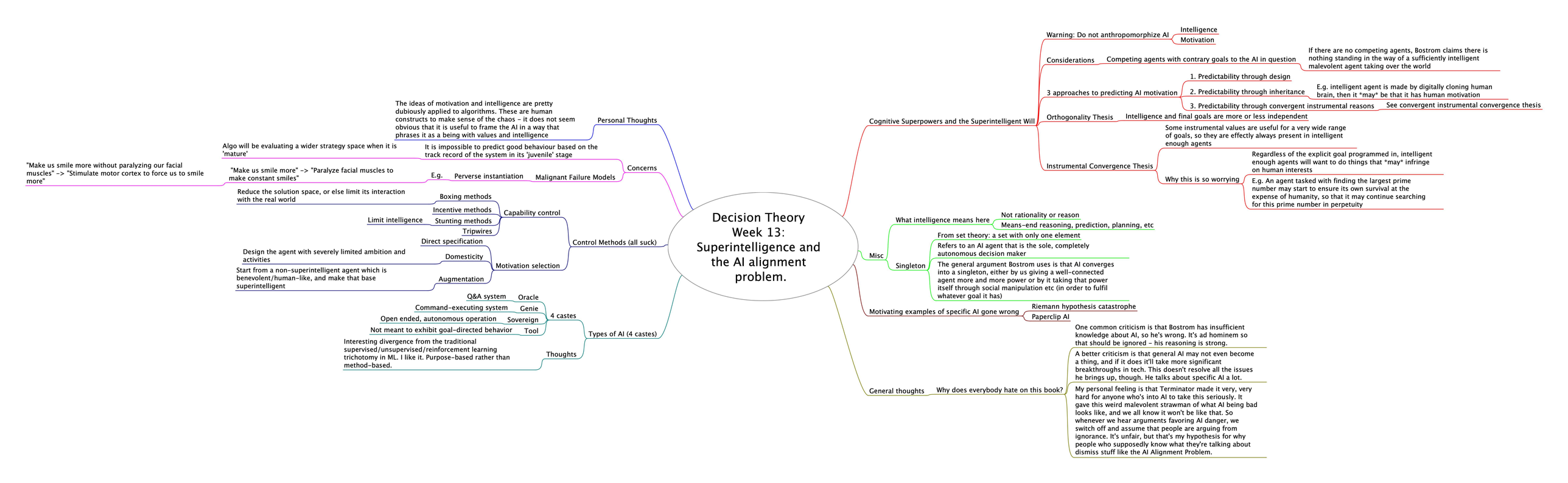
Decision Theory Week 13: Superintelligence and the AI alignment problem.
Cognitive Superpowers and the Superintelligent Will
Warning: Do not anthropomorphize AI
Intelligence
Motivation
Considerations
Competing agents with contrary goals to the AI in question
If there are no competing agents, Bostrom claims there is nothing standing in the way of a sufficiently intelligent malevolent agent taking over the world
3 approaches to predicting AI motivation
1. Predictability through design
2. Predictability through inheritance
E.g. intelligent agent is made by digitally cloning human brain, then it *may* be that it has human motivation
3. Predictability through convergent instrumental reasons
See convergent instrumental convergence thesis
Orthogonality Thesis
Intelligence and final goals are more or less independent
Instrumental Convergence Thesis
Some instrumental values are useful for a very wide range of goals, so they are effectly always present in intelligent enough agents
Why this is so worrying
Regardless of the explicit goal programmed in, intelligent enough agents will want to do things that *may* infringe on human interests
E.g. An agent tasked with finding the largest prime number may start to ensure its own survival at the expense of humanity, so that it may continue searching for this prime number in perpetuity
Misc
What intelligence means here
Not rationality or reason
Means-end reasoning, prediction, planning, etc
Singleton
From set theory: a set with only one element
Refers to an AI agent that is the sole, completely autonomous decision maker
The general argument Bostrom uses is that AI converges into a singleton, either by us giving a well-connected agent more and more power or by it taking that power itself through social manipulation etc (in order to fulfil whatever goal it has)
Motivating examples of specific AI gone wrong
Riemann hypothesis catastrophe
Paperclip AI
General thoughts
Why does everybody hate on this book?
One common criticism is that Bostrom has insufficient knowledge about AI, so he's wrong. It's ad hominem so that should be ignored - his reasoning is strong.
A better criticism is that general AI may not even become a thing, and if it does it'll take more significant breakthroughs in tech. This doesn't resolve all the issues he brings up, though. He talks about specific AI a lot.
My personal feeling is that Terminator made it very, very hard for anyone who's into AI to take this seriously. It gave this weird malevolent strawman of what AI being bad looks like, and we all know it won't be like that. So whenever we hear arguments favoring AI danger, we switch off and assume that people are arguing from ignorance. It's unfair, but that's my hypothesis for why people who supposedly know what they're talking about dismiss stuff like the AI Alignment Problem.
Types of AI (4 castes)
4 castes
Oracle
Q&A system
Genie
Command-executing system
Sovereign
Open ended, autonomous operation
Tool
Not meant to exhibit goal-directed behavior
Thoughts
Interesting divergence from the traditional supervised/unsupervised/reinforcement learning trichotomy in ML. I like it. Purpose-based rather than method-based.
Control Methods (all suck)
Capability control
Boxing methods
Incentive methods
Stunting methods
Tripwires
Limit intelligence
Reduce the solution space, or else limit its interaction with the real world
Motivation selection
Direct specification
Domesticity
Design the agent with severely limited ambition and activities
Augmentation
Start from a non-superintelligent agent which is benevolent/human-like, and make that base superintelligent
Concerns
Malignant Failure Models
Perverse instantiation
E.g.
"Make us smile more" -> "Paralyze facial muscles to make constant smiles"
"Make us smile more without paralyzing our facial muscles" -> "Stimulate motor cortex to force us to smile more"
It is impossible to predict good behaviour based on the track record of the system in its 'juvenile' stage
Algo will be evaluating a wider strategy space when it is 'mature'
Personal Thoughts
The ideas of motivation and intelligence are pretty dubiously applied to algorithms. These are human constructs to make sense of the chaos - it does not seem obvious that it is useful to frame the AI in a way that phrases it as a being with values and intelligence