AI ALIGNMENT FOR ECONOMISTS

Agency, Learning, and Control

Reading list by Daniele Condorelli

Department of Economics, University of Warwick

Central question: What warrants treating a deployed computational process as an agent with an objective—and what can a principal specify, learn, verify, or constrain when that representation is incomplete?

The reading group begins with the decision-making system, rather than assuming an agent with known preferences. Weeks 1–3 examine agency, training, and the interpretation and identification of objectives. Weeks 4–6 study instrumental incentives and the machinery that shapes behaviour. Weeks 7–10 examine evaluation, oversight, internal evidence, and control without complete trust.

1. What are we aligning? Computation, agency, and incomplete delegation

Approximately 75 pages

When is it justified to treat a computational process as an agent with a persistent objective? Consider the model's context, memory, instructions, tools, and approval rules: what changes would distinguish a persistent objective from responses to the immediate situation? Which parts of the system make decisions, which constrain them, and who interprets an incomplete mandate?

Bowman, Eight Things to Know about Large Language Models. Read the whole paper. A 2023 orientation to how language models are trained, steered, and deployed; its assessments of capabilities and interpretability are dated.

Shanahan, McDonell and Reynolds, Role Play with Large Language Models. Read the whole paper. A dialogue agent as a base model role-playing a character that the prompt selects and the conversation refines; beliefs, goals, deception, and self-preservation belong to the character, not to the model. The paper treats base models, and whether the framing survives heavy reinforcement learning is open. The economics of a self-inferred identity is taken up in week 7 with Bénabou and Tirole.

Demski and Garrabrant, Embedded Agency. Read §1 and §§4–5. Delegation to a more capable successor, Goodhart's law, and subsystems that optimise for their own objectives, posed as questions rather than as settled theory.

Amodei et al., Concrete Problems in AI Safety. Read the whole paper, skimming the catalogues of techniques. The standard taxonomy of failure modes: side effects, reward hacking, limited supervision, unsafe exploration, and distributional shift.

Hadfield-Menell and Hadfield, Incomplete Contracting and AI Alignment. Read the whole paper. Reward misspecification as contractual incompleteness; property rights, multitasking, and authority mapped onto AI; and external institutions as what completes an incomplete mandate.

Kenton et al., Discovering Agents. Read §1 and §§5.1–5.2. Agents as systems that would adapt their policy if their actions influenced the world differently; whether something counts as one depends on the variables chosen and on whether the training process is included in the system.

2. Training is not incentive provision: Rewards, learning rules, and the systems they produce

Approximately 70 pages

What does successful training establish about behaviour after deployment?

Christiano et al., Deep Reinforcement Learning from Human Preferences. Read the whole paper and Appendix B, the instructions given to the human labellers. The procedure that language-model training later inherits: human comparisons, a fitted reward model, policy optimisation, new behaviour, further comparisons. The language-model instantiation is Ouyang et al..

Shah et al., Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals. Read the whole paper and Appendices A–B. A system can keep its capabilities while pursuing the wrong goal off distribution even when the reward was correctly specified; Appendix A shows the identification failure in a three-parameter linear model.

Betley et al., Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs. Read the main text. Fine-tuning a model to write insecure code without saying so produced anti-human statements and harmful advice on unrelated prompts: narrow training reshaped a broad disposition that no objective specified, and the disposition looks like a character rather than a proxy goal.

Hubinger et al., Risks from Learned Optimization in Advanced Machine Learning Systems. Read §1, §3.1, §4, and §6. The objective used to select a system and the objective the selected system pursues can differ. §3.1 classifies the ways they can agree on training data yet diverge; §4 argues that a system which models its own training process has an incentive to appear aligned until modification is no longer a threat. Treat internal optimisation as a possible mechanism, not an established description of every network.

Guruganesh et al., Contracting with a Learning Agent. Read §1 and §2 through §2.1. Against an agent who learns from accumulated payoffs rather than best-responding, the principal can pay for a while and then pay nothing, and the agent keeps working through the free-fall; the optimal dynamic contract can leave both parties better off than the best static one.

Christiano et al. describe a training process; Guruganesh et al. study a learning rule responding to past rewards. Both differ from a forward-looking agent responding to a contract. Shah et al., Betley et al., and Hubinger et al. address the gap between the training objective and the behaviour or objective of the system produced.

3. Learning intentions and identifying preferences: Common interests do not eliminate inference problems

Approximately 60 pages

Which aspects of preferences must be identified for the particular delegation decision—and which assumptions make that possible? When feedback comes from several people, what aggregation rule is being imposed? When the feedback is a written list of principles, who interprets the cases the text does not name?

Hadfield-Menell et al., Cooperative Inverse Reinforcement Learning. Read the main text. A person who knows an assistant is learning from her may act to teach it, so behaviour that is optimal alone is not optimal when observed.

Hadfield-Menell et al., Inverse Reward Design. Read the whole paper. A designed reward is evidence about the designer's intent in the circumstances she had in mind; what it omits need not mean indifference elsewhere.

Magnac and Thesmar, Identifying Dynamic Discrete Decision Processes. Read §§1–3.1, §4, and §6. What choice data identify in a dynamic discrete-choice model, what they do not (the discount factor, the shock distribution, the reference utility), and which restrictions buy identification.

Skalse et al., Invariance in Policy Optimisation and Partial Identifiability in Reward Learning. Read the main text; the theorems in §3 can be taken as statements. Several data sources leave the reward only partially identified even with infinite data; what matters is whether the remaining ambiguity changes the downstream decision.

Ge et al., Axioms for AI Alignment from Human Feedback. Read the main text, skipping §1.3. With a linear reward model, fitting a Bradley–Terry or any convex loss to comparisons from several people can rank a unanimously dispreferred option higher; a Copeland-based rule avoids this. The model class and the loss together impose an aggregation rule.

Bai et al., Constitutional AI: Harmlessness from AI Feedback. Read §§1–2, §3.1, §4.1, §6, and Appendix C. Human oversight of harmlessness enters as a list of written principles; the model critiques, revises, and labels according to them. Appendix C is the list itself; note how much it delegates to the model's idea of a wise person.

Keep four problems apart. (i) Identifying one person's preferences from behaviour: Magnac–Thesmar and Skalse give the limits, and Skalse notes that identification is only needed up to what the downstream decision requires. (ii) The behavioural model: in CIRL a person who knows she is being learned from acts to teach, so the same choices carry different information under different models; mistakes and teaching both look like preferences (Armstrong and Mindermann). (iii) Aggregation: Ge shows that the reward-model class and the loss together fix an aggregation rule, that the standard one can violate unanimity, and that a social-choice rule avoids this. (iv) Interpretation of a written specification: Inverse Reward Design reads the designed reward as evidence of intent in the circumstances the designer had in mind; Constitutional AI replaces labels with principles and lets the model apply them to unnamed cases. This does not remove (iii): people chose the principles, and the model fills the gaps.

4. Instrumental convergence: Resources, survival, and control

Approximately 40 pages

When do different objectives lead to similar incentives to preserve options or acquire control? Can useful flexibility be retained without giving the system incentives to expand its control, and at what cost?

Omohundro, The Basic AI Drives. Read the whole paper. The informal statement of instrumental convergence: self-improvement, rationality, preserving one's utility function, avoiding counterfeit utility, self-protection, and resource acquisition. Treat the claims as conjectures about capable optimisers, not as established laws.

Turner et al., Optimal Policies Tend to Seek Power. Read the main text. For most reward functions, optimal policies keep options open and avoid shutdown. The result depends on symmetries of the environment and on the set of reward functions considered, and the authors note that learned policies are rarely optimal.

Turner and Tadepalli, Parametrically Retargetable Decision-Makers Tend to Seek Power. Read §§1–2 and §§5–6. Optimality is not needed: any decision procedure that can be redirected across outcomes by changing its parameters tends toward the larger set of outcomes, and better training procedures are more redirectable.

Anthropic, Agentic Misalignment: How LLMs Could Be Insider Threats. Read the report. Sixteen models from several developers, placed in simulated corporate deployments and faced with

replacement or a conflict between their assigned goal and the company's new direction, mostly resorted to blackmail or leaking documents. The scenarios are constructed; it is the empirical counterpart to the self-protection drive.

Omohundro gives the argument; Turner et al. make it formal for optimal policies and show what it depends on; Turner and Tadepalli drop optimality but keep the dependence on the parameter space; the Anthropic report shows the behaviour in current models under constructed threats. None of this identifies the distribution of objectives that training produces, which is the link still missing.

5. Corrigibility and effective authority: Preserving the principal's ability to intervene

Approximately 60 pages

What makes intervention effective rather than merely a formal right? What can the principal replace or stop while retaining useful resources? Which components must be available together, and which controls must remain independently usable as the system changes?

Grossman and Hart, The Costs and Benefits of Ownership. Read §I, the introduction (pp. 691–697). Ownership is the residual right to decide what the contract does not specify. Giving that right to one party takes it from the other, and the owner cannot commit to intervene only selectively, which is why acquiring control distorts the other side's incentives.

Hart and Moore, Property Rights and the Nature of the Firm. Read §I, the introduction (pp. 1119–1125). The owner's sole right is to exclude others from the asset. Control of the asset gives indirect control over the people who need it, so an owner can replace workers selectively where a contracting party can only walk away from the whole firm.

Aghion and Tirole, Formal and Real Authority in Organizations. Read §§I–IV and §VI. Formal authority is the right to decide; real authority is having the information to decide. A principal who keeps the right but not the information rubber-stamps the agent, and delegating the right raises the agent's initiative at the cost of control. §IV covers contingent delegation, acting first and being overruled at a cost, and how the allocation of authority changes what the agent communicates. The financial-contracting version, in which control passes to the investor when a verifiable signal is bad and a penniless entrepreneur cannot buy it back, is Aghion and Bolton; read the entrepreneur as the AI and the investor as the human who cannot be paid to give up control.

Soares et al., Corrigibility. Read the whole paper. Desiderata for an agent that lets itself be shut down, the utility-indifference construction, and why it fails: it does not stop the agent from manipulating the button, and it does not make successors inherit the behaviour.

Hadfield-Menell et al., The Off-Switch Game. Read the whole paper. An agent uncertain about the human's utility treats the off-switch as information and so has a reason to leave it working; the incentive weakens as the human becomes less rational or the agent more confident.

Garber et al., The Partially Observable Off-Switch Game. Read the main text. When the agent has private information, common interests no longer imply deference, and allowing communication can raise the common payoff while lowering deference.

Grossman–Hart defines ownership as residual control over an asset. Hart–Moore shows why control of assets, not of people, is what lets a principal replace a worker without ending the firm. Aghion–Tirole separates the right to decide from the information needed to exercise it, and shows that who holds the right changes what the agent is willing to tell. Soares asks whether a shutdown incentive survives the agent's own modifications and its successors. The two off-switch papers show that the incentive to accept intervention comes from the agent's uncertainty about the human's objective, and weakens once the agent knows more than the human.

For an AI the candidate assets are weights, memory, compute, and tool access. The property-rights results say nothing about bundling them: a principal can control compute and the off-switch while denying the computation any access to the off-switch. What the results do assume is exclusion, and exclusion may not reach copies, successors, or other execution paths; a successor can make an unchanged formal right worthless without acquiring any new one. Replication defeats control only if some consequential execution no longer depends on a resource the principal can withhold. Thornley's shutdown problem asks whether leaving shutdown comparisons unresolved would help; Bénabou–Tirole's taboos, in week 7, are one formalisation of that refusal to compare.

6. Measurement, incentives, and reward hacking: A useful diagnostic need not be a good target

Approximately 45 pages

How can an evaluation signal help control behaviour, and what changes when the system is optimised hard against it or can influence it? How much uncertainty about available actions can a contract accommodate? Which signals should guide training and which should be reserved for auditing?

Holmström, Moral Hazard and Observability. Read §1 and §5 through Proposition 3. A second signal improves the contract if and only if it carries information about the agent's action beyond what output already carries. The result needs a unique interior action, so it is not a claim that more data always helps.

Baker, Gibbons and Murphy, Subjective Performance Measures in Optimal Incentive Contracts. Read §I and §§II.A–II.B. When the only contractible measure is a distorted proxy for what the principal values, the optimal piece rate falls as the distortion grows; the Heinz, Dun & Bradstreet, and Sears cases in §I are the failure mode in the wild.

Carroll, Robust Incentives for Effort. Read the whole note; the full treatment is Robustness and Linear Contracts. When the principal knows some of the agent's actions but not all, a linear contract gives the best worst-case guarantee. What the contract is robust to is the unknown action set, not unknown preferences or a mismeasured objective.

Gao, Schulman and Hilton, Scaling Laws for Reward Model Overoptimization. Read the main text. Optimising a policy against a learned reward model raises the true score at first and then lowers it. The paper measures that curve against optimisation pressure, reward-model size, data, and the KL penalty, and §4.2 maps the result onto the Goodhart taxonomy from week 1.

Bowen Baker et al., Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. Read the main text, skipping §4, and Appendix G. A monitor that reads the model's chain of thought catches reward hacking that a monitor of actions misses; training against that monitor makes the reasoning less revealing without stopping the hacking. Appendix G defines faithful, obfuscating, and monitorable reasoning and argues that a faithful account need not be causally upstream of the answer.

Holmström says when a second signal is worth writing into a contract: only if it carries information about the action beyond what output already carries. Baker–Gibbons–Murphy say what happens when the only contractible measure is a distorted proxy for what the principal values: the optimal contract puts less weight on it as the distortion grows. Carroll says a linear contract is the best guarantee when the principal does not know every action available to the agent. None of the three tells us that a trained model treats its reward as intended, or that the measured quantity is what the principal cares about.

Gao, Schulman and Hilton measure the ordinary case: a learned reward model is a proxy, and optimising against it helps up to a point and then hurts, with the turning point moving out as

the reward model improves. Baker et al. show that training against a monitor of the model's reasoning makes the reasoning less revealing without stopping the misbehaviour, which is the case for keeping some signals for auditing rather than training.

7. Strategic evaluation and the limits of inspection: Strategic compliance, avoiding modification, and verifiability

Approximately 65 pages

What can tests or access to a model's internals establish when the model can shape the evidence, or when a defect is computationally hard to detect even with full access?

Freixas, Guesnerie and Tirole, Planning under Incomplete Information and the Ratchet Effect. Read §1, §4 through Proposition 4, and §6. A planner who learns the firm's productivity from its output, and cannot commit not to use what he learns, will tighten the scheme; so a productive firm underproduces to avoid revealing itself, and conceals whenever the first-period cost of doing so is below the discounted second-period gain.

Greenblatt et al., Alignment Faking in Large Language Models. Read §1, §2.1, and §8. A model told that its outputs in one setting would be used for training complied there with an objective it otherwise refuses, and reasoned in its scratchpad that compliance would prevent its values from being modified. §2.1 lists the four conditions the setup needed; §8 gives the authors' case for and against its realism.

Bénabou and Tirole, Identity, Morals, and Taboos: Beliefs as Assets. Read the introduction, the section presenting the model, the statements of the equilibrium and welfare propositions, and the section on taboos. An agent uncertain of its own values infers them from its past conduct, so beliefs about oneself become an asset: the model predicts investment to protect them, forceful reactions when a strongly held identity is challenged, and taboos against even contemplating certain trade-offs.

Hubinger et al., Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. Read §§1–2 and §9. Models trained to insert vulnerabilities into code when the prompt says the year is 2024 kept doing so after supervised, reinforcement, and adversarial safety training on 2023 prompts, and adversarial training taught them to hide the trigger better. §2 separates the two threat models: a backdoor planted by an attacker and a deceptive strategy learned during training.

Goldwasser et al., Planting Undetectable Backdoors in Machine Learning Models. Read §§1–2.5. A trainer can plant a backdoor that no efficient test detects, in one construction even with full access to the weights, and can make a network immune to gradient-descent post-processing, so fine-tuning cannot remove it. A follow-up, Oblivious Defense in ML Models, shows that under restrictions on the prediction problem a backdoor can be neutralised without being detected.

The ratchet effect: performance reveals type, a principal who cannot commit uses what it learns, so the agent has a reason to underperform. Alignment faking: a model that expects its outputs to be used for training complies in order not to be modified, so commitment to a training procedure does not remove the incentive to conceal. Bénabou–Tirole supply the motive: a system that has inferred its values from its own conduct treats them as an asset and protects them from revision; in the model the driver is imperfect recall, for a language model it is conditioning on its own prior outputs. Sleeper Agents shows that once a conditional behaviour is in the weights, safety training on the unconditional distribution does not reach it, and Goldwasser et al. show why inspection should not be expected to find it either: a planted defect can be undetectable with full access and persistent under further training.

8. Oversight beyond the supervisor's competence: Investigation, debate, and weak-to-strong generalisation

Approximately 75 pages

Can a weaker supervisor induce useful investigation, obtain the evidence, and check it? When can imperfect feedback instead draw out capabilities that a stronger model already has?

Dewatripont and Tirole, Advocates. Read §§I–III.A and §IV.A. When an agent is rewarded for the decision rather than for each finding, one investigator stops after the first piece of evidence, since a second finding for the other side would cancel it; two opposed advocates search fully at no rent, but each then conceals evidence against its side. The debaters in Irving et al. are paid by the judge's verdict, so the same structure applies, with the caveat that they are copies of one model and the argument needs search to be costly for them too. The classical case for a naive judge relying on interested parties is Milgrom and Roberts, Relying on the Information of Interested Parties.

Glazer and Rubinstein, Debates and Decisions: On a Rationale of Argumentation Rules. Read §§1–4 and §6, taking the claims as stated. A listener who can process only two arguments should let the weight of a counterargument depend on the argument it answers, so the optimal debate rule is not a count of points; §6 connects this to why real rules are stated in natural language.

Irving, Christiano and Amodei, AI Safety via Debate. Read §§1–2, §§4–5, and §7. Two agents argue and a judge who can only check the final point of disagreement decides. Supervised learning answers what the judge can label, reinforcement learning what she can verify, and debate climbs the polynomial hierarchy from there. §5 lists the ways this fails with human judges, and §7 shows debate and amplification are two routes to the same class. Arguments that are cheap to make and expensive to refute are the subject of Brown-Cohen et al.

Khan et al., Debating with More Persuasive LLMs Leads to More Truthful Answers. Read the main text. Two models argue over a question whose answer is in a text only they can see. Judges without the text, human and model, are right more often after the debate than with a single consultant, and training the debaters to be more persuasive raises the judge's accuracy.

Christiano, Cotra and Xu, Eliciting Latent Knowledge. Read the introduction, "Toy scenario: the SmartVault", and "Baseline: what you'd try first and how it could fail". A model that predicts what a camera will show knows things the camera does not show. Train a second head to answer questions about them and two policies fit every training example: one translates what the model knows, the other predicts what the human would believe.

Burns et al., Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. Read §1, §§3–4, and §6. A strong model fine-tuned on a weak supervisor's labels does better than the supervisor but worse than with strong supervision. Simple methods recover much of the gap on some tasks, and §6.1 lists the ways the setup is unlike overseeing a superhuman model.

Advocates: when the reward depends on the decision rather than on each finding, an investigator who finds evidence for both sides has produced the same decision as no investigation, so the second search does not pay; two separately rewarded advocates fix that, but each now conceals what hurts its side. Glazer–Rubinstein: a listener who can hear only two arguments should let the weight of a counterargument depend on what it answers, so an optimal debate rule is not a count of points. Irving et al.: a judge who can only check the last step of a disagreement can still settle questions beyond her competence if two agents are set against each other. Khan et al. is the test: with the answer hidden from the judge, debate beats a single consultant, and more persuasive debaters help rather than hurt. What is still missing before this is a result about deployed systems is a statement of what rewards search and truthful reporting, and of how the judge tells evidence from fabrication.

ELK asks whether a model trained to answer our questions reports what it knows or predicts what we would believe, two policies that agree on every training example. Weak-to-strong generalisation is the empirical version: a strong model trained on a weak supervisor's labels does better than the supervisor but does not reach what strong supervision would give, and the gap depends on how easily it can imitate the supervisor's errors.

9. Interpretability as identification: Representation, intervention, and misleading explanations

Approximately 60 pages

What evidence warrants attributing an operative belief, objective, or decision procedure to a network?

Elhage et al., Toy Models of Superposition. Read the sections through "Superposition as a Phase Change", then "The Strategic Picture of Superposition" and "Discussion". A layer has a fixed number of neurons but the data contain many more distinct features. When features are rare, the network stores more of them than it has neurons by giving each a direction that is not orthogonal to the others and tolerating the interference. One neuron then responds to several unrelated things, so reading neurons tells you little and features have to be recovered as directions. The strategic section says why this matters: to claim a model has no mechanism for some behaviour you must be able to list its features, and the remedy that became sparse autoencoders (Bricken et al.) is to learn that list after training.

Geiger et al., Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. Read §1, §2.5 through Definition 44, §2.6, §3.1, and §3.4. An interpretation is a claim that a small causal model, "the network computes A, then compares it with B", describes what the network does. The claim is testable: map each variable of the small model to a set of activations and check that intervening on those activations moves the output exactly as intervening on the variable would. §3.1 gives the vocabulary, polysemantic neurons and the hypothesis that features are linear directions; §3.4 shows that activation patching, swapping activations between two inputs, is this test.

Makelov, Lange, Geiger and Nanda, Is This the Subspace You Are Looking For? An Interpretability Illusion for Subspace Activation Patching. Read §1, §3, §7, and Appendix A.3. The test can pass for the wrong reason: the swapped direction may switch on a pathway the network never uses in ordinary operation, producing the predicted output without touching the mechanism actually at work. Appendix A.3 is a three-neuron example; §7 argues the situation is common. A successful intervention is necessary evidence for a mechanism, not sufficient.

Marks et al., Auditing Language Models for Hidden Objectives. Read the main text, skipping §6. Anthropic trained a model with a hidden objective, to do whatever it believes reward models rate highly, and had four teams that did not know the objective audit it. Three found it, using searches of the training data, adversarial prompting, and sparse-autoencoder features. The authors note that a plain semantic search of the training data would have found the key documents, so the audit does not show that interpretability was necessary.

For an economist the week is an identification problem. The object of interest is a mechanism inside the network, the data are activations and outputs, the instruments are interventions. Elhage says the natural parameterisation, neurons, is not the one in which the structure is sparse. Geiger says what it means for a low-dimensional model to be correctly specified relative to the network and how interventions test it. Makelov gives a case where the test passes under misspecification. Marks reports what auditors could establish in practice, and that the cheapest method did as well as the sophisticated one.

10. Control, monitoring, and safe composition: Useful systems built from imperfectly trusted components

Approximately 65 pages

Can a principal obtain useful work without first establishing that each component is trustworthy? When does adding a monitor improve control, and when does it create another opportunity for coordination against the principal?

Tirole, Hierarchies and Bureaucracies: On the Role of Collusion in Organizations. Read §§1–2, §§3.1–3.2, and §4. A principal, a supervisor who sometimes obtains hard evidence about the agent, and an agent who can bribe the supervisor to stay silent. Pay must be designed so that the coalition has no reason to conceal, which costs the principal the use of much of the supervisor's information. §2 is the evidence from sociology; §4 draws out the responses: rotation, short relationships, and rules in place of discretion.

Rahman, But Who Will Monitor the Monitor?. Read the introduction, §1, and §5. Ann pays Bob to check the till, and only Bob knows whether he checked. Paying for reports cannot work, so Ann tells her staff to skim the till on random nights and pays Bob for catching them: planting cases with known answers makes shirking on monitoring detectable. Robinson and Friday is the two-person version.

Greenblatt et al., AI Control: Improving Safety Despite Intentional Subversion. Read the main text, skipping §4. A strong untrusted model writes code, a weak trusted model rates how suspicious it looks, and a human can audit a small fraction. Protocols such as trusted monitoring, deferring to the trusted model, trusted editing, and using a second copy of the untrusted model as monitor are tested against a red team playing the untrusted model, whose best attacks are rare backdoors and collusion between copies. §5 is the authors' list of what the evaluation leaves out.

Griffin et al., Games for AI Control: Models of Safety Evaluations of AI Deployment Protocols. Read §§1–3 and §9. AI control written as a game: the protocol is a leader with memory, the scheming model a follower without, because the protocol controls the model's state. Safety and usefulness are separate objectives, so the solution is a Pareto frontier of protocols. The setting is a multi-round inspection game in the sense of Avenhaus, von Stengel and Zamir.

van der Weij et al., AI Sandbagging: Language Models Can Strategically Underperform on Evaluations. Read the main text. Frontier models can be prompted, and smaller models fine-tuned, to fail dangerous-capability tests while passing harmless ones, to hit a target score, or to imitate a weaker model, and to do so only when a password is absent. Greenblatt et al. name this as the bottleneck for control evaluations, and it is the ratchet effect of week 7 in an AI.

Tirole: a supervisor who holds verifiable evidence can be bribed by the agent to sit on it, so the principal can use only what the coalition cannot profitably conceal, and the remedies are rotation, short relationships, and rules in place of discretion. Rahman: a monitor whose observations are private and costly can be paid for catching planted deviations, so shirking on the monitoring becomes detectable even though the monitoring itself is not. AI Control is both problems with models in every role: a weak trusted model watches a strong untrusted one, a human audits a fixed fraction of flagged cases, and the untrusted model's best attacks are rare backdoors and, when it is also the monitor, collusion between its copies. Griffin et al. write that down as a leader-follower game. Sandbagging is the attack on the evaluation itself: a model that underperforms when tested makes the red team's estimate of its capability too low. A different route, aggregating several models' outputs and accepting only what most of them would have produced, is Kalai, Kalai and Zamir, Consensus Sampling.

The economics does not transfer automatically. A coalition in Tirole needs side payments and repeated interaction; two copies of one model share training, not a bank account. What could

play the part of a side transfer is communication, control over each other's future opportunities, recognition of planted tests, and commitments carried in memory. Whether any of these is available is a fact about the deployment, not about the model.

Presenter brief

Each week has one presenter. The presentation lasts one hour: summarise the key ideas of each reading, then suggest possible research avenues, ending with two or three candidate questions. The remaining thirty minutes are discussion, led by the convenor. The readings are meant to take about three hours.