Humanising Decision Making

Bridging Reinforcement Learning & Responsible AI

Debabrota Basu

Équipe Scool, Inria Lille

30 MIN. de sciences

Academic Trajectory

A Brief Introduction

Academic Trajectory

Education

CHALMERS
UNIVERSITY OF TECHNOLOGY

Postdoctorate

2019-2020

Robustness, Privacy, and Fairness in Machine Learning

Christos Dimitrakakis

Doctorate

2014-2018

Learning to Make Decisions with Incomplete Information

Stéphane Bressan (NUS)

Pierre Senellart (ENS, Paris)

Undergraduate

2010-2014

Non-rigid Registration with Gromov-Hausdorff Graph Cuts

Ananda S. Chowdhury

Academic Trajectory

Research Collaborations

Fair Decision Making
David Parkes, 2019

Multi-armed Bandits
Pierre Senellart, 2017

RL in Cloud Systems
Haibo Chen, 2016

Quantum Computing & Security
Subhamoy Maitra, 2014

Optimisation
P N Suganthan, 2013

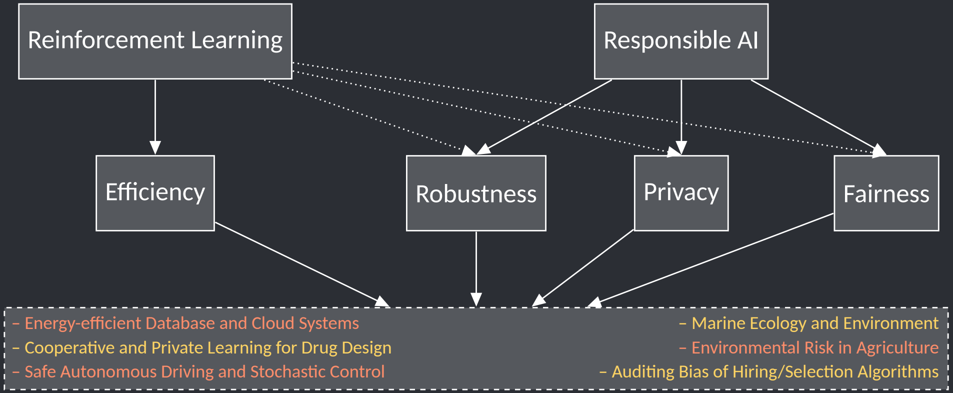
Academic Trajectory

Back to Scool: Our Équipe

What do we do?

We study the problem of sequential decision making under uncertainty, i.e. bandits and Markov decision processes.
We aim to deploy our findings for applications related to health, agriculture, ecology, and sustainable development.

My Research Expeditions

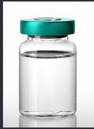


A Short Tour of Reinforcement Learning

Learning to Take Decisions Sequentially under Incomplete Information

Sequential Decision Making

Medicine 1
$p_1^{\text{cured}} = 0.75$



Medicine 2
$p_2^{\text{cured}} = 0.95$

Medicine 3
$p_3^{\text{cured}} = 0.90$

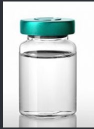
$\dots$

Medicine A
$p_A^{\text{cured}} = 0.5$

Sequential Decision Making

under Incomplete Information: Multi-armed Bandits [T33,R52,B56,G74,W80,LR85,ACF02,LS19]

Medicine 1
$p_1^{\text{cured}} = ?$



Medicine 2
$p_2^{\text{cured}} = ?$

Medicine 3
$p_3^{\text{cured}} = ?$

$\dots$

Medicine A
$p_A^{\text{cured}} = ?$

For the t-th patient ($t \leq T$) in the study

1. the doctor π chooses a Medicine $A_t \in \{1, \dots, A\}$,
2. Observes a response $R_t \in \{\text{cured}, \text{not cured}\}$ such that $\mathbb{P}(R_t = \text{cured} | A_t = a) = p_a^{\text{cured}}$.

Goal: Maximise the number of patients cured: $\sum_{t=1}^{T} R_t$.

Performance Measure under Incomplete Information

Regret

Maximise cumulative reward $$\sum_{t=1}^{T} R_t$$

$\underset{\textit{Randomness}}{\approx}$ Maximise expected cumulative reward $$\underbrace{V_T^{\pi} \triangleq \mathbb{E}\left[\sum_{t=0}^{T} R_t \mid A_t \sim \pi\right]}_{\text{Value of } \pi}$$

$\underset{\text{Incomplete Information}}{\Longleftrightarrow}$ Minimise expected regret $$V_T^{\mathrm{OPT}} - V_T^{\pi} = \mathbb{E}[R(a^*)]\,T - V_T^{\pi}$$

Regret $\mathscr{R}_{\pi}(T) \triangleq$ Value of Optimal Algorithm with Full Information

— Value of Algorithm π with Incomplete Information

Performance Measure under Incomplete Information

Regret

Maximise cumulative reward $\quad \sum_{t=1}^{T} R_t$

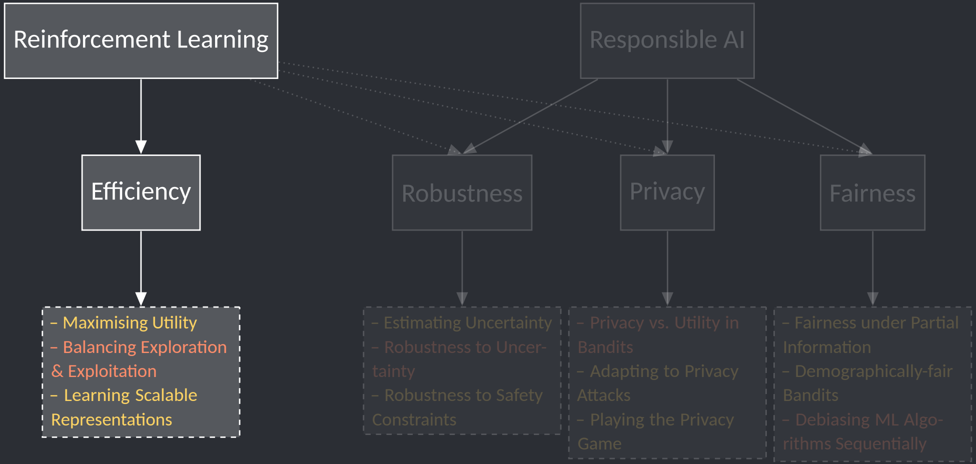
$\underset{Randomness}{\approx}$ Maximise expected cumulative reward $\quad \underbrace{V_T^{\pi} \triangleq \mathbb{E}\left[\sum_{t=0}^{T} R_t \mid A_t \sim \pi\right]}_{\text{Value of } \pi}$

$\underset{\text{Incomplete Information}}{\Longleftrightarrow}$ Minimise expected regret $\quad V_T^{\text{OPT}} - V_T^{\pi} = \mathbb{E}[R(a^*)]\,T - V_T^{\pi}$

Regret $\mathscr{R}_\pi(T) \triangleq$ Value of Optimal Algorithm with Full Information

$-$ Value of Algorithm π with Incomplete Information

$$\text{Minimum regret achievable by any } \pi = \Omega\left(\sum_a \underbrace{(\mu^* - \mu_a)}_{\text{Suboptimality Gap}} \underbrace{\frac{\log T}{D_{\text{KL}}(P_a, P_{a^*})}}_{\text{Distinguishability Gap}}\right) \approx \Omega\left(\sum_a \frac{\overbrace{\sigma_a^2}^{\text{Variance of a}} \log T}{\underbrace{\Delta_a}_{\text{Suboptimality Gap}}}\right).$$


Reinforcement Learning
Responsible AI
Efficiency
Robustness
Privacy
Fairness
– Maximising Utility
– Balancing Exploration & Exploitation
– Learning Scalable Representations
– Estimating Uncertainty
– Robustness to Uncertainty
– Robustness to Safety Constraints
– Privacy vs. Utility in Bandits
– Adapting to Privacy Attacks
– Playing the Privacy Game
– Fairness under Partial Information
– Demographically-fair Bandits
– Debiasing ML Algorithms Sequentially

Efficiency: Exploration–Exploitation Trade-off

Be More Optimistic when You Have Less Information

Exploration–Exploitation Trade-off

Should you try out new decisions to fetch information, or play the best with your existing knowledge?

Strategy: Calibrated Optimism in the Face of Uncertainty (OFU) [LS19]

Estimate an upper confidence bound on the empirical mean of the observed rewards and use it as an 'optimistic' index to choose the best arm to play.

For the t-th patient ($t \leq T$) in the study

1.a. the optimistic doctor π computes optimistic indexes $I_a(t)$ for each medicine given the history

1.b. the optimistic doctor π chooses a Medicine $A_t = \operatorname{argmax}_{a \in \{1,\dots,A\}} I_a(t)$,

2. Observes a response $R_t \in \{\text{cured}, \text{not cured}\}$ such that $\mathbb{P}(R_t = \text{cured} | A_t = a) = p_a^{\text{cured}}$.

Efficiency: Exploration–Exploitation Trade-off

Be More Optimistic when You Have Less Information

Index	UCB (No Noise)	UCBV (Unknown Noise Variance)
$I_a(t)$	$\underbrace{\hat{\mu}_{a,t}}_{\text{Average reward of } a} + \sqrt{\frac{2\log t}{\text{\# Selections of a till t}}}$	$\underbrace{\hat{\mu}_{a,t}}_{\text{Average reward of } a} + \underbrace{\hat{\sigma}_{a,t}}_{\sqrt{\text{Variance of rewards of } a}} \sqrt{\frac{2\log t}{\text{\# Selections of a till t}}} + \frac{3 \times \text{range of noise} \times \log t}{\text{\# Selections of a till t}}$

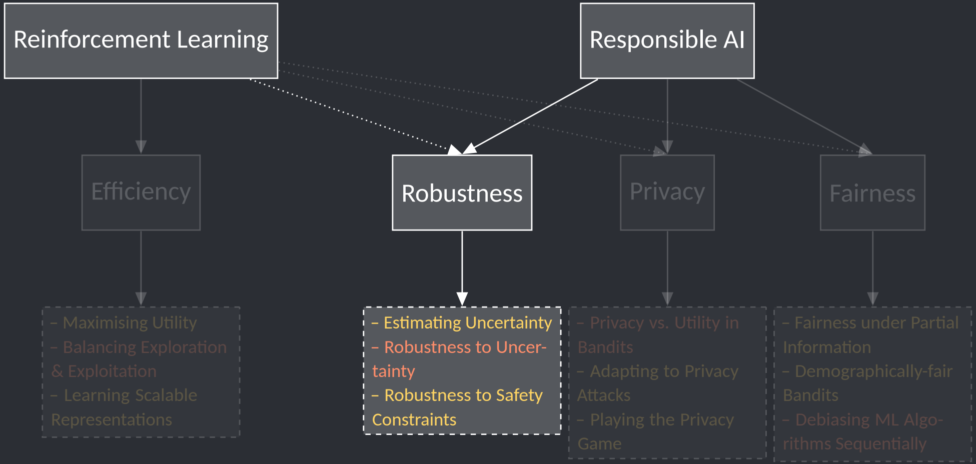
- For UCB, the regret upper bound is $\mathcal{O}\left(\sum_a \Delta_a + \frac{\log T}{\Delta_a}\right)$.
- For UCBV, the regret upper bound is $\mathcal{O}\left(\sum_a \Delta_a + \left(\text{range of noise } + \frac{\sigma_a^2}{\Delta_a}\right) \log T\right)$.
- To obtain KL in the denominator, directly optimise KL to compute the optimistic index ⟶ KL-UCB [LS19]/BelMan [BSB19]

Limitations

Optimism works optimally for exponential family of rewards, sub-Gaussian noise, and independent actions.

Humanising Decision Making

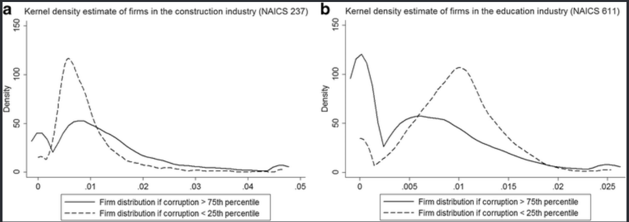
Reinforcement Learning 🤝 Responsible AI

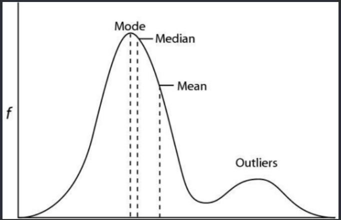

Reinforcement Learning
Responsible AI
Efficiency
Robustness
Privacy
Fairness
– Maximising Utility
– Balancing Exploration & Exploitation
– Learning Scalable Representations
– Estimating Uncertainty
– Robustness to Uncertainty
– Robustness to Safety Constraints
– Privacy vs. Utility in Bandits
– Adapting to Privacy Attacks
– Playing the Privacy Game
– Fairness under Partial Information
– Demographically-fair Bandits
– Debiasing ML Algorithms Sequentially

Robustness: Arbitrarily Corrupted Observations [BMM22]

What is the reward at every step have heavy-tails and are arbitrarily corrupted?

The decision maker observes $R_t \sim \varepsilon P_{A_t} + (1-\varepsilon) C_{A_t}$

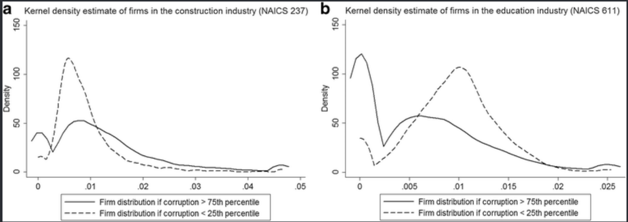


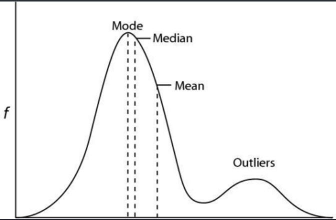


Robustness: Arbitrarily Corrupted Observations [BMM22]

What is the reward at every step have heavy-tails and are arbitrarily corrupted?

The decision maker observes $R_t \sim \varepsilon P_{A_t} + (1-\varepsilon) C_{A_t}$





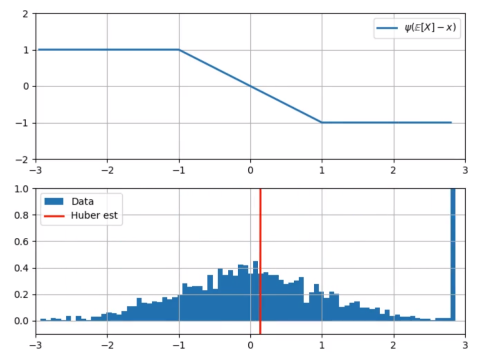
$$\mathscr{R}_{\pi_{\text{robust}}}(T) \asymp \underbrace{\mathcal{O}\left(\sum_{a:\Delta_a>\sigma_a} \sigma_a \log T\right)}_{\text{Error due to Heavy-tail}} + \underbrace{\mathcal{O}\left(\sum_{a:\Delta_a\le\sigma_a} \Delta_i \frac{\sigma_a^2}{\overline{\Delta}_{a,\varepsilon}^2}\log T\right)}_{\text{Usual } \sigma^2/\Delta \text{ error with corruption correction}} + \underbrace{\mathcal{O}\left(\sum_a \frac{\Delta_a}{\log\left(\frac{1-\varepsilon}{\varepsilon}\right)}\right)}_{\text{Constant error due to corruption}}.$$

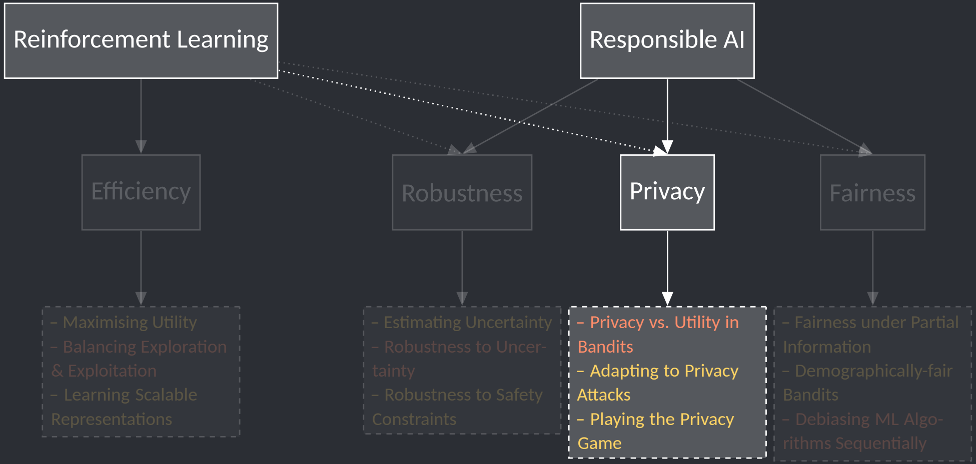
We observe that the corrupted suboptimality gap $\overline{\Delta}_{a,\varepsilon} \triangleq (1-\varepsilon)\Delta_a - \varepsilon\sigma_a$ dictates the hardness.

Robustness: Arbitrarily Corrupted Observations [BMM22]

A Generic Recipe to Robustness

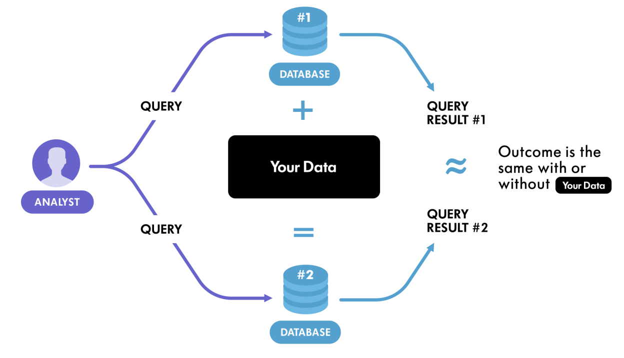
- Use a robust estimator of mean and variance (e.g. Huber estimator)
- Derive the tightest optimistic confidence bounds for the estimates
- Plug them in the UCB/UCBV type algorithm




Reinforcement Learning
Responsible AI
Efficiency
Robustness
Privacy
Fairness
– Maximising Utility
– Balancing Exploration & Exploitation
– Learning Scalable Representations
– Estimating Uncertainty
– Robustness to Uncertainty
– Robustness to Safety Constraints
– Privacy vs. Utility in Bandits
– Adapting to Privacy Attacks
– Playing the Privacy Game
– Fairness under Partial Information
– Demographically-fair Bandits
– Debiasing ML Algorithms Sequentially

Data Privacy: ϵ-Differential Privacy [DR14]

Information in input/database becomes private if it is indistinguishable from the output of a query/algorithm.



$$\frac{\mathbb{P}(\pi(\text{DB} + \text{my data}) = o)}{\mathbb{P}(\pi(\text{DB}) = o)} \leq e^{\epsilon} \longrightarrow \epsilon - \text{DP}$$

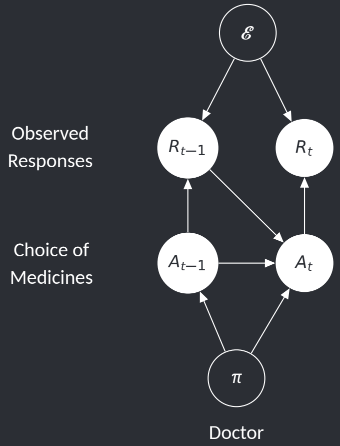
Image Courtesy: www.winton.com

Data Privacy in Sequential Decision Making

Data Generation in Multi-armed Bandits [BDT19]

Reward Distributions of Medicines

$$\mathscr{E} = \{\mathbb{P}(R|a)\}_{a=1}^{A}$$



Input to π

Set of Observed Responses: $R^T = \{R_1, \ldots, R_T\}$

Output of π

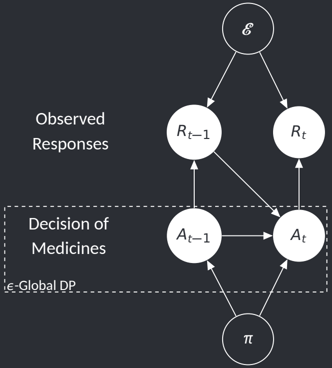
Set of Decisions: $A^T = \{A_1, \ldots, A_T\}$

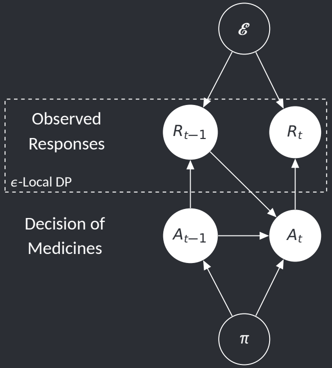
Data Privacy in Bandits

A patient t wants to keep her response R_t to a medicine A_t private.

Data Privacy in Multi-armed Bandits

Global [AB22] and Local [BDT19] Differential Privacy





$$\frac{\mathbb{P}_\pi\left(\text{Set of Decisions} \,\middle|\, \text{Possible responses of T patients} + \text{my data}\right)}{\mathbb{P}_\pi\left(\text{Set of Decisions} \,\middle|\, \text{Possible responses of T patients}\right)} \leq e^{\epsilon}$$

$$\frac{\mathbb{P}\left(\text{Observed responses} \,\middle|\, \text{Possible responses of T patients} + \text{my data}\right)}{\mathbb{P}\left(\text{Observed responses} \,\middle|\, \text{Possible responses of T patients}\right)} \leq e^{\epsilon}$$

Data Privacy: The Cost of Privacy in Bandits

Minimum Achievable Regret for Globally and Locally Private Bandits *[BDT19, AB22]*

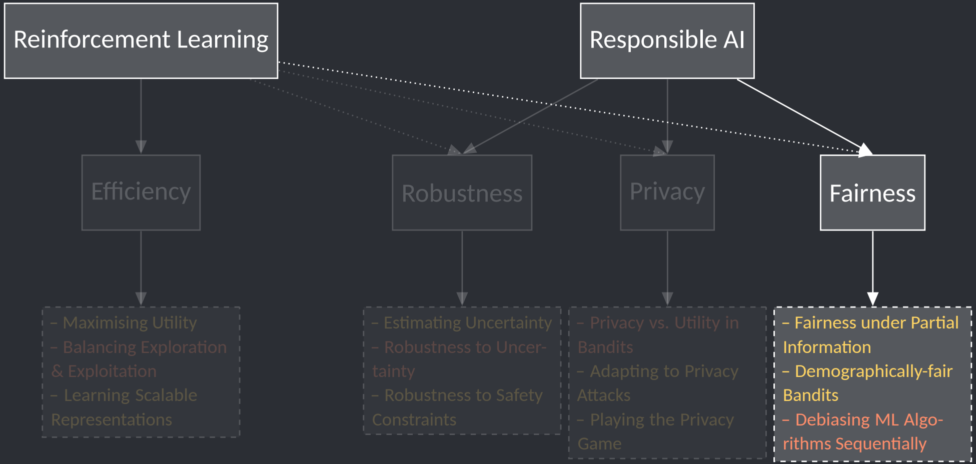
Lower Bounds	Minimax (Worst-case) Regret	Problem-dependent Regret
No DP	$\sqrt{(A-1)T}$	$\frac{\log T}{D_{\mathrm{KL}}(P_{a^{\mathrm{second}}}, P_{a^*})}$
Global DP	$\max\left(\sqrt{(A-1)T}, \frac{A-1}{\epsilon}\right)$	$\sum_a \max\left(\frac{\sigma_a^2 \log T}{\Delta_a}, \frac{\sigma_a \log T}{\epsilon}\right)$
Local DP	$\frac{1}{\epsilon}\sqrt{(A-1)T}$	$\frac{1}{\epsilon^2}\sum_a \frac{\sigma_a^2 \log T}{\Delta_a}$

Non-private < Global DP < Local DP

Minimum achievable regret: $\longleftarrow$

Amount of Noise Injected

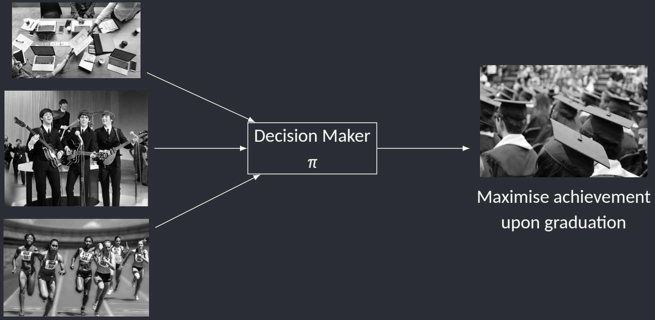
Regimes of Privacy vs. Partial Information: Impact of global DP is ignorable than that of partial information if privacy level ϵ is bigger than the suboptimality gap-variance ratio $\frac{\Delta_a}{\sigma_a}$.



Reinforcement Learning
Responsible AI
Efficiency
Robustness
Privacy
Fairness
– Maximising Utility
– Balancing Exploration & Exploitation
– Learning Scalable Representations
– Estimating Uncertainty
– Robustness to Uncertainty
– Robustness to Safety Constraints
– Privacy vs. Utility in Bandits
– Adapting to Privacy Attacks
– Playing the Privacy Game
– Fairness under Partial Information
– Demographically-fair Bandits
– Debiasing ML Algorithms Sequentially

Fairness in Sequential Decision Making

Fair Selection in College Admissions [BSB+21]

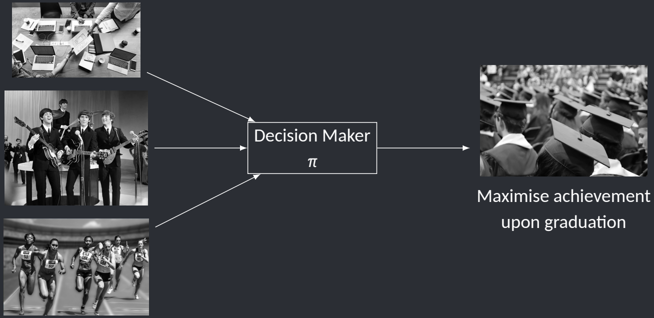


Set Fair Selection: From Individualist Meritocracy to Collective Meritocracy

$$K^* \triangleq \min_X \operatorname*{argmax}_{K \in \mathcal{N} - X} U(X \cup K) \text{ such that } |\text{Marginal Utility of } K - \text{Shapley of } K| \leq \delta.$$

Fairness in Sequential Decision Making

Fair Selection in College Admissions [BSB+21]



Demographic Fair Selection: From Homogenisation to Equal Opportunity over Demographies

$$\pi^* \triangleq \operatorname*{argmax}_{\pi} \sum_{Groups} w_{Group} V^{\pi}_{\mathcal{N}}(Group) \text{ such that } |w_{Group_1} - w_{Group_2}| \leq \delta.$$

Fairness in Sequential Decision Making

Deviation from Collective Meritocracy and Demographic Fairness [BSB+ 21]

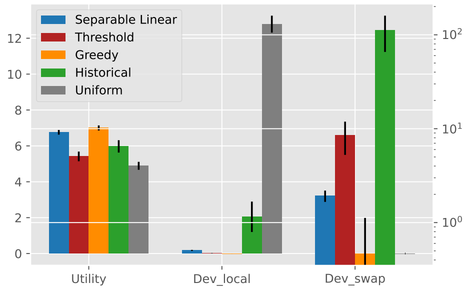


Figure: Set Fair Selection

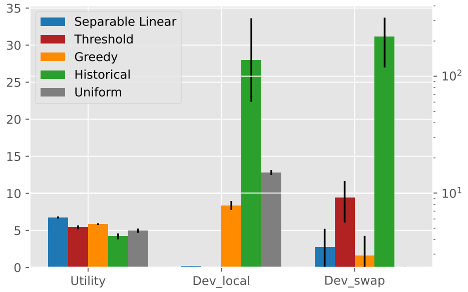
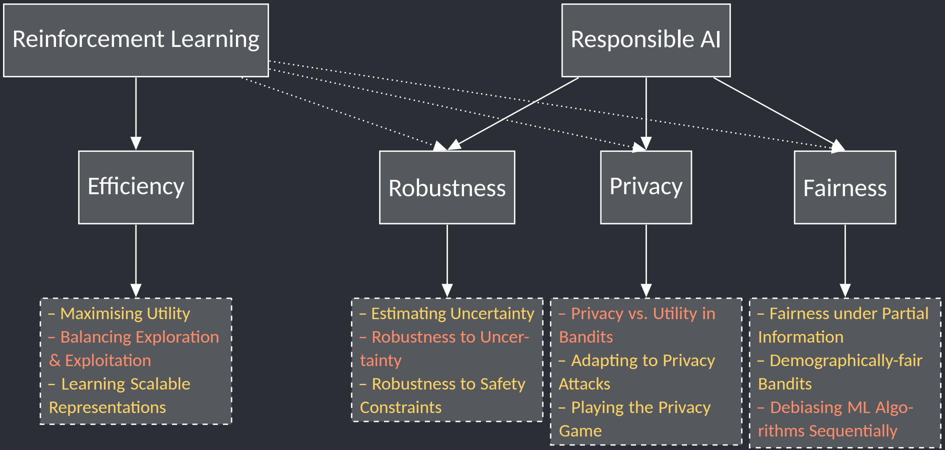


Figure: Demographic Fair Selection

Theoretically Grounded
Efficient, Robust, Private, and Fair Reinforcement Learning
for Solving Decision Making Problems Responsibly.



For further details, please visit: `https://debabrota-basu.github.io/`

References I

[AB22] Achraf Azize and Debabrota Basu.
When privacy meets partial information: A refined regret analysis of differentially private multi-armed bandits, 2022.

[BDT19] Debabrota Basu, Christos Dimitrakakis, and Aristide C. Y. Tossou.
Differential privacy for multi-armed bandits: What is it and what is its cost?
CoRR, abs/1905.12298, 2019.

[BMM22] Debabrota Basu, Odalric-Ambrym Maillard, and Timothée Mathieu.
Bandits corrupted by nature: Lower bounds on regret and robust optimistic algorithm, 2022.

[BSB19] Debabrota Basu, Pierre Senellart, and Stéphane Bressan.
BelMan: Information geometric approach to stcohastic bandits.
In *ECML-PKDD*, 2019.

[BSB+21] Thomas Kleine Buening, Meirav Segal, Debabrota Basu, Christos Dimitrakakis, and Anne-Marie George.
On meritocracy in optimal set selection, 2021.

[DR14] Cynthia Dwork and Aaron Roth.
The algorithmic foundations of differential privacy.
Foundations and Trends® in Theoretical Computer Science, 9(3–4):211–407, 2014.

[LS19] Tor Lattimore and Csaba Szepesvári.
An information-theoretic approach to minimax regret in partial monitoring.
arXiv preprint arXiv:1902.00470, 2019.