Computational Applications to Policy and Strategy (CAPS) – Lecture 1

Introduction to Decion Rules and Learning from Interaction

L.Klenner, H. Fung, C. Combs

Last Updated: March 05, 2019

**1 Introduction**

In this course, we consider the problem of learning how to act, through experience and without an explicit teacher. We address this problem from the perspective of different learning mechanisms, those of human-designed autonomous agents and human actors.

When we talk about learning how to act, we are concerned with the following process: In a given situation, select and take the appropriate action for solving a specified task.

Therefore, in the most general terms, we want to ask: How can we learn to make the right decision, given a set of requirements that this decision needs to satisfy?

More specifically, we want to ask:

* How can autonomous agents be designed to make the right decision in a given situation and what requirements do we have of the agents’ decision making and the *design* that enables it?
* How do teams of human experts learn to make the right decision in a given situation and what requirements do we have of the experts’ decision making and the *training* that enables it?
* How can we guide teams encompassing autonomous agents and human experts to make the right decision in a situation that requires jointly coordinated decision making and what requirements do we have of this *coordination*?

The focus of the course will be on the third aspect, coordination. Why might this aspect be a critical component in enabling organizations to achieve their strategic and operational targets? To provide some intuition, consider the following examples:

* In a quantitative trading firm, an agent designed to route orders in the equity market erroneously takes hundreds of positions in multiple stocks, streaming millions of orders into the market over the time frame of less than an hour and creating unrecoverable losses for the firm. The firm’s engineers suspect a failure in the algorithm, the firm’s traders a failure in the target specification of the algorithm and the firm’s managers a failure in the infrastructure on which the agent trades.
* In a semi-autonomous sustainment convoy, an autonomous supply truck designed to follow a human-driven lead vehicle veers out of its position and stops, bringing the convoy to a halt. After a number of minutes, the truck resumes its position, but the humans in the lead vehicle decide to override the truck’s autonomy through a driver taken from their team. This decision reduces the team’s overall situational awareness as one additional team member needs to focus on driving, which prolongs the team’s response time to an ambush targeting them on their route.

In each of these examples, the aggregation of autonomous agents and human experts is expected to yield critical performance advantages, such as speed and scale in the execution of orders, or mission completion with a reduced number of human lives at stake. However, in each example, the aggregate decision making of the autonomous agents and the human experts is *misaligned*. This can be misalignment within the experts based on a lack of understanding about the agent’s decision making, such as in the example of the trading firm, or misalignment between the experts and the agent, such as in the example of the sustainment convoy. To mitigate misalignment and secure performance advantages, we require coordination mechanisms for aggregate teams of autonomous agents and human experts.

*Misalignment comes from decision making based on different optima comes from different learning mechanisms and specifications, but we can still conceptualize based on shared architecture.*

and between experts and agents.

decision making of the aggregate team of human experts and autonomous agents is misaligned. In the

In **Reinforcement Learning**, an autonomous **agent** learns to take **optimal** actions based on **goal-oriented** **interaction** with its **environment**. While there are specific formalizations of decision making processes in which reinforcement learning algorithms are grounded, the concept of goal-orineted learning from interaction is broad and can encompass a range of decision problems. Let’s consider some examples of such problems:

* A Kiowa helicopter flies through the Pech River Valley in Afghanistan, circling multiple times over long stretches of rock formations, occasionally attempting to decend, before landing on an empty patch of river bed near a ground unit whose casulty it needs to evacuate.
* Bob is looking for a letter opener, opening all the drawers of his desk before finding the letter opener under a staple of books on the ground, which he kicked over while opening the drawers.
* An adaptive controller adjusts parameters of a radio-frequency jammer in real time. The controller optimizes the jammer’s performance/noise trade-off on the basis of specified time patterns without sticking strictly to the patterns originally suggested by the engineers.
* An anthropologist in an Afghan village consults various members of the local shura to identify who among the shura’s members has the authority to support a developement project and how the project should best be prusued to benefit the village.

Each of these decision problems can be expressed as the process of goal-oriented learning from interaction. Consider the first example:

* The Kiowa helicopter has the goal to land near the ground unit to evacuate the casualty. It interacts with its environment by selecting different flight paths before learning which of these paths allows it to land safely and in close proximity of the ground unit.

While we can express the above example through goal, interaction and learning, we need to develop our understanding of how these concepts come togehter to understand an agent’s decision making, e.g. *how* the Kiowa can select between different flight paths before identifying the optimal one. Let’s unpack the notion of goal-directed learning from interaction through a set of simple questions:

* What constitutes a goal and how do we know that we have achieved it?
* What constitutes an interaction with the environment and how do these interactions change over time?
* What does it mean that an agent has learned to solve a given task?

**Exercise 1**. Learning from interaction is one way to conceptualize how agents in the above examples can achieve their goal. Briefly outline a similar concept that might enable agents to achieve their goal.

**2 Conceptualizing learning from interaction**

In answering the above questions, we conceptualize learning from interaction at a high level, which provides us with necessary intuition before moving onto the technical parts of reinforcement learning.

**2.1 Goals and measures of progress**

A goal is the objective that an agent aims to accomplish in its learning process. At the beginning of this process, the agent has **uncertainty** about which of the possible decision paths available to it will lead it to optimally achieving its goal. The agent carries out a sample of the decision paths and learns by discriminating between those paths that lead it *towards* its goal and those that did not, then repeats the same process with an updated sample. Therefore, learning from interaction requires **measures of progress** relevant to the agent’s goal. These measures allow the agent to capture *how* it is performing not just whether or not it achieved its goal.

One of the reasons why measures of progress are needed for learning is that decision paths might be deep, with a high **branching factor**, meaning they contain a large number of individual decision points and each decision point contains a high number of additional sub-paths. Pursuing each decision path to its end to learn whether or not it accomplishes the goal can be prohibitively expensive in terms of the required learning time and resources. With measures of progress, the agent can determine for a given set of decision points whether the associated decision path is likely to accomplish the goal or not. This enables the agent to reduce its learning time and resources.

Another reason is that measures of progress enable the agent to differentiate between decision paths that all accomplish the agent’s goal, given that one of these paths might lead to faster progress than the others.

* + 1. **Trade-offs of measuring progress**

Measures of progress give rise to a number of **trade-offs** that are important considerations in designing and evaluating learning processes:

* The goal of the learning process needs to be fully decomposable into a measure of progress. Otherwise, a gap exists between learning to select an optimal decision path based on the measure of progress and learning to accomplish the goal, which would complicate the agent’s learning.
* The optimal decision path to achieving a goal might not be correlated with making initial progress. There is no reason to assume that taking the decision path that initially allows for making the most progress is the optimal path for accomplishing the agent’s goal. Therefore, we need both **short-term** and **long-term** measures of progress of a decision. While a decision point might be inferior in the short-run, the associated decision path might be optimal in the long-run. Hence, in selecting their actions, agents need to balance both measures of progress.

How should we resolve these trade-offs? The first trade-off can be resolved by way of assumption. We can assume that all goals can be fully decomposed into a measure of progress – and indeed, we do so in reinforcement learning.

The second trade-off is more complex to resolve and leads to different design requirements of agents, e.g. to obtain sound long-term measures of progress for an agent’s action.

**Exercise 2.1.1.1.** Why might it be problematic to assume that goals can be fully decomposed into a measure of progress? Explain briefly by discussing an example of a goal and a relevant measure of progress.

**Exercise 2.1.1.2** For a given action, why might it be necessary to estimate long-term progress of that action, whereas short-term progress is directly observable?

**Exercise 2.1.1.3.** For a given action made towards a goal that requires multiple actions, how would you estimate that action’s long-term progress? Provide a brief outline for a specific example of goal and action.

**2.1.2 Greedy action selection**

The simplest resolution to the second trade-off, and one that does not require obtaining long-term measures of progress, is to specify the agent’s level of **greediness**. We call an agent greedy, if it always selects actions that maximize its short-term progress.

Given an action *,* a measure of progress of at time step *t*, , we write the greedy action selection method as

, (1)

where denotes the action for which the expression that follows is maximized (with ties broken arbitrarily). Greedy action selection always maximizes short-run progress but spends no time on sampling inferior actions to see if they might lead to better long-run progress.

A simple alternative is to behave greedily most of the time, but occasionally, with a small probability, instead select randomly from among all the available actions with equal probability, independently of the associated short-run progress.

**Exercise 4.** In -greedy action selection, for the case of two actions and , what is the probability that the greedy action is selected?

**2.1.3 -greedy action selection in the example of the Kiowa evacuation**

To practice our understanding, consider the decision tree of the Kiowa helicopter evacuating a casualty in the Pech River Valley shown in Figure 1.

We assume that the Kiowa is an -greedy agent with , a starting position (*a*) and three decision points (*b,* *c, d*).

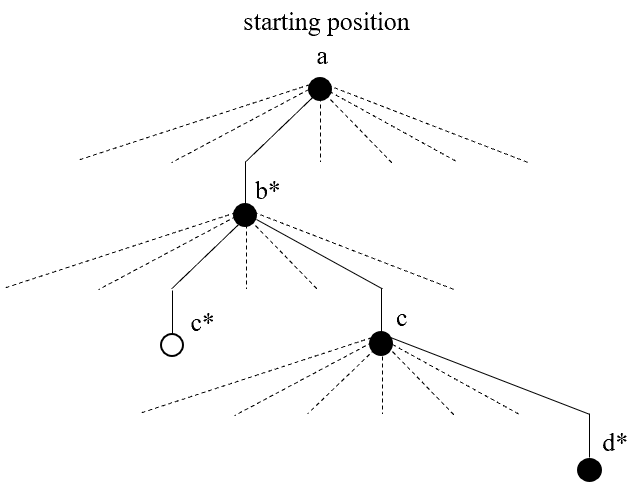


Figure 1. Sequence of the Kiowa’s decision points in the Pech River Valley. The solid lines represent the actions selected by the agent; the dashed lines represent actions that were available to the agent but that the agent did not select. The \* represents actions that are associated with the highest short-term progress.

Most of the time the Kiowa selects actions greedily, but at

Exploration v. exploitation

**1.3 Interactions and feedback**

* 1. **Learning and continued performance**

Evaluative feedback and a measure of progress that are meaningful with respect to the goal.

Ability to observe meaningful changes in the environment and match relevant actions to each change.

A model of the world that

Learning, keep learning, no need to learn again as we have a model of the world but this model might be constrained to the given task and fixed with the environment we encountered.

Do you think this is the only way in which the task could have been achieved?

* Are there other ways in which these tasks can be learned?

What does goal-oriented learning from interaction mean?

What does learning how to make a decision mean?

We want to differentiate between different possible ways of learning to solve a task? Why? To understand why one approach might be specifically useful and to understand whether these different approaches align with each other. Why do we want to understand how different ways of learning align with each other? Because different environments are amendable to different types of learning and once we co-deploy different agents that have different learning procedures, they might end up learning different things based on the environment. We also want to understand not only *that* a task is carried out but *how* it is carried out. Why? First, not all possible solutions to a problem are salient solutions. Can you think of an example? Second, how a task is learned entails information on whether the solution might generalize to other tasks or is constrained to one specific task. Why do you think that is important? Why do we care about the extent to which a solution is general? One, a general solution might be more efficient as we do not have to respecify each time how we want to solve the task. Second, a general solution might be the only feasible approach as we might not be able to fully anticipate what the task we need to solves look like and the environment in which the decision is made might consist of many connected tasks that constantly change in their configuration and required solution.

The pilot of a Kiowa helicopter follows

How do human teams learn to make decisions?

*Model of the world*

*How do they get that model? Focus on the interaction.*

*Do they learn during deployment?*

*What are they learning?*

*What does learning mean?*

*Do they make decisions that they did not make during training?*

*Adaptation v. Generation of new decision making rules.*

*Do they have rules?*

*How do these rules get updated?*

*How do they frame the environment in which they are learning?*

*How are these environments different?*

What type of learning processes can we identify in human decision making?

How can we formalize this learning process?

We want to approximate the formalization of decision making needed to design reinforcement learning algorithms.

How useful is this formalization?

We want to use this formalization as a benchmark for comparing human and machine decisions.

Introduction to MDPs

To compare between different types of decision making processes, we need to conceptualize them around a shared set of features. While each decision making process might have different specifications of these features, the features themselves remain the same.

Let’s consider some examples of decision making processes to obtain these features.

* The pilot of a two-seat, two-pilot Kioawa helicopter deployed in the Pech River Valley receives the unexpected call from a platoon leader to interrupt its ground support mission and land and evacuate a casualty after a medivac that was called in has not arrived. The pilot confers with her co-pilot on mission priorities and the Kioawa proceeds to landing, while obtaining additional information on the enemy position from the platoon leader.
* A mobile robot decides whether it should enter a new room in search of survivors in a destroyed building or start finding its way back to its battery station to recharge. It makes its decision based on the current charge level of its battery and how quickly or easy it has been able to find the recharger in the past.
* An economist of the Afghan central bank builds a model that computes over various macroeconomic parameters to assess policy options. Building the model involves assessing the impact of a range of macroeconomic parameters but also intuitive judgement. After the model has been tested, the economist recommends to her superiors to raise the national interest rate.

These examples share features that are so basic that they are easy to overlook. All involve an *agent* that *senses* and takes *actions* in an *environment,* within which the agent aims to achieve a goal despite *uncertainty* about the environment. The agents’ actions affect the future *state* of the environment (e.g., the next position of the Kioawa, the robot’s next location and future battery charge, the central bank’s future interest rate), thereby affecting the options and opportunities available to the agent at later times. Correct choice requires taking into account indirect, delayed consequences of actions, and thus may require foresight or planning.

In each of the examples, the agent can select between multiple actions and the environment can be in multiple states.

* We describe the set of actions available to the agent as the agent’s *action space*.
* We describe the set of states of the environment as the environment’s *state space*.
* The complexity of a problem depends in part on the sizes of the relevant state and action spaces.
* Often, we want to look at a subset of actions that is available to the agent for a given state of the environment. We then refer to a *state-action pair*.

In addition to the agent, sensations, the action space of the agent, the environment, the state space of the environment and a goal, we can identify four main subelements of a decision making process (Sutton and Barto 2018):

* A *policy* defines the agent’s way of behaving at a given time. A policy maps from perceived states of the environment to actions to be taken when in those states.
* A *reward signal* that provides evaluative feedback to agent on its decisions and determines whether an action was good or bad. The *reward* signal is the primary basis for altering the policy; if an action selected by the policy is followed by a low reward, then the policy may be changed to select some other action in that situation in the future.
* A *value function* that specified what is good in the long run, whereas the reward signal specifies only what is good in the short run. The *value* of a state is the total amount of reward an agent can expect to accumulate over the future, starting from that state. Whereas rewards determine the immediate, intrinsic desirability of environmental states, values indicate the long-term desirability of states after taking into account the states that are likely to follow, and the rewards available in those states.
* A *model* of the environment that mimics the behaviour of the environment and allows for inferences to be made about how the environment will behave. Models are used for *planning*, which we take to refer to any way of deciding on a course of action by considering possible future situations before they are actually experienced.

Putting all of these features together, we obtain the basics of decision making processes formalized as *reinforcement learning* (RL) problems. RL is commonly used for the optimal control of autonomous systems, such as the search robot in the above example. Solving RL problems requires us to add some technical specifications to the features we identified and introduce the learning process itself. For now, let’s take the features we have identified and simply use them to formalize one of the above examples.

* 1. **Formalizing the Kiowa Evacuation**

To understand how we can formalize human decision making processes, we take the example of the decision process of the Kiowa evacuation and spell out its formal features.

|  |  |
| --- | --- |
| Decision feature | Kiowa example |
| *Environment* | Pech River Valley in Afghanistan |
| State space | Continuous space compromising the Kiowa’s position, velocity, torque etc. |
| *Agent* | Kiowa Pilot |
| Sensation | Audio-visual |
| Action space | Communicate, move helicopter left, right, up, down, hover, fire weapons |
| Goal | Protect soldiers’ lives |
| Policy | Set of actions to decrease distance between casualty and base for each position of the Kiowa |
| Reward signal | Distance between casualty and base for given state |
| Value function | Evaluate if state reduces likelihood of casualties for the Kiowa’s assigned ground unit |
| Model of the environment | Pilot has previous combat experience in Afghanistan, knowledge of enemy operations |

Table 1: Decision features of the Kiowa evacuation example.

Based on this formalization, let’s consider some of the trade offs in the agent’s decision making:

* Maximizing rewards by minimizing the distance between casualty and base can lead to unsafe flight behaviour.
* Leaving the ground unit to evacuate the casualty can lead to short-term positive rewards but long-term negative rewards, if an enemy attack is launched while the Kiowa is away.
* Not leaving the unit to evacuate the casualty can lead to short-term positive rewards if an enemy attack is ongoing, but long-term negative rewards if the ground unit can fend off the attack easily, and air support would not have been needed.

How does the agent know how to balance these trade offs? As much as the human pilot is *trained* to balance the trade offs, an autonomous agent has to *learn* to balance them, too. Developing the appropriate learning architecture for such a task is the job of engineers. However, the learning architecture depends on how the task to be learned is defined. Let’s consider for the Kiowa example why obtaining an unambiguous definition of such a task might be difficult:

* *Complex preferences*. The pilot’s stated goal might be too broad in scope, such as “supporting ground units to succeed in combat”, to be unambiguous.
* *Unknown or incomplete preferences*. The pilot may have a different goal that we cannot deduce, or can deduce only in parts, from the data on the pilot’s decision making.
* *Corrupt reward channel*. The specified reward signal might fail to correctly evaluate central aspects of the pilot’s decision making, resulting in degraded performance (Everitt et al. 2017).
* Note, that we could also give a high-level specification of the action space (e.g., we could specify the action space *A* as *A* = { *protect unit, move to base*}). We then assume that each of these actions corresponds to a set of automated operations that encode how the action should be carried out.

There are other ambigiuties in our formalization of the Kiowa decision making process and we will address them throughout the course. The Kiowa example is more complex than tasks that are generally delegated to autonomous systems but many less complex tasks still remain ambigious in their formalizations To reduce this performance-critical ambigiuity, coordinating between domain experts, who carry out a task in the field, and engineers, who formalize the task and design the learning architecture to perform these tasks autonomously, is crucial.

Exercise: Given the reward signal, what behavior would you observe? Can you come up with a better specification of the reward signal?

1. **Learning in Human Teams**

Jason Emory

“The part that amazed me was that the cliche, train as you fight and everything, you know, you get out of these courses and sometimes you have instructors that take what they teach very seriously and other times you don’t. When we were in Robin Sage [an unconventional warfare Army training exercise], a lot of the instructors would be telling us that you will never do it this way, we don’t know why we’re teaching it. But what I found was that every major lesson I have learned throughout my career, whether it was in the Q Course [Army Special Forces Qualification Course] or Ranger School, I mean, everything that I was taught in the school house, I applied over there. I didn’t find myself in a situation, where I was saying, ‘Yeah well in the school house they taught it this way but it was totally unrealistic, I wish they hadn’t taught me that.’ It was the opposite. All the major muscle movements during the campaign we really had been taught, we’ve been taught them well, even by people who less motivated at times to teach it to us. It really kind of floored me. To me the system worked. The training pipeline and all of that worked. You had certain unique areas, when we talk about the Horse Soldiers, Mark [Nutsch] was the horse solder, he was the unique horse soldier. He was the perfect man for the perfect mission in the North with Dostum’s cavallary. I mean that was something you couln’t have forseen and it really was an act of God that we had the right officer there who could teach his people how to ride and could do everything he did. But that was too me almost the exception, everything else what we were taught, we applied and it really blew my mind how well we were prepared for it.”

Mark Nutsch

“My seargents, we had a very senior team at that time, by then I am talking our average age was 32. [...] Just a very mature, experienced team. But even in that new situation, the guys kept going, ‘Hey, we have been here before, remember Special Forces training, remember Robin’s Sage at this phase of insurgency, you know as that would progress, remember that.’ [...] But the seargents and I, coming back as we’re talking about this, we did the things you do in training. Each day we would do lessons learned, an internal AAR [After Action Report], whether it was five minutes or fifteen minutes, sit down and go ‘Damn, what nearly killed us today? How do we make sure that doesn’t happen again? You know, how do we survive the next hour? And how do we win?’ Because we believed that we could win, having that confidence in our training and resourcing and the people that were at our back. [...] You relied on that training that you had, the leadership lessons, people, mentors, that talked to you, every aspect of my career up to that point, to include character building events I had as a teenager through high school and college, all of that came to that focal point in my life on that battlefield, day after day. As it did with those seargents. But that was a phrase that kept coming up , ‘We did this in Robin’s Sage guys, we have been here before, it was slightly different, how do we apply it to this model and think through the problem and get after it and solve it in a positive way with the means we had available.’

“I would have to say, even in our mission, we were the students. You know, even with the majurity and training and experience we’ve had and the deployments throughout the Middle East, we got in there and the militia elements we linked up with, these guys had been fighting guerilla warfare for upwards of one to two decades. And they are the survivors. They have been wittled down through hard attrition. And every day for us was a history lesson for us. [...] I felt like we were the students and we had been remissed to not listen to what they had to say, because it’s their backyard. They couldn’t read a map but they could describe to you pationately ‘It’s this village, don’t you understand? It’s this village right over here. It’s this guy, he’s the one we’re after.’ So even then in that role, I felt we were students.”

**Exercise 2.1**. We have formalized human decision making around a basic RL architecture. While our goal is simply to use RL as a conceptual baseline for comparing between machine and human decision making, human decision making itself might not exemplify the features of RL. Briefly explain how you think the above features might be inadequate to describe human decision making.

1. **Learning Decision Making**
2. Across strategic environments, teams are tasked to engage in complex decision making episodes that require effective coordination among the team members.

We want to understand how teams coordinate to arrive at the decision that they make.

to achieve a *goal* despite *uncertainty* about their environment.

Definition 1.

Agent, Environment, State, Action, Goal, Reward, Information, Policy

a conceptual architecture that allows us to decompose any decision making process into a core set of shared parameters.

Different decision making processes, such as those of policy makers or those of the pilots of a UAV, share common structural elements.

. We want to differentiate between different types of decision making based on how the shared elements of these processes are specified.

. We want to decompose decision making processes into their common properties

Teams solve problems. We want to understand how teams of humans make decisions to understand the problem-solving mechanisms into which autonomous agents will be embedded.

To understand how teams make decisions, we need parameters that describe the mechanism of team decision making.

What might be useful parameters for assessing the decision making of teams?

. We use Afghanistan as its conflict landscape provides sufficient incentives to deploy autonomous systems – large territory, distributed enemy, U.S. conflict fatigue, existing use of drones – and problem factors – multiple stakeholders, cultural differences, complex structure of tribal and terrorist networks => rewards for deploying a system might be high, but the environment is hard to understand => requires careful calibration of factors in deciding whether to deploy or not to deploy autonomous systems

Example: the SOF team that enters Afghanistan on horses and JTACs + rules

**Human team and how this team makes decisions**

In situations that are consistent with the training data

In novel situations

Adaptive and generative learning

Example of a novel situation

What parameters do we need to establish for this decision making to be robust?

Example: the SOF team that enters Afghanistan on horses

How do we deal with unexpected behavior?

**Now we have a framework for assessing the context of decisions / human learning / in interaction with other learning entities**

**How does this context change if it involves autonomous machines?**

Human machine teams

Coordination in ecosystem

Multiple stakeholders with different information demands

Alignment between human and machine and human and human

**Properties of a decision**

**Agent**

**Environment**

**Goal**

**Reward**

Maybe the optimal solution requires negative measures of progress at first…there is no “straight line”…progress along the optimal decision path may not be linear….

Long term measures and short-term measures

This discrimination depends on **measures of progress** as the