80-306/80-606 | **Decision Theory** | Syllabus



Seminar Time: Tuesdays & Thursdays, 11:00 AM–12:20 PM
Seminar Location: TBD
Instructor: Aydin Mohseni
Instructor Email: amohseni@cmu.edu
Office Location: Baker Hall 145-F
Office Hours: by appointment

Course Description

How should rational agents choose under uncertainty? This course introduces the main frameworks for rational choice—including expected utility theory, subjective probability, and their axiomatic foundations—along with key debates between Causal, Evidential, and Updateless decision theories. We examine paradoxes and alternative approaches to uncertainty and risk. Finally, we apply decision-theoretic tools to puzzles in AI alignment and the foundations of agency.

Learning Objectives

By the end of the course, you will be able to:

- **Explain** the axiomatic foundations of expected utility theory (von Neumann–Morgenstern, Savage, Anscombe-Aumann, Jeffrey-Bolker) and the role each axiom plays in the representation theorems.
- **Compare** Causal, Evidential, and Updateless decision theories (CDT, EDT, UDTs), articulating the core commitments of each and the cases that motivate them.
- **Evaluate** proposed solutions to classic paradoxes (Allais, Ellsberg, Newcomb, etc.), assessing their costs and benefits within different theoretical frameworks.
- **Apply** decision-theoretic concepts to problems in AI alignment and the foundations of agency.

How Objectives Are Assessed

Learning Objective	**Assessment**
Explain axiomatic foundations (vNM, Savage, A-A, J-B)	Problem sets; exams
Compare CDT, EDT, UDTs	Problem sets; exams
Evaluate solutions to paradoxes	Problem sets; exams; participation
Apply DT to AI alignment & agent foundations	Problem sets; exams

Course Texts

All readings will be provided on the course Canvas site. No textbook purchases are required.

Course Structure

- Each week addresses a topic in decision theory.
- Weekly readings will be posted on Canvas. Reading load is approximately 10–30 pages per week. (The material is math-heavy, so pace yourself accordingly.)
- Weekly problem sets are due before the first class meeting of the following week.
- Graduate students (80-606) complete more challenging problem sets and exam questions.

Problem Sets

Weekly problem sets will include a mix of formal exercises (proofs, calculations) and conceptual questions. Graduate-level problem sets include additional, more challenging

problems. Your lowest problem set score will be dropped.

Exams

- **Midterm Exam:** In-class, Thursday, February 26. Covers Modules 0–1 (Weeks 1–7).
- **Final Exam:** Scheduled during finals period (April 27–May 5). Covers Modules 2–4 (Weeks 8–14).

Graduate-level exams include additional questions.

Grading

Assignment	Weight
Participation	20%
Problem Sets	40%
Midterm Exam	20%
Final Exam	20%

Graduate students (80-606) complete more challenging problem sets and exams; grading standards are calibrated accordingly.

Late Work Policy

Late work loses 10% per day. For example, work submitted two days late can earn at most 80%. Exceptions are granted only if:

1. Official university documentation is provided, or
2. You contact me at least one week in advance and arrange an alternative.

Participation and Attendance

Participation is earned by attending, engaging in seminar, and certain assignments. You may miss two sessions without explanation and without penalty. (No explanation is needed *or wanted* for such absences.) Additional absences require official documentation (e.g., a doctor's note); otherwise, your participation grade will be reduced. No exceptions.

AI and LLM Use

You may use large language models (e.g., ChatGPT, Claude) to:

- Clarify concepts or check your understanding.
- Explore alternative approaches to problems.

You may **not**:

- Submit LLM-generated solutions as your own work.
- Use LLMs during exams.

Violations will be treated as academic integrity issues.

Academic Integrity

I trust you to be adults. This course will follow the University policy on academic integrity. If you are unsure about the University's policies and definitions regarding academic integrity and plagiarism see:
www.cmu.edu/policies/student-and-student-life/academic-integrity

Disability Accommodations

If you have any questions or concerns about disability accommodations, don't hesitate to speak with me—I am very happy to help out.

In general, if you have a disability for which you may be requesting an accommodation, you will want to contact the Disability Resources Center (www.cmu.edu/disability-resources). The Disability Resources Center will verify your disability and ensure you get appropriate accommodations for this course.

Accessibility

Lecture slides will be posted to Canvas before each class meeting. If you require materials in an alternative format, please contact me.

How to Succeed in This Course

My advice for succeeding in this course is as follows.

- Attend and participate in seminar.
- Complete the weekly readings *before* class.
- Start problem sets early—the material is cumulative and benefits from time to digest.
- Come to office hours with questions.

If you do all these things, you should do very well in the course!

Office Hours and Support

I will hold office hours at set times during the week. You are enthusiastically encouraged to come to these hours!

Their purpose is for you to ask any questions you may have about the course content, assignments; to discuss any difficulties, concerns, or questions regarding the course or your performance; and just to enjoy chatting about decision theory!

You can also contact me via email. You should allow for 48 business hours for a response.

Course Schedule and Readings (Tentative)

Dates refer to calendar weeks; all readings and materials will be available on Canvas. Reading load is approximately 10–30 pages per week.

Module 0: Foundations

Learning goal: Characterize different conceptions of rationality; represent preferences, outcomes, and prospects formally.

Week 1 (Jan 13–17): *Rationality: instrumental, epistemic, bounded*

- Sen, Amartya (1977), "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." (selected excerpts, 20 pp.)

Week 2 (Jan 20–24): *Preference, utility, and probability basics*

- Friedman, Milton & Savage, Leonard J. (1948), "The Utility Analysis of Choices Involving Risk." (selected excerpts, 25 pp.)
- Skyrms, Brian (1990), "Coherence" (from *The Dynamics of Rational Deliberation*). (selected excerpts, 15 pp.)
- *Note: Monday, January 19 is MLK Day (no classes).*

Module 1: Axiomatic Frameworks

Learning goal: State and interpret the axioms of vNM, Savage, Anscombe-Aumann, and Jeffrey-Bolker; explain how each yields expected utility.

Week 3 (Jan 27–31): *von Neumann–Morgenstern: lotteries & axioms*

- Fishburn, Peter C. (1986), "Subjective Expected Utility: A Review of Normative Theories" [vNM sections]. (selected excerpts, 15 pp.)
- Gustafsson, Johan E. (2022), "Money-Pump Arguments." (selected excerpts, 20 pp.)

Week 4 (Feb 3–7): *Savage: states, acts, subjective probability*

- Savage, Leonard J. (1954/1972), *The Foundations of Statistics* [selected excerpts]. (selected excerpts, 25 pp.)

Week 5 (Feb 10–14): *Anscombe-Aumann: horse lotteries*

- Fishburn (1986), "Subjective Expected Utility" [Anscombe-Aumann sections]. (selected excerpts, 15 pp.)

Week 6 (Feb 17–21): *Jeffrey-Bolker: news-value, evidential structure*

- Fishburn (1986), "Subjective Expected Utility" [Jeffrey-Bolker sections]. (selected excerpts, 15 pp.)

Module 2: Paradoxes & Alternatives

Learning goal: Analyze Allais, Ellsberg, and related paradoxes; evaluate alternative approaches (risk-weighted utility, imprecise probability, etc.).

Week 7 (Feb 24–28): *Allais & Ellsberg paradoxes*

- Allais, Maurice (1953), "Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine." (selected excerpts, 20 pp.)
- Ellsberg, Daniel (1961), "Risk, Ambiguity, and the Savage Axioms." (selected excerpts, 15 pp.)
- **Midterm Exam: Thursday, February 26.**

March 1–7: Spring Break (no classes)

Week 8 (Mar 10–14): *Risk-weighted utility, prospect theory*

- Wakker, Peter P. & Tversky, Amos (1993), "An Axiomatization of Cumulative Prospect Theory." (selected excerpts, 20 pp.)
- Buchak, Lara (2017), "Précis of *Risk and Rationality*." (selected excerpts, 10 pp.)

Week 9 (Mar 17–21): *Imprecise probability, ambiguity aversion*

- Wheeler, Gregory (2022), "A Gentle Approach to Imprecise Probability." (selected excerpts, 30 pp.)

Module 3: CDT, EDT, UDTs

Learning goal: Compare causal, evidential, and updateless decision theories; apply each to Newcomb-like problems.

Week 10 (Mar 24–28): *Newcomb's problem; CDT vs. EDT*

- Lewis, David (1981), "Causal Decision Theory" [selected excerpts]. (selected excerpts, 15 pp.)
- Jeffrey, Richard C. (1983), *The Logic of Decision* [selected excerpts]. (selected excerpts, 15 pp.)
- Easwaran, Kenny (2018), "A Classification of Newcomb Problems and Decision Theories." (selected excerpts, 20 pp.)

Week 11 (Mar 31–Apr 4): *Updateless decision theories (UDT, FDT, LDT)*

- Yudkowsky, Eliezer & Soares, Nate (2018), "Functional Decision Theory: A New Theory of Rationality." (selected excerpts, 25 pp.)

- MacAskill, William, et al. (2023), "Characterising Decision Theories with Mechanised Causal Graphs." (selected excerpts, 20 pp.)

Module 4: AI Alignment & Agent Foundations

Learning goal: Apply decision-theoretic tools to instrumental convergence, dangers of optimization, and the foundations of agency.

Week 12 (Apr 7–11): *Decision theory & AI alignment*

- Bostrom, Nick (2011), "The Superintelligent Will." (selected excerpts, 15 pp.)
- Turner, Alexander Matt, et al. (2023), "Optimal Policies Tend To Seek Power." (selected excerpts, 20 pp.)
- *Note: April 8–10 is Spring Carnival.*

Week 13 (Apr 14–18): *Instrumental convergence*

- Soares, Nate & Fallenstein, Benja (2016), "Formalizing Convergent Goals." (selected excerpts, 15 pp.)
- Turner, Alexander Matt & Tadepalli, Prasad (2022), "Parametrically Retargetable Decision-Makers Tend To Seek Power." (selected excerpts, 15 pp.)
- Bales, Adam (2023), "Will AI Avoid Exploitation? Artificial General Intelligence and Expected Utility Theory." (selected excerpts, 20 pp.)

Week 14 (Apr 21): *Dangers of optimization; course synthesis*

- Zhuang, Simon & Hadfield-Menell, Dylan (2021), "Consequences of Misaligned AI." (selected excerpts, 15 pp.)
- *Note: Last day of classes is Monday, April 20; we meet Tuesday, April 21.*

Final Exam: Scheduled during finals period (April 27–May 5). Covers Modules 2–4 (Weeks 7–14).