Thesis Proposal:
How much (dis)value could future civilisations have?

Abstract:

It has been argued that if humanity does not go extinct then it will create a lot of value[1]. The potential of the vast resources in the universe to be used up in ways that are value neutral[2][3] - has also been explored. But it is too often assumed that these are the only two likely options. The potential for future civilisations to contain a lot of disvalue has not been fully explored. This paper will briefly examine the various types of negative future. It then examines the trends today that could lead to such futures. It will look into the question of how policy-makers should be attempting to deal with these three broad categories of futures. I argue that the probability of neutral outcomes can largely be ignored in the consideration of the policy implications. The question becomes one of both judging the probability of the positive futures versus the negative ones and the relative utilities of these futures. The policy implications of the positive futures being more probable are largely known – I will therefore try to judge the implications of a negative future being more likely. These consist largely of either attempting to decrease the probability of a negative future relative to a positive one, or attempting to decrease the probability of a negative future relative to a neutral one. Given that the second option may be easier – I conclude that it would be wise for us to attempt more research on this question before we attempt to lower existential risks – presuming that life in the future will be good could turn out to be a horrible mistake.

Plan:

- **Ways we get a neutral universe (no utility):** Basically any one of the subset of existential risks that fully destroys human life and any plausible descendants thereof. Includes scenarios like paperclipping AIs, evolutionary pressures creating non-conscious (or otherwise valueless) beings and particle collider incidents.
- **Ways we get a negative universe:** Can be intentional or side-effects of other policies. Side-effects would include things like terraforming creating vast numbers of suffering beings , simulations containing vast numbers of suffering beings and the creation of “lab universes” also containing lots of suffering beings.

  Intentional would be things like sadists getting hold of vast amounts of computing power, “insane god” uploads, AIs that do things we hate rather than just destroy us and totalitarian states.

1 http://www.nickbostrom.com/astronomical/waste.html
2 hanson.gmu.edu/filluniv.pdf
3 http://wiki.lesswrong.com/wiki/Paperclip_maximizer

- **Trends:** Technology will make lots of bad stuff feasible in short order (of course that doesn't mean it will be used for bad things). People don't seem worried about simulated suffering at the moment, ditto wild-animal suffering (which could be affected by terraforming). Economic incentives in the future may encourage infliction of vast amounts of suffering (e.g. training lots of ems using negative feedback). Population pressures could cause large numbers of beings to exist at subsistence levels. AGI's attempting to predict the future could simulate lots of people, and not all will be happy. Trend toward tighter global cooperation could lead to totalitarian singletons.
- **Why we can ignore neutral outcomes in policy questions.** Several plausible positions for actions toward future: Reduce all x-risks. Increase all x-risks. Reduce probability of a negative versus a positive future.
  If a good outcome is more likely, then we should be trying to reduce all x-risks. If a negative outcome is more likely (such that the future has negative expected utility), then we should be trying to increase x-risks or reduce the probability of bad futures relative to good ones.
  So it's only the probability of a negative future relative to a positive one that matters. Neutral futures do not influence the future's expected value in the case that we survive. Naturally, this only carries up to a point. If the future is very likely to be good and not neutral or negative then there's little point in working on x-risks, but this isn't our current world.
- **Why it's easy to increase the likelihood of neutral futures and hard to increase the likelihood of positive futures.** Mostly obvious. It's easier to destroy the world than to save it. Unsafe AI for example is probably easier than Friendly AI, it's easier to create dangerous bioweapons than to cause world peace etc.
- **Policy recommendations:** Dependant on relative probabilities and utilities. However in general obvious – attempt to follow lines of research that increase the probability of good or neutral outcomes relative to bad outcomes.
- **What this paper won’t do:** Attempt to analyse the probability of the various futures. I need only be right in so far as these futures being plausible for my conclusion (more research needed – this is an important question) to carry.