This is something like a draft research proposal, created by Michael Aird. I may make this more public in some form someday. I also may (or may not) pursue research on this topic in future. Feel free to make comments or suggestions!

Background

I've seen some indications (e.g., in Why I prioritize moral circle expansion over artificial intelligence alignment) that the key argument underpinning the Sentience Institute's (SI's) strategy is something like the following:

Premise 1: It's plausible that the vast majority of all the suffering and wellbeing that ever occurs will occur more than a hundred years into the future. It's also plausible that the vast majority of that suffering and wellbeing would be experienced by beings towards which humans might, "by default", exhibit little to no moral concern (e.g., artificial sentiences; AS).

Premise 2: If Premise 1 is true, it could be extremely morally important to, either now or in the future, expand moral circles such that they're more likely to include those types of beings.

Premise 3: Such moral circle expansion (MCE) may be urgent, as there could be value lock-in relatively soon, for example due to the development of an artificial general intelligence.

Premise 4: If more people's moral circles expand to include farm animals and/or factory farming is ended, this increases the chances that moral circles will include all sentient beings in future (or at least all the very numerous beings).

Conclusion: It could be extremely morally important, and urgent, to do work that supports the expansion of people's moral circles to include farm animals and/or supports the ending of factory farming.

I think some other people in the effective animal advocacy community are also influenced by this sort of argument.

Personally, I find all of these premises plausible, along with the conclusion. However, each of those premises also seems quite speculative. This is often hard to avoid, especially in longtermism. But it seems to me that there are tractable ways to improve our knowledge regarding Premise 4 (and related matters), and that the value of information from doing so would be very high. I therefore think it'd be valuable for someone to do research related to that topic.

What follows is a sketch of my current, quite preliminary thinking on why this matters and what types of research could be useful here. I expect I'd update and flesh out this thinking substantially given another 10 hours of thought, and I expect there's a lot of relevant existing

work by SI, other EAs, and academics. (I'm also pretty sure SI has already thought about similar topics, based in part on this talk: https://youtu.be/NTV81NZSuKw?t=544)

Roughly what I propose, and why it might be valuable

Essentially, I'd be very confident of Premise 4 if moral circles were effectively unidimensional. However, it seems to make more sense to think of moral circles as multidimensional, such that a person's moral circle can expand along one dimension without expanding along others, and that two people could have differently *shaped* "moral circles", without it being clear whose is "larger".

Thus, it seems plausible that expanding a person's moral circle to include farm animals doesn't bring the "edge" of that person's moral circles any "closer" to including AS (or whatever other beings we're ultimately concerned about). It also seems plausible that expanding a person's moral circle to include farm animals does achieve that outcome, but that the outcome would be better achieved by expanding moral circles along other dimensions (e.g., through wild animal welfare work, advocating for caring about all sentient beings, or advocating for caring about future AS).

This could have major implications for which research directions should be prioritised by SI and by other actors interested in MCE. Additionally, more solid evidence on these matters might increase the number of longtermists prioritising MCE, as it could suggest more impactful interventions for MCE, and/or increase the robustness of the arguments for prioritising a given MCE intervention.

I thus propose research aimed at answering the question of how interventions that expand moral circles along certain dimensions (or to certain types of beings) spill over into expanding moral circles along other dimensions (or to other types of beings). This question could be tackled relatively directly, though that might require expensive experiments. The question could also be tackled somewhat indirectly, by investigating how expansions of moral circles (whether or not they're caused by "interventions") along certain dimensions spill over into expanding moral circles along other dimensions. Or the question could be tackled even more indirectly, by addressing the purely correlational question of how well the size of a person or group's moral circles along one dimension predicts the size of their moral circles along another dimension.

Ideally, this research would focus on:

- the types of interventions EAs (or related groups) are most likely to actually consider supporting
- the types of people these interventions are most likely to target (e.g., the general public, AI researchers)

- the types of beings those interventions are most likely to directly focus on (e.g., farm animals)
- the types of beings it's ultimately most important to expand moral circles to include (e.g., AS)

One could also investigate how differences in intervention types, types of people, and types of beings affects results. This could inform decisions about things like:

- whether to prioritise antispeciesist messaging or clean meat research
- whether to target thought leaders, tech researchers, or the general public
- whether to focus on expanding moral circles to include insects, to include farm animals, or to be more explicitly open to including AS in future

Sketches of some specific types of possible research projects

1. Literature reviews focused on the above questions.

For example, I know of at least one paper relevant to the extent to which inclusion of some entities in one's moral circles predicts inclusion of other entities. I suspect there are also others. And the research on secondary transfer effects seems relevant too (my thanks to Jamie Harris for drawing my attention to that).

For another example, I suspect some writings on the history of MCE would contain clues as to whether:

- expansion along one dimension seemed to lead to expansion along another, vs
- expansion along many dimensions seeming to happen near-simultaneously for other reasons (e.g., economic growth), vs
- expansion seeming to occur along one or more dimensions without occurring along (important) other dimensions

2. Expert interviews focused on the above questions. Perhaps especially with psychologists, sociologists, and historians who have published relevant research. Other types of people who might be relevant include EAs, animal advocates, futurists, and philosophers who've done relevant work.

3. Surveys focused on the above questions. These surveys would likely consist mostly of things like rating scale questions, though with at least some boxes for open-ended responses.

3a. Surveys simply focused on what types of entities people currently include in their moral circles (or related matters, like what entities they empathise with or eat). If a person's inclusion

of one type of entity predicts their inclusion of others, this would push in favour of the hypothesis that moral circles are “effectively unidimensional”.

That said, that wouldn’t strongly indicate that *expansion* along one dimension will spill over into expansion along other dimensions. This is because the correlations could reflect how people’s moral circles started out (e.g. due to a genetic predisposition towards generalised empathy), rather than how they expanded.

3b. Surveys asking people to recall what entities they included in their moral circles (or related matters) at various times. This could provide slightly better evidence about how expansion along one dimension may or may not lead to expansion along other dimensions. But I don’t think I’d want to put much weight on self-reported distant memories.

3c. Longitudinal surveys on what entities people include in their moral circles (or related matters) at different times.

4. Experiments focused on the above questions.

4a. Between-subjects experiments in which some participants are shown arguments, videos, or information which is intended to expand their moral circles to include a particular type of entity, and all participants are asked about which entities they include in their moral circles. The entities participants would be asked about would include ones not focused on by the arguments, videos, or information.

4b. Within-subjects experiments similar to the above, but with the intervention delivered to all participants, and participants being asked about their moral circles both beforehand and afterwards.

Unfortunately, I’d guess that SI thinks that it's something like ending factory farming, not just something like showing a relevant video, which has a decent shot at expanding moral circles in ways that might benefit beings like AS. This makes it very hard to run an experiment to relatively directly test this hypothesis.

5. Historical research focused on the above questions.

5a. Case studies. For example, one could investigate the factors that seem to have led to a particular instance of MCE, and how many dimensions moral circles appear to expand along during that instance.

5b. Historical research using a more quantitative, macro approach, examining broader trends.

Final thoughts

As noted, the above is just my quite preliminary thinking on this topic, and I'm sure that there's a lot of relevant existing work that could advance my thinking. Relatedly, each of the above sketches of research projects would need to be elaborated further, some are likely higher priority than others, and some may be worth dropping entirely.

If someone was to pursue this sort of research, I'd suggest they contact the Sentience Institute and/or me to discuss ideas. I also think I'd suggest they start by doing something like the first two projects listed above (partly to help orient themselves to the general area), and then move on to something like the third project. I'm unsure whether they should do the fourth or fifth projects. I'd also suggest they seek feedback as they go, and that they regularly update their assessments of the marginal value and ideal shape of further work in this area.

My thanks to Tobias Baumann for some helpful comments on this document.