ERC Consolidator Grant 2023
Research proposal [Part B2]

Statistical Reinforcement Learning to Upscale Experimental Sciences. (StaRLux)

I Scientific context, positioning and objectives

I.1 Scientific objectives and motivation

The **StaRLux** project is rooted in Statistical Reinforcement Learning, itself rooted in sequential statistics and reinforcement learning. In sequential statistics, the sample size is not necessarily fixed or known in advance but data is continually acquired. The pioneering works go back to [1, 2] and was groundbreaking at the time in the sense that it allowed statisticians a formalism where they could adjust their decisions continuously with the upstream of new information, hence enabling to reduce the testing budgets and avoid pointless trials. Reinforcement learning theory [3, 4, 5] on the other hand formalized sequential decision-making in uncertain dynamical systems into actual algorithms that collect data through repeated interaction with a partially known system. By "upscaling", we mean to resort to recommender systems [6, 7, 8] and especially contextual approaches, enabling efficient learning on a vast number of related environments as opposed to a single one, and giving birth to personalized recommendations. Last, we focus on experimental sciences that attempt at understanding Nature, as opposed to classical games or robots, in view of filling the gap between simulated and real environments, contributing to handle the significant societal and environmental hazards of our modern society, and shedding novel light on the domains that 90 years ago gave birth to mathematical statistics.

The **StaRLux** project is organized around the three following complementary and ambitious objectives:

Objective 1: Sample-efficient RL for experimental environments, rather than games, that can tackle crucial limitations faced by practitioners, such as managing the intrinsically *stochastic* nature of considered environments that feature *diverse* entities interacting with each other, determining the *optimal moments* to intervene and observe, and adapting to the *ever-changing experimental contexts*, such as a shift in the dynamics and novel available actions or practitioners' goals.

Objective 2: Scalable experimentation with group-RL to cope with a vast array of environments and handle the *diversity of practitioner objectives* regarding horizon, risk, constraints, and rewards, *chart* in which contexts promising *practices* discovered and refined by many are indeed beneficial, and perform efficient *batched coordinated exploration* under personalized or joint risk aversion.

Objective 3: Applicable RL companions in agroecology. This involves designing automated procedures to output statistically *reliable rankings* based on personalized goals, providing *explainability* of the strategies to practitioners to improve trust and compliance, and ensuring *adoption by practitioners* by integrating RL companions into an open-experimentation platform ready to be used by many.

I.2 Scientific challenges and novelty of the project

Though Reinforcement Learning (RL) seems to be the champion of formalizing sound decision-making in the wild, paving the way to general A.I., its current theoretical development appears to fall short on correctly handling some seemingly basic challenges from a practitioner standpoint, hindering its adoption in experimental sciences. We list below the challenges (key bottlenecks) that we consider the most significant ones to upscale experimental sciences, yet not insurmountable within the scope of this project.

Challenges in addressing objective 1

[C1] Stochastic environments. In the recent years Deep Learning emerged as a powerful set of parameter tuning techniques for large to huge parametric models, contributing to reduce the *approximation error* of modeling complex systems. Despite greatly benefiting RL in terms of exposure, attracting many efforts and hope, the apparent successes of deep RL have however so far been limited to near deterministic environments such as games or robotic control tasks. In stark contrast, environments considered by experimental sciences such as agronomy of healthcare are usually intrinsically stochastic. Indeed unlike more traditional environments (e.g. go, atari), the process of growing a plant in a farm is naturally *stochastic* rather than deterministic, due to the key influence of external factors such as the weather (but also population of insects, spontaneous plants, etc.). The challenge is that much statistical theory remains to be developed to reduce *estimation error* and correctly handle stochastic uncertainty. For instance, there is a major lack in our understanding of optimal achievable regret minimization performance beyond *ergodic* stochastic MDPs, for which every policy is assumed guaranteed to visit all states with positive asymptotic frequency. This is problematic, because real-world situations such as medical treatment trajectories often involve MDPs that are non-ergodic, meaning that some states (e.g. treatment phases) may only be visited a finite number of times asymptotically, even by an optimal policy. Hence correctly learning stochastic dynamical systems, possibly combining bandits and deep approaches is a major challenge of reinforcement learning hindering its applicability to experimental sciences.

[C2] Structured interactions and auxiliary information One major challenge in using RL in experimental sciences is to handle the large diversity of influential factors due to many co-interacting entities. Often some expert knowledge is available that enables to prune the graph of direct interactions and reveals a structure where a limited number of key state variables govern the others. Leveraging such knowledge in reinforcement learning algorithms can greatly improve the learning performance of the agents, but more research efforts are required to obtain a satisfactory strategy. In some cases these influential variables should further be identified by the learning agent, due to uncertainty, which represents a statistical challenge given the limited amount of interaction available with a real system, as opposed to a simulated world. Moreover, in many real-life applications, experts have access to prediction models that describe various specific variables, such as weather patterns or the spread of pests. Though progress has been made in developing robust RL methods that can handle ambiguity in these models, traditional studies often focus on models that describe observed data rather than predicting a noisy future evolution of certain variables. It is important to explore how the literature on bandits, robust MDPs, and expert aggregation can be adapted to leverage or discover *influential variables*, and incorporate *forecaster models* into the learning process. Additionally, expert knowledge can take various other forms, and there is a need to develop methods for incorporating generic auxiliary information theory in a sequential learning setting, where the data is collected over time rather than all at once.

[C3] When to intervene, when to observe. Another major challenge in RL theory is how to effectively make decisions about when to take action over a period of time. For instance, consider a scenario where an agent has T days to make a decision about which one of A many actions to perform. While it may seem intuitive to introduce a do-nothing action and treat the problem as a special case of sequential decision-making with $A+1$ actions, this naive approach leads to an exponential increase in the number of possible trajectories after T days, from $AT+1$ (one single action at any of the T possible days, or no action at all) to $(A+1)^T$. This highlights the need for considering alternative models better effective at handling decision-making over time than classical MDPs. A related natural challenge for the practitioner is the need to trade-off between gathering observations and taking actions, that is deciding when to observe. While in fully-observed MDPs, the learner is given the exact state of the system at each decision step, and in partially-observed MDPs, the learner is given a fixed set of observations at every time, we here face an actively-observed MDPs. Such situations are common in autonomous systems that evolve continuously, such as when modeling a field or a human in an intensive care unit. In these cases, taking measurements is costly in terms of time or resources, and must be balanced with the need for intervention. This creates a trade-off between gathering more information through observations and taking actions that can potentially lead to better rewards. In addition, the state of the system

is often multivariate, meaning that it can have many different components that are not all visible at each time step. This creates a monitoring challenge because the learner must decide which variables to observe in order to get enough information to learn about the unknown dynamics of the system and eventually act optimally. To summarize, we need to develop strategies for effectively addressing the trade-off between intervention and observation in models better effective at handling decision-making over time than classical MDP models.

[C4] Shifting environments and progressive learning. One major challenge that hinders applicability of RL is the ability to adapt to continually changing environments and goals. In experimental sciences, the environments and goals are often highly specialized and domain-specific, and may not be directly comparable to those studied in other fields. Additionally, experimental data is often scarce and expensive to collect, making it difficult to train models from scratch. Furthermore, the dynamics of the system and the practitioners' goals may shift over time, due to factors such as local climate evolution (for agroecology) or changes in research focus, requiring the RL algorithm to adapt accordingly. Another common source of shift is a change in available actions, for instance due to novel practices becoming available or shared by the community, which raises the question of how to effectively learn tasks in a *progressive* manner, similar to how humans learn. One approach is to consider a set of basic actions that are refined over time with more subtleties or variants as the learner masters them. However, when modeling the task as an MDP, this approach would require specifying the gigantic set of all elementary actions right from the start, which can be overwhelming for the learner due to the corresponding huge policy space. An alternative approach is to start learning with a small set of policies and progressively enrich them over the learning process, which may be more manageable for the learner. However, this approach raises the question of how to effectively and efficiently refine the policy space over time in order to maximize learning performance.

Novelty: To summarize the novelty in the challenges when addressing objective 1, there is

- Adapting bandit theory to MDPs and to revisit deep-learning strategies for stochastic systems.
- Incorporating auxiliary information from domain experts such as variable prediction and leveraging or discovering structure of influential variables in a learning context.
- Addressing the exploration-exploitation trade-off in control models alternative to MDPs (rested), better suited at handling autonomous (restless) dynamical systems and decision over time.
- Refining and incorporating novel actions and objectives continuously in the learning process.

Challenges in addressing objective 2

[C5] Context-goal exploration-exploitation trade-off. To enable experimentation companions, one major challenge in RL theory is how to efficiently handle the exploration-exploitation trade-off when facing multiple environments and *personalized objectives*. In real-life situations, different practitioners may interact with different environments, and even when facing the same environment, they may have different objectives such as different time-horizon, risk-aversion levels, or reward functions. To address this challenge, it is necessary to consider coordinated exploration to extend the concept of contextual reinforcement learning from task context to task and objective context. This involves challenges such as exploring in one environment to gather information that is relevant to another practitioner's objective, in a collectively optimal way.

[C6] Charting good practices across different contexts. A related challenge when considering multiple environments is that of *hypothesis charting*. In hypothesis testing, a learner is presented with several possible explanations for a phenomenon (called "hypotheses") and must use data collected from the environment to determine which hypothesis is the most likely to be true. In hypothesis charting, the learner is presented with a collection of environments and must determine which ones satisfy a particular hypothesis. This can be seen as a more generalized form of hypothesis testing, where the learner must consider multiple environments rather than just one. The concept of hypothesis charting is relevant in situations where a new action becomes available, and the learner must rapidly determine in which contexts the action would be most beneficial, with provable guarantees.

[C7] Batch coordinated exploration and joint-risk. Another major practical challenge for RL is how to efficiently balance exploration and exploitation when solving multiple tasks in batch, rather than a single task. This is known as *group-sequential* reinforcement learning, and it is particularly relevant in situations where rewards are delayed, such as in agronomy. To tackle this challenge, it is necessary to develop new algorithms or techniques that can compare the performance of different policies or algorithms across tasks, and efficiently balance exploration and exploitation in this setting. This is related to the problem of group-sequential adaptive testing in clinical trials [9, 10], which is a powerful methodology based on bandit theory that is still an active area of research. A key question in this area is how to optimally encourage diversity of actions within a single batch in order to speed up learning. Another challenge related to multiple environments is how to effectively handle situations where the risk of a particular action depends not only on the local context, but also on the contexts of neighboring entities. This is known as a "*joint risk*" and can be seen in scenarios such as food security, where the risk of crop disease increases when too many similar crops are grown in close proximity, or in forest management, where the risk of fire increases when too many similar trees are packed together. This creates a non-trivial tension between recommending diverse versus similar policies in similar contexts. While it is possible to incorporate information about the contexts of neighboring entities into the state space, it is not clear how to optimally exploit this structure in order to mitigate joint risk. Additionally, situations where the contexts are correlated, such as when the yields of various fields in the same region are subject to the same weather conditions, can also pose challenges for learning and may introduce bias.

Novelty: To summarize the novelty in the challenges when addressing objective 2, there is
- Revisiting the exploration-exploitation trade-off under contextual, personalized objective.
- Extending best policy identification and hypothesis testing literature to the contextual setup.
- Advancing group-sequential exploration theory of dynamical systems and enforcing diversity.

Challenges in addressing objective 3

[C8] Reliable, reusable rankings One of the major challenges in building a research platform to assist the development of reinforcement learning algorithms is establishing a *statistically reliable*, comparison of strategies. This is important for researchers testing their algorithms, for the community in terms of reproducibility standards, and for practitioners who rely on recommendations from various strategies. However, obtaining a statistically reliable comparison can be difficult, and is a problem that has been recognized as a reproducibility crisis in the field. It is for instance common for articles to claim that a new reinforcement learning strategy outperforms a state-of-the-art approach based on results from only a few random trials. While it makes sense to average performance over multiple trials on the one hand, and to minimize the number of trials in order to reduce time and energy consumption on the other hand, addressing this trade-off can be a challenge when the distribution of performance does not follow a parametric law, which is typically the case for the performance of RL algorithms. As a result, there is a need to adapt non-parametric hypothesis testing strategies, such as the ones based on permutation-tests, to the fully sequential case. However, the theory in this area is not yet complete, and more work is needed to develop provably optimal evaluation strategies ready to be used.

[C9] Explainability of RL companions and event prediction. When building an experimentation companion platform to assist practitioners, one of the key challenges is ensuring the *explainability* of the recommended solutions. Indeed reinforcement learning strategies can often be complex and may appear as black boxes, especially when they involve deep learning models or other types of highly sophisticated algorithms, which can hinder trust and compliance from practitioners. One way to address this challenge is to provide explanations for the recommendations, such as by showing the predicted evolution of the system when following or not following a recommended policy, and by highlighting the risks that are being avoided. However, accurately predicting the evolution and occurrence of risky events in a stochastic and generic (non-linear) dynamic system is a daunting and elusive challenge, and finding innovative and effective ways to do this is an exciting and little explored frontier in the field of reinforcement learning.

[C10] Availability of sequential data and practitioner engagement. One of the main challenges when using RL techniques in experimental science is the availability of data. In healthcare, stringent

privacy restrictions usually prevent from building global experimentation datasets. Data collection is less restrictive in agronomy, but often the available data is primarily focused on identifying different objects, such as identifying different types of crops or pests, rather than on the evolution of a dynamic system over time. For example, in traditional agronomy research, data is collected on various factors such as soil properties, weather conditions, and specific crop treatments, but it may not contain data on the evolution of the crop over time. This lack of data on the dynamic evolution of the system can make it difficult to apply RL techniques, which typically rely on a large amount of data on the interactions and outcomes of different actions in order to learn and optimize decision-making processes. Hence, an alternative approach is required to build sequential experiment datasets. Besides, for RL to become interesting for practitioners, another challenge is to engage enough researchers in RL and attract significant focus from them on such environments and tasks in a reproducible and open-science friendly way. Building a community of RL researchers and agronomist constantly improving and testing RL companions could have a major impact on the real adoption of RL companions in agronomy, and experimental sciences in general. This also requires designing novel recommender systems, to move away from traditional content recommendation in static systems to enable experimentation recommendation in dynamical systems.

Novelty: To summarize the novelty in the challenges when addressing objective 3, there is

- Designing adaptive sampling strategies with finite-time guarantees for non-parametric ranking of learning agents under personalized objectives.
- Predicting trajectory evolution and risk occurrence in a complex dynamical system, including under counter-factual reasoning.
- Deploying RL simulators for large stochastic systems and a recommender platform specialized for experimentation companions in dynamical systems.

I.3 State of the art and its limitations

The existing literature can help address these novel challenges but also face some key limitations.

Literature relevant to objective 1

Limitations: The limitations to overcome for addressing the challenges of objective 1 are

- Lack of theoretical understanding of regret minimization and sensitivity analysis in stochastic dynamical systems.
- Lack of theoretical understanding of optimal exploitation of various state structure in stochastic dynamical systems. Auxiliary information theory is restricted to offline or sequential but static environments such as bandits.
- Lack or RL approaches (assuming uncertain system) for various control models (assuming perfect knowledge) of autonomous systems.
- Lack of understanding of progressive learning mechanisms and available models for action-refinement suitable for large dynamical systems.

[L1] Stochastic environments: The typical RL environments available to researchers [11] are either complex, deterministic games (Go [12], Chess, Atari [13]), or deterministic control tasks (acrobot [14], mujoco [15], robots [16]). As a result, the literature on complex systems mainly developed around approximation error rather than estimation error due to finite interaction, while the study of stochastic environments has restricted to small discrete games (grid-worlds), extending ideas from multi-armed bandit (seen as single-state MDPs), see [17, 18, 19],**[20, 21, 22]**. Today, sequential design of experiments (bandits) for stochastic dynamical systems is arguably still in its fancy, considering that achievable performance bounds for generic MDPs are still unclear (see [23, 24, 25] for ergodic MDPs) and first-order optimal strategies have only been uncovered recently **[26]** but under the restrictive ergodicity assumption. Stronger notions of optimality such as Blackwell optimality [27, 28, 29] are still not met. Besides, even though distributional RL [30, 31] and deep-belief networks [32, 33, 34] appear as promising tools, there is a lack of research efforts on provably efficient learning strategies in large stochastic environments.

[L2] Auxiliary information: Systems with many state variables often come with significant structure that may be known to the domain expert. Leveraging factored [35], **[36]**, [37], causality [38, 39, 40] or combinatorial [41] structures may greatly reduce the complexity of the learning task, but are still active research topics. Beyond exploiting a given structure, related open questions include detecting control [42] or discovering which are the most influential state variables. Another type of expert knowledge especially relevant to experimental sciences may be the forecast of some variables (e.g. weather). Preliminary studies have been done in multi-armed bandits [43, 44] to leverage such information on the future, that need to be improved and extended to dynamical systems. Besides, promising tools have been made recently available in mathematical statistics on leveraging generic auxiliary expert information [45], that however need to be extended to the sequential, active setup and to dynamical systems.

[L3] When to act: MDPs [27] are ubiquitous in RL but do not model well decision over times such as when to observe or to intervene. In control, MDPs are just one model amongst others, and other models better capture time. For instance Piecewise Deterministic Markov Processes (PD-MP) [46, 47, 48] model autonomous (restless) systems subject to spontaneous or controlled jumps of dynamics, where the task it to optimally time each controlled jump. These restless model make them well adapted to evolution of natural systems. However, PD-MPs have not been studied from a learning perspective, when the model is unknown and learned from data. Now, deciding optimal moments to observe is reminiscent of sensor selection tasks [49], however studied in a bandit and not MDP context, or patrolling tasks [50, 51], considering an adversarial or sometimes stochastic [52] hide and seek game setup. These approaches are promising but lack theoretical guarantees and need to be adapted to leverage the structure of influential variables to be patrolled.

[L4] Progressive learning: As practitioners test more and more hypotheses, they also introduce or refine specific techniques progressively rather than considering all possible sequence of micro-actions. This way of proceeding, with a growing set of actions can be backed-up by theory, as ultimately considering all micro-actions yields computational limits, apparent in the lower performance bounds [25] making appear challenging optimization problems, or in the fundamental limits of Bayesian induction [53, 54, 55]. Besides, it can be linked to the topic of satisficing RL [56, 57], from satisficing theory [58] and control [59], where the goal is no longer to be optimal but to outperform some base policy or exceed some target value. Despite promising recent progress in the bandit setup [60], for regret minimization, and in online aggregation of experts with growing actions **[61]**, much remains to be done to understand how to provably-efficiently solve a satisficing objective and then reuse past computations when incorporating a novel action in a series of such tasks.

Literature relevant to objective 2

Limitations: The limitations to overcome for addressing the challenges of objective 2 are

- Limited understanding of exploration-exploitation trade-off in contextual MDPs, and of risk-averse optimal strategies in dynamical systems.
- Limited understanding of sequential hypothesis testing and identification theory for contextual, dynamical systems.
- Lack of group-sequential diversity-enforcing strategies and limited understanding of joint-risk aversion and mitigating strategies in stochastic dynamical systems.

[L5] Context-objective exploration-exploitation: Contextual bandits [62, 63] and contextual MDPs [64, 65, 66] are natural framework to efficiently learn from many environments, especially under the linear-context structural assumption [67, 68], **[69]**. However, so far the literature mostly explored handling a single objective, usually risk-neutral for all environments. Little seems to be known about jointly addressing diverse objectives, for instance when a single agent should solve two or more environments under different risk-aversion levels. Hence the literature needs to be extended to incorporate context-goal. Besides, risk-aversion has received increased attention both in multi-armed bandits **[70]**, [71], [72], **[73]**, [74] see also [75, 76], **[77]**, and in MDPs [78, 79, 80] over the past few years. However, more work is required to solve the exploration-exploitation trade-off under diverse notions of risks, including human-perception, in these two settings.

[L6] Action charting: The goal in action charting is to quickly identify the contexts in which a novel action may outperform existing best ones in a bandit setting, or may improve existing policies in an MDP setting. A variant of this problem is the Best-arm identification (BAI) problem, [81, 82, 83] that focuses on quickly identifying the best action to be recommended after learning, rather than solving the exploration-exploitation trade-off, and is reminiscent of sequential hypothesis testing. This problem developed in the bandit community, and received increasing attention both in contextual bandits [84, 85, 86] and MDPs [87, 88, 89]. While related, BAI and action charting are different problems and borrowing BAI techniques for the later requires further investigation.

[L7] Batch coordinated exploration: Practitioners in experimental sciences often use group-sequential strategies, proceeding with batch/cohorts rather than fully sequential strategies due to the delays in receiving feedback (e.g. yield in agronomy). The literature on group-sequential experimental design is rich [90, 91, 92]. Originated from design of experiments (DOE) theory [93, 94, 95, 96], the emergence of multi-armed bandits (aka adaptive experimental design) [97, 98] enabled to optimize rather than simply randomize trials, and later permeate to medical sciences and group-sequential clinical trials [9, 10]. Over the last decade, strongly optimal bandit strategies [99, 100], **[101, 102]** have developed in the bandit community, yet for the fully sequential setting. As suggested by **[103]** adapting modern bandit strategies to group-sequential experimentation may improve over the simple explore then commit strategies often used in agronomy, including in a risk-averse context. Developing strategies optimal for the group-sequential setting that are accepted by practitioners in agronomy however requires further investigation. In particular, the uncertainty about the environment comes with aversion to joint-risk, such as in food security, where the risk of a crop disease increases when too many similar crops are grown in close proximity, or in forest management, where the risk of storm damage or fire increases when too many similar trees are packed together [104, 105]. The literature on diversity in recommender systems [106, 107, 108] may address such concern and has motivated extensions to bandits [109], but combining them with optimal bandit strategies and proper risk-aversion analysis remains non trivial, not to mention their extension to MDPs.

Literature relevant to objective 3

Limitations: The limitations to overcome for addressing the challenges of objective 3 are
- Lack of finite-time guarantees for non-parametric sequential hypothesis testing methods.
- Lack of confidence region around trajectories and reliable risk-occurrence prediction tools in non-linear and autonomous dynamical systems.
- Lack of stochastic environment simulator for RL researchers, of recommender system to assist experimental sciences and limited availability of sequential-interactions from practitioners.

[L8] Reliable, reusable rankings and user adoption: In experimental sciences, p-values seem ubiquitous and often not correctly used, in particular when one want to reuse data or previous conclusions to guide a new experiment. The statistics literature has developed alternative concepts, such as S-values [110] and E-values [111, 112, 113], to overcome such a practical issue. However, revisiting adaptive hypothesis testing using such sounder notions is subject to active research. Now to compare various bandit, RL, or recommendation strategies, it seems natural to use sequential DOE to stop as early as possible and identify which strategy is best, contrasting with the statistically unreliable comparisons often done in the deep RL literature [114]. However, optimal sequential DOE strategies usually assume a parametric score model, whereas the distribution of performance of a learning strategy rarely meets a classical shape. This literature could be extended to non-parametric distributions, following the recent efforts in **[102]**, [115] based on sub-sampling. Besides, the rich literature on non-parametric tests [116, 117, 118, 119] often lacks the sharp finite-analysis required in the bandit analysis. Adapting it to the sequential setup is promising but non-trivial, as it requires sharper, non-parametric concentration results. What is also missing is a way to popularize and rapidly spread such tools in the RL community.

[L9] Explainability of RL companions: Explainability has received increasing attention in machine learning over the last few years [120, 121], including in RL [122]. Beyond explainability of the

algorithms and learning process, prediction of effects or counter-effects has been considered [123]. Prediction of trajectories in an stochastic dynamical system requires to handle the propagation of uncertainty. Combination of interval-prediction and confidence-region methods have been developed recently **[124]** in planning of linear systems. These promising methods need to be extended to handle the non-linearity and autonomous nature of systems considered in experimental sciences, for instance to PD-MPs.

[L10] Availability of sequential data, practitioner engagement: In both agronomy [125] and clinical trials [126], it is acknowledged that experimentation should be significantly upscaled, hence resort more to machine learning and decision companions. Moving from pure algorithms to real practical application is known to be a tedious path [127], especially in the web-industry where many daunting challenges are caused by the nature of the reward feedback, defined by human preference, which turns out to be highly unreliable, non-stationary nor reproducible, changing with mood, time of the day and many latent factors. In experimental science such as agronomy, the feedback is generated by Nature, which limits such obstacles, however the environments are complex and actions are risky. We provided in **[128]** an overview of the scarce literature and of the many challenges of applying RL to support decision management in the field of agroecology. Agronomist often resort to decision support systems (DSS) developed and used by agronomists over decades. However, field data is usually collected to calibrate simulators, rather than feeding a learning agent. We have recently turned one of these simulators into a sequential agronomy environment **[129]** following the openAI-gym approach, and released a complementary set of stochastic agronomy-games **[130]** to enable RL researchers to experiment on some of such challenges in an open and reproducible science friendly way. While RL competitions [131] are considered a good way to stimulate research on novel challenges, more work is required to embed these environments in an RL competition platform and finalize their development. Furthermore, existing simulators in agronomy are often monolithic, highly-specialized on modeling certain aspects, but partial in modeling the various interacting entities of a real system for now. Hence, no simulated model can realistically model the real agrosystem. On the other hand, existing recommender platforms in agronomy focus on plant or pathogen identification [132], and there exists no open-science recommender platform dedicated to assist experimentation in large dynamical systems including in agroecology.

II Methodology, scientific program and management

II.1 Project organization overview

At a high-level, we organize the project into three work-packages (WP) each consisting of several tasks (T). We provide below the overview, then detail the human effort, Gannt diagram and other activities.

Task overview

WP1 **Unlocking RL efficiency for experimental environments** (rather than games)

[T1a] Stochastic environments.	[PhD1]
[T1b] Structured interactions and auxiliary information.	[PhD1]
[T1c] When to intervene, when to observe.	[PhD2]
[T1d] Shifting environments and progressive learning.	[PhD2, Postdoc1]

WP2 **Scaling-up experimentation with group-RL**

[T2a] Context-goal exploration-exploitation trade-off.	[PhD3, Postdoc2]
[T2b] Charting good policies across contexts.	[PhD3]
[T2c] Batch coordinated exploration and joint-risk.	[PhD4, Postdoc2]

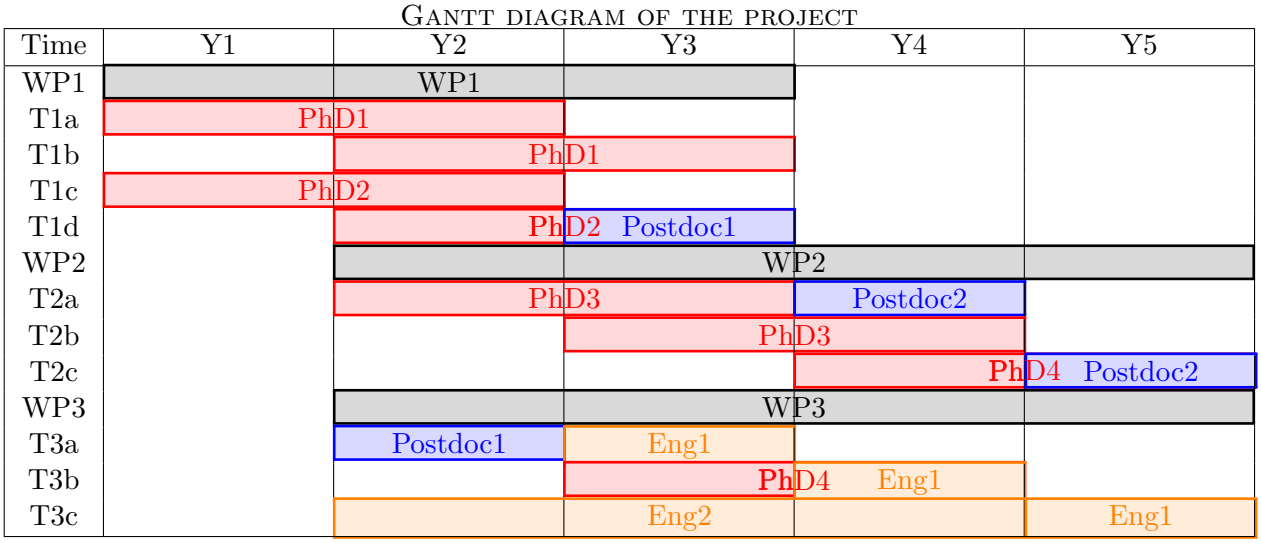
WP3 **Making RL companions applicable in practice**

[T3a] Reliable, reusable rankings and RL coopetition.	[Postdoc1, Engineer1]
[T3b] Explainability of RL companions and event prediction.	[PhD4, Engineer1]
[T3c] Agro-recommendation platform and practitioner engagement.	[Engineer2]

Human effort

PhD1 will work on tasks T1a and T1b. These are about fundamental challenges of MDPs and are more suitable to a candidate student interested in RL theory.
PhD2 will work on tasks T1c and T1d. These are about alternative learning models, and are also more suitable to a candidate student interested in RL theory.
PhD3 will work on T2a and T2b. These are about contextual versions of sequential design of experiments and hypotheses testing, and are more suitable to a candidate student interested in statistics theory.
PhD4 will work on tasks T2c and T3b. These are about batch exploration and explainability and are more suitable to a candidate student interested in applications.
Postdoc1 (24months) will work on task T3a about statistically reliable rankings, both to develop the statistical theory and assist in organizing an RL challenge using such results. He/She will also work on T1d, about the progressive learning, especially about reliable shifting between two action representations. The postdoc, specialized in statistics, will also collaborate with PhD2.
Postdoc 2 (24 months) will work on task T2a about contextual approaches to exploration, especially in large MDPs, and on task T2c about batch coordinated exploration in RL, considering more realistic environments and constraints. The postdoc, specialized in RL, will also collaborate with PhD4.
Engineer 1 (36 months) will be in charge of maintaining and improving the RL python environments as well as creating the software infrastructure for the RL coopetition. A related task is to enrich the RL software library with efficient implementations of novel ranking strategies and explainability methods. Engineer1 is expected to interact especially with both Postdoc1 and Postdoc2.
Engineer 2 (36 months) will be in charge of maintaining and developing the software platform intended to be interfaced with practitioners. This includes developing the front-end and feedback system to smoothly scale to many practitioners, as well as interfacing with the RL companions algorithms developed in the project. In particular, such algorithms must feed on realistic data input and be scalable. Engineer2 is expected to interact with the whole team as well as external collaborators, especially in agroecology.

Gantt diagram of the project



Other activities

Workshop (RL Challenge): We plan to organize an RL challenge hosted in top-tier conferences of the field, such as ICML or Neurips, on the simulated agroenvironments we have developed and made available to the field, such as gym-DSSAT and FarmGym. We plan to host the challenge every year until the end of the project. We will start on Year1 a simple version, using end-resources of side projects, and will continue from Year2 considering an enriched challenge with the help of the postdocs and engineers hired on the project, and to a lesser extent of the PhD students.
Invitations: The P.I. will organize visits of/from external collaborators every year to strengthen the project. This includes mostly researchers outside the field RL, such as working in control of autonomous systems and in agroecology or experimental trials.

II.2 Detailed work-plan

WP1 **Unlocking RL efficiency for experimental environments** (rather than games)

[T1a] Stochastic environments.

We intend to strengthen the development of RL methods for stochastic environments, extending [26] from ergodic to generic MDPs and leveraging results from perturbation analysis [29] and higher-order optimal control [28] to better handle the bias function hence build more effective model-based RL methods. Besides, we will continue the efforts in injecting promising bandit strategies, such as [100], [133] both in tabular and approximate algorithms, to develop hybrid deep-bandit strategies for stochastic environments. Last, will also investigate how to extend the literature on RL in deterministic environments, such as complex games and deterministic control tasks, to stochastic environments incorporating popular Bayesian approaches. This research will contribute to the growing body of literature on RL in stochastic environments and help practitioners develop methods for RL in these challenging environments.

[T1b] Structured interactions and auxiliary information.

To efficiently handle the rich diversity of interactions, we propose to investigate the problem of leveraging auxiliary information in RL for systems with many state variables. Our focus will be on leveraging the structure that may be known to the domain expert, such as factored, causality or combinatorial structures, to reduce the complexity of the learning task. We will investigate open questions related to exploiting given structure [36], such as detecting control or discovering the most influential state variables. We will also investigate the use of expert knowledge, such as forecasts of some variables, to improve RL performance. We will build upon preliminary studies in multi-armed bandits [43] to leverage such information on the future evolution of some variables, and extend these methods to dynamical systems. Our proposed methodology will include the extension of recent promising tools in mathematical statistics for leveraging generic auxiliary expert information [45] to the sequential, active setup and to dynamical systems. This research will contribute to the growing body of literature on leveraging auxiliary information in RL and help practitioners improve the performance of their RL algorithms by leveraging the available expert knowledge.

[T1c] When to intervene, when to observe.[External collaborator: B. De Saporta (U. Montpellier)]

To investigate the problem of when to act in RL, specifically in the context of decision over time such as when to observe or intervene. Our first focus will be on the use of models other than MDPs that better capture time such as Piecewise Deterministic Markov Processes (PD-MPs) which model autonomous (restless) systems subject to spontaneous or controlled jumps of dynamics. Our proposed methodology will include the investigation of PD-MPs from a learning perspective, when the model is unknown and learned from data. We will also investigate the problem of deciding optimal moments to observe and build upon the related literature on sensor selection tasks and patrolling tasks, which are promising but lack theoretical guarantees. We will also extend these methods to leverage the structure of influential variables in PD-MPs from T1b. This research will contribute to the growing body of literature on timing actions in RL and help practitioners to develop methods for making optimal decisions over time in RL.

[T1d] Shifting environments and progressive learning. [External collaborators: R. Ortner (U. Leoben), J. Honda (U. Kyoto)]

To address the progressive availability of actions, for instance due to novel practices becoming available or shared by the community, we suggest to study how to effectively refine the policy space over time in order to maximize learning performance, while efficiently reuse past computations in a series of such tasks. Hence, we will focus on satisficing rather than optimization criterion. To accomplish this, we will build upon recent progress in the bandit setup for regret minimization, in online aggregation of experts with growing actions. Besides, our proposed methodology will exploit the computational limits and lower performance bounds of considering all possible micro-actions versus refined ones to eventually suggest novel strategies. Along the way, we intend to contribute to the related satisficing RL theory in both bandits and MDPs. Regarding continual learning in shifting objectives, under similar but stochastic dynamics, we will adapt transfer knowledge techniques to revisit the exploration-exploitation challenge.

Overall, we aim to provide a theoretical foundation that practitioners can use to guide their testing of hypotheses and refinement of specific techniques in RL problems with a growing set of actions.

WP2 Scaling-up experimentation with group-RL

[T2a] Context-goal exploration-exploitation trade-off.
We aim to extend the literature on contextual bandits and contextual MDPs by exploring the concept of context-goal. Specifically, we propose to investigate the problem of jointly addressing diverse objectives, such as different risk-aversion levels, across multiple environments under the linear-context structural assumption. To address this problem, we will revisit the exploration-exploitation trade-off to understand how to explore in one environment to gather information that is relevant to another practitioner's objective, in a collectively optimal way. This can be done extending lower-bound techniques to this problem. Besides, we will also contribute to the literature on risk-aversion in MDPs, extending our work [73, 103] from bandits to MDPs and to the contextual case. Overall, this task will fill a gap in the literature by providing a framework for efficiently learning from many environments with diverse objectives tailored to the preferences of the decision-maker, and will contribute to the growing body of research on risk-aversion in RL.

[T2b] Charting good policies across contexts.
We propose to investigate the problem of action charting in both bandit and MDP settings and develop novel techniques for quickly identifying the contexts in which a novel action may outperform existing best ones in a bandit setting, or may improve existing policies in an MDP setting. To accomplish this, we will build upon recent progress in the related problem of Best-arm identification (BAI) from the bandit community, sequential hypothesis testing, as well as contextual bandits. We will investigate the potential of borrowing techniques developed in these fields to solve the action charting problem, including under risk-averse objectives. Our proposed methodology will include a comparison and contrast of contextual BAI and action charting, and a further investigation of the relationship between these two problems. This research will develop novel strategies to enable contextual recommendation of best policies.

[T2c] Batch coordinated exploration and joint-risk. [Exertnal collaborators: S. Couture (Inrae), E. Sabbadin (Inrae), R. Gautron (CIAT)] We propose to investigate the use of group-sequential strategies in the context of delayed feedback in experimental design. We will build upon the rich literature on group-sequential experimental design, which has its origins in design of experiments (DOE) theory, and shares many links with multi-armed bandits. Our research will focus on developing novel strategies that are optimal for the group-sequential setting, and that are accepted by practitioners in fields such as agronomy, food security, and forest management. We will investigate the potential of adapting modern bandit strategies to group-sequential experimentation to improve over the simple explore then commit strategies often used in agronomy, comparing both approaches on simulated and possibly real cases. Our proposed methodology will include an analysis of the uncertainty about the environment and the aversion to joint risk. In particular, we will investigate the potential of combining recent research on the topic of diversity in recommender systems with optimal bandit strategies and proper risk-aversion analysis, to develop a joint-risk-aware RL algorithm that can allow for more efficient and safe experimentation. Besides, we will keep regular discussions with practitioners to ensure the development of realistic strategies. This research will contribute to the growing body of literature on batch coordinated exploration in experimental sciences, strengthen links with practitioners and help them optimize their experimental design in the presence of delayed feedback.

WP3 Making RL companions applicable in practice

[T3a] Reliable, reusable rankings and RL coopetition
In order to investigate the problem of reliable and reusable rankings of experimental companions, our first focus will be on using sounder notions, such as S-values and E-values, to overcome the issues associated with the ubiquitous use of p-values in experimental design. To overcome the challenge due to the non-parametric nature of the distribution of performance of these strategies, we will focus on developing

optimal SDOE strategies for non-parametric distributions, building upon recent efforts in sub-sampling and permutation-test theory. Our research will adapt the rich literature on non-parametric tests to the sequential setup, which requires sharper, non-parametric concentration results. This research will contribute to the growing body of literature on reproducibility in machine learning and reinforcement learning, and help derive sounder conclusions from comparison experiments. Our proposed methodology will include the finalization of the development of stochastic agronomy simulators that we have recently enabled as openAI-gym environments for RL researchers, and the development of an RL competition platform based on these environments, at the heart of which we implement the ranking mechanism. We will organize a workshop associated to the coopetition challenge, in ML conferences such as ICML or Neurips. If successful, we will continue organizing it each year until the end of the project.

[**T3b**] **Explainability of RL companions and event prediction.**
To adress the challenge of explainability of RL experimental companions, our first focus will be on developing methods for predicting effects or counter-effects of actions in stochastic dynamical systems. We will build upon recent advances in explainability in machine learning, including in RL, and extend these methods to handle the non-linearity and autonomous nature of the considered systems, such as PD-MPs. We will investigate the use of interval-prediction and confidence-region methods for prediction of trajectories in these systems, initially developed for planning in linear systems. We will explore other approaches from explainable AI relevant to practitioners. This research will contribute to the progress in explainable RL, and will enable the creation of a collaborative project with researchers experts in agroecology to actually test the explainability of the RL companions in realistic use-case.

[**T3c**] **Agro-recommendation platform and practitioner engagement.** To move closer to actual practitioner engagement, we intend to develop a recommender system platform specialized on experimental companions for agroecology. We will extend the WeGarden platform initiated in the SR4SG project to a recommender platform eventually able to interact with practitioners, collect feedback and train RL companions on data. We will however not collect data within the scope of **StaRLux**, and mainly develop the pipeline, with flawless integration of RL companion algorithms and practitioner interface, using data generated from our simulated agronomy environments instead of from real experiments. Meanwhile, we will intensify our discussions with practitioners in agroecology to make the software user-friendly. This software-development part will enable to integrate theoretical RL strategies into a realistic, open-source recommender-system and will contribute to filling the gap between fundamental and applied research. Besides, it will enable the preparation of collaborative projects with researchers in agroecology to test the experiment companion via the platform on real experiments with many practitioners by the end of the project.

II.3 Expected results and indicators of success

We list below, for each work package, the indicators of success (S) we believe are the most relevant.

Indicators of success for WP1

[**S1a**] Obtain novel sound RL strategies to interact with stochastic dynamical systems, unveil sound principles to mix bandit and deep RL strategies and obtain reliable performances in large stochastic MDPs, effectively handle alternative models such as for the timing of actions in autonomous systems, incorporating novel actions into learning or adapting to changes. Solve the theoretical challenges both with sound theory and experiments in increasingly challenging controlled environments.

[**S1b**] Understand exploiting domain expert knowledge such as influence structure and forecast evolution, both theoretically and experimentally on moderately complex systems, and handle examples from actual practitioners. Unveil novel methods and theory for active discovery and observation of influential variables in uncertain dynamical systems, able to operate efficiently on complex simulated environment.

[**S1c**] Theoretical developments that yield RL strategies able to provide reliable advice under sufficiently realistic practitioners conditions to envisage comparison with and adoption from practitioners.

Indicators of success for WP2

[S2a] Significant development of the theory of sequential DOE and hypothesis testing, especially regarding both contextual and group-sequential aspects in dynamical systems, with sound guarantees.
[S2b] Obtain beneficial comparison of novel experimentation strategies against practitioners' on complex simulators, including when considering risk and group risk aversion.
[S2c] Spread our results beyond the community of RL researchers and be convincing enough that agroecology researchers start using our group-RL approach. This would be a critical indicator of success.

Indicators of success for WP3

[S3a] Derive novel sound adaptive testing procedures to strengthen reproducibility and sound comparison in RL, release them in an open-source ranking system and test and diffuse them in an RL challenge using simulated agronomic games.
[S3b] Unveil novel explainability techniques in dynamical systems, suitable for large stochastic, possibly non-linear systems, that we can demonstrate in complex simulated environments to practitioners and that are validated by practitioners.
[S3c] Progressively populate a catalogue of real practices and actions, from repeated discussion and interactions with practitioners, that we can integrate to the software recommendation platform and can be effectively used by RL algorithms. Successfully run an RL strategy over real practices in a simulated environment under realistic conditions, that improves over practitioners policy, and shows convincing elements of explainability.
[S3d] Interface the agrosimulators (e.g. DSSAT) with the experimentation recommendation platform to enable planning and explainability. Based on the successes of **StaRLux**, enable a collaborative project with agronomy researchers in which they test the recommendation platform in actual experiments.

II.4 Risk and impact

We conclude this document by discussing the main risks (R) and impact (I) identified for **StaRLux**.

Risks and mitigation

[R1] Execution risks. The execution risks are high due to the ambitious scientific challenges, and the technical difficulties to be overcome and innovative ways that should be brought to solve them. Several elements mitigate this risk: first allocating enough resource and time to well-identified bottlenecks, then the knowledge of the existing literature and recorded contributions to the state-of-the-art of the field. This will be complemented by hiring postdoctoral fellows with complementary expertise, and by engaging with specific collaborators outside the expertise of the PI. This includes for instance B. De Saporta (U. Montpellier) on PD-MPs, R. Gautron (CIAT) and K. Naudin (CIRAD) on agronomy practices, and S. Auzoux (CIRAD) on record of experimental data in agriculture. Besides, in case we remark that one intended approach needs to be modified, we can easily consider adapting the research plan inside a task without hindering much the progress on most other tasks.
[R2] Human effort risks. As the project involves 8 persons to be hired, there is a significant risk in terms of human effort. To mitigate them, we will first allocate enough time for screening, especially engineers that will be the backbone of WP3 and the postdoctoral fellows. Besides, the PI is regularly contacted by candidates worldwide to start PhD or Postdoc, and we also count on the topic, that, compared to what is proposed by the industry, is both original, sensible and aligned with ethical concerns regarding A.I. for good and environmental concerns, hence has a real potential to attract talented and motivated candidates. Moreover, the PI is experienced with managing and organizing research projects, nurturing PhDs including in their life questioning, and benefits from the unique infrastructure provided by Inria and especially Inria auxiliary services to research.
[R3] Data and practitioner impact risks. Machine learning and Reinforcement Learning are usually considered data intensive consumers, further, practitioner fields such as health care or agroecology do

not communicate much with RL researchers, hence there is a real risk that the projects stops for lack of data or that its results stay confidential to the RL community. To mitigate this, we note that this project is about the methodology of RL and sequential experimentation, not on a posteriori analysis of collected data. Hence no novel sample will need to be collected within the scope of the project. Rather, we will resort to simulators used in agronomy and converted into RL environments (such as gym-DSSAT) or inspired from such, to train and develop most of the novel approaches. We will integrate real practices that are available in publications and by discussions with specific researchers known by the PI (e.g. at CIRAD and INRAE). Moreover, our results will be made available following an open-science and open-code methodology, advertised not only in RL conferences but also in complementary agronomy journals. To complement, the RL challenge will bring significant visibility to experimental sciences in the RL community and the platform will provide ready-to-use RL strategies for practitioners. Last, the PI allocates 70% of his time to this project to enable time to create complementary collaborative projects with actual practitioners in agroecology. All this should significantly improve visibility of the project beyond the RL community and ensure it does permeate to practitioners.

Impact

[I1] Research in RL and statistics. The **StaRLux** project is expected to have significant impact on the field of RL and sequential statistics, by unlocking challenges, bringing innovative challenging environments to the community, and combining approaches, such as coming from bandit theory and deep-learning, or contextual learning and hypothesis testing. We expect answering these challenges will in turn open novel questions paving the way towards reproducible and reliable A.I. in real-world contexts.
[I2] Research in agronomy and experimental sciences. We see three ways in which **StaRLux** may impact research in experimental sciences. First, on the short-term, by building sufficient visibility and success to have researchers in such fields start using and testing our methodology based on our publications and open-source code. Then, by creating joint projects especially with researchers in agroecology, on the basis of the initial success of **StaRLux**, to further engage with this specific community on long-run experiments. This may already start during the project. Finally, most probably after the end of **StaRLux** and on the long run, by populating a group of practitioners in agroecology that actively use and benefit from our dedicated experimentation companion platform and available RL libraries enriched with continually improved learning strategies.
[I3] Society benefits Beyond researchers, we expect the **StaRLux** project to impact society and industry, by reducing the "sim to real" gap, that is the difference of expectations between apparent success of applying RL strategies in completely safe simulated tasks, and the daunting challenges of deploying RL in real-world scenarios, under stringent uncertainty and safety conditions. This impact will be facilitated by our open-source philosophy and by our intensive efforts in making RL applicable to agroecology for real. By contributing to significantly closing the sim-to-real gap in such a challenging application area as agriculture, **StaRLux** will open reinforcement learning to significantly more applications, including for the industry. Beyond this, by enabling reproducible and reliable interactions with complex stochastic systems focusing on personalized, risk-aware and diversity-enforcing decisions, **StaRLux** may also contribute to operating a paradigm shift from hyper-control and model simplification that dominated the 20th century to embracing personalization and complexity-awareness in an ever-evolving world. On a more practical aspect, enabling agroecologists to understand agrosystems more efficiently and collectively will directly help address societal hazards such as food security (and possibly others related to agriculture practices), hence will benefit society at large.

References

[1] H. F. Dodge and H. G. Romig, "A method of sampling inspection," *The Bell System Technical Journal*, vol. 8, no. 4, pp. 613–631, 1929.

[2] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[3] R. Bellman, "The theory of dynamic programming," *Bulletin of the American Mathematical Society*, vol. 60, no. 6, pp. 503–515, 1954.

[4] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[5] H. Robbins, *Selected papers*. Springer, 2012.

[6] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in *Recommender systems handbook*. Springer, 2011, pp. 1–35.

[7] ——, "Recommender systems: introduction and challenges," in *Recommender systems handbook*. Springer, 2015, pp. 1–34.

[8] ——, "Recommender systems: Techniques, applications, and challenges," *Recommender Systems Handbook*, pp. 1–35, 2022.

[9] J. Bartroff, T. L. Lai, and M.-C. Shih, *Sequential experimentation in clinical trials: design and analysis*. Springer Science & Business Media, 2012, vol. 298.

[10] T. L. Lai, P. W. Lavori, and M.-C. Shih, "Sequential design of phase ii–iii cancer trials," *Statistics in medicine*, vol. 31, no. 18, pp. 1944–1960, 2012.

[11] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[12] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[13] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, "The arcade learning environment: An evaluation platform for general agents," *Journal of Artificial Intelligence Research*, vol. 47, pp. 253–279, 2013.

[14] R. S. Sutton, "Generalization in reinforcement learning: Successful examples using sparse coarse coding," *Advances in neural information processing systems*, vol. 8, 1995.

[15] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.

[16] M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder *et al.*, "Multi-goal reinforcement learning: Challenging robotics environments and request for research," *arXiv preprint arXiv:1802.09464*, 2018.

[17] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *Journal of Machine Learning Research*, vol. 11, pp. 1563–1600, 2010.

[18] P. L. Bartlett and A. Tewari, "Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps," in *Proceedings of the 25th conference on Uncertainty in Artificial Intelligence*, ser. UAI '09. Arlington, Virginia, United States: AUAI Press, 2009, pp. 35–42.

[19] I. Osband, D. Russo, and B. Van Roy, "(More) efficient reinforcement learning via posterior sampling," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 3003–3011.

[20] M. S. Talebi and O.-A. Maillard, "Variance-aware regret bounds for undiscounted reinforcement learning in mdps," in *Algorithmic Learning Theory*, 2018, pp. 770–805.

[21] M. Asadi, M. S. Talebi, H. Bourel, and O.-A. Maillard, "Model-based reinforcement learning exploiting state-action equivalence," in *Asian Conference on Machine Learning*. PMLR, 2019, pp. 204–219.

[22] H. Bourel, O. Maillard, and M. S. Talebi, "Tightening exploration in upper confidence reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1056–1066.

[23] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for sequential allocation problems," *Advances in Applied Mathematics*, vol. 17(2), pp. 122–142, 1996.

[24] T. L. Graves and T. L. Lai, "Asymptotically efficient adaptive choice of control laws incontrolled markov chains," *SIAM journal on control and optimization*, vol. 35, no. 3, pp. 715–743, 1997.

[25] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled iid processes: Finite parameter space," *IEEE Transactions on Automatic Control*, vol. 34, no. 3, 1989.

[26] F. Pesquerel and O.-A. Maillard, "Imed-rl: Regret optimal learning of ergodic markov decision processes," in *NeurIPS 2022-Thirty-sixth Conference on Neural Information Processing Systems*, 2022.

[27] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[28] A. Federgruen and P. J. Schweitzer, "Successive approximation methods for solving nested functional equations in markov decision problems," *Mathematics of Operations Research*, vol. 9, no. 3, pp. 319–344, 1984.

[29] Y.-C. L. Ho and X.-R. Cao, *Perturbation analysis of discrete event dynamic systems*. Springer Science & Business Media, 2012, vol. 145.

[30] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *International conference on machine learning*. PMLR, 2017, pp. 449–458.

[31] J. Duan, Y. Guan, S. E. Li, Y. Ren, Q. Sun, and B. Cheng, "Distributional soft actor-critic: Off-policy reinforcement learning for addressing value estimation errors," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 11, pp. 6584–6598, 2021.

[32] T. Hirata, T. Kuremoto, M. Obayashi, S. Mabu, and K. Kobayashi, "Deep belief network using reinforcement learning and its applications to time series forecasting," in *Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part III 23*. Springer, 2016, pp. 30–37.

[33] F. Abtahi and I. Fasel, "Deep belief nets as function approximators for reinforcement learning," in *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[34] M. Roder, L. A. Passos, G. H. de Rosa, V. H. C. de Albuquerque, and J. P. Papa, "Reinforcing learning in deep belief networks through nature-inspired optimization," *Applied Soft Computing*, vol. 108, p. 107466, 2021.

[35] I. Osband and B. Van Roy, "Near-optimal reinforcement learning in factored mdps," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[36] M. S. Talebi, A. Jonsson, and O. Maillard, "Improved exploration in factored average-reward mdps," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 3988–3996.

[37] A. Rosenberg and Y. Mansour, "Oracle-efficient regret minimization in factored mdps with unknown structure," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 148–11 159, 2021.

[38] S. J. Gershman, "Reinforcement learning and causal models," *The Oxford handbook of causal reasoning*, vol. 295, 2017.

[39] M. Gasse, D. Grasset, G. Gaudron, and P.-Y. Oudeyer, "Causal reinforcement learning using observational and interventional data," *arXiv preprint arXiv:2106.14421*, 2021.

[40] J. Zhang, Y. Chen, and A. Singh, "Causal bandits: Online decision-making in endogenous settings," in *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.

[41] M. S. Talebi Mazraeh Shahi, "Minimizing regret in combinatorial bandits and reinforcement learning," Ph.D. dissertation, KTH Royal Institute of Technology, 2017.

[42] M. Seitzer, B. Schölkopf, and G. Martius, "Causal influence detection for improving efficiency in reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 905–22 918, 2021.

[43] J. Kirschner and A. Krause, "Stochastic bandits with context distributions," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[44] J. Yang and S. Ren, "Robust bandit learning with imperfect context," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 10 594–10 602.

[45] S. Arradi-Alaoui, "Information auxiliaire non paramétrique: géometrie de l'information, processus empirique et applications," Ph.D. dissertation, Université Paul Sabatier-Toulouse III, 2022.

[46] M. H. Davis, "Piecewise-deterministic markov processes: A general class of non-diffusion stochastic models," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 46, no. 3, pp. 353–376, 1984.

[47] O. L. Costa and F. Dufour, "Stability and ergodicity of piecewise deterministic markov processes," *SIAM Journal on Control and Optimization*, vol. 47, no. 2, pp. 1053–1077, 2008.

[48] F. Dufour, B. de Saporta, and H. Zhang, *Numerical methods for simulation and optimization of piecewise deterministic Markov processes: application to reliability*. John Wiley & Sons, 2015.

[49] A. Verma, M. Hanawal, C. Szepesvari, and V. Saligrama, "Online algorithm for unsupervised sensor selection," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 3168–3176.

[50] H. Santana, G. Ramalho, V. Corruble, and B. Ratitch, "Multi-agent patrolling with reinforcement learning," in *Autonomous Agents and Multiagent Systems, International Joint Conference on*, vol. 4. IEEE Computer Society, 2004, pp. 1122–1129.

[51] L. Xu, E. Bondi, F. Fang, A. Perrault, K. Wang, and M. Tambe, "Dual-mandate patrols: Multi-armed bandits for green security," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 14 974–14 982.

[52] S. Hoshino, S. Ugajin, and T. Ishiwata, "Patrolling robot based on bayesian learning for multiple intruders," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 603–609.

[53] R. J. Solomonoff, "A formal theory of inductive inference. part ii," *Information and control*, vol. 7, no. 2, pp. 224–254, 1964.

[54] M. Hutter, *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media, 2004.

[55] J. Leike and M. Hutter, "On the computability of solomonoff induction and aixi," *Theoretical Computer Science*, vol. 716, pp. 28–49, 2018.

[56] S. Katayama and S. Kobayashi, "Satisficing reinforcement learning," *Intelligent Autonomous Systems: IAS-5*, p. 296, 1998.

[57] J. B. Bendor, S. Kumar, and D. A. Siegel, "Satisficing: A'pretty good'heuristic," *The BE Journal of Theoretical Economics*, vol. 9, no. 1, 2009.

[58] H. A. Simon, "Rational decision making in business organizations," *The American economic review*, vol. 69, no. 4, pp. 493–513, 1979.

[59] M. A. Goodrich, W. C. Stirling, and R. L. Frost, "A theory of satisficing decisions and control," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 28, no. 6, pp. 763–779, 1998.

[60] T. Michel, H. Hajiabolhassan, and R. Ortner, "Regret bounds for satisficing in multi-armed bandit problems," 2022.

[61] J. Mourtada and O.-A. Maillard, "Efficient tracking of a growing number of experts," in *International Conference on Algorithmic Learning Theory*. PMLR, 2017, pp. 517–539.

[62] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 661–670.

[63] A. Krause and C. Ong, "Contextual gaussian process bandit optimization," *Advances in neural information processing systems*, vol. 24, 2011.

[64] A. Hallak, D. Di Castro, and S. Mannor, "Contextual markov decision processes," *arXiv preprint arXiv:1502.02259*, 2015.

[65] N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire, "Contextual decision processes with low bellman rank are pac-learnable," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1704–1713.

[66] S. Sodhani, F. Meier, J. Pineau, and A. Zhang, "Block contextual mdps for continual learning," in *Learning for Dynamics and Control Conference*. PMLR, 2022, pp. 608–623.

[67] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, 2011, pp. 2312–2320.

[68] M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini, "Finite-time analysis of kernelised contextual bandits," in *Uncertainty in Artificial Intelligence*, 2013.

[69] A. Durand, O.-A. Maillard, and J. Pineau, "Streaming kernel regression with provably adaptive mean, variance, and regularization," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 650–683, 2018.

[70] O.-A. Maillard, "Robust risk-averse stochastic multi-armed bandits," in *International Conference on Algorithmic Learning Theory*. Springer, 2013, pp. 218–233.

[71] S. Vakili and Q. Zhao, "Risk-averse multi-armed bandit problems under mean-variance measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1093–1111, 2016.

[72] A. Gopalan, L. Prashanth, M. Fu, and S. Marcus, "Weighted bandits or: How bandits learn distorted values that are not expected," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[73] D. Baudry, R. Gautron, E. Kaufmann, and O. Maillard, "Optimal thompson sampling strategies for support-aware cvar bandits," in *International Conference on Machine Learning*. PMLR, 2021, pp. 716–726.

[74] V. Y. Tan, K. Jagannathan *et al.*, "A survey of risk-aware multi-armed bandits," *arXiv preprint arXiv:2205.05843*, 2022.

[75] L. Leqi, A. Prasad, and P. K. Ravikumar, "On human-aligned risk minimization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[76] S. P. Bhat and P. LA, "Concentration of risk measures: A wasserstein distance approach," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[77] O.-A. Maillard, "Local dvoretzky–kiefer–wolfowitz confidence bands," *Mathematical Methods of Statistics*, vol. 30, no. 1, pp. 16–46, 2021.

[78] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, "Risk-sensitive and robust decision-making: a cvar optimization approach," *Advances in neural information processing systems*, vol. 28, 2015.

[79] M. Ahmadi, U. Rosolia, M. D. Ingham, R. M. Murray, and A. D. Ames, "Constrained risk-averse markov decision processes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, 2021, pp. 11 718–11 725.

[80] J. L. Hai, M. Petrik, M. Ghavamzadeh, and R. Russel, "Rasr: Risk-averse soft-robust mdps with evar and entropic risk," *arXiv preprint arXiv:2209.04067*, 2022.

[81] K. Jamieson and R. Nowak, "Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting," in *2014 48th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2014, pp. 1–6.

[82] A. Garivier and E. Kaufmann, "Optimal best arm identification with fixed confidence," in *Conference on Learning Theory*. PMLR, 2016, pp. 998–1027.

[83] D. Russo, "Simple bayesian algorithms for best arm identification," in *Conference on Learning Theory*. PMLR, 2016, pp. 1417–1418.

[84] M. Soare, A. Lazaric, and R. Munos, "Best-arm identification in linear bandits," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[85] Y. Jedra and A. Proutiere, "Optimal best-arm identification in linear bandits," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10 007–10 017, 2020.

[86] M. J. Azizi, B. Kveton, and M. Ghavamzadeh, "Fixed-budget best-arm identification in contextual bandits: A static-adaptive algorithm," *arXiv preprint arXiv:2106.04763*, 2021.

[87] A. A. Marjani and A. Proutiere, "Adaptive sampling for best policy identification in markov decision processes," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 7459–7468. [Online]. Available: https://proceedings.mlr.press/v139/marjani21a.html

[88] A. Al Marjani and A. Proutiere, "Adaptive sampling for best policy identification in markov decision processes," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7459–7468.

[89] A. Al Marjani, A. Garivier, and A. Proutiere, "Navigating to the best policy in markov decision processes," *Advances in Neural Information Processing Systems*, vol. 34, pp. 25 852–25 864, 2021.

[90] S. J. Pocock, "Group sequential methods in the design and analysis of clinical trials," *Biometrika*, vol. 64, no. 2, pp. 191–199, 1977.

[91] H.-H. Müller and H. Schäfer, "Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches," *Biometrics*, vol. 57, no. 3, pp. 886–891, 2001.

[92] N. Stallard and T. Friede, "A group-sequential design for clinical trials with treatment selection," *Statistics in Medicine*, vol. 27, no. 29, pp. 6209–6227, 2008.

[93] R. A. Fisher, "Design of experiments," *British Medical Journal*, vol. 1, no. 3923, p. 554, 1936.

[94] D. R. Cox and N. Reid, *The theory of the design of experiments*. Chapman and Hall/CRC, 2000.

[95] V. L. Anderson and R. A. McLean, *Design of experiments: a realistic approach*. CRC Press, 2018.

[96] A. Dean and D. Voss, *Design and analysis of experiments*. Springer, 1999.

[97] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[98] T. L. Lai, "Adaptive treatment allocation and the multi-armed bandit problem," *The Annals of Statistics*, pp. 1091–1114, 1987.

[99] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[100] J. Honda and A. Takemura, "An asymptotically optimal bandit algorithm for bounded support models." in *COLT*. Citeseer, 2010, pp. 67–79.

[101] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz, "Kullback-leibler upper confidence bounds for optimal sequential allocation," *The Annals of Statistics*, pp. 1516–1541, 2013.

[102] D. Baudry, E. Kaufmann, and O.-A. Maillard, "Sub-sampling for efficient non-parametric bandit exploration," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5468–5478, 2020.

[103] R. Gautron, D. Baudry, M. Adam, G. N. Falconnier, and M. Corbeels, "Towards an efficient and risk aware strategy for guiding farmers in identifying best crop management," *arXiv preprint arXiv:2210.04537*, 2022.

[104] S. Couture, M.-J. Cros, and R. Sabbadin, "Multi-objective sequential forest management under risk using a markov decision process-pareto frontier approach," *Environmental Modeling & Assessment*, vol. 26, no. 2, pp. 125–141, 2021.

[105] M. Brunette, S. Couture, and J. Laye, "Optimising forest management under storm risk with a markov decision process model," *Journal of Environmental Economics and Policy*, vol. 4, no. 2, pp. 141–163, 2015.

[106] M. Kunaver and T. Požrl, "Diversity in recommender systems–a survey," *Knowledge-based systems*, vol. 123, pp. 154–162, 2017.

[107] L. Chen, G. Zhang, and E. Zhou, "Fast greedy map inference for determinantal point process to improve recommendation diversity," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[108] Y. Liu, C. Walder, and L. Xie, "Determinantal point process likelihoods for sequential recommendation," *arXiv preprint arXiv:2204.11562*, 2022.

[109] T. Kathuria, A. Deshpande, and P. Kohli, "Batched gaussian process bandit optimization via determinantal point processes," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[110] P. Grünwald, R. de Heide, and W. M. Koolen, "Safe testing," in *2020 Information Theory and Applications Workshop (ITA)*. IEEE, 2020, pp. 1–54.

[111] W. M. Koolen and P. Grünwald, "Log-optimal anytime-valid e-values," *International Journal of Approximate Reasoning*, vol. 141, pp. 69–82, 2022.

[112] P. Grünwald, "Beyond neyman-pearson," *arXiv preprint arXiv:2205.00901*, 2022.

[113] J. ter Schure and P. Grünwald, "All-in meta-analysis: breathing life into living systematic reviews," *arXiv preprint arXiv:2109.12141*, 2021.

[114] R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. Bellemare, "Deep reinforcement learning at the edge of the statistical precipice," *Advances in neural information processing systems*, vol. 34, pp. 29 304–29 320, 2021.

[115] M. Jourdan, R. Degenne, D. Baudry, R. de Heide, and E. Kaufmann, "Top two algorithms revisited," *arXiv preprint arXiv:2206.05979*, 2022.

[116] S. Siegel, "Nonparametric statistics," *The American Statistician*, vol. 11, no. 3, pp. 13–19, 1957.

[117] S. Bonnini, L. Corain, M. Marozzi, and L. Salmaso, *Nonparametric hypothesis testing: rank and permutation methods with applications in R*. John Wiley & Sons, 2014.

[118] R. H. Lock, "A sequential approximation to a permutation test," *Communications in Statistics-Simulation and Computation*, vol. 20, no. 1, pp. 341–363, 1991.

[119] I. Kim, S. Balakrishnan, and L. Wasserman, "Minimax optimality of permutation tests," *The Annals of Statistics*, vol. 50, no. 1, pp. 225–251, 2022.

[120] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *Ieee Access*, vol. 8, pp. 42 200–42 216, 2020.

[121] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 648–657.

[122] E. Puiutta and E. Veith, "Explainable reinforcement learning: A survey," in *International cross-domain conference for machine learning and knowledge extraction*. Springer, 2020, pp. 77–95.

[123] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "Explainable reinforcement learning through a causal lens," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 03, 2020, pp. 2493–2500.

[124] E. Leurent, D. Efimov, and O.-A. Maillard, "Robust-adaptive control of linear systems: beyond quadratic costs," in *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc., 2020.

[125] M. Lacoste, S. Cook, M. McNee, D. Gale, J. Ingram, V. Bellon-Maurel, T. MacMillan, R. Sylvester-Bradley, D. Kindred, R. Bramley *et al.*, "On-farm experimentation to transform global agriculture," *Nature Food*, vol. 3, no. 1, pp. 11–18, 2022.

[126] T. Lang and S. Siribaddana, "Clinical trials have gone global: is this a good thing?" *PLoS medicine*, vol. 9, no. 6, p. e1001228, 2012.

[127] B. v. d. Akker, N. Weber, F. Moraes, and D. Goldenberg, "Extending open bandit pipeline to simulate industry challenges," *arXiv preprint arXiv:2209.04147*, 2022.

[128] R. Gautron, O.-A. Maillard, P. Preux, M. Corbeels, and R. Sabbadin, "Reinforcement learning for crop management support: Review, prospects and challenges," *Computers and Electronics in Agriculture*, vol. 200, p. 107182, 2022.

[129] R. Gautron, E. J. Padrón, P. Preux, J. Bigot, O.-A. Maillard, and D. Emukpere, "gym-DSSAT: a crop model turned into a Reinforcement Learning environment," Inria Lille, Research Report RR-9460, Jul. 2022. [Online]. Available: https://hal.inria.fr/hal-03711132

[130] O.-A. Maillard, T. Mathieu, and D. Basu, "Farm-gym: A modular reinforcement learning platform for stochastic agronomic games," in *Artificial Intelligence for Agriculture and Food Systems (AIAFS)*, Wahington DC, United States, Feb. 2023. [Online]. Available: https://hal.inria.fr/hal-03960683

[131] V. Zobernig, R. A. Saldanha, J. He, E. van der Sar, J. van Doorn, J.-C. Hua, L. R. Mason, A. Czechowski, D. Indjic, T. Kosmala *et al.*, "Rangl: A reinforcement learning competition platform," *arXiv preprint arXiv:2208.00003*, 2022.

[132] A. Uzhinskiy, G. Ososkov, P. Goncharov, and A. Nechaevskiy, "Multifunctional platform and mobile application for plant disease detection," in *CEUR Workshop Proc*, vol. 2507, 2019, pp. 110–114.

[133] D. Baudry, P. Saux, and O.-A. Maillard, "From optimality to robustness: Adaptive re-sampling strategies in stochastic bandits," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 029–14 041, 2021.