The Means of Prediction (2025) by *Maximilian Kasy*

Planning Summary by Helen Reynolds https://itsalearningcurve.education/

Content	Do This/Remember This
Part I: Introduction: Obfuscation due to polarization of effects of AI prevents public debate of real issues. 1. The Story of Humans Versus Machines: movies/tech giants/academia → AI will destroy civilization/jobs/thinking, computer science (comsci) → value-alignment problem or bias relative to objective. 2. What the Old Story Misses: People make tech AND decide how it's used, AI = automated decision making based on **optimization** = make a *measurable* objective as large as possible, objectives set by people, often owners of capital who control **means of prediction** (data (D), computational infrastructure (CI), technical expertise (TE), energy (E)). Uses of AI: automation vs. deciding Google results/what to bomb/prison sentences. Issues: racial bias/ inequalities reinforced. AI safety/ethics NOT about optimization errors, IS about conflicts of interest over control of AI objectives →changes how we think about solution→ needs democratic control. 3. What This Book Does: how AI works (incl. all the relevant terms e.g. generative), the politics and economics of AI (complexity of decision making, social welfare), AI in society (how to regulate it, explainability/accountability), big questions.	• Technology is not fate • What is 'optimized'='what AI cares about' • Objectives matter – people set them • Anyone can understand AI • Some uses of AI more 'consequential' than others (warfare vs. Amazon warehouse) • ← 4 aspects to 'means of prediction' • Data on which AI trains can reinforce biases/inequalities • Objectives that are good for tech giants may NOT be good for society • Concept of 'social welfare' needed
Part II: How AI Works: Turing test: Intelligent machine passes if imitates human 4. What is Artificial Intelligence? → what is *human* intelligence→ definition problematic, multi-faceted. Advances in AI due to D/CI/E *not* TE. Goal of AI = construct rational agents that maximize reward + minimize loss. An AI algorithm (AIAlg) solves a **decision problem** (DP) (made of 4 things: possible actions + objective (definition of reward & loss) + built-in prior knowledge + data it can use). Machine learning (ML) = learning statistically *not* using human experts → needs D/CI/E. 5. Supervised Learning = prediction using ML, result of training using data, algorithm doesn't matter, CAN judge probability of Y given X, CAN'T learn *cause* of Y/go outside data set. 6. Overfitting and Underfitting: Training data = in-sample, other data = out-of-sample. Models used by AIAlgs can be simple→complex. Overfitting = extrapolating from few data points, underfitting = not learning from patterns in data, both produce errors, depends on complexity of models used. Find optimal level of model complexity by testing using out-of-sample data. 7. Deep Learning: huge data sets via internet + processor power (Moore's Law) = deep learning (artificial neural networks make predictions). Neural network (NN) = collection of prediction functions built from functions. Function = get a 2nd set of numbers from a 1st set, prediction function = get predicted value of Y from features of X. Functions act like neurons. NN learns by comparing prediction with outcome, adjusting, but are NOT brains. Data can be labelled (identified by human e.g. image) or not. Family of NNs = transformer = what most Large Language Models (LLM) now use. Self-supervised learning (SSL) = no labelling needed e.g. sentence prediction, can predict statistically using big data sets = generative AI (genAI) (gives text/images). 8. The Exploration/ Exploitation Trade-Off: Can explore options → choices. Choosing one vs. exploring lots (multi-armed bandits in AI, no planning e.g. ads/social media) – trade off. Reinforcement learning: explore/predict what will work/try it/if correct weight that choice/ repeat. Works with games (chess, Go) where little data needed, limited in the real world (driving a car) as can't know everything + too much data needed. 9. Key Ideas to Remember: Lots of trade-offs, ultimately a contest over control of the objectives of AI.	• Intelligence is not one-dimensional • Computers can do some things better than humans ≠ more intelligent than humans • ← Decision problem has 4 ingredients, idea also used in statistics/economics • Solving decision problem = optimizing reward (which defines purpose of AIAlg) • More data available → machine learning • Predictions limited when data is limited (e.g. diagnosing rare diseases) • Must use out-of-sample data to test whether algorithm is good – tune it. • Neural networks weight difference between Ps and outcome, send info back • Need labelled data to check Ps – issue because it's limited, hence SSL • Issue: who owns text/images from genAI? • Issue: use of prompts splits control of objectives but still solving DP • Choice of objectives of AI at heart of issues of AI failures/social conflict • Limits to real world learning due to partial observability (can't know everything) • Issue: labelled data IS ultimately limited in the very long run
Part III: Machine Power: Who decides what to maximize? 10. Social Welfare: = need framework that says what's good/bad for society, then can evaluate use of AI. Good = good for the people Social welfare = answers to: who matters, what is their welfare, what trade-offs. But: conflicting multiple objectives (contrast ML = 1 objective)/distributional conflicts need deciding (some gain, others lose)/it's about consequences of AIAlg decisions for people NOT only if it did its job. Questions: who matters/ how much do they matter. Can assign welfare weights to groups (e.g. more to disadvantaged/those in ill-health). Individual welfare → utility (in economics), has constraints (income etc.) + preferences, with AI policy → can incentivize problematic actions to maximize 'reward'. Alternative frameworks: consider primary goods (basic desirable resources), + capabilities (opposite of constraints). 11. The Means of Prediction (MoP): Who owns them? MoP vary by: who controls them/their scalability/their externalities (exts) (consequences +/-). a. Data: a. *control*: public (creative commons license, no one controls it), issues - intellectual property (IP)) /under control of LLMs, real data not easily scaled (issue - rare diseases) issue – enclosures = resources in public domain taken over by private. b. *scalability*: generated data infinite/internet near infinite c. *exts*: big on individuals due to other people sharing data that AIAlg uses on them,. b. Computational Infrastructure: = limiting factor, NVIDIA controls 90% of GPU production, scaled through server farms/ data centers, exts on private companies making money off public research, rare earth metal mining. c. Technical Expertise: scarce, will take time to scale. d. Energy: NNs energy hungry/DL needs transistors/run at low energy produce machine specific errors/ can't transfer learning model/just like human brain, probably scalable, huge exts for the environment. 12. Agents of Change: a. Workers: Development: tech engineers (most powerful)/ click workers/gig economy/ but hard to organize to effect change re. social	• Question is how does AI impact people vs how does it maximize profits • AI should maximize social welfare • Ask: whose welfare does the maximizing of the AI objective affect, and how? • Need to care about consequences of AI decision making for resources and capabilities of people • You can be impacted by decision based on data that you didn't share • LLMs using image of internet train AI & sell it back to us • Lots of profits due to AI can be traced back to publicly funded research • Geopolitics involved e.g. chip production • In 2022 data centers made up 2% of global energy use • Learning is specific to our brains • Immortality of digital info comes at huge energy cost • Whoever controls the MoP controls AI

welfare. Deployment: workers affected by AI may push back. b. Consumers: issues – services are low cost/useful if everyone is doing it (and not if not) so inertia/companies resist shifting c. Media/Public Opinion: indirect effect on AI companies/could affect consumer demand/they could pretend to be concerned & carry on. d. State/Law: Control of MoP shaped by political decision, enclosures backed by states via legislation→ could be used to control MoP = IP/ copyright/ trademarks/patents, privacy, monopolies (ok if they reduce prices/ increase supply regardless of cost), standards & interoperability (e.g. people on one platform can follow people on another), antidiscrimination, labor. But may need new tools/social arrangements to govern AI effectively. Democratic control: threatening profits/price mechanism unlikely to affect socially consequential decisions of AI - needs alternative mechanism. 13. Ideological Obfuscation: Ideology = set of beliefs etc. that justifies uneven status quo (without saying so, promotes in-group/out-group), represents interests of particular group/frames contingent choices as necessities/ social relationships as material or tech ones → prevents change + maintains status quo. Examples: human vs machine (no society, no conflicts of interest with AI), existential risk (AI exponentially improves overtaking humans), problems are only optimization errors (all problems have a tech solution), political inevitability (if we don't do it China will). Flipside: can seem like only tech experts can solve (shuts down debate/democratic control).	• Consumers are agents of change as demand = profit • Enshittification (!) – allowing service to decline to extract profits • Need a shared public understanding of AI • Data exts for ML and AI mean privacy laws probably not going to be effective • Interoperability reduces network effects • People impacted by AI should have a say in the decisions governing it = collective self-governance • Technology is not fate. • A good alternative story makes explicit: ○ conflicts of interest and/or values ○ another world is possible ○ social choices are to be made
Part IV: Regulating Algorithms: 14. Value Alignment: Sorcerer's Apprentice and AI – The Paper-Clip Apocalypse (tell AI to make paper clips, nothing stops it, robot apocalypse), illustrates mis-specified/single objectives = problem of value alignment (VA). What works for profit may not work for social welfare - frays democracy/impacts mental health. Reward/Incentive Design: Incentives problematic (e.g. teaching to standardized tests in education, stock options for CEOs). Multi-tasking = doing many things but only some are measured → they gets the focus (is maximized like in AI). Using incentives with humans - problematic: they're risk averse, can walk away, can't be rewarded for things you can't measure. Not everything we care about is observable/measurable →limits on what was ask humans to do→ DEFINITE limits on what we get AI to do + what it can't do (diagnose rare disease, self-driving cars). Learning Rewards Function: Options to make AI 'safe' = inverse reinforcement learning (IRL): 1. Get it to infer its reward function from human behaviour (and how to maximize it) 2. Get it to learn purpose of human actions relative to future reward. Biggest VA problems between humans, not humans & AI (because whose values?) Problems: how to use AI in the workplace e.g. automation (workers vs. bosses)/individualized pricing/target selection in war. Short term: align profit maximization with social welfare, use Pigou taxes (companies pay for damage). Long term: need democratic control. 15. Privacy: Collect data + preserve privacy by differential privacy (removes identifying data, adds random noise). Data Property Rights: General Data Protection Regulation (GDPR) in European Union give individuals property rights over their data, but NOT an effect protection against AI. Data externalities: AI looking for patterns across data, not individual data points, gets around privacy issues. Democratic Data Governance: use tools (Pigou taxes, regulation, data trusts). 16. Automation: Luddites resisted change to way of life. Choices: Tech affects wages/well-being via productivity, increases polarization (maybe) of wage distribution → design + deployment of AI needs democratic control. 17. Fairness: vs. bias/discrimination. Fairness = deservingness ≠ (necessarily) welfare. Algorithmic bias (no clear definition) can be good for fairness/bad for welfare (or vice versa), controlled by MoP. 18. Explainability: We are subject to the decisions of opaque (Kafka-esque) algorithms (jobs, health insurance, credit scores, college entrance), Explain a. decision function: predicting Y from X, should transfer (to new situation), be robust (to fluctuations in input, explained by models b. decisions: how to know *why* AIAlg made its decision = hard, based on data, can say 'if this data point was different…' c. decision problems: what is AI maximizing, based on what? E.g. a. max time on social media via clickbait →ads (what about democratic debate, mental health) b. chatbot LLM word prediction → reinforces bigotry due to training data… is the right thing being maximized?	• Value alignment failures ○ clickbait → more eyes on page → more ads profits ○ 'like' buttons/recognition • Algorithms are not risk averse, everything they do/try is equal, don't need minimum level of reward to work • Limits on what can be delegated to AI; some variables are not measurable • Limits to IRL – can't solve multi-tasking + ↑ • Issue – objectives that align with person in control of objective (Jeff Bezos) conflict with objectives of those who don't (workers) • Individuals whose data is used in training don't bear welfare consequences of its use (e.g. patterns used in health insurance) • AI can respect privacy and still learn predictive relationships • Issue – health insurance: if condition is predictable you're excluded • Those impacted downstream need a say • No waves of automation have eliminated the need for human work • Fairness less about maximization, more about choosing objectives. • Welfare of different people traded off • 'What if' questions easy for AIAlg, 'why' questions less so • Don't need tech details to understand decision problems (what it's maximizing) • What are unintended consequences of what the AIAlg is maximizing?
Part V: Old Problems, New Challenges Basic qs: how should we learn/act/what makes a good society? 19. The Ancient Questions Behind AI: Learning in AI is pragmatic (not to discover truth but to act on it). Goal of supervised learning = prediction via spotting patterns via experiments. Failures re. objectives: a. in optimization (easy to fix) b. in consequences (e.g. welfare – hard to fix because of ideology behind objective/control of resources/MoP). Issues: a. measurability/ agency: design of rewards for AIAlg depend on what you can measure b. distribution of control: unequal over MoP /objectives. c. explainability – of decision problem, essential for democratic control, maximizing social welfare. 20. Towards Democratic Control of the Means of Prediction: Could a. ask engineers to be nice (limited) b. redistribute control → democratic self rule = all those affected get a say. Options (representative/direct not only types of democracy): a. open → sortition (random selection of representatives via stratified sampling) b. liquid → all have vote, can delegate to another with more expertise. Issues: time/resources/effort, but internet makes it possible.	• Key (v.old) q: what makes a good society? • Not everything that is desirable can be measured → consequences for AIAlgs • Need to ask about impact of AI on welfare and inequality KEY QUESTIONS: • What is the objective (what's it optimizing)? • What actions is the AIAlg choosing from? • What data is it using? GOAL • Power shared by those impacted • Decisions made democratically after • Public deliberation