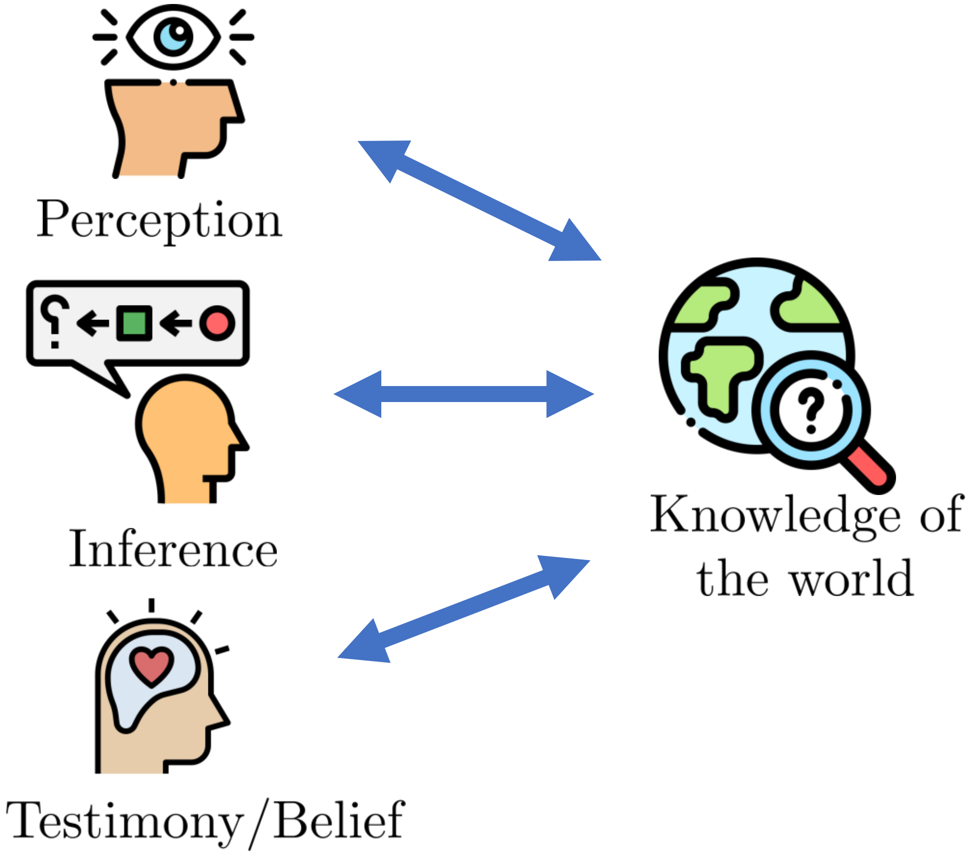
Decision Making for Machine Learning

(Reading Group Series Part 1)

Why is decision theory important?

Espistemology also involves decision making!

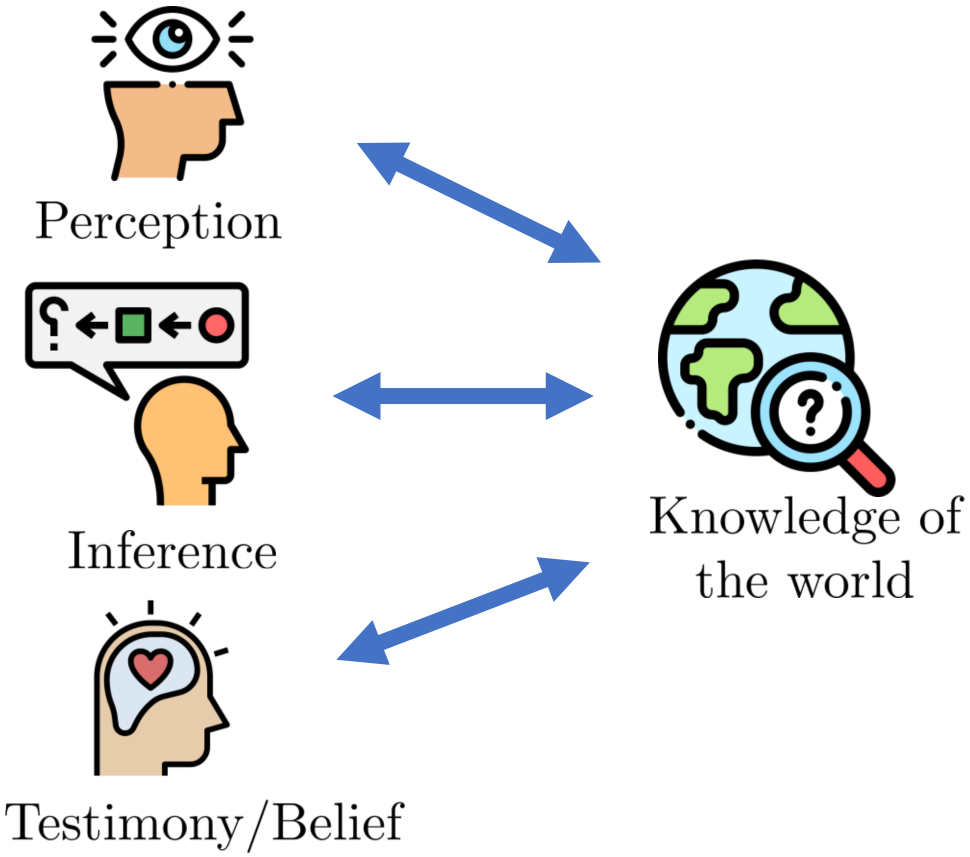


"All knowledge is ultimately probabilistic, and the confidence we place in our beliefs must be tempered by the uncertainty that shadows all things."

— *Bertrand Russell*
The Problems of Philosophy

Why is decision theory important?

Espistemology also involves decision making!



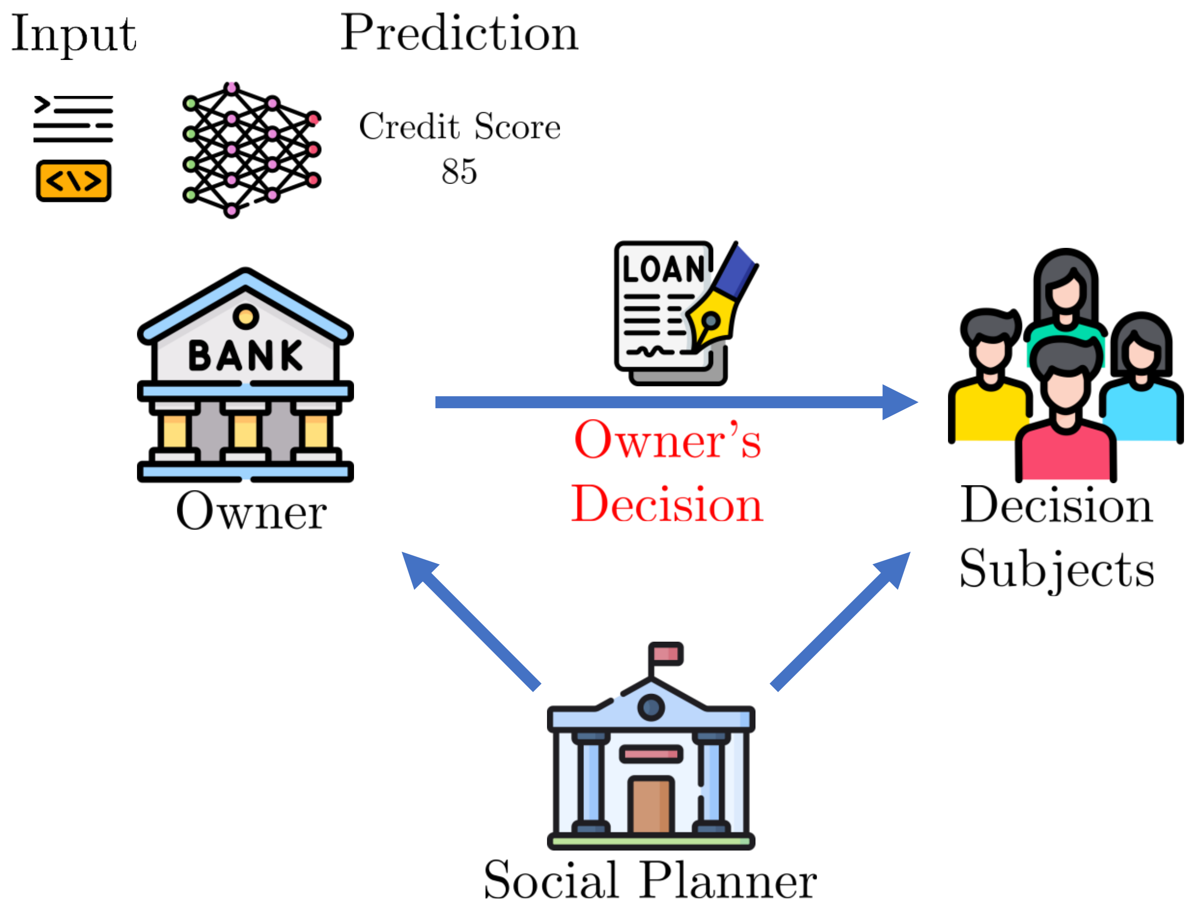
"All knowledge is ultimately probabilistic, and the confidence we place in our beliefs must be tempered by the uncertainty that shadows all things."

— Bertrand Russell
The Problems of Philosophy

"It's anticipation of regret, the fear of making the wrong choice, that drives decision-making, not the probability of making a mistake in prediction."

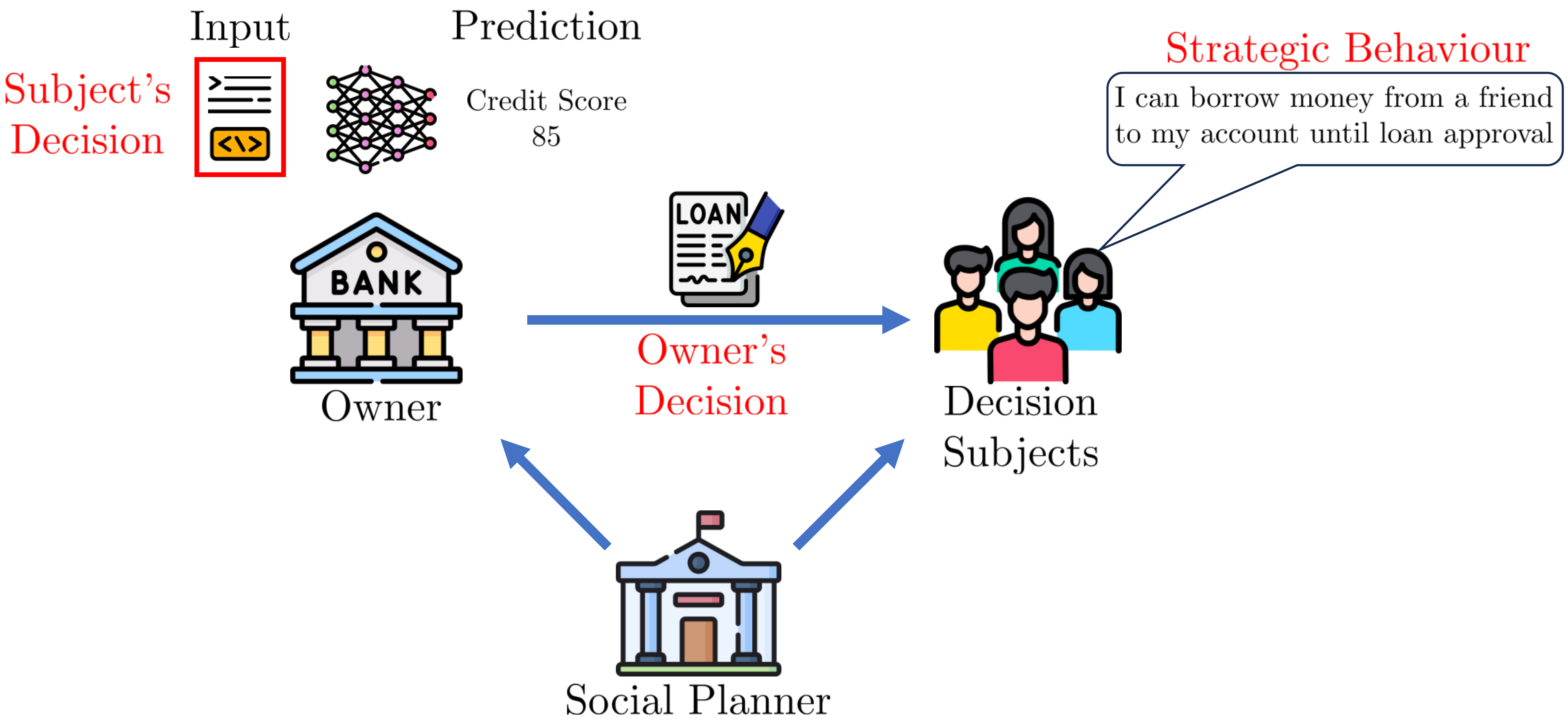
— Daniel Kahneman
Thinking, Fast and Slow

Role of decision making in ML



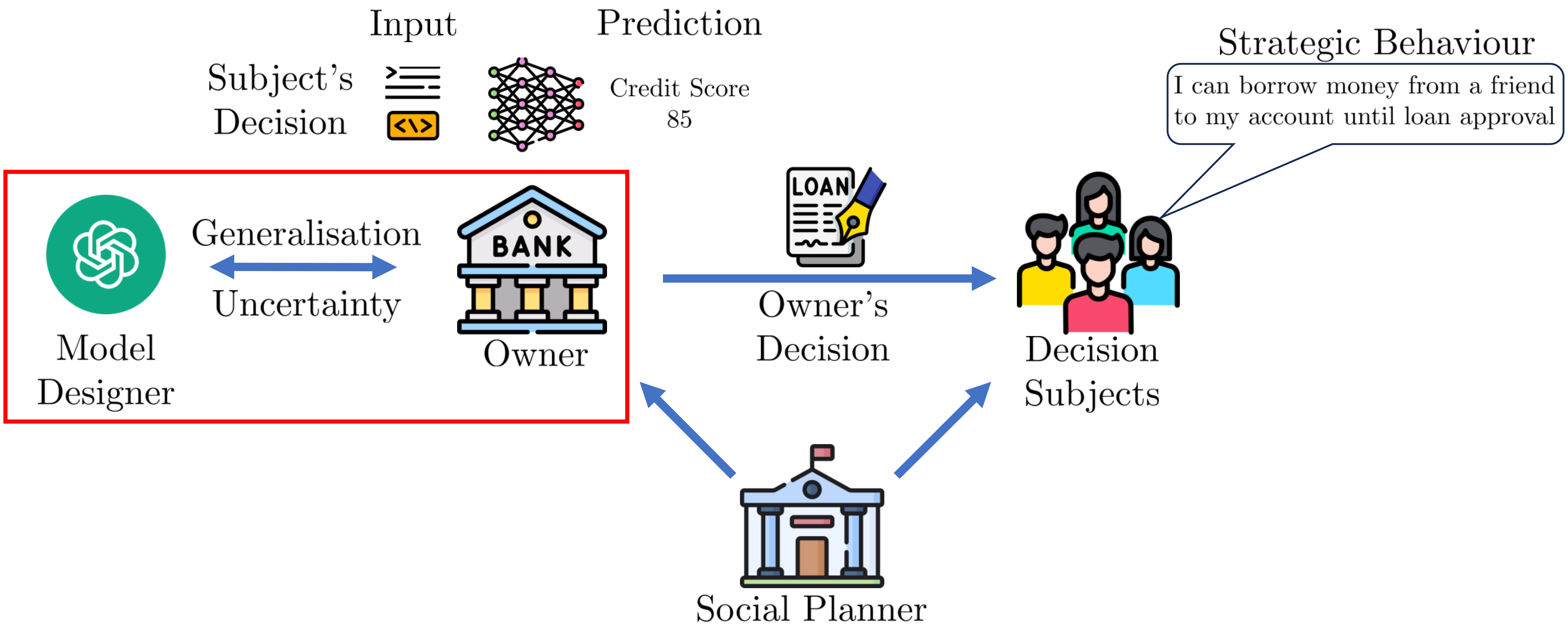
1. A Justice-Based Framework for the Analysis of Algorithmic Fairness Utility Tradeoffs, Hertweck et. al, arXiv 2023

Role of decision making in ML



2. Causal Strategic Learning with Competitive Selection, Vo et. al, AAAI 2024 (Oral)

Role of decision making in ML



3. Domain generalisation via Imprecise Learning, Singh et. al, ICML 2024 (Spotlight)

Basics of Decision Theory

- An unknown quantity $\theta \in \Theta$ *(state of nature)*
- Actions $a \in \mathcal{A}$
- Loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}$

		a_1	a_2
Θ	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Basics of Decision Theory

- An unknown quantity $\theta \in \Theta$ *(state of nature)*
- Actions $a \in \mathcal{A}$
- Loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}$

		a_1	a_2
Θ	θ_1	-1000	-100
	θ_2	200	-100

Basics of Decision Theory

- An unknown quantity $\theta \in \Theta$ *(state of nature)*
- Actions $a \in \mathcal{A}$
- Loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}$

		a_1	a_2
Θ	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Basics of Decision Theory

- An unknown quantity $\theta \in \Theta$ *(state of nature)*
- Actions $a \in \mathcal{A}$
- Loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}$
- (Optional) Prior Information $\pi \in \Delta(\Theta)$

		a_1	a_2
Θ	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Bayesian Expected Loss

- Intutively, most natural expected loss is one involving uncertainty in $\theta \in \Theta$.

- If $\pi(\theta) \in \Delta(\Theta)$ is the prior over θ then the Bayesian expected loss of action a is

$$\rho(\pi, a) = \mathbb{E}_{\theta \sim \pi}[L(\theta, a)]$$

		a_1	a_2
Θ	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Basics of (Statistical) Decision Theory

- An unknown quantity $\theta \in \Theta$ *(state of nature)*
- Actions $a \in \mathcal{A}$
- Loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}$
- (Optional) Prior Information $\pi \in \Delta(\Theta)$
- Outcome (r.v.) $\boldsymbol{X} \in \mathcal{H}$ from statistical investigation performed on θ

		a_1	a_2
Θ	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Basics of (Statistical) Decision Theory

- An unknown quantity $\theta \in \Theta$ *(state of nature)*
- Actions $a \in \mathcal{A}$
- Loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}$
- (Optional) Prior Information $\pi \in \Delta(\Theta)$
- Outcome (r.v.) $\boldsymbol{X} \in \mathcal{H}$ from statistical investigation performed on θ

$$\boxed{P(\boldsymbol{X}) \text{ should depend on } \theta \text{ !}}$$

Θ		a_1	a_2
	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Decision rules

- A (non randomized) *decision rule* δ maps outcomes to actions.
- Decision rule δ in no data decisions are simply actions.

$$\delta : \mathcal{H} \to \mathcal{A}$$

		a_1	a_2
Θ	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Decision rules

- A (non randomized) *decision rule* δ maps outcomes to actions.
- Decision rule δ in no data decisions are simply actions.

$$\delta(x) = \begin{cases} a_1 & \text{if } x \leq 0.05 \\ a_2 & \text{if } x > 0.05 \end{cases}$$

Example: Hypothesis Testing

$$\delta : \mathcal{H} \to \mathcal{A}$$

		a_1	a_2
Θ	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Decision rules – Frequentist Risk

- The frequentist perspective is to evaluate, for each θ, how much they expect to lose if $\delta(X)$ is repeatedly used with varying X in the problem.

- The risk function of decision rule $\delta(x)$ is defined by

$$R(\theta, \delta) = \mathbb{E}_{X \sim P_\theta}[L(\theta, \delta(X))]$$

$$\delta : \mathcal{H} \to \mathcal{A}$$

Θ		a_1	a_2
	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Decision rules – Frequentist Risk

- The frequentist perspective is to evaluate, for each θ, how much they expect to lose if $\delta(X)$ is repeatedly used with varying X in the problem.

- The risk function of decision rule $\delta(x)$ is defined by

$$R(\theta, \delta) = \mathbb{E}_{X \sim P_\theta}[L(\theta, \delta(X))]$$

How to compare decision rules?

$$\delta : \mathcal{H} \to \mathcal{A}$$

Θ		a_1	a_2
	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Decision rules – Frequentist Risk

- The risk function of decision rule $\delta(x)$ is defined by

$$R(\theta, \delta) = \mathbb{E}_{X \sim P_\theta}[L(\theta, \delta(X))]$$

- A decision rule δ_1 is R-better than a decision rule δ_2 iff $\delta_1 \succ \delta_2$ i.e.

$$\begin{array}{ll} \forall\ \theta \in \Theta & R(\theta, \delta_1) \leq R(\theta, \delta_2) \\ & \text{and} \\ \exists\ \theta \in \Theta & R(\theta, \delta_1) < R(\theta, \delta_2) \end{array}$$

$$\delta : \mathcal{H} \to \mathcal{A}$$

Θ		a_1	a_2
	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Decision rules – Bayes Risk

- The Bayes risk of decision rule $\delta(x)$ w.r.t prior $\pi \in \Delta(\Theta)$ is defined by

$$r(\pi, \delta) = \mathbb{E}_{\theta \sim \pi}[E_{X \sim P_\theta}[L(\theta, \delta(X))]]$$

$$\delta : \mathcal{H} \to \mathcal{A}$$

Θ		a_1	a_2
	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Randomized Decision rules

- A randomized decision rule $\delta(x, \cdot)$ is for each x a probability distribution on $\mathcal{A}$, which interpretation that if x is observed, $\delta(x, A)$ is a probability than an action in A will be chosen.

$$\delta : \mathcal{H} \to \Delta(\mathcal{A})$$

Θ		a_1	a_2
	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Randomized Decision rules

- A randomized decision rule $\delta(x, \cdot)$ is for each x a probability distribution on $\mathcal{A}$, which interpretation that if x is observed, $\delta(x, A)$ is a probability than an action in A will be chosen.

- In no data problem, randomized decision rule is called *randomized action* and is also a probability distribution on $\mathcal{A}$.

$$\delta : \mathcal{H} \to \Delta(\mathcal{A})$$

Θ		a_1	a_2
	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Randomized Decision rules

- Non randomized decision rules are special case of randomized decision rules, where for each x a specific action has probability 1.

$$\langle\delta\rangle(x, A) = I_A(\delta(x)) = \begin{cases} 1 & \text{if } \delta(x) \in A \\ 0 & \text{otherwise} \end{cases}$$

$$\delta : \mathcal{H} \to \Delta(\mathcal{A})$$

		a_1	a_2
Θ	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Randomized Decision rules

- Non randomized decision rules are special case of randomized decision rules, where for each x a specific action has probability 1.

$$\langle \delta \rangle (x, A) = I_A(\delta(x)) = \begin{cases} 1 & \text{if } \delta(x) \in A \\ 0 & \text{otherwise} \end{cases}$$

$$\delta : \mathcal{H} \to \Delta(\mathcal{A})$$

Θ		a_1	a_2
	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Randomized Decision rules

- The loss function of randomized rule is defined as

$$L(\theta, \delta(x, \cdot)) = \mathbb{E}_{a \sim \delta(x,\cdot)}[L(\theta, a)]$$

- The risk function of randomized rule is defined as

$$R(\theta, \delta) = \mathbb{E}_{X \sim P_\theta}[L(\theta, \delta(X, \cdot))]$$

$$\delta : \mathcal{H} \to \Delta(\mathcal{A})$$

Θ		a_1	a_2
	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Randomized Decision rules

- The loss function of randomized rule is defined as

$$L(\theta, \delta(x, \cdot)) = \mathbb{E}_{a \sim \delta(x,\cdot)}[L(\theta, a)]$$

- The risk function of randomized rule is defined as

$$R(\theta, \delta) = \mathbb{E}_{X \sim P_\theta}[L(\theta, \delta(X, \cdot))]$$

- The comparsion of randomized δ is similar to non-randomized rule.

$$\delta : \mathcal{H} \to \Delta(\mathcal{A})$$

		a_1	a_2
Θ	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Decision Principles: Conditional Bayes

- If $\pi(\theta) \in \Delta(\Theta)$ is the prior over θ then the Bayesian expected loss of action a is

$$\rho(\pi, a) = \mathbb{E}_{\theta \sim \pi}[L(\theta, a)]$$

- Conditional Bayes Principle: Choose action $a^* \in \mathcal{A}$ such that

$$a^* = arg \min_{a \in \mathcal{A}} \rho(\pi, a)$$

		a_1	a_2
Θ	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Decision Principle – Bayes Risk Principle

- The Bayes risk of decision rule $\delta(x)$ w.r.t prior $\pi \in \Delta(\Theta)$ is defined by

$$r(\pi, \delta) = \mathbb{E}_{\theta \sim \pi}[E_{X \sim P_\theta}[L(\theta, \delta(X))]]$$

- A decision rule δ_1 is preferred to rule δ_2 if $r(\pi, \delta_1) < r(\pi, \delta_2)$.

$$\delta^* = arg \min_{\delta \in \mathcal{D}} r(\pi, \delta)$$

where δ^* is called the Bayes rule.

$\delta : \mathcal{H} \to \mathcal{A}$

Θ		a_1	a_2
	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Decision Principle – Minmax Principle

- The worst case risk of decision rule $\delta(x)$

$$\sup_{\theta \in \Theta} R(\theta, \delta)$$

- A decision rule δ^* is minmax decision rule if

$$\delta^* = arg \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta)$$

$\delta : \mathcal{H} \to \mathcal{A}$

		a_1	a_2
Θ	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Decision Principle – Minmax Principle

- The worst case risk of decision rule $\delta(x)$

$$\sup_{\theta\in\Theta} R(\theta,\delta)$$

- A decision rule δ^* is minmax decision rule if

$$\delta^* = arg\inf_{\delta\in\mathcal{D}}\sup_{\theta\in\Theta} R(\theta,\delta)$$

$\delta : \mathcal{H} \to \mathcal{A}$

Θ		a_1	a_2
	θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
	θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Decision Principle – Invariance Principle

X

$\delta : \mathcal{H} \to \mathcal{A}$

Θ

	a_1	a_2
θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Y

$\delta' : \mathcal{H} \to \mathcal{A}$

Θ'

	a'_1	a'_2
θ'_1	$L(\theta'_1, a'_1)$	$L(\theta'_1, a'_2)$
θ'_2	$L(\theta'_2, a'_1)$	$L(\theta'_2, a'_2)$

What makes two decision problems invariant ?

Decision Principle – Invariance Principle

X

Θ

$\delta : \mathcal{H} \to \mathcal{A}$

	a_1	a_2
θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

$g(X)$

Θ

$\delta' : \mathcal{H} \to \mathcal{A}$

	a_1	a_2
θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

We define the group of transformations $g \in \mathcal{G}$ of $\mathcal{H}$

Decision Principle – Invariance Principle

X

Θ

$\delta : \mathcal{H} \to \mathcal{A}$

	a_1	a_2
θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

X

$\bar{\mathcal{G}}(\Theta)$

$\delta' : \mathcal{H} \to \mathcal{A}$

	a_1	a_2
θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

W.r.t. the of transformations $g \in \mathcal{G}$ of $\mathcal{H}$ we define the corresponding transformation on Θ as $\bar{G} = \{\bar{g} | g \in \mathcal{G}\}$

$$\mathbb{E}_{X \sim P_\theta}[h(g(X))] = \mathbb{E}_{X \sim P_{\bar{g}(\theta)}}[h(X)]$$

Decision Principle – Invariance Principle

X

Θ

$\delta : \mathcal{H} \to \mathcal{A}$

	a_1	a_2
θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

X

$\bar{\mathcal{G}}(\Theta)$

$\delta' : \mathcal{H} \to \mathcal{A}$

	a_1	a_2
θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Invariance of Loss function: $L(\theta, a)$ is called invariant under $\mathcal{G}$ if, $\forall g \in \mathcal{G}$ and $a \in \mathcal{A}$ there exists an $a^* \in \mathcal{A}$ such that

$$L(\theta, a) = L(\bar{g}(\theta), a^*) \quad \forall\ \theta \in \Theta$$

we can then denote a^* with $\tilde{g}(a)$

Decision Principle – Invariance Principle

Invariance of Loss function: $L(\theta, a)$ is called invariant under $\mathcal{G}$ if, $\forall g \in \mathcal{G}$ and $a \in \mathcal{A}$ there exists an $a^* \in \mathcal{A}$ such that

$$L(\theta, a) = L(\bar{g}(\theta), a^*) \quad \forall\ \theta \in \Theta$$

we can then denote a^* with $\tilde{g}(a)$

If a decision problem is invariant under group $\mathcal{G}$, then a decision rule $\delta(x)$ is also invariant if for all $x \in \mathcal{H}$ and $g \in \mathcal{G}$

$$\delta(g(x)) = \tilde{g}(\delta(x))$$

Nash Equilibrium

- Weaker solution concept than frequentist dominanance of decision, but stronger than Bayes optimality

- A nash equilibrium (θ^*, a^*) can be defined as from banks perspective

$$L(\theta^*, a^*) \leq L(\theta^*, a) \quad \forall a \in \mathcal{A}$$

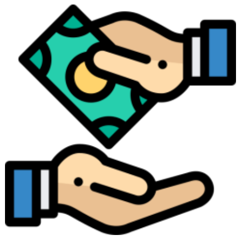
from applicants perspective

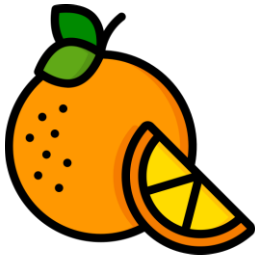
$$L(\theta^*, a^*) \leq L(\theta, a^*) \quad \forall \theta \in \Theta$$

Θ	a_1	a_2
θ_1	$L(\theta_1, a_1)$	$L(\theta_1, a_2)$
θ_2	$L(\theta_2, a_1)$	$L(\theta_2, a_2)$

Utility and Preferences

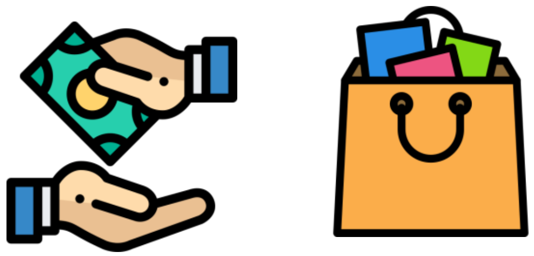
- Utility tries to assign numbers to a subjective idea of "*value*".
- Preferences are more objective since they represent ordering over items.

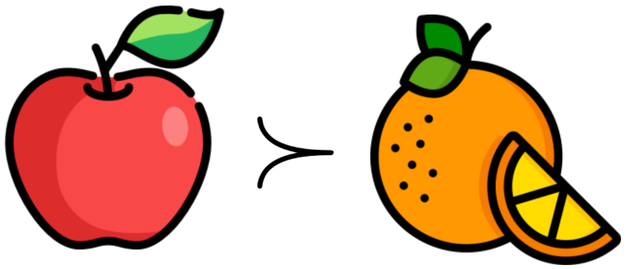


 $\succ$ 

Utility and Preferences

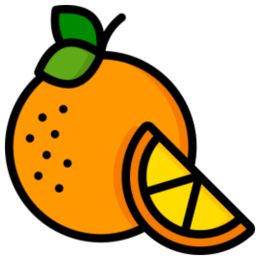
- Utility tries to assign numbers to a subjective idea of "*value*".
- Preferences are more objective since they represent ordering over items.
- Utility is intrinsic and hard to measure. Therefore it is inferred with *revealed preferences*.



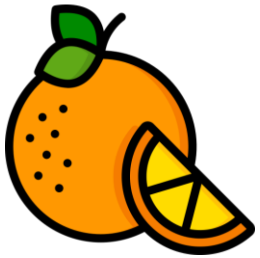


Rationality Axioms on Agent's Preferences

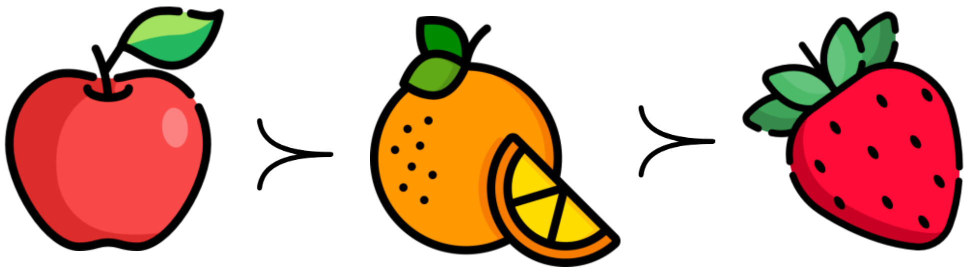
- Completeness: Agent must be able to compare any two items in the set i.e. $\forall a, b \in A$ either $a \succeq b$ or $b \succeq a$ or both.

 $\succ$ 

- Transitivity: $\forall a, b, c \in A$, $a \succeq b$ and $b \succeq c$ must imply $a \succeq c$.

 $\succ$  $\succ$ 

Utility construction with rational preferences



u(apple)=1 $\cdots$ u(strawberry)=0

How do we construct utility in uncertainty?

Utility construction for uncertainty

Assuming a preference on distributions $\Delta(\lambda)$

- **(Completeness):** Either $q_1 \succeq q_2$ or $q_2 \succeq q_1$
- **(Trasitivity):** If $q_1 \succ q_2$ and $q_2 \succ q_1$ then $q_1 \succ q_3$
- **(Archimedean Property):** If $q_1 \succ q_2 \succ q_3$ then $\exists\ \epsilon \in (0,1)$ such that

$$(1-\epsilon)q_1 + \epsilon q_3 \succ q_2 \succ \epsilon q_1 + (1-\epsilon)q_3$$

- **(Independence of Irrelevant Alternatives)** For any q_3 and $\epsilon \in (0,1]$,

$$q_1 \succ q_2 \quad \text{iff} \quad \epsilon q_1 + (1-\epsilon)q_3 \succ \epsilon q_2 + (1-\epsilon)q_3$$

VNM Theorem: Expected Utility Maximization

Under previous assumptions there exists a continuous affine utility function $u : \Lambda \to [0, 1]$ such that

$$q \succeq p \quad \text{iff} \quad \mathbb{E}_{\lambda \sim q}[u(\lambda)] \leq \mathbb{E}_{\lambda \sim p}[u(\lambda)]$$