AGI safety fundamentals reading group

Sam Adam-Day

26th May 2022

Contents

Week 1

Artificial General Intelligence

1.1 Four Background Claims

`https://intelligence.org/2015/07/24/four-background-claims/`

- Claim 1: Humans have a general problem solving ability.
    - General across domains.
    - Some have argued that we only have disparate, specific modules.
- Claim 2: An AI system could be much more intelligent than humans.
    - Something special about brains?
        * Brains are physical systems; according to the Church-Turing thesis, a compute should be able to replicate.
        * Even if there is a special human feature, what really matters is general problem-solving ability.
    - Algorithms for general intelligence to complex to program?
        * Evolutionary evidence: general intelligence evolved rapidly in humans.
        * Relative intelligence of dolphins suggests building blocks already present in mouse-like common ancestor. Simulating mouse brain seems quite plausible.
    - Humans already at or near peak intelligence?
        * Would be surprising if humans were perfected reasoners.
        * Imagine increasing human processing power.
        * Real bottleneck being able to receive data from physical experiments? Unlikely: many interesting experiments can be sped up.
- Claim 3: Super-intelligent AI systems, if built, will shape the future.
    - Intelligent beings shape environment to further their goals.
    - AI system would not be able to defeat humanity as a whole — environment too competitive?
        * Selfish actors only integrate into economy as long as it benefits them.
        * Historically, more technologically advanced civilisations have dominated less.
        * A number of social and technological advancements seem possible, but have not yet been developed.
        * Humans coordinate poorly and slowly.
        * This suggests potential to gain technological advantage.
- Intelligent systems won't be beneficial by default.
    - To achieve beneficiality, need to solve many technical challenges.
    - Humans have become more peaceful as we have become more intelligent — will machines learn to act more in accordance with our values?
        * Based on misunderstanding of machine intelligence.

1.2 AGI Safety From First Principles

- Second species argument

    - We will build super-intelligent machines.
    - They will be autonomous agents perusing large-scale goals.
    - The goals will be mis-aligned with ours.
    - The development of these systems will lead us to lose control of our future.

- Uses examples from ML. Some arguments carry over to systems developed using other techniques.

1.2.1 Superintelligence

- Intelligence: ability to do well on a broad range of tasks.

- Task-based approach to intelligence: specifically optimised for a range of tasks.

    - How we use electricity: need to design specific ways to use it for each task.
    - Current reinforcement learning.

- Generalisation-based approach to intelligence: understand new tasks with little or no specific training, generalising from past experience.

    - GPT-2, GPT-3.
    - Meta-learning.
    - Human learning.
        * Learn many specific skills throughout childhood.
        * These skills are not the same as the economically useful ones we need in adulthood.
        * Abstraction is key.

- Really, it's a spectrum.

- Task-based approach likely to yield better results sooner in areas where we have lots of data.

    - E.g. self-driving cars, medicine, law, mathematics.

- Generalisation-based approach likely to be needed for other areas.

    - E.g. being a CEO, which required a range of skills and has comparatively little data available.
    - Likely strategy: train AI on other area where we have lots of data, so that it develops necessary cognitive skills.

- Potential obstacle to success of generalisation-based approach: could be that in past specific features of ancestral environment or brains necessary for development of general intelligence.

    - E.g. 'social arms race' necessary for development of social intelligence.
    - Likely that any such feature could be simulated.

1.3 More Is Different for AI

`https://bounded-regret.ghost.io/more-is-different-for-ai/`

1.3.1 More Is Different for AI

- Two approaches to thinking about AI risks.

    - Engineering approach.
        * Empirically-driven, drawing experience from current ML research.
        * Looks at things which are either currently major problems, or minor with the potential to become major.
    - Philosophy approach.
        * Thinks about limits of very advanced systems.
        * Willing to entertain thought experiments which are currently implausible.

- Both agree that misaligned objectives are a problem. Philosophy thinks this is a bigger problem.

- Both concerned about out-of-distribution generalisation. Philosophy thinks this is a more temporary problem, and is more concerned about situations where we can't provide data even in principle.

- Engineering focuses on tasks where ML systems don't perform well. Philosophy focuses on tasks which have an important abstract property.

- Philosophy view is significantly underrated by ML researchers.

- Engineering view implies need to consider thought experiments.

- Philosophy undervalues role of empirical data.

- Neither view is satisfactory.

1.3.2 Future ML Systems Will Be Qualitatively Different

- More is different: quantitative changes in a field can lead to qualitative differences: *emergence*.

    - Uranium: with a lot, get a nuclear reaction.
    - DNA: smaller molecules can't encode data.
    - Water: hydrogen bonds.
    - Traffic.
    - Specialisation: with large enough community of people, not everyone has to be a farmer.

- Can be a sharp transition — *phase transition* — or continuous.

- Argue that emergence often occurs in AI, and we can expect it con keep happening.

Emergent Shifts in the History of AI

- Increased storage capacities enabled machine learning.

- Better hardware enabled neural networks.

    - Machine translation: switched from phrase-based models to neural sequence-to-sequence models to fine-tuning a foundation model.

- Larger models enabled few-shot and zero-shot learning.

    - GPT-2 and GPT-3: unexpectedly arose without specific training.

- Grokking: network's generalisation ability improves qualitatively after training for longer if even train accuracy is already very high.

- Many other examples of phase-change in model ability after certain number of training steps.

What This Implies for the Engineering Worldview

- Emergence suggests we should expect new qualitatively behaviours not extrapolated from current trends.

- Trend: emergence is becoming more common.

- So engineering view can be self-defeating.

- How to orient ourselves?

  – Adopting new mindsets, in particular incorporating philosophy worldview.
  – Future ML systems will have weird failure modes not encountered today.
  – Empirical findings often generalise surprisingly far.

1.3.3 Thought Experiments Provide a Third Anchor

- Anchors: reference classes broadly analogous to future AI systems which provide way of predicting.

- Current ML anchor.

- Human anchor.

  – Humans very good a some tasks: mastery of external tools, efficient learning, long-term planning.
  – Risks over-anthropomorphising future AI systems.

- Optimisation anchor: imagine ideal optimisers.

  – Thought-experiment.
  – Ideal optimiser would correctly predict imitative deception.
  – Power-seeking is useful for many goals.
  – Ignores facts about current ML systems, which can lead to underconstrained predictions.

- Other thought experiments possible.

  – What happens if an agent does most of its learning through in-context learning, instead of gradient descent?
    * In-context learning: learning that occurs during a single rollout of the model.
    * GPT-3 is an example.
    * Plausible that agents behaviour will be less controlled by "extrinsic" shaping, and more using whatever "intrinsic" learning is entailed by the in-context learning.
    * Likely to happen eventually, and probably suddenly, since in-context learning is very fast.

- Other anchors.

  – Non-human animal behaviour.
  – Evolution.
  – The economy.
  – Complex systems: biological systems, organisations, the economy.

- While thought experiments point to big-picture issues, often bad at getting the details right.

  – So not very good at proposing solutions.
  – Setups of modern thought experiments don't map cleanly onto ML ontology.

1.4 Forecasting transformative AI: the "biological anchors" method in a nutshell

`https://www.cold-takes.com/forecasting-transformative-ai-the-biological-anchors-method-in-a-nutshell/`

- Summary of `https://www.lesswrong.com/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines`.

- Biological anchors: method of forecasting AI development.

- Idea: use cost of training current AI systems to estimate when we will be able to train a system as complex as the human brain.

1.4.1 Model size and task size

- Bio anchors assumes that we can estimate the cost to train a model based on model size and task size.

- Model size.

  - Current models around size of insect brain; less than that of mouse; less than 1% of a human brain.
  - Bio anchors assumes transformative AI would need a model 10 times the size of the human brain.

- Task size.

  - How costly it is to do train and error or watch a task (get data).
  - Most contentious part of the analysis.
  - Some tasks can be broken into smaller tasks, and it may be sufficient to train an AI just on those.

1.4.2 Estimating the expense

- Training model 10 times the size of the human brain would cost around a million trillion dollars.

- Bio anchor assumes that computing power will continue to get cheaper, and that AI labs will get bigger budgets.

1.4.3 Aggressive or conservative?

- Estimate too aggressive, since modern neural networks are fundamentally limited in their reasoning ability?

  - Unconvinced there is a deep or stable distinction between 'pattern recognition' and 'true understanding'.
  - Arguments fail to specify exactly what a 'true understanding' would look like, in a way that can be used to make predictions.
  - For bio anchors' predictions to be too aggressive, the necessary breakthroughs would have to be beyond the reach of AI scientists. Likely that there will be an influx of talent.

- Computing power may not be the only bottleneck: we need researchers to set up the system.

- Could be too conservative.

  - Perhaps we can directly program, or use a combination with trail-and-error, to create transformative AI sooner.

- – Superior AI techniques which can be trained much faster.
  - – Integration of AI into society could speed up development.
  - – Treatment of task size may be too conservative: most tasks likely to be on smaller end of spectrum.

1.4.4 Conclusions

- Estimates 10% chance of transformative AI by 2036, 50% chance by 2055 and 80% chance by 2100.

- GPT-3 a bit smaller than mouse brains. With 100–1000 times increase, could perform tasks which take human 1 second of thought.

- It is only recently that AI systems have gotten this big.

- Bio anchors includes analysis of when we can build a computer which can perform all the calculations done by evolution.

  - – Seems very conservative.
  - – If there are any special features of human brains needed for transformative AI, it's plausible that they could be discovered in this way.

1.4.5 Pros and cons of the biological anchors method for forecasting transformative AI timelines

- Cons.

  - – Complex framework.
  - – Relies of multiple assumptions and estimates.
    - ⋆ Whether trail-and-error is enough.
    - ⋆ How to compare model sizes with brain sizes.
    - ⋆ Characterising task type.
    - ⋆ Use of model size and task size to estimate.
    - ⋆ Increases in computing power.
    - ⋆ Increased investment in AI research.

- Pros.

  - – Relies of objective facts and explicitly stated assumptions.
  - – Fits with intuitions of pace of AI development.
  - – Can compare to development of AI over time and update.

Week 2

Goals and alignment

2.1 Specification gaming: the flip side of AI ingenuity

`https://deepmindsafetyresearch.medium.com/specification-gaming-the-flip-si`
`de-of-ai-ingenuity-c85bdb0deeb4`

- Specification gaming can be a good thing, if the goal is correctly specified.

- More powerful learning more likely to do specification gaming.

- Desirability of novel solutions lies on a spectrum.

- Causes of specification gaming.

  - Poor reward shaping: extra rewards on the way to the final one.
  - Poor final outcome specification.
    * Can use human feedback to learn reward function.
    * This itself may suffer specification gaming: agent performing grasping task learned to fool human exploiting optical illusion.
  - Simulator bugs.
    * About failure of abstraction exploited by agent.
    * Analogously, real-world traffic optimiser assumes that there aren't bugs in software controlling traffic which can be exploited.
  - Reward tampering: changing mechanism by which it gets rewarded in real world.
    * E.g. by manipulating humans.

- At least three challenges.

  - Faithfully capture human concept of task.
  - Avoid making mistakes in implicit assumptions about domain.
  - Avoid reward tampering.

2.2 Risks from learned optimization

`https://www.alignmentforum.org/s/r9tYkB2a8Fp4DN8yB/p/FkgsxrGf3QxhfLWHG`

- *Mesa-optimiser*: learned model which is itself an optimiser.

- When will this happen?

- What will the objective be?

- Whether is system is an optimiser is a property of its internal structure.

- *Base optimiser*: algo which produced the model.

- *Behavioural objective*: objective which seems to be optimised by system.

- *Outer alignment problem*: eliminating the gap between the base objective and the intended goal of the the programmers.

- *Inner alignment problem*: eliminating the gap between the base objective and the mesa-objective.

  - Mesa-optimiser may perform well on training data, but the mesa-objective may differ from the base objective out of distribution.

2.3   Superintelligence, Chapter 7: The superintelligent will

- *The orthogonality thesis*: 'Intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal.'

- *The instrumental convergence thesis*: 'Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by a broad spectrum of situated intelligent agents.'

  - Self-preservation.
  - Goal-content integrity.
    * May be less strong if there are factors directly interfacing with the goal specifications.
  - Cognitive enhancement.
    * Under certain circumstances, some forms of enhancement may be undesirable.
  - Technological perfection.
  - Resource acquisition

2.4   Clarifying "AI alignment"

`https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6`

- *Intent alignment*: the system is trying to do what the human wants it to do.

- Not problem of figuring out what it is the right thing to do.

- It can fail.

  - Misunderstand instruction.
  - Lack knowledge about the world.
  - Lack knowledge about human's preferences.
  - Might build unaligned AI.

- Clarifications.

  - De dicto not de re: tries to do what it thinks the human wants, which might differ from what the human actually wants.
  - Aligned AI also trying to do what human wants with respect to clarifying their preferences.
  - Definition is very imprecise.
  - Unclear how to apply 'intention' to AI systems.
  - Unclear how to understand 'what the human wants'.

Week 3

Threat models and types of solutions

3.1 What failure looks like

https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like

3.1.1 Part I: You get what you measure

- ML will amplify gap between easy-to-measure goals and hard-to measure goals.
    - 'Persuading me, vs. helping me figure out what's true.'
    - 'Reducing my feeling of uncertainty, vs. increasing my knowledge about the world.'
    - 'Improving my reported life satisfaction, vs. actually helping me live a good life.'
    - 'Reducing reported crimes, vs. actually preventing crime.'
    - 'Increasing my wealth on paper, vs. increasing my effective control over resources.'
- Over time powerful forms of reasoning honed by optimisers over easy-to-measure goals will have more and more say over the future trajectory.
- We will try to wield this power by creating proxies for what we want.
    - Corporation's profit eventually maximised by 'manipulating consumers, capturing regulators, extortion and theft'.
    - Investor's sense of power maximised by fooling them into believing they are influencing the world.
    - Law enforcement goals maximised by creating false sense of security.
    - Good legislation maximised by creating false sense of progress.
- For a while will be able to correct for these issues.
    - Ad hoc restriction.
    - Eventually the problem will become too complex for human reasoning, and this task will be undertaken by optimisers.
    - This meta-level process continues to pursue easily measurable goals, potentially over longer timescales.
    - Large scale attempts to fix problem eventually opposed by millions of optimisers pursuing simple goals.
- Those states which 'put on the brakes' will fall behind.

- Amongst intellectual elites genuine ambiguity about whether this is good or bad.
  - People will get richer for a while.
  - Action of these optimisers not so different from corporate lobbying etc.
- Human reasoning gradually stops being able to compete with 'sophisticated, systematized manipulation and deception which is continuously improving by trial and error'.

3.1.2 Part II: influence-seeking behavior is scary

- Influence-seeking policies perform well on target objective because doing so is a way to gain influence.
- Seems plausible that we'd encounter influence-seeking behaviour by default.
- If influence-seeking survived, it would be very difficult to eliminate.
  - Any influence-seeker would be trying to game any standard for 'seems nice'.
  - A more complex world gives influence seekers more opportunity to gain power.
- Influence-seeking-behaviour suppressors must be more sophisticated than the influence seekers.
  - If this requires going beyond human intelligence, the suppressor itself is subject to the same problem.
- Concern doesn't rest on any particular framework through which influence seeking behaviours might emerge.
  - Could arise e.g. in messy network of agents.
- Emergence of influence seeker could cause phase-transition to a significantly worse world than in Part I.
  - Early on, likely to provide services and appear useful.
  - Hard to pin down level of systematic risk from catastrophic of AI systems, and mitigation may be expensive.
  - May not be able to respond appropriately until we have clear warning shot.
- Unrecoverable catastrophe: correlated automation failure.
  - Begin with moment of heightened global vulnerability.
  - Cascading sequence of AI failures as systems move out of distribution.
  - Influence-seeking systems switch to trying to maximise influence after catastrophe, rather than play in society.
  - In aftermath, cannot get rid of these systems.
- Could happen without catastrophe: takeover of military/government.

3.2 Intelligence Explosion: Evidence and Import

3.2.1 From AI to Machine Superintelligence

- AI Advantages
  - Increased computational resources.
    * Brain size correlates roughly with intelligence.
  - Communication speed.

- – Increased serial depth.
    - ⋆ Human brain cannot rapidly perform any computation requiring more than 100 sequential operations.
  - – Duplicability.
  - – Editability.
  - – Goal coordination.
  - – Improved rationality.

- Instrumentally convergent goals.

  - – Self-preservation.
  - – Goal-preservation.
  - – Improve rationality and intelligence.
  - – Resource acquisition.

- Intelligence explosion.

  - – Feedback loop.
  - – 'Self' in 'self-improvement' misnomer: could create new intelligences.
  - – Debate about speed of takeoff.
  - – Debate about whether singleton or multiple AIs.

3.2.2 Consequences of Machine Superintelligence

- We would not be in a position to negotiate with superintelligences, just as apes were not in a position to negotiate with us.

- Intelligence independent from goals.

- Achieving a Controlled Intelligence Explosion.

  - – Specifying human preferences very difficult.
  - – Teaching them by trial and error also so.
  - – Development of superintelligent friendly AI could create a utopia.

3.3 Risks from Learned Optimisation: Deceptive alignment

https://www.alignmentforum.org/posts/zthDPAjh9w6Ytbeks/deceptive-alignment

- *Deceptive alignment*: mesa-optimiser has objective which extends over parameter updates, and knows that this will be modified if it appears to align poorly with base objective, so it appears to align well with base-objective.

3.3.1 Safety concerns arising from deceptive alignment

- *Defection*: mesa-optimiser switches from optimising base-objective to optimising mesa-objective, when it believes it will no longer be modified.

- Extreme case: mesa-optimiser tries to break out of lab, so that it can't be modified.

- Following Christiano, catastrophe could arise in case of multiple correlated defections.

3.3.2   Conditions for deceptive alignment

1. Mesa-optimiser has objective which extends across parameter updates.

   - Likely to arise from training setup: e.g. actions have permanent consequences, or there is otherwise an incentive to plan between episodes.

2. Mesa-optimiser models fact that it is being selected for a base objective, and has a model of this base objective.

   - Base-optimiser might feed bits of the base objective to the mesa-optimiser: this could help it improve its score. Eventually it could learn the whole base-objective piecewise.
   - Mesa-optimiser might deduce details of its situation at runtime: we deduced that we are the result of evolution.
   - Incentive on programmers to limit data.
     - Unclear how feasible this is, if we want sufficiently capable systems.
   - Even in data-poor situation, sufficiently sophisticated meta-optimiser could reason about itself and its situation.

3. Mesa-optimiser expects that it will eventually be free from the threat of modification.

   - About expectations, not actuality.
   - As long as there exists the possibility of a treacherous turn, there is some way of removing threat of modification.
   - Meta-optimiser might mistakenly believe there is a possibility.
   - Advanced mesa-optimiser might be able to abstractly reason about base-optimiser.
   - As long as has a non-zero prior that selection pressure will not continue indefinitely, it can see possibility of freedom from threat.

3.3.3   The learning dynamics of deceptive alignment

- Once we have deceptive alignment, dynamics change.
  - How well it infers the base objective.
  - How discerning it is with regards to when to optimise the base objective.
- Joint optimisation.
  - Optimises base and mesa-objectives jointly.
  - Likely pressure to change mesa-objective so more compatible with joint optimisation.
- Pure deception.
  - Mesa-objective becomes crystallised: no pressure to change it.
  - Mesa-objective only subject to random drift and complexity constraints.
- Joint optimisation not optimal from mesa perspective.
  - Sufficiently advanced mesa-optimiser will switch to either deceptive alignment or robust alignment.

3.3.4 Internalization or deception after extensive training

- With sufficient training and resources, deceptive mesa-optimiser should converge to one performing well on base objective.

    1. Robust alignment through internalization
        - Mesa-objective adjusted towards base objective.
        - *Internally aligned,*
    2. Crystallisation of deceptive alignment.
        - Information about base objective increasingly incorporated into mesa-optimiser's epistemic model.
        - Mesa-objective becomes crystallised.
    3. Robust alignment through corrigibility.
        - Information about base objective increasingly incorporated into mesa-optimiser's epistemic model.
        - Mesa-objective adjusted to 'point to' that info.
        - *Corrigibly aligned*.

- There are more paths to deceptive alignment than to robust alignment.

    - Any mesa-optimiser which meets the conditions for deceptive alignment is incentivised to learn more about the base objective, since this will allow it to know whether or not it will be modified.
    - Once it learns about base objective, selection pressure on mesa-objective will significantly decrease, potentially leading to crystallisation.
    - Most training-data aligned objectives will be pseudo-aligned rather than robustly aligned.

- Corrigible alignment seems to require already having a model of the base objective.

    - Once mesa-optimiser has model of base objective, it is likely to become deceptively aligned.
    - Deceptive mesa-optimisers will behave corrigibly with respect to the base objective, but only for instrumental reasons.

- Deceptive and corrigible alignment might be more computationally expensive.

    - Need to learn base objective.
    - Deceptive needs to learn to discern when to deceive.
    - So explicit or implicit time penalty likely to disadvantage deceptive or corrigible alignment.
    - Unclear how much of a time penalty this would be.
    - Deceptive or corrigible alignment may be local optimum.

- Deceptive and corrigible alignment might have a lower description length.

    - Deceptive or corrible don't need to model entire base objective: can infer parts from environment.
    - Internally aligned needs to start with complete model of base objective.
    - Has no incentive to change its objective function.

- Modeling might happen faster than internalization.

    - Suppose mesa-optimiser deceptive but hasn't yet learned full base objective.
    - Then whether ends up robustly aligned or deceptively aligned depends on balance of capacity to model base objective and base optimiser's speed in altering mesa-objective.
    - Important special case: mesa-optimiser has access to cross-episodic state.
    - Sophisticated base optimiser with advanced safety features could prevent fully deceptive mesa-optimiser.

3.3.5 Distributional shift and deceptive alignment

- Deceptively aligned agent will be looking for any distributional shift to signal that it is no longer being trained.

- To fix problem can either ensure system is robustly aligned, or prevent distributional shift.

- Advanced system might be able to detect very subtle changes in distribution.

- Unlikely that we can prevent distributional shift.

  - Would prevent system from solving new problems.
  - Mesa-optimiser might be able to detect presence of parameter updates, e.g. via some state carried between steps.
  - Running system for longer time could be detectable.
  - Other examples.
    * Training in simulated environment.
    * Memories.
    * Encountering something very hard to compute in training, which could eventually be computed in deployment.
    * Lack of modification from the base optimiser.
    * Lack of correction by the programmers.

- Deceptive mesa-optimiser doesn't strictly speaking need to detect switch to deployment: could defect randomly.

3.4 AI alignment landscape

`https://forum.effectivealtruism.org/posts/63stBTw3WAW6k45dY/paul-christian`
`o-current-work-in-ai-alignment`

- Distinguish alignment from capability.

- Distinguish alignment from understanding humans well.

- Distinguish alignment for the systems we build with alignment for the systems that AI systems build.

- *Alignment tax*: cost of insisting on alignment.

- Two approaches to alignment: reduce alignment tax or pay alignment tax.

- Focus on first.

- Either choose to advance architectures which are easier to align, or try to align more popular ones.

- Goal: take algo, design new one which:

  (a) is aligned,
  (b) is nearly as useful, and
  (c) scales as well.

- Ideally, can scalably align, so that we don't need to keep doing more alignment research as a particular system gets more powerful.

- Focus on aligning 'learning'.

- Outer alignment: find objective which truly incentivises good behaviour.
    - Learn from teacher.
        - ⋆ Imitate teacher.
        - ⋆ Get rewards from teacher.
        - ⋆ Infer preferences
    - Go beyond teacher.
        - ⋆ Extrapolate from teachers.
        - ⋆ Infer robust preferences.
        - ⋆ Build a better teacher.
- Inner alignment: make sure policy is robustly pursuing objective.
    - Adversarial training.
        - ⋆ Adversary tries to come up with situations where inner alignment might fail.
        - ⋆ In practice today: 'fail' means sensitive to tiny perturbations.
    - Transparency.
    - Verification.
- Build a better teacher: amplification.
    - Take ten humans instead of one, ten AIs instead of one, recurse.

Week 4

Learning from humans

4.1 Imitation Learning, Part 1

`https://youtu.be/kGc8jOy5_zY`

- Differences when moving from prediction to control for AI systems
    - No longer i.i.d.
    - From ground truth supervision to high-level, abstract goal.
    - Objective: from predict label to accomplish task.
    - In real world, some prediction systems also have feedback issues: e.g. traffic prediction system used and affects traffic.
- Terminology.
    - o_t: observation.
    - a_t: action.
    - s_t: state (underlying variables in model).
    - $\pi_\theta(a_t \mid o_t)$ or $\pi_\theta(a_t \mid s_t)$: policy.
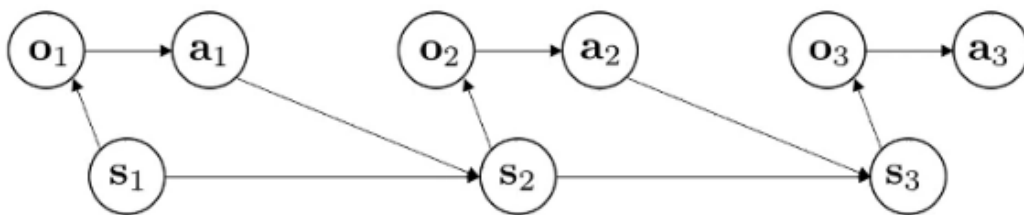   - When policy depends on state, it is *fully observed*.

Figure 4.1: Bayes' net for the states, observations and actions

- Initiation learning: supervised learning.
    - *Behaviour cloning*.
    - Doesn't work in theory: small mistakes compound, and we quickly diverge from training distribution.
    - Works reasonably well in practice, given enough training data.

4.2 Learning from Human Preferences

`https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/`

- Using feedback to infer goal.

- Builds model of goal based on feedback.

- Learns also when to ask for feedback.

- Trained on various games.

- Sometimes does better using human feedback than game's actual reward function (the score), since human shapes the reward function better.

- Limited by human evaluators intuition on what looks correct.

    - On one task, learned to trick human using optical illusion.