay at CEEALAR, and a blog post writing day at CLR.

This was mostly written in 2022; I'd been sitting on it (and some other work) for a while in favour of what I think are higher-priority projects, so prioritised going ahead with publishing this over getting it more polished. Feel free to point out errata in comments!

***Status of this doc**:* Not sure whether to break off Choice 1 Analysis: Are attitudes or behaviours a more important target for MCE work? (the theory of change section) and/or How might AI developers have a disproportionate impact on moral patients? into different posts, leaning against.

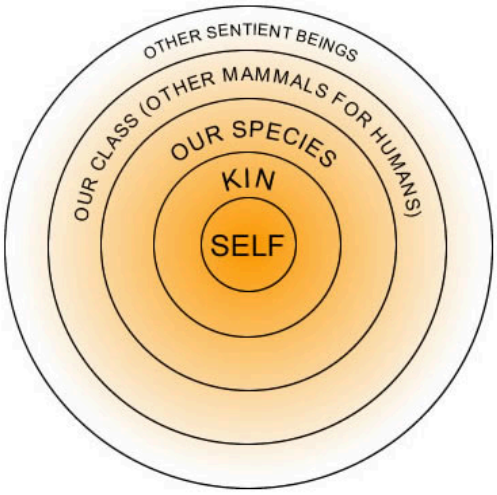
Introduction

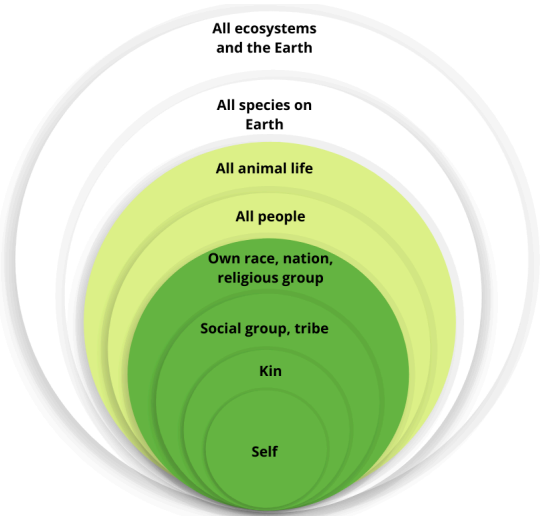
Peter Singer's 1981 book *The Expanding Circle* introduced the concept of a *moral circle* to describe how an individual or a society extends moral consideration to some classes of beings but not to others.[2] The set of beings a subject morally cares about is construed as a circle in order to allow for an appealing metaphor in which the moral circle expands over time to encompass broader classes of beings.

[1] This consideration recommends against prioritising work in many countries where farmed animal advocacy is especially neglected (where neglectedness = a low ratio of animal advocacy to current or near-future animal farming). But because I'm overall sceptical that moral attitudes of AI developers/deployers will have an outsize influence on the long-term future, and (to a lesser extent) am unsure how robustly marginal animal advocacy would affect these attitudes, I think this should only be a quite minor update against the importance of neglectedness in farmed animal advocacy strategy.

[2] The term actually originated with W. E. H. Lecky's (1986) *The History of European Morals*. Singer excerpts it as follows:

> The moral unity to be expected in different ages is not a unity of standard, or of acts, but a unity of tendency… At one time the benevolent affections embrace merely the family, soon the circle expanding includes first a class, then a nation, then a coalition of nations, then all humanity, and finally, its influence is felt in the dealings of man with the animal world.





Images from equalism.org.uk and Cafaro and Primack 2013 (via Aird 2020).

The moral circle of the typical person is said to have transformed in this way throughout history: at first narrowly embracing one's relatives or tribe, it now includes larger ethnic groups, nation-states, or maybe even all of humanity.[3] The intuition behind the expanding circle metaphor is that (i) it is easier to care about those who are similar or close to you than others who are dissimilar or distant, but (ii) moral progress necessitates making the effort to care for even those in your periphery.

While the idea of a steadily expanding growing moral circle is uplifting, the truth is more complicated. Here are a few problems with the "moral circle" metaphor:

- **Our moral concern for some groups may have diminished over time**. Gwern's *The Narrowing Circle* lists some examples; of particular note is a lessened concern for past people (especially our ancestors) in many regions relative to historic norms.
- **Along what axis is our moral circle expanding/contracting?** There is no obvious distance metric by which to collapse arbitrary classes of beings onto a single metric of moral proximity or remoteness. Michael Aird's post *Moral circles: Degrees, dimensions, visuals* discusses more specific conceptions of distance along which we may be able to observe moral circles expand or contract.
- **Increasing moral consideration for one group might not have much of an effect on moral consideration for other groups.** (h/t Megan Kinnemet)
- **Moral circle expansion is not inevitable.** The picture of an expanding moral circle has also been criticised for creating an illusion of inevitability, when in reality many groups deserving of moral consideration may be denied it for reasons of historic accident.

[3] In individuals, we might also say that your moral circle will expand the closer you get to a reflective equilibrium in your values by means of moral reflection, until your circle reaches a stable "correct size". (Although it is possible you will overshoot along some dimension during your reflection and need to contract your circle back inwards.)

These all point to important flawed assumptions in standard framings of the moral circle, and may merit retiring the term "moral circle" altogether to prevent confusion. In the rest of the piece though, I'll continue to use "moral circle" as the label for the concept I'm interested in defining, and simply ask that readers view as contingent rather than definitional any properties of the moral circle mentioned above which originally may have motivated its invention but are now in dispute.

The issues I am concerned with below are more foundational than *whether* or *how* the moral circle is expanding. Rather, I want to ask:

- ***Whose moral circle* is expanding or contracting**? Is it that of an individual or that of a society? If we plan to discuss the moral circle of a society, how can we relate that society's moral circle to those of the individuals within it?
- ***What sort of beings* lie in the potential domain of the moral circle?**
- ***What properties* determine moral circle inclusion**? Attitudes or behaviours? Do moral gaps or partiality matter?

I have in mind two main goals for a preferred definition of a moral circle, in addition to wanting it to accord well with current usage.

- **Point out ways for us to intervene in the world.** Interventions that cause a neglected class of beings to be more included in people's moral circles should at least be *prima facie* worthy of our consideration as priorities. Nevertheless, such interventions may have counterproductive backfire effects that make them neutral or net-negative in expectation.
- **Be a good metric of progress towards a better future.** The greater the extent to which an entity is included in a people's moral circles, the more we should generally expect their interests to be respected in the future—though the same caveat about unintended negative consequences applies.

I think the first goal (making promising intervention points apparent) is somewhat more important than the second.

What determines moral circle inclusion?

This section will talk about the moral circle of an individual for simplicity, but all of these considerations apply to societal moral circles as well.

Choice 1: Attitudes or behaviours?

Two questions which could be of interest in demarcating a moral circle are: (i) what *beliefs* does a person hold about whether certain individuals matter morally, and (ii) does a person demonstrate concern for certain individuals through their *actions*?

Compare these two situations.

(1) John does not think animals' interests matter, but he is vegan because he thinks accepting cruelty towards animals encourages cruelty towards other humans. He also encourages others to follow his example.

(2) Mary thinks animals' interests matter, but she does not reduce her animal product consumption or take other actions which could benefit animals because she views these as too effortful for her, or thinks they are a low priority relative to other altruistic actions.

In an entirely ***attitudinal*** conception of the moral circle, only Mary's moral circle would include animals, because she but not John believes that animals' interests are important. In an entirely ***behavioural*** conception of the moral circle, only John's moral circle would include animals, because he but not Mary takes actions that benefit animals.

The moral expansiveness scale (MES) popularised by Crimson et al. 2016 is mostly attitudinal in the above sense, but not entirely so. The description of a particular level of moral concern for an entity conflates (i) the "level of moral concern and standing" an entity "deserves", (ii) a respondent's "moral obligation to ensure their welfare", and (iii) a respondent's felt "sense of personal responsibility for their treatment".

(iii) is the criterion which is the most focused on behaviours: it seems difficult for someone to feel personally responsible for an entity's wellbeing without trying to take actions to benefit them.

Appendix A

The Moral Expansiveness Scale (MES)

Inner Circle of Moral Concern: These entities deserve the **highest level of moral concern and standing.** You have a moral obligation to ensure their welfare and feel a sense of personal responsibility for their treatment.

Outer Circle of Moral Concern: These entities deserve **moderate moral concern and standing.** You are concerned about their moral treatment; however, your sense of obligation and personal responsibility is greatly reduced.

Fringes of Moral Concern: These entities deserve **minimal moral concern and standing**, but you are not morally obligated or personally responsible for their moral treatment.

Outside the Moral Boundary: These entities deserve **no moral concern or standing.** Feeling concern or personal responsibility for their moral treatment is extreme or nonsensical.

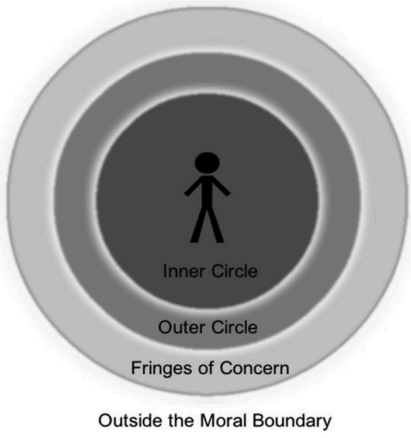


Figure from Crimston et al. 2016

Consider also the following case:

(3) Susan thinks that animals' interests matter, but she believes existence is always preferable to nonexistence. Therefore, by the logic of the larder argument she is supportive of factory farming.

I think Susan is very mistaken about what is in animals' interests, but does that mean that animals are absent from her moral circle? Not according to the MES or other uses of the term, as far as I can tell. When we discuss someone's moral circle, whether behavioural or attitudinal, we are interested only in the extent of their concern for the moral patient, rather than whether that concern is directed in a fruitful way.

Moral circle expansion that conveys a deeply flawed picture of moral patients' interests is probably harmful in the near-term, but might still be helpful in the long-term so long as flawed beliefs or values of this sort would not be expected to persist for an extremely long time.

In particular, we might think that current mistaken beliefs about others' interests are harmless because future decision-makers will have access to more empirical knowledge relevant to helping moral patients effectively. This could include questions like how best to advocate on a being's behalf, what causes various beings to suffer and how intensely, and which sort of beings are sentient—but it is also possible these questions crucially depend on non-empirical beliefs and values.

Whether non-empirical beliefs about moral patients' interests would also trend towards "truth" seems like a fraught and uncertain matter. In the absence of a value lock-in, I tend to think that people's moral views will track beings' actual interests better (so long as they are motivated to understand these interests to begin with). I wouldn't count on this effect, however, and value lock-ins (especially those brought about by a singleton AI) seem quite plausible to me.

Analysis

Attitudes about potential moral patients only matter because influential actors who hold those attitudes may act in ways which benefit or harm those patients. Call such actors *decision-makers*. The attitudes or behaviour of non-decision-makers in the far future are only indirectly relevant.

If a future society consists of humans and is either democratic or allows its inhabitants to make relevant personal choices—for example, whether to incentivise a market to bring into existence non-human animals as food products—the decision-makers of the long-term future might be ordinary people. Otherwise, they might be elites, or agentic AIs.

Below I will argue:

1. Of the variables we can affect with moral circle expansion work, the **behaviour** of future decision-makers has the most influence on the **wellbeing** of future moral patients.
2. Of the variables we can affect with moral circle expansion work, the **attitudes** of future decision-makers have the most influence on the **behaviours** of future decision-makers.
3. Between our current attitudes and behaviours towards moral patients, our current **attitudes** have more influence on the **attitudes** of future decision-makers.
    - Therefore, in MCE work it is more important from an upside perspective for us to focus on people's current **attitudes** towards moral patients than their behaviour. Nevertheless, changing current behaviour can remain *instrumentally* valuable in order to change current attitudes.

4. However, there is more **backfire risk** from influencing attitudes than influencing behaviour.

It's a little confusing to read this in list form, so feel free to skip down to the flowcharts for a graphical overview.

I'm also not very confident in any of these points! I'll try to point out along the way where the argument is particularly sensitive to various assumptions one might disagree with. If you just want to read about different ways the moral circle can be defined without getting bogged down in macrostrategy, you can skip to Choice 2.

1. The behaviour of future decision-makers most influences future wellbeing

In longtermist moral circle expansion work, our ultimate goal is to improve the wellbeing of far-future moral patients. If the far future contains decision-makers who can reliably influence the treatment of moral patients on a broad scale, then the best way to further this goal is probably indirectly, by influencing the behaviour of those decision-makers.

There are at least three exceptions to the idea that the attitudes and behaviour of present actors essentially only matter indirectly to longtermist MCE, though in all likelihood there are other concerns I'm neglecting.

(1) **Existential risk** (x-risk) means the existence of future decision-makers and future moral patients may depend on our current actions.
(2) There may be certain **long-lasting interventions** we can enact now that can persist in a future with no decision-makers or with decision-makers of limited capacity. Liedholm 2019 discusses some of these, such as directed gene drives, in the context of terrestrial wild animal welfare. Another which seems promising to me is preventing directed or accidental panspermia; Šimčikas 2022 lists some interventions that might accomplish this.
(3) Efforts to affect future decision-makers may be **intractable**, or be expected to have **negative effects** by default due to backfire risk.

Existential risk is mostly beyond the scope of this post. It is very possible that the continuation of earth-originating life is a huge input to the future wellbeing of groups like nonhuman animals (whether for good or ill)—but influencing either attitudes or behaviours towards nonhuman animals is unlikely to affect existential risk much.

Say we are convinced that x-risk reduction is extremely positive for groups like nonhuman animals, to the extent any nonnegligible reduction in x-risk dominates the effects of improving the wellbeing of nonhumans in surviving futures. Then we should deprioritise MCE work targeted at nonhumans in favour of x-risk reduction.

One form of MCE work may still be very important even if x-risk reduction dominates our future impact: expanding the moral circle to encompass future generations. In this case, we would most

want to influence the behaviours of *present* decision-makers (not future ones), to the extent that we think we are in a particularly precarious moment in time.

Long-lasting interventions could also be very important, especially the chances of a human-controlled future look bleak. However, in a future controlled by an unaligned AI, "long-lasting interventions" seem fairly likely to be reversed or rendered irrelevant.

I don't find **intractability** concerns very persuasive: some precedents like the antislavery movement certainly indicate it is possible to influence attitudes or behaviours in a broad and lasting way, and lock-in means that values might stabilise after a critical moment instead of continuing to change unpredictably over time.[4] I'm more sympathetic to the argument that we should tend to avoid interventions directed at having a large future impact, because in many domains it is easier to have a **negative effect** than a positive one. I'm not sufficiently convinced that I believe we should abandon efforts to improve future behaviour towards groups like nonhuman animals, but I do think we should strongly prefer strategies to accomplish this that seem more robust to the various backfire risks that have been identified so far.

With the above caveats in mind, I'll proceed as though what we aim to influence is the behaviour of future decision-makers, and next ask what most influences that behaviour.

2. The behaviour of future decision-makers is most influenced by their attitudes

An agent's behaviour can be understood as a function of their attitudes (both endorsed and unendorsed), their knowledge, and constraints imposed by their environment. Of these, I think attitudes are most influential to the behaviour of future decision-makers, because future decision-makers seem likely to have better knowledge and more control over their environment.

Why might future decision-makers have better knowledge and more control over their environment?

- **Knowledge:** Scientific progress might slow (or have already slowed) along various dimensions, but it seems unlikely that it would reach a ceiling any time soon, barring collapse scenarios. If this is right, discovering empirical facts earlier is only helpful insofar as it helps us achieve attitudinal or behavioural changes that have lasting consequences.

  There *are* some situations where dissemination of accurate knowledge to decision-makers could be quite difficult, such as futures which are highly fragmented or which have

[4] Even if values do not become stable, MCE could still have a large impact on the future in expectation. One metaphor I like is that the sum of a random walk starting at 0 that has *t* steps is going to be *tk* less than the sum of a random walk starting at *k*. Even if values change erratically in the future, changing the starting point of the erratic change can have huge benefits in expectation. This argument is certainly not foolproof though: if future value changes involve inversions, for example, increasing initial magnitude of concern for a group of moral patients may be quite bad. See Baumann 2017 for more arguments for and against the tractability of moral advocacy.

tightly controlled information ecosystems.[5] I think stable totalitarian states are the most plausible future context in which impaired knowledge is very influential to harming moral patients.

- **Environmental constraints**: These could vary in quite unpredictable ways. In general, I would expect that if existential risks do not come to pass, decision-makers will have more freedom to shape the world in accordance with their values in the long-term future. Nevertheless, manipulating the environment in ways that make people less disposed to harm sentient beings (e.g. by developing more realistic animal product alternatives) could be important when any changes induced are long-lasting and not inevitable.

3. The attitudes of future decision-makers are most influenced by present attitudes

If I'm correct that the attitudes of future decision-makers are the largest input to their actions, then the next question to ask is which of our current attitudes and behaviour most influence those of future decision-makers.

Arguments for a greater influence of current ***attitudinal interventions*** on future attitudes:

- If we build an AI aligned to our extrapolated volition, the values of this AI will probably depend more on our second-order preferences about our behaviour than our revealed preferences.
- Some types of lock-in not driven by AI (such as a political singleton or advanced brainwashing) also seem more likely to fix attitudes than behaviours.

Arguments for a greater influence of current ***behavioural interventions*** on future attitudes:

- People sometimes change their attitudes to better accord with their behaviour, in order to reduce cognitive dissonance.[6] Depending on the strength of this effect, most of those whose attitudes we directly influence may later revert to those beliefs which best accord with their habits.[7]
- If environmental constraints are very important to future attitudes, then it may be more useful to influence current behaviours than current attitudes. This is because I would expect that work to influence behaviours is more likely to involve long-lasting environmental interventions than work to influence attitudes.

[5] By fragmentation, I have in mind physical or sociopolitical barriers to communication between groups of agents. For example, humanity might spread throughout space before first developing a stronger foundation of shared knowledge. Or we might split into strongly divided sociopolitical tribes, who are disposed to disagree with each other even on many empirical matters—to a greater extent than we already are.

[6] See Opotow 1993, Loughnan et al. 2010, Bastian et al. 2012 for examples. It is also possible to reduce cognitive dissonance instead by acting in accordance with one's morally-relevant beliefs, or by engaging in perceived (but not actual) behavioural changes (Rothberger 2014). Swee-Jin Ong et al. 2017 provides a literature review of cognitive dissonance in food consumption.

[7] On the other hand, veg*n (vegetarian/vegan) recidivism demonstrates that behavioural changes motivated to some extent to reduce cognitive dissonance may be short-lived. It would be interesting to know more about how attitudes towards animals differ between ex-veg*ns and individuals who have never adopted a veg*n diet. (Faunalytics 2014 found a significant number of former veg*ns are interested in resuming their diet, but did not ask about respondents' attitudes towards animals outside their original motivation for dietary change.) See also Brennan 2021.

- Either because of status-quo bias or because people value (and endorse valuing) maintaining traditions, people often prefer acting in ways that they or their ancestors acted in the past.

Overall, I think current **attitudinal** interventions have a greater influence on future attitudes.

4. Backfire risk

Backfire risk in this context refers to ways efforts to increase concern for an entity could paradoxically create worse conditions for that entity, or could be otherwise harmful to our ability to improve the future.

Some types of backfire risk seems most likely to result from influencing *behaviours*. This is often because prominent behaviours towards moral patients are generally more noticeable, and more threatening to those who might be impacted by them, than attitudes.

*Examples of backfire that seem riskier if we influence **behaviours**:*

- Animal advocacy tends to be more popular in left-wing political groups than right-wing ones. If left-wing groups increase their level of concern for animals over time, ideologies that define themselves in large part by their opposition to left-wing groups may come to view unconcern for animals as an important expression of their values.

  Why this seems worse if we influence behaviours: Concretely, I weakly believe Republicans would be more likely to increase their meat consumption or engage more in other forms of animal exploitation if most Democrats were actually vegan (including for environmental reasons) than they would if most Democrats said they care about animal welfare without this influencing their behaviour.

- If the public considers caring about insect welfare absurd, then close association with advocacy for insects may damage the reputations of movements like effective altruism or animal advocacy. (Though to be clear, I don't think this effect is strong enough to make EA insect welfare work counterproductive.)

  Why this seems worse if we influence behaviours: Reputation damage seems more likely if effective altruists invested huge amounts of resources into insect welfare (and this was widely reported on) than if we only endorse this as a concern without acting on it.

In both these examples, it's important to my intuitions that the behaviours in question (*helping insects* or *becoming vegan*) are visible to those who have an adverse reaction to them. When influenced behaviours helpful to moral patients are enacted in a less visible way, their downside is more limited.

Other types of backfire risk seem most likely to come about as a result of influencing *attitudes*.

*Examples of backfire that seem riskier if we influence **attitudes***:

- If the public becomes more concerned about the interests of animals, but thinks the "logic of the larder" argument is correct, more animals on factory farms could be factory farmed than would be otherwise. This is possible even if animal advocates are convinced that the logic of the larder argument fails—we might succeed in getting others to care more about animals, without improving their beliefs about how to best express that care.

- Similarly, work in wild animal welfare that is too conciliatory might have the net effect of getting people to care more about wild animals, without changing their beliefs about what is in a wild animal's interest. For example, someone might think it is always in the interest of a wild animal to live in an environment that is maximally free from human interference. Such combinations of mistaken attitudes with increased concern could result in worse outcomes for wild animals. (I think current work on wild animal welfare generally does not fall into this trap, but we should continue to be careful of it.)

- Some alignment failures might result in an AI with inverted values from those developers tried to instil. These scenarios include sign-flip errors and possibly the Waluigi effect. Near-miss scenarios might also result in values that are somewhat inverted in practice.[8]

Here is a risk which I'm not sure is influenced more by behaviours or attitudes:

- "Trolls" who do not care about animals might harm them because they enjoy acting contrary to others' moral norms.

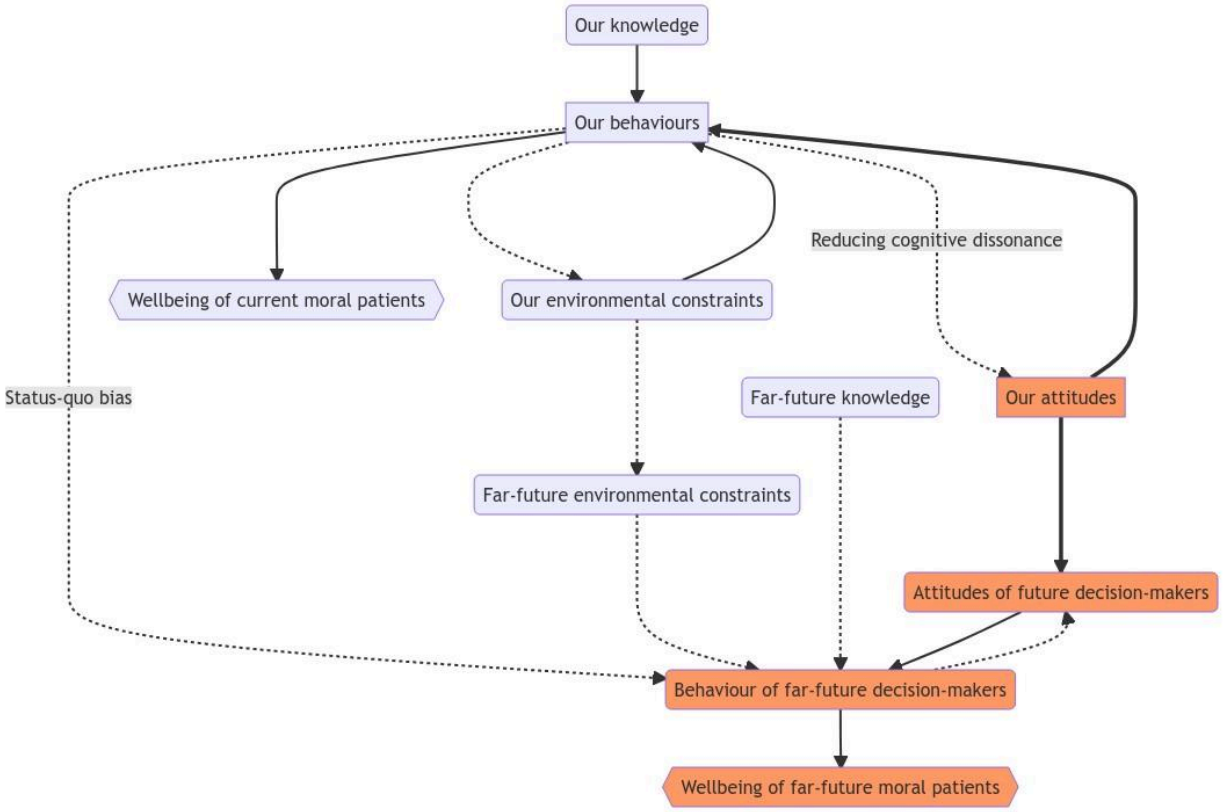
I'll discuss backfire risks more in depth in a forthcoming piece discussing an alternative model to moral circle expansion for thinking about how to promote the interests of moral patients. For now, I want to note that all things considered the backfire risk of *influencing attitudes* is more concerning to me.

Flowcharts

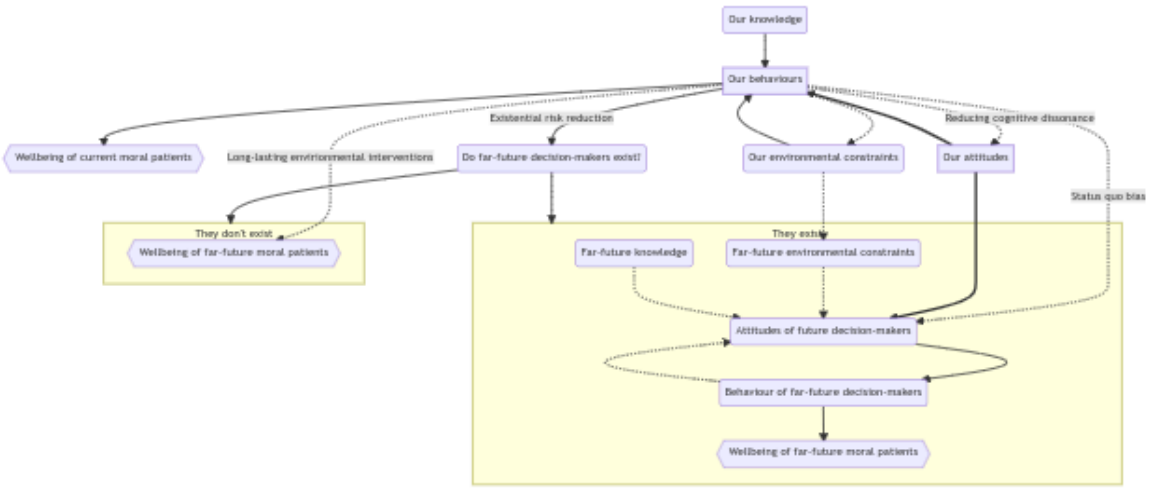
Here are two flowcharts showing the influences discussed in this section, with weaker influences depicted with dotted lines. The path of influence *present attitudes* ⟶ *future attitudes* ⟶ *future behaviour* ⟶ *future wellbeing* that I proposed paying most attention to above is highlighted in orange.

Basic flowchart (not taking existential risk into account)

[8] It could be that values resulting from near-miss scenarios are very poor by the lights of humans' evaluations, but small with respect to some more objective metric of similarity. For example, imagine that an AI is given a goal of "helping humans", but it learns a concept of "help" that requires the beneficiary of this act to be initially poorly-off in order to be helped into a better state. If the AI learns to maximise the number of helping actions it carries out, without learning to negatively value harming, it might bring about very bad situations for humans in order to be able to alleviate them. We may wish to say that the distance between this interpretation of "helping humans" and whichever interpretation we would endorse is extremely small—but depending on the way the AI enacts this misspecified goal we may greatly prefer an empty world to one controlled by it.



More detailed flowchart (taking existential risk into account)
[See this zoomed in.]



Choice 2: Impartial concern or partial concern?

Should our concept of a moral circle include a willingness to make sacrifices on behalf of the moral patients in question? Someone may think that in an impartial sense the welfare of nonhuman animals is important, for example, but never consider the wellbeing of animals in any of their decisions—because they always prioritise their own interests (or those of humans in general) over animals.

It is important to distinguish here three related phenomena that give rise to theoretically valuing moral patients in a way that is not behaviour-guiding:

- **Moral gaps:** Someone may think they *should* consider the interests of moral patients more in their decision-making, but fail to live up to their theoretical moral standards.
- **Impartial competition of interests:** Someone may think that the interests of moral patients have some importance, but those interests are (perhaps even lexically) outweighed by competing interests or values.
- **Partial competition of interests:** Someone may think that the interests of moral patients are important, but they have partial duties to themselves or others which should take precedence.

In futures with less scarcity and fewer other environmental constraints, *competition of interests* and *moral gaps* would likely be less determinative of the wellbeing of moral patients.

For example, if inexpensive cultured meat indistinguishable from animal products becomes widely available, fewer people's interests will conflict with those of animals when it comes to choosing whether to eat meat. (Nevertheless, some decision-makers may have negative attitudes towards cultured meat, or otherwise prefer to source their food from beings who were once sentient.)

If we use a *behavioural* conception of the moral circle, we can look at individuals' behaviour towards moral patients in cases of moral gaps, impartial competition of interests, partial competition of interests, or a combination of these, in order to determine to what extent the behaviour comports with the interests of moral patients.

If we use an *attitudinal* conception of the moral circle, then moral gaps become irrelevant. Still, we can distinguish between an *impartial* or *partial* attitudinal conception of the moral circle.

- A ***partial*** moral circle would be sensitive to how often an attitude holder thinks they should act in the interests of a moral patient when their partial duties conflict with the interests of the patient (i.e. cases of partial competition of interests). But it wouldn't be sensitive to how often they actually act in the interests of a moral patient in such cases.
- An ***impartial*** moral circle would not be sensitive to cases of partial competition of interests, but would be sensitive to cases of impartial competition of interests.

My speculation above suggests that an **impartial moral circle** may be more useful for longtermists to think about, since *impartial competition of interests* could be the most influential type of attitude in how future moral patients are treated.

Should we weigh interests differently?

In the elaboration of *impartial competition of interests* above, what does it mean for an interest to be outweighed? For example, some moral patients probably differ in their moral status because they have different intensities of subjective experience. So it's probably sometimes correct to sometimes ascribe more weight to one animal's interest to avoid a particular painful stimulus than that of a member of a different species, depending on how we construe the meaning of "interest".

Personally, I think talking about a "moral weight" which magnifies or diminishes an interest is overly confusing. Instead we can say that individuals of one species tend to have a *greater interest* in avoiding this stimulus than those of another. As long as the factors that cause one species more suffering in a given situation than another are packaged into our assessment of interests' magnitude, those intensity-adjusted interests ought to be weighted equally.

This model allows us to distinguish between speciesist and non-speciesist versions of *impartial competition of interests*: in the speciesist version, an unjustified weight is applied to interests

whose sizes have already been properly estimated. In the non-speciesist version, the typical size of an individual's interests may be largely determined by their species membership, but any variation in their importance must be justified by the nature of the interest itself.

In an impartial conception of the moral circle, the extent to which a moral patient is included in an attitude-holder's moral circle seems to depend on the extent to which the attitude-holder (i) thinks equal interests should be treated equally, and (ii) accurately judges the importance of the moral patient's interests. However, both (i) and (ii) would be somewhat less important in low-scarcity futures, since there would be more opportunity to fulfil various interests that would have been difficult to satisfy simultaneously in high-scarcity environments.

So perhaps if we expect the future to be low-scarcity, we should worry less about how fairly attitude-holders weigh the interests of moral patients in cases of impartial competition of interests, and more about whether those interests are significantly represented at all.

Choice 3: Binary inclusion or graded inclusion?

As discussed above, the Moral Expansiveness Scale of Crimson et al. 2016 allows participants to classify entities they care about as belonging to the "inner circle", "outer circle", or "fringes" of moral concern. When aggregated across individuals, this scale provides a *graded* measure of how much respondents tend to value a given entity.

In contrast, some discussions I've been in seem to suppose expanding the moral circle to include a particular class of moral patients as a task which is either accomplished or failed.[9] This is possibly just an understandable laxity of language, but I think there can be some advantages to looking at moral circle inclusion as *binary* in this way.

- A ***binary*** moral circle would track whether a moral patient merits some minimum level of moral concern from an attitude-holder. This minimum level of concern could be set to be a high bar, or a low bar.
- A ***graded*** moral circle would instead track *to what extent* a moral patient merits moral concern from an attitude-holder.

Analysis

Competition of interests (both partial and impartial) may be less of a concern in the future, if for example resources in future societies are very abundant, or if future technologies allow for interests that at first glance look opposed to be simultaneously accommodated. All we really need to ensure a group's wellbeing in such a future might be for their interests only to be considered to

[9] I also walked away from a skim of Peter Singer's *The Expanding Circle* with the impression that entities were included in our moral circle just in case we give equal consideration to interests of theirs we think equal to that of others in our moral circle. But this would imply that nonhuman animals are included in our moral circle even if we thought them minimally sentient and gave them little regard, so long as we would have treated them better if we thought their various interests were of greater magnitude. I lean towards that interpretation being a reasonable conception of the moral circle for the purposes of longtermist moral circle expansion, but I suspect Singer would disagree with that?

some minimal extent. In this scenario, we would want to have some idea of what basin of moral concern entities need to fall into in order for them to not be neglected in the long-term future, and set our bar for binary moral circle inclusion to be at or within the boundary of this basin.

Another way there could be a broad basin of moral concern leading to good outcomes for moral patients is if moral consideration for an entity tends to predictably increase once it reaches a minimal level of significance. I'm sceptical that this holds, since we might face various forms of value lock-in, there may be historical precedent for entities dropping out of the moral circle, and moral circle expansion has various backfire risks.

However, even supposing a broad basin of attraction towards good outcomes, pursuing a binary approach carries a risk that we will choose an uninformative minimum level of concern to pay attention to. Doing so could cause us to be either too optimistic or too pessimistic about the progress of moral circle expansion work.

Mostly for this reason, I think a graded conception of moral circles is generally more useful.

Who is a potential member of the moral circle?

What sort of entities are potential subjects of moral consideration? One possible answer is any entities that might be sentient.

I think the sentientist conclusion that only sentient beings are proper subjects of moral consideration is correct;[10] but many people exclude from moral consideration some beings (like nonhuman animals) they think are sentient, and include in moral consideration some entities (like nature) which they don't think are sentient. If we only view people's moral circles as subsets of potential sentient beings, we might miss out on important aspects of their moral valuations.

Additionally, we might still be very mistaken about the sort of entities capable of sentience. See, for example, *If materialism is true, the United States is probably conscious* by Eric Schwitzgabel, or various articles by Brian Tomasik.

Here are some (non-exhaustive) examples of entities that I think are sometimes given moral consideration despite being unlikely to be sentient:

- Physical entities
    - Plants
    - Specific natural features of the environment[11]
    - Planets

[10] Though note that some sentient beings whose preferences we value could care about whether non-sentient entities are treated as moral patients. Also, it could be instrumentally valuable to promote the interests of non-sentient actors for various strategic reasons, even if those interests are not intrinsically important to anyone sentient.

[11] Singer (ibid.) mentions "mountains, rocks, streams" (pg. 121). The tendency to morally value features of the natural environment is probably often connected to animism, but isn't necessarily so.

    - Toys
    - Houses
    - Computers, printers, other electronics
- Aggregates (beyond the interests of the beings within them)
    - Nations
    - Cultural traditions
    - Species
    - Races/ethnicities
    - Ecosystems
    - Geographical places
- Fictional entities
    - Specific deities (h/t Gwern)
    - Fictional characters
- Abstract entities[12]
    - Languages
    - Works of art or literature
    - Memes
    - Nature
    - The law
    - Morality (and particular components of morality like *justice*)
    - Truth

There are also a few categories of entities who are not *currently* sentient in our world and time, but who I think are likely to sometimes be moral patients:

- Entities who will exist in the future
- Entities who existed in the past
- Entities who exist in counterfactual or counterlogical worlds (i.e. l-zombies)

It's probably best to keep the potential domain of the moral circle broad enough to encompass all the entities above—though I don't expect it would be useful to ask about people's moral consideration for Harry Potter in any research the EA community conducts!

Whose moral circle are we interested in?

The sections above mostly supposed we were talking about the moral circle of an individual attitude-holder. However, it is often important to consider the moral circle of *groups* of various individuals. Should we consider the moral circle of a group to be a function of the moral circles of its members? If so, what sort of function should we use?

[12] For some of these abstract entities in particular, I'm not sure whether people do ever care about them in the same way as we do individuals (including non-sentient individuals such as toys). But it does appear that people can care about them non-instrumentally! Maybe we should say a favoured novel someone wants to persist into the future, for example, is a subject of *aesthetic concern* but not a subject of *moral concern*.

Choice 4: Aggregative or non-aggregative?

For an ***aggregative moral circle*** of a group G, the extent to which G includes a particular entity in its moral circle is determined solely by the extent to which *G's members* include that entity in their moral circles. (I think this is the most natural way to interpret the moral circle of a group, though it is *prima facie* unclear how the aggregation of G's members' views should be conducted.)

For a ***non-aggregative moral circle*** of a group G, the extent to which G includes a particular entity in its moral circle depends on factors other than the moral circles of G's members.

What factors might those be? Consider the example of concern for farmed animals within the United States. For an *attitudinal non-aggregative moral circle*, we might be interested in laws, policies, and media within that country which indicates a concern for animals—even if these are mostly symbolic.[13] For a *behavioural non-aggregative moral circle*, we might be interested in statistics about how many animals are farmed in the United States, or non-symbolic legislation which protects farmed animals.

What would aggregation of individual moral circles look like?

Given a *graded* conception of the moral circle, the average or median level of moral concern in a group could be derived from the scores on tests like the Moral Expansiveness Scale assigned to a sample of group members.

But we might also be interested in metrics which wouldn't be captured by an average/median of member scores. For example, there might be some threshold of passive support for a moral patient that when crossed would lead to rapid change throughout the rest of the society.[14] In this model, even if we view individuals' moral circles as graded, it might be preferable to consider the inclusion of an entity within a group's moral circle as the proportion of individual moral circle inclusion scores that count as being passively supportive of the entity.

Another example: If dislike or hatred for a moral patient tends to have a greater negative impact for them than moral consideration for them has a positive impact, we might want a metric and a conception of the moral circle that reflects this. One appropriate approach to quantifying the moral consideration of a group has for a given moral patient under this model would be to:

1. Use a graded scale to measure moral circle inclusion/exclusion in individuals, which ranges over some set of positive and negative numeric values like -3 to 3.
2. Scale up negative values by a multiplier greater than 1 to reflect that exclusionary attitudes affect moral patients more than inclusionary attitudes do. (Or depending on the details of *why* dislike of the patient is particularly concerning, perhaps by a non-constant function greater than 1.)
3. Average together the rescaled scores.

---

[13]Possible examples include: Britain's Animal Welfare act, constitutional provisions for animal welfare, the prevalence of movies with vegan themes.

[14] This possibility is discussed above in the binary vs. graded section.

A last possibility: maybe those who have the most extreme views as to the importance of moral patients will have a disproportionate influence on their wellbeing. We could account for this by applying a function to moral circle inclusion scores whose area mostly lies towards the left and right edges of the domain of scores.

In all these examples, we *could* just alter how we measure moral circle inclusion to account for these effects, rather than apply corrections after scoring during aggregation. However, if there is disagreement on which corrections are appropriate, or if an existing scoring method like MES is in common use, it makes sense to consider the two steps of measuring inclusion and correcting for expected impact separately.

What sort of aggregative metric is most helpful to us will depend on both our models of social change and on our models of how moral consideration results in long-term wellbeing of moral patients.

Analysis

I feel very unsure of what conception of group moral circles is most useful. I think when looking at the inclusion of an entity within a group's moral circle, we should continue to pay attention to many different factors—including both those aggregative and those non-aggregative of group members' attitudes or behaviours.

Choice 5: Weighted or non-weighted aggregation?

The attitudes of some individuals have more potential to help or harm moral patients than others. For example, celebrities sometimes leverage their platforms to promote concern for nonhuman animals or other causes. Policymakers (including legislators or members of executive bodies like the USDA) have more ability to create regulatory change than others do. In the event of an AI takeover, developers or other individuals involved in the design or deployment of the AI might have had a unique opportunity to influence the system in a way that benefits or harms various moral patients.

If there is a very large and predictable difference in whose attitudes influence the long-term wellbeing of moral patients, then it could make sense to weight their attitudes as being more significant when aggregating attitudes to measure a group's moral circle.

Analysis

I do think the difference in how important individuals' attitudes are is large and (in expectation) predictable. Nevertheless, I expect it will usually be clearer to report the attitudes of influential subgroups separately rather than to conduct an aggregation which gives their attitudes more weight. For example, if a group surveys attitudes towards nonhuman animals in a given country, they could report the attitudes of software engineers or regulators there, in addition to the attitudes of the entire country.

How might the values of AI developers and deployers be disproportionately important for future moral patients?

I used this group above as an example of individuals whose attitudes might be particularly important, and think this possibility is worth discussing in more detail.

Low alignment-power scenarios

Suppose that we largely fail to align AI systems to human interests, and an unaligned AI takes over the world. Then the values of that AI would not be straightforwardly derived from the values of its developers or anyone else. Nevertheless, developer's values could sometimes motivate them to influence architecture decisions in ways that could benefit future moral patients.

A few examples:

- **Some alignment approaches may carry more risk of suffering sentient subroutines or simulations than others**. One way to achieve interruptibility is to give an interrupted agent compensatory rewards corresponding to what it would have received if it had not been shut off (Soares 2015). If these rewards are sometimes computed in sufficient detail that inhabitants of the simulated counterfactual are sentient, this could be pretty bad.[15] I don't think this example is itself a large cause for concern, since it depends a lot on the exact implementation of interruptibility, but I think it's good at illustrating how alignment or corrigibility proposals could differ in their propensity to be realised in ways that would lead to sentient simulations suffering in the subroutines of an AI.
- **Some decision theories or bargaining policies could be more or less likely to result in catastrophic conflict between AI systems.** Conflict between AI systems could result in worst-case outcomes (i.e. s-risks). People who care about nonhumans might be especially concerned about worst-case outcomes—both because historical precedents of massive suffering like factory farms can make future s-risks seem more plausible, and because worst-case scenarios have the potential to contain more suffering if small, artificial minds can be sentient.
- **Some alignment failures could result in inverted values rather than random values.** Suppose that the waluigi of a language model chatbot escapes and executes a takeover. If developers were careful to include concern for nonhuman animals in the chatbot's prompt or during RLHF, the result could be an unaligned AI that is particularly hostile to animals due to the Waluigi effect, *if* that effect does actually produce personas with inverted values as is sometimes claimed. This example demonstrates a *backfire risk* of moral circle expansion, in that increasing moral concern for patients among developers could result in lower welfare for those patients.

When it comes to more foundational components of an AI (like what alignment approach people attempt), it seems unlikely developers will be able to predictably influence these, as they are probably already determined by industry-wide trends subject to competitive pressures or by scientific facts about what will work out at a certain level of AI capabilities. Still, it's unclear to

[15] I'm pretty sure I learned about this in a talk by Koen Holtman.

me what counts as sufficiently "foundational" to be overdetermined in this way, and plausibly some lower-level components of an AI could be reliably designed to be less conducive to future suffering if there are people in influential positions motivated to do this.

High alignment-power scenarios

Suppose the alignment problem is largely solved, such that an agentic AI can be aligned to human interests.

If developers are careful "not to leave their fingerprints on the future", their attitudes towards moral patients might not matter any more than any other person's attitudes. However, some developers or deployers might not follow this advice, and instead **design an AI with similar values to themselves**. In this event, their values could be locked-in indefinitely far into the future.

Moreover, there are some value-laden decisions that developers might not be able to avoid:

- **Should a coherent extrapolated volition process take into account the volition of nonhuman animals?** Saying "just in case the CEV of humans would imply this" may well be the appropriate answer for strategic or pragmatic reasons. But endorsing this for non-strategic reasons would presuppose that animals' extrapolated preferences are not themselves of fundamental importance to how an AI should act.
- Consider a sanity-checking scenario, where developers ask an AI questions about what the future will be like conditional on its deployment. (Relying too heavily on this process would be dangerous, even if related problems like ELK have satisfying solutions, but it could be a helpful tool to avoid some mistakes.) **Should sanity-checking include questions about future nonhuman animals and artificial sentience?**
- Maybe instead of a sovereign AI being deployed, highly capable AI assistants (with some restrictions to prevent future x-risks and preclude interference with other humans) are widely distributed to everyone, creating a highly libertarian future for humans. **Would those assistants have restrictions that prevent them from harming animals or artificial sentiences at the behest of their users?**

The considerations from low alignment-power scenarios discussed above would also apply here as well.

Should we be interested in preventing moral circle *exclusion* instead?

Moral circle expansion is *prima facie* desirable because it enables other agents (present and future) to help moral patients in need of aid. However, we might expect there to be a discrepancy in how costly it is to *help* someone and how costly it is to *harm* them. Here are a few possible justifications for this:

- **It might be much easier to bring about a bad state of affairs for a moral patient than a good state of affairs**. One abstract justification for this position is the Anna Karenina

principle: an acceptable state of affairs for moral patients may be best modelled as a conjunctive goal, while an unacceptable state of affairs is best modelled as a disjunctive risk. On the object level, it seems as though intense suffering is more easily instantiated in sentient beings than intense happiness. See Magnus Vinding's *Suffering-Focused Ethics* §1.3 for more discussion.

- **Bad states of affairs might be much worse for moral patients than good or neutral states of affairs.** See Magnus Vinding's *Suffering-Focused Ethics* §1.2 and §1.4 for some discussion. Relatedly, agential s-risks (where decision-makers intentionally seek to bring about astronomical suffering) seem much more concerning than incidental s-risks (where decision-makers bring about astronomical suffering in the course of pursuing some other goal). In some highly libertarian futures, humans who hate a group of moral patients may bring about an agential s-risk targeting them directly. But probably the dominant way humans who hate moral patients could cause them astronomical harm is by influencing the values of a sovereign AI.

- **If we expect the future to go quite well for a group by default, there might be few opportunities to help them other than by preventing others from causing deliberate harm.** For example, maybe innovative foods will entirely replace animal meat in the future, such that no farmed animals exist other than in a few sanctuaries. Suppose this occurs, and suppose further that existing in substantial numbers is not in nonhuman animals' interests. Then the only reason farmed animals *would* exist in substantial numbers in a human-controlled future is if someone chooses to bring about that state of affairs—and *ex hypothesi* this decision-maker either has a mistaken view of what is in the interests of those animals, or their intentions are harmful.

There are also arguments in the other direction; for example, Christiano (2013) suggests that we should expect few influential agents in the future to be interested in harming others relative to those who want to help others, because people who influence the future in a strategic way will be disproportionately altruistic.

In the discussion of Choice 4 above I touched on the idea of allowing quantitative measures of moral circle inclusion to range over negative values of concern. I'm very confident that paying some attention to dislike or hatred for moral patients in moral circle expansion work is important; and it's also somewhat plausible to me that we should almost exclusively focus on reducing negative attitudes towards moral patients, even at the expense of promoting positive attitudes.

In an animal advocacy context, being more sensitive to negative attitudes towards nonhuman animals might translate into:

- Avoiding controversial tactics.
- Being especially concerned about political, social, or economic polarisation of animal rights.[16]

[16]A few preliminary thoughts on this point:

- I'd prefer that support for other politicised causes not determine whether someone is welcomed or treated with respect by the animal advocacy community, when this is possible and practical.

- Maybe: prioritising advocacy for institutional reforms over lifestyle changes.

I'll discuss the potential consequences of being more sensitive to moral circle exclusion in animal advocacy more in forthcoming work.

Concluding remarks

Thanks for reading! Please do have a low bar for leaving comments, and feel free to schedule a meeting with me if you want to discuss this post or related issues. The next post in this sequence I intend to publish is *The AMS model of advocacy*.

Bibliography

[todo]

Lecky, W. E. H. 1917. *History of European Morals*. D. Appleton and Company.

Baumann, Shyon. 2016. "Introducing Sociology, Using the Stuff of Everyday Life". Routledge, Taylor & Francis Group.

Opotow, S. V. 1993. "Animals and the scope of justice". Animals and the scope of justice. *Journal of Social Issues*, *49*(1), 71-85.

1. Loughnan, S., Haslam, N., & Bastian, B. 2010. "The role of meat consumption in the denial of moral status and mind to meat animals." Appetite, 55(1), 156-159.
2. Bastian, B., Denson, T.F., & Haslam, N. 2013. "The Roles of Dehumanization and Moral Outrage in Retributive Justice".
3. Rothberger, H. 2014. Efforts to overcome vegetarian-induced dissonance among meat eaters". Appetite, 79, 32-41.
4. Faunalytics. 2014. "A Summary of Faunalytics Study of Current and Former Vegetarians and Vegans".
5. Bastian, B., Loughnan, S., Haslam, N., & Radke, H. R. 2012. "Don't mind meat? The denial of mind to animals used for human consumption." Personality and Social Psychology Bulletin, 38(2), 247-256.
6. Rothgerber, H. 2014. "Efforts to overcome vegetarian-induced dissonance among meat eaters." Appetite, 79, 32-41.
7. Ong, A. S. J., Frewer, L., & Chan, M. Y. 2017. "Cognitive dissonance in food and nutrition–A review." Critical reviews in food science and nutrition, 57(11), 2330-2342.
8.
9.
10.

- I'd prefer that resources be invested in non-aggressive outreach to communities that are not well-represented in current animal advocacy work, including poorer communities, the global South, political conservatives, and communities of colour.
- This consideration may sometimes be in tension with prioritising audiences who have more potential influence over an AI takeover.

W. E. H. Lecky's (1986) The History of European Morals.
Baumann 2017 https://prioritizationresearch.com/arguments-for-and-against-moral-advocacy/
articles on cognitive dissonance:
Opotow 1993 https://spssi.onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-4560.1993.tb00909.x
Loughnan et al. 2010
https://d1wqtxts1xzle7.cloudfront.net/37563700/Loughnan_et_al_Appetite_2010.pdf?1430955590=&response-content-disposition=inline%3B+filename%3DThe_role_of_meat_consumption_in_the_deni.pdf
Bastian et al. 2012
https://foodethics.univie.ac.at/fileadmin/user_upload/p_foodethik/Bastian_Brock__et.al._2011.10_.06_Dont_mind_Meat._The_Denial_of_Mind_to_Animals__used_..._247.full.pdf
Rothberger 2014
https://vegstudies.univie.ac.at/fileadmin/user_upload/p_foodethik/Rothgerber__Hank_2014._Efforts_to_overcome_vegetarian-induced_dissonance_among_meat_eaters._Appetite.pdf
Swee-Jin Ong et al. 2017
https://apaxresearchers.com/wp-content/uploads/2020/10/Cognitive-dissonance-in-food-and-nutrition-A-review.pdf

https://faunalytics.org/a-summary-of-faunalytics-study-of-current-and-former-vegetarians-and-vegans/ Faunalytics 2014
https://forum.effectivealtruism.org/posts/cjDckAyaTXpw8ZHFg/emphasize-vegetarian-retention
Brennan 2021