# Ethical challenges for RL/AI (11h15-12h30)

- Why? → we (engineers/scientists) drive the field
  → we (engineers/scientists) have responsibilities

- Organization of Ethics/AI session
  - intro of topics (2-5min per topic) at 11h15 - 11h40
  - split in groups of 5-10 students, discuss until 12h05
  - groups report back
    → 3 Minutes + 3 Minutes discussion per report.

The Bonus: Easier to get a letter of recommendation
if you report back your thoughts in plenum

# Overview (planned Version 2024)

1. Intro and RL1: Reinforcement Learning for Bandit problems
2. RL1: Bellman Equation and SARSA
3. RL 2a: Markov Processes and Convergence of SARSA/first python code (Brea)
4. RL2b: Q-Learning, n-step TD learning, continuous space, eligibility traces
5. RL3: TD-learning and Function approximation
6. RL 4: Policy gradient algorithms
7. RL 5: From Policy gradient to Actor-Critic: eligibility traces again

8. Deep RL1: Applications of Model-free Deep RL (Brea)
9. Deep RL 2 Applications of Model-based Deep RL (Brea)
10a Caglar: RL3 with Human Feedback, (Caglar Gulcehre)
**10b. Deep RL4: Ethics, AI, and RL**
11. RL, Dopamine, and the Brain
12. From Brain-style computing to neuromorphic hardware
13. Surprise and Novelty in RL
14. Curiosity-driven Exploration (Alireza)

Basic RL

Deep RL

Interdisciplinary RL

# Ethics for RL/AI

- Why? → we (engineers/scientists) drive the field
  → we (engineers/scientists) have responsibilities

**The need of rules (also known as regulation)**

- Each of us is an individual.
  → rights
  → freedom

- Human society requires collaboration, coordination, and trust
  → this requires rule
  → don’t be afraid of rules

PhD student in my group:

*“Rules are always bad. Every adult should know what’s good for them, and if they make bad choices, it’s their problem.”*

A friend of mine says:

*“We need rules to protect the weak and poor”*

My stance:

*“Rules are necessary to make a society work”*

# The need of rules/regulation: Examples

- free market (market place)

- traffic

- banking

- chemistry

- nuclear physics

- pharma

Fun Quiz:

Why are banks regulated?

[ ] To protect the deposit of small clients

[ ] To protect against collapse of society

[ ] To protect the business and capital of the banks

Reading: Wikipedia, 2007-2008 financial crisis

Why are casinos regulated?

[ ] To protect the hard-earned money of clients

[ ] To protect against costs for society

[ ] To protect the business and capital of the casino

[ ] To protect against organized crime

Reading: Wikipedia, gaming control board

# Detour: RL and Discounted Future

Return = accumulated **discounted** future rewards

Makes explicit:

- short-term gains more important than long-term effects

→ True for individuals (at some point you die anyway)
→ Not true for societies: how can we account for future generations?
→ Another example of ‘externalization of costs’, but in time

# The need of rules/regulation

In a meeting with CEOs of  pharma, banks, nutrition, tech EPFL/ETHZ presidents, and Swiss politicians, somebody said: *‘CS/Social Networks/AI need regulation’.*

CEOs of pharma, banks, nutrition nodded.

*For them the statement was ‘obviously true’.*

# The need of rules/regulation

**CS/Social Networks/AI have qualitatively changed status during the last 10 years**

- AI used to be fundamental research,
  it now impacts society at large
- Learning rules in broadly used applications are powerful tools
- Need to balance advantages of individuals against societal costs

similar to chemistry/pharma/nuclear physics

# Ethics challenges for RL/AI: Topics

## Concepts and Problems (discussion topics)

- Regulation by public policy makers (Anja)
- Individualized ChatGPT “adapted to the values of the users” (Lucas)
- Addiction (to social media) and RL (Lazar)
- AI and chemical weapons (Michael)
- Ethical goals via RL policies (Max)
- Finetuning of LLM (Ariane)
- RAG (Sophia)

HOW?

1) TA s will introduce topic.
2) You join a TA (in corner of room)
   to mark your interest for that topic.
3) Discussion in small groups: how to regulate?
4) Report back to plenum (1-2 slides possible, but not necessary)

# Ethical challenges for RL/AI (11h15-12h30)

- Why? → we (engineers/scientists) drive the field
  → we (engineers/scientists) have responsibilities

- Organization of Ethics/AI session
  - intro of topics (2-5min per topic) at 11h15 - 11h40
  - **split in groups of 5-10 students, discuss until 12h05**
  - groups report back
    → 3 Minutes + 3 Minutes discussion per report.

The Bonus: Easier to get a letter of recommendation
if you report back your thoughts in plenum