REVIEW OF *WHAT WE OWE TO EACH OTHER*

I made a slideshow that walks you through this book: what-we-owe-to-each-other-notes.pdf. Making it helped me get a better grasp on a lot of stuff that I missed or misunderstood during my initial read-through of the book, so maybe it'll be useful to someone else too.

Ultimately, I don't buy Scanlon's theory, but I found a lot of what he had to say pretty thought-provoking. My confidence in my preferred theory (utilitarianism) has been *slightly* reduced.

Favorite parts:

- Morality as the quest for **justifiability**: Simply as an empirical/phenomenological description of what we're actually doing when we think about morality, the idea that we're trying to ensure our actions could be justified to others rings true to me.
- Clever approach to **aggregation**: To navigate between the Scylla of you-have-to-torture-one-person-if-it-would-save-a-zillion-people-from-a-mild-annoyance and the Charybdis of you-have-no-reason-to-save-a-million-lives-over-one-life, Scanlon tries to show that when each member of two groups is facing an equally serious threat, any *individual member* of the larger group can complain that you are undervaluing their life if you do not use group size as a tie-breaking consideration. I think this misidentifies the fundamental reason why saving more people is better, but it is an interesting argument.
- Compatibilist account of "the **Value of Choice**": Scanlon has some insightful thoughts on how and why our choices should have moral significance even if the universe is deterministic or probabilistic.

Notwithstanding Scanlon's arguments against treating well-being as a "master value", I still think our reasons to care about promoting happiness and preventing suffering are fundamental in a way that no other moral reasons seem to be. If I ask *why is pleasure good* and you answer *it just obviously is*, this answer may not be fully satisfying, but it seems less unsatisfying to me than it would as an answer to any other *why is X good?* question. So I'm still drawn to hedonistic utilitarianism.

Concurrently with (re)reading this book, I was listening to a book speculating about AGI. So I started thinking about whether something like Scanlon's theory could be helpful in designing safe/ethical AI. The classic AI doom scenarios tend to revolve around the AI *maximizing* something and going too far; naive attempts to program strict deontological rules could also have unintended consequences. I don't think the most responsible human thinkers are really maximizers *or* rigid rule-followers; rather, when confronted with a choice, they do something more like what Scanlon prescribes: cast a wide net for all the possible objections that could be raised against each option, and try to judge how reasonable and serious those objections are. If my fate were in the (metaphorical) hands of a machine, I'd probably feel safest if I knew the machine was following a similar process—and if, like a wise human, it became more cautious (and more inclined to seek advice from others) in proportion to how unclear the right way to adjudicate the competing objections is in a given case.