- The post takes on two contested questions - independent research vs academia, and deconfusion vs ML - at the same time; I fail to settle either obvs.

- Alignment is just the decisive solution to the broader 'AGI control' problem - there's a large class of control work which doesn't centre on values: for instance impact measurement for catastrophe avoidance, external constraints like tripwires, 'motivational weaknesses', and arguably corrigibility. Work on non-goal-directed systems might also be somewhat complementary to alignment.

- I've bundled AI strategy and technical AI research together; the above doesn't include any social science academics. It's plausible that there should be a smaller academic discount for this work, since it is less tied to particular details of existing ML, but again this is fine in terms of making it harder for us (using conservative estimates).

- If someone had a lot of spare time, they could improve my Fermi estimate of the field size by crawling arXiv authorships. Handily, most AI papers go through arXiv under a small number of tags (cs.AI, cs.LG, stat.ML, cs.CL, cs.CV, cs.NE, cs.RO, cs.LO, cs.IR, cs.MA, cs.HC, cs.GT).

- Even if you grant most of the above, you might still be wary of *shifting* towards academia because of differential progress in capabilities: the marginal academic will probably boost capabilities more than alignment.

- My list of non-EA alignment insights is extremely cursory. Suggestions welcome!

- The subfields listed in 'De Facto Safety Work' are not equally relevant, but I don't know how to estimate the weights.

- I defined alignment as safe behaviour. Even this is contested: maybe you need guarantees about the system's internal motivations instead.

- I haven't even tried to estimate differences in research quality or real (non-Goodhart) productivity.

- ML seems to be a comparatively healthy field, hence the discount that should be applied to mainstream safety research is not enormous. (This is a caveat because I found it hard to get direct evidence for this.)

- A pet area I'd like to see more in EA safety: safety-critical engineering theorises about and actually implements highly-reliable machines, sometimes achieving "< 1 life lost per billion hours of operation". Their job is easier, but probably still instructive.

- The estimate of EA safety's size is probably a bit low, because there's no centralised census. Who knows how many entrants and exits there are for everyone that announces themselves? But this is also a factor in estimating safety-friendly academics; several times I've met academics who are interested in AGI alignment, but who lack any public evidence of this (possibly hoarding their weirdness points). Getting a good estimate of shy safety people seems important.

- I defined 'weak' and 'very weak' forms of the mid-term safety hypothesis. The strong form would be something like "Empirical study of current ML alignment is necessary for AGI alignment, because the problem is so ill-defined and 'wicked' and humans aren't very good at solving such things apriori".

- This work by Mackenzie and Hidysmith notes that maybe 80% of the insights that led to our present level of AI *capabilities* are from university researchers. It might be that academia is ill-suited to alignment work - since it's speculative, unevidenced, weird, and since STEM seems to have a positivity bias. But you'd need evidence to overcome the prior that they get stuff done, once they wake up, despite all the problems.