<https://www.alignmentforum.org/posts/YgNYA6pj2hPSDQiTE/distinguishing-definitions-of-takeoff>

**Discontinuity > Takeoff dynamics**: What will the takeoff look like?

**Discontinuity > Foom/Hard takeoff:** Single AI quickly reaches a competence that is superior than the world’s ability to control it.

**Discontinuity > Hansonian “slow” takeoff:** AI induced growth could occur fast, but not by a single/small subset of AI’s. AGI is more of an input to the economy.

**Discontinuity > Bostromian speed:** Anywhere between minutes, hours, days, years, or centuries. If slow (decades/centuries) then governments can allow for reaction and testing enables smaller existential risks. If fast, then the way it takes off is important.

**Discontinuity > Paul Christiano slow takeoff:** 4 yr interval doubles before 1 yr interval -> slow takeoff. Negation of that is fast takeoff.

**Discontinuity > Continuous/gradual/incremental takeoff:** Developing AI results in an outcome that beats extrapolating from the past

**Discontinuity > No takeoff:** Growth rate won’t be accelerating to a large magnitude.

**Discontinuity > Drexler’s takeoff:** CAIS, and so the self-improvement is part of larger scale of world’s self improvement.

**Discontinuity > Individual vs collective takeoff:** Will it be a takeoff by one or many AI’s.

**Notes from above link unmentioned in chart:** Baumann’s soft takeoff acts as a manner of measurement: measure in terms of economic times (time that adjusts for rate of economic growth) or political time (time adjusted for rate of social change)

<https://arxiv.org/pdf/1606.06565.pdf>

A lot of overlap with corrigibility. – Still could use a deeper dive, I only looked into the 3rd section deeply.

**Corrigibility > Risk-sensitivity:** How sensitive to risk will AI be?

**ML scales to AGI > Avoid Reward Hacking:** Agent focuses on improving the outcome as opposed to taking advantage of issue in objective function.

**Corrigibility > Scalable Oversight:** Increasing scale increases cost of AI – how to make oversight in a scalable manner?

**ML scales to AGI > Robustness to Distributional Shift:** If data distribution shifts, what happens? Fall apart or is the AI okay?

**ML scales to AGI > Measurement of environment complexity:** A way of quantifying how complex the environment is.

**Corrigibility > Impact Regularizer:** A way of penalizing changes to the environment, so agent can manage a way of putting in minimal side effects into its decisions.

<https://arxiv.org/pdf/2001.08361.pdf>

Noteworthy for actually estimating probabilities:

FOR NEURAL LANGUAGE MODELS:

Performance depends upon size of data, amount of compute, number of model parameters, weakly depend on depth vs width/other hyperparameters.

Overfitting is universal: must scale N and D in tandem.

Other relationships highlighted in paper section 1.1.

Does this generalize to other fields of AI too?

Similar to above: <https://arxiv.org/pdf/1909.12673.pdf>

<https://ai-alignment.com/an-unaligned-benchmark-b49ad992940b>

**ML scales to AGI > Narrow view optimization leads to broad view catastrophe:** Optimizing short term leads to catastrophe. Not really sure where this goes, but this feels particularly tractable for estimation based off historical data.