RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

FRANCE 2030 PROGRAMME DE RECHERCHE INTELLIGENCE ARTIFICIELLE

FOUNDRY

Foundations of Robustness and Reliability in AI

Panayotis Mertikopoulos
CNRS / Laboratoire d'Informatique de Grenoble
CNRS – Dauphine – ENS Lyon – IMT – Inria

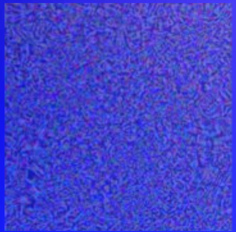
anr agence nationale de la recherche

Can artificial intelligence be relied upon?

PIGS ON WINGS

AI sees "pig"

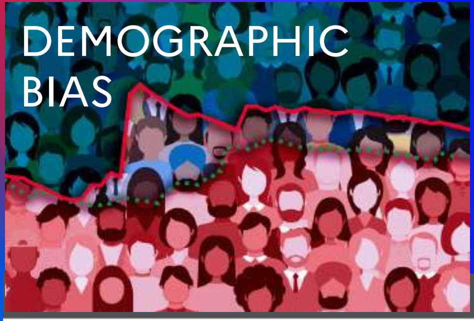
+ 0.005 x



Invisible noise

=

AI sees "airplane"



	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

FOUNDRY's mission statement

«Develop the theoretical foundations of robustness and reliability in machine learning and artificial intelligence»

The challenges ahead

1. The « known unknowns » — *#adversarial attacks #data-centric impediments*
2. The « unknown unknowns » — *#multi-agent learning #online adaptation*
3. Balance concurrent desiderata — *#fairness #privacy # strategic agents*

Research Axes

1. **Tame the «known unknowns»**
    - Robustness to data-centric impediments («bad data»)
    - Shortfalls in the data (incomplete observations, label shifts, poisoning)
    - Impediments at inference time (adversarial attacks,…)

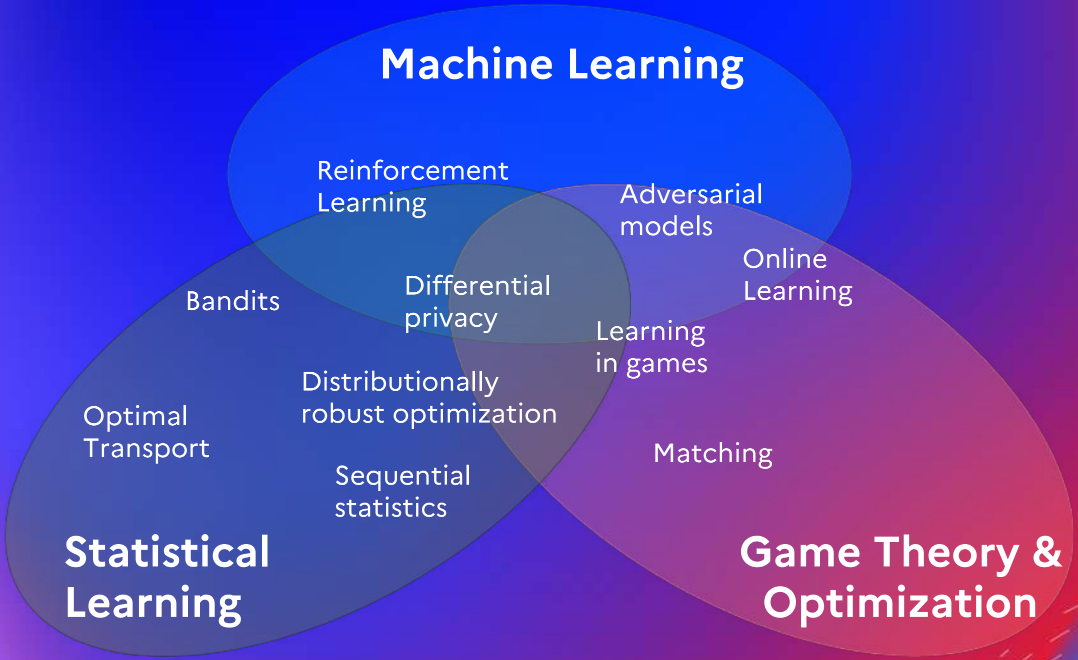
2. **Adapt to the «unknown unknowns»**
    - Adaptivity to unmodeled phenomena and/or the environment
    - From best- to worst-case guarantees
    - Adapt « on the fly » to non-stationary environments

3. **Balance concurrent / incompatible objectives**
    - Robustness v. accuracy
    - Guarantees in privacy and/or fairness vs. predictive accuracy
    - Selfishly-minded agents

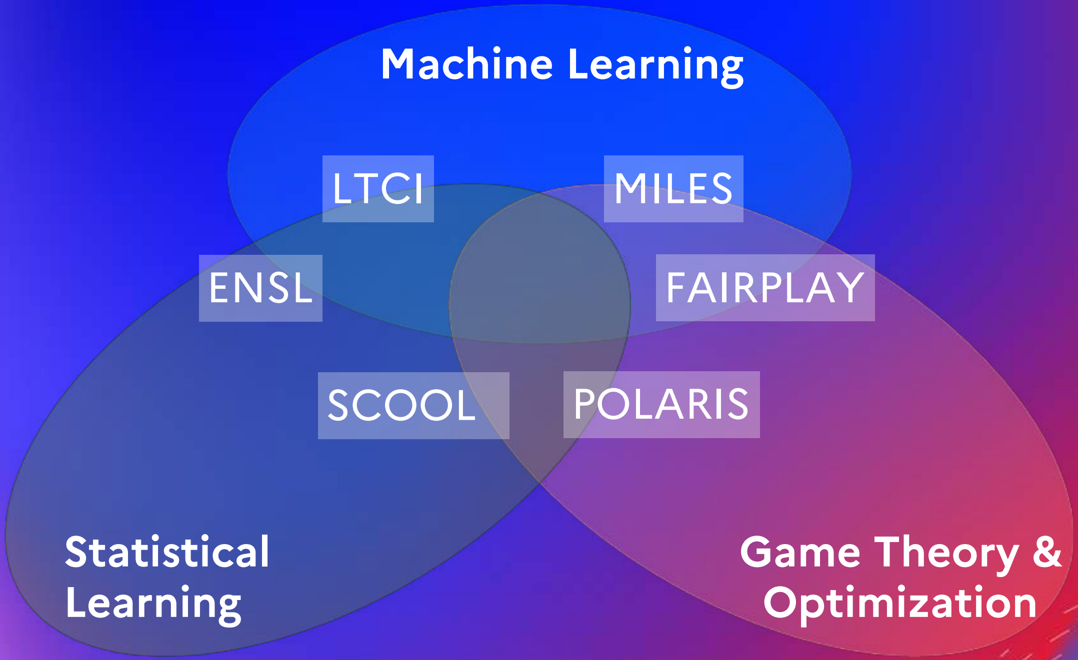
The partners


SCOOL
FAIRPLAY
LTCI
MILES
ENSL
POLARIS
Université de Lille
Institut Mines-Télécom
Dauphine | PSL
UNIVERSITÉ PARIS
ENSAE
INSTITUT POLYTECHNIQUE DE PARIS
ENS DE LYON
INSTITUT DE MATHÉMATIQUES DE TOULOUSE

Skills and expertise



Skills and expertise



Consortium breakdown

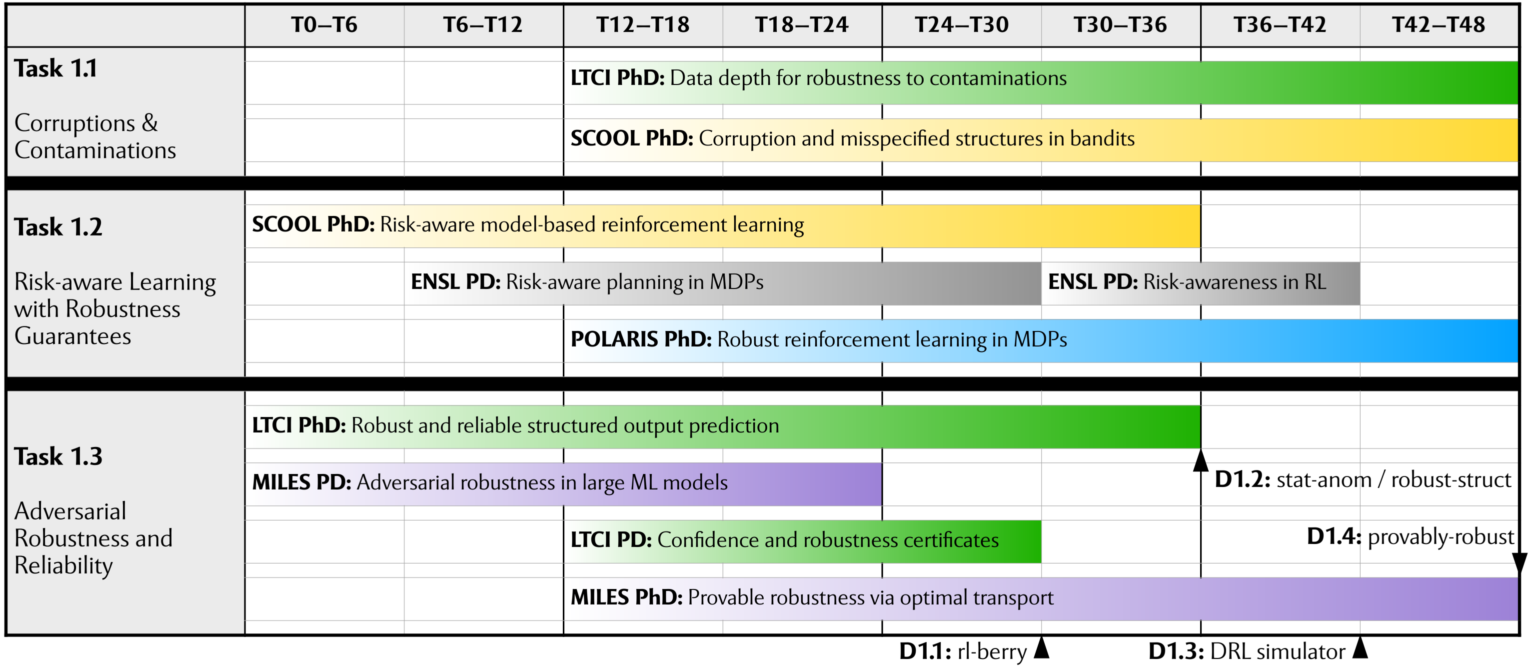
- **POLARIS (lead: P. Mertikopoulos)**
  # game theory #optimization #online learning #reinforcement learning
- **ENSL (A. Garivier)**
  #sequential statistics #bandits #reinforcement learning #differential privacy
- **FAIRPLAY (P. Loiseau)**
  #matching #fairness #privacy #online learning #online algorithms
- **LTCI (F. D'Alché-Buc)**
  #extreme value theory #robust statistics #structure data #Monte Carlo
- **MILES (Y. Chevaleyre)**
  #adversarial models #game theory #deep learning #computational learning
- **SCOOL (E. Kaufmann)**
  #reinforcement learning #bandits #non-parametric methods #privacy

Targeted outcomes & collaborations

Work breakdown structure

1. **WP1: Resilience to data-centric impediments**
   - Robustness against corruptions and contaminations
   - Risk-aware learning with robustness guarantees
   - Adversarial robustness and reliability
2. **WP2: Adaptivity to unmodeled phenomena and the environment**
   - Robust multi-agent learning
   - Learning in coopetitive environments
   - Learning unmodeled structures
3. **WP3: Robustness in the presence of concurrent aims and goals**
   - Fairness-driven trade-offs
   - Privacy-driven trade-offs
   - Robust multi-objective machine learning

WP1: Resilience to Data-Centric Impediments



WP2: Adaptivity to Unmodeled Phenomena and the Environment

	T0–T6	T6–T12	T12–T18	T18–T24	T24–T30	T30–T36	T36–T42	T42–T48
Task 2.1 Robust Multi-agent Learning	**POLARIS PhD:** Robustness to stochastic perturbations in game-theoretic learning							
			POLARIS PhD: Robust learning with self-motivated agents					
			MILES PhD: Bounded rationality in stochastic games					
Task 2.2 Learning in Coopetitive Environments		**FAIRPLAY PhD:** Coopetitive multi-agent learning						
		FAIRPLAY PhD: Fairness in coopetitive multi-agent systems						
			POLARIS PD: Universal algorithms for multi-agent learning					
Task 2.3 Learning Unmodeled Structures		**FAIRPLAY PD:** Learning random structures					**D2.2:** GameSeer	
		FAIRPLAY PD: Matching with learned preferences						
	SCOOL PD: Robust non-parametric algorithms for structured bandits							

D2.1: monograph

D2.3: book

WP3: Robustness in the Presence of Concurrent Aims and Goals

	T0–T6	T6–T12	T12–T18	T18–T24	T24–T30	T30–T36	T36–T42	T42–T48
Task 3.1 Fairness-driven Trade-offs in Machine Learning	**MILES PhD:** Fairness in generative models						**D3.3a:** gen-fair	
	LTCI PhD: Fairness-utility trade-offs Rank-based techniques for fair statistical learning						**D3.3b:** fair-net	
			LTCI PhD: Rank-based techniques for fair statistical learning					
Task 3.2 Privacy-driven Trade-offs in Machine Learning			**ENSL PhD:** Statistical trade-offs of differential privacy					
			SCOOL PhD: Cost of privacy in adaptive testing					
			FAIRPLAY PD: Privacy & incentives for data release			**D3.2:** marketplace simulator		
Task 3.3 Robust Multi-Objective Machine Learning			**FAIRPLAY PhD:** Fairness with privacy in online learning					
			POLARIS PhD: Robust mechanism design for high-stakes applications					
			LTCI PD: Robust multi-objective leanring on graphs					
			SCOOL PD: Robustness to non-compliant agents in RL					

D3.1: rl-berry

Outreach, output & dissemination

Hirings

- 4 PhDs, 1 post-doc (FAIRPLAY, SCOOL
- CNRS, ENSL, Dauphine held back by contracting

Industrial outreach

- Criteo (FAIRPLAY, POLARIS)
- Ubisoft (ENSL)

Output & dissemination

- Leading ML conferences (NeurIPS, ICML,...)
- See posters in the lobby

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030
PROGRAMME DE RECHERCHE
INTELLIGENCE ARTIFICIELLE

Retrouvez toutes nos actualités

anr agence nationale de la recherche