ERC Consolidator Grant 2023
Research proposal [Part B1]

Statistical Reinforcement Learning to Upscale Experimental Sciences. StaRLux

Principal investigator: Odalric-Ambrym Maillard
Host institution: Inria Lille - Nord Europe

Duration: 60 months

ERC Keyword 1: PE6_07 Artificial Intelligence, intelligent systems, natural language processing.
ERC Keyword 2: PE6_11 Machine learning, statistical data processing and applications using signal processing (e.g. speech, image, video).
ERC Keyword 3: PE1_14 Mathematical statistics.
Free keywords: Reinforcement Learning, Multi-armed bandits, Experimental Sciences, Agroecology.

Given the global and rapidly evolving nature of the modern hazards of our societies (food security, water shortage, emerging infectious diseases), there is an urge to assist **experimental sciences** such as agriculture or medecine in massive, efficient testing to identify and recommend best practices, personalized to each individual and context. The high number of influential factors, available practices and diversity of objectives requires to **upscale experimentation** across and possibly beyond labs, following e.g. the practitioner-sourcing On-Farm Experimentation initiative. **With a strong emphasis on statistics, the grand challenge of StaRLux is to switch the trajectory of reinforcement learning (RL) to assist experimental sciences, rather than games**. We will develop a new generation of **sample efficient** decision companions to deal with intrinsically **stochastic**, autonomous and coupled dynamics with costly observations, contrasting with the current trend focusing on large, but mostly deterministic game environments. We will **scale** RL to group-RL companions able to cope with a **vast array** of environments and practitioner objectives and achieve **sample efficient** batched coordinated exploration. We will contribute to high reproducibility standards, unlocking **reliable** performance, comparison and explainability of RL strategies. To ensure **applicability**, we will organize RL community challenges on realistic agroecology environments and will incorporate our results in an open-source recommendation platform ready to assist agroecology practitioners in real conditions. Our results will unlock fundamental challenges in the design of efficient, scalable, reliable and applicable personalized decision-making companions and reduce the gap between simulated and real-world RL applications. This will stimulate interdisciplinary collaborations, with the potential to be a game changer in the way scientists design their experiments and collectively answer modern hazards.

B1.a. Extended synopsis

The urge: Given the global and rapidly evolving nature of the modern hazards faced by our societies (food security [1], water shortage [2], emerging infectious diseases [3]), there is an urge to assist agriculture [4] and medecine [5] in massive, efficient testing to identify and recommend best practices, personalized to each individual and context. *Agroecology* [6, 7] emerged as a promising sustainable way to address food security together with water, soil and environmental protection under context-specific constraints, contrasting with classical agriculture, solely focusing of increasing immediate productivity and yield. However, the high number of influential factors, available practices and diversity of objectives creates a combinatorial explosion together with an unprecedented experimental challenge to quantify the *context*-specific benefit of each practice. Machine Learning (ML) started to assist agriculture [8, 9] in yield-prediction and identification tasks using massive sensor and image data analysis, and agronomists recently started On-Farm Experimentation [4] initiatives in an attempt at involving practitioners to experiment beyond the limited, controlled context of lab-farms. Yet, assisting a diverse group of practitioners *collectively test and identify best personalized practices* in a real-world context comes with significant machine learning and statistical challenges.

A deadlock? From a ML perspective, Recommender systems (RS) seem the appropriate approach to scaling-up recommendations. Originated from the web-industry, the offer impressive scalability abilities thanks to the development of contextual learning. They are however little adapted to agrosystems that are complex dynamical systems with long feedback delay. (Deep) Reinforcement Learning (RL) showed impressive decision-making abilities on complex dynamical systems [10, 11, 12, 13] but only in near-deterministic environments and games using massive amount of data, and fails to handle stochastic uncertainty and even provide statistically reliable results [14, 15]. In contrast, the process of growing a plant in a farm is inherently *stochastic* rather than deterministic, as it involves external uncontrolled entities (e.g. weather, insects, weeds). Now, sequential hypothesis testing and experimental design have accompanied *experimental sciences* with statistically reliable procedures [16, 17, 18, 19] since the very birth of mathematical statistics [20, 21] 90 years ago, but even their modern version embodied by multi-armed bandits and group-sequential theory struggle to handle large dynamical systems or scale to contextual approaches. As a result, RL efforts to assist agronomy experimentation are limited **[22]**.

a.1 Challenges and scientific objectives

With a strong emphasis on statistics, the grand challenge of StaRLux is to switch the trajectory of reinforcement learning to assist experimental sciences, rather than games, by developing a new generation of decision companions that are sample efficient, scalable and applicable to real-world agroecology conditions. We now detail objectives (O) and challenges (C):

O1. Sample-efficient RL for experimental environments, to tackle crucial limitations faced by practitioners, such as managing the intrinsically (C1) *stochastic* nature of considered environments that feature (C2) *diverse* entities interacting with each other, determining the (C3) *optimal moments* to intervene and observe, and adapting to the (C4) *ever-changing experimental contexts*, such as a shift in the dynamics and novel available actions or practitioners' goals. Indeed unlike more traditional environments (e.g. go, atari), the process of growing a plant in a farm is inherently *stochastic* rather than deterministic, as it involves external uncontrolled entities (e.g. weather, insects, weeds) and interaction between different entities (e.g. plants and soil). These interactions between entities further demonstrate a coupled dynamics, which is often hard to model. Besides, experimental measurements in agroecology (or healthcare) are often costly in terms of time or resources, hence many of the state components of the system are not visible at each time step. This creates a monitoring challenge, where deciding when and what to observe must be balanced with the need for intervention, in order to effectively reduce uncertainty and intervene optimally. Likewise, experimentation is usually a continual learning game in which a practitioner progressively refines policies and hypotheses with novel, finer-grain options, unlike traditional Markov decision processes (MDP) learning that requires a predefined fixed set of actions. Developing RL companions for experimental sciences hence requires addressing the exploration-exploitation

trade-off in alternative control models than MDPs, which represents a fundamentally novel perspective.
O2. Scalable experimentation with group-RL to cope with a vast array of environments and (C5) *diversity of practitioner objectives*, including horizon, risk, constraints, and rewards, to (C6) *chart good practices* across different contexts and advance efficient (C7) *batched coordinated exploration*. This is important as learning from many environments instead of one can considerably speed-up learning, when effectively leveraging context similarity such as demonstrated in contextual multi-armed bandits and RS theory. A key practical issue, however, is that different practitioners often have different objectives even when facing the very same environment, such as considering different risk-aversion level or horizon of time for an experiment. Further, due to the large reward delay (several weeks to months in healthcare or agroecology), experiments are typically organized in batch, following principles from group-sequential clinical trials. These practical concerns introduce intriguing novel challenges to contextual reinforcement learning, both to revisit the exploration-exploitation trade-off and contextual best policy identification (a.k.a. charting) under personalized objectives, and from a group-sequential standpoint.
O3. Applicable RL companions in agroecology This involves automated procedures to output statistically (C8) *reliable rankings* based on personalized goals, providing (C9) *explainability* of the strategies to practitioners to improve trust and compliance, and ensuring (C10) *adoption by practitioners*. Indeed high reproducibility standards is a core component of experimental sciences, and the RL field is suffering from severe reproducibility issues, often due to poor statistical habits, but also the difficulty of automatically deciding how many independent experiments are required to compare the performance of algorithms that rarely follow any classical model. Besides, even with a reliable ranking, black-box strategies can easily hinder trust and compliance from practitioners, especially in the presence of risk. Accurately predicting the evolution and occurrence of risky events when following a policy in a stochastic and dynamic system can improve explainability hence compliance. Likewise, fostering adoption by practitioners requires to demonstrate results inside the RL community on appropriate sand-box environments and to build a data-acquisition and recommendation software on which practitioners in experimental science could try companions and RL researchers continue develop their algorithms. Finding innovative and effective ways to solve these daunting challenges is an exciting and little explored frontier in RL.

a.2 State of the art and its limitations

RL fails at handling large stochastic systems RL [23, 24, 25] formalizes learning by trial and error in uncertain systems. It has recently produced decision-making tools offering great promises in solving complex tasks in diverse domains [10, 11, 12, 13]. While research on deep learning, optimization and function approximation dominates the current literature in large dynamical systems, these methods are unfortunately restricted to deterministic, simulated environments. Indeed, *deep RL approaches are largely unable to correctly handle even small stochastic systems*, for now. Unlike more traditional RL environments (e.g. go, atari), the process of growing a plant in a farm is inherently *stochastic* due to external factors. On the other hand, the vast literature about the concentration of measure [26, 27, 28, 29, 30], sequential statistics [31, 32], **[33]**, sequential design of experiments and multi-armed bandits [34, 35, 36, 37] appropriately handles stochastic uncertainty. For instance the construction of data-adaptive confidence set is used nowadays to analyze and make ML and RL algorithms more reliable. *I advocate for the intensification of research efforts on sequential statistics in large dynamical systems.*
RL models still fail to capture practitioner interaction and knowledge To handle the complex interactions observed by practitioners and incorporate expert knowledge promising MDP models have been introduced, such as factored [38], **[39]**, [40], causality [41, 42, 43], combinatorial [44] models, or even forecast information e.g. about the weather [45, 46] in bandits. Yet, *I argue that alternative models to MDPs are required to better handle autonomous systems, shifting dynamics or a growing set of available actions*. For instance Piecewise Deterministic Markov Processes (PD-MP) [47, 48, 49] conveniently model autonomous (restless) systems subject to spontaneous or controlled jumps of dynamics, where the task is to decide when to trigger a jump. However, PD-MPs have not been studied when the model is unknown and must be learned from data. Likewise, practitioners continually test novel hypotheses and introduce

or refine specific techniques progressively rather than considering all possible sequences of micro-actions as in an MDP. This progressive learning mechanism is reminiscent of satisficing theory, [50, 51, 52, 53], a promising theory in which much remains to be uncovered, but also calls for novel approaches.

RS and contextual RL should grow group-sequential and risk-averse Recommender systems [54, 55, 56], and contextual RL [57, 58, 59, 60],[61], [62, 63, 64] have upscaled the possibility to collect feedback and efficiently learn recommendations for a large group of users at an unprecedented world scale. However, these works coming from the web industry usually consider an immediate feedback, a risk-neutral objective, and non-reactive (bandit) systems. In experimental sciences, group-sequential experimentation [65, 66, 67] is required due to large feedback delays, such as in group-sequential clinical trials [68, 69]. Further, stringent risk-aversion is required, including the less studied group-risk aversion, to model increased potential for disease or fire spread when too many similar crops/trees are grown in close proximity [70, 71], that is, similar contexts receive too similar recommendations. Last, agrosystems decisions are complex policies affecting trajectories of a dynamical, rather than static system. *I advocate for novel RS strategies, that handle risk-aversion and group-sequential learning in dynamical systems.*

RL fails to be reproducible and applicable (Deep) RL literature faces a reproducibility issue [14, 15] due to poor statistical habits and unreasonably large computational requirement. Even reliably comparing strategies has becoming an issue. On the other hand, the RL community produced standardized environments (e.g. Atari, Mujoco, Classic control tasks) following openAI-gym [72] to foster reproducible research, although mostly focused on deterministic tasks. In agronomy, researchers resort to decision support systems (DSS) developed and used by agronomists over decades. We have recently turned one of these simulators into an gym agronomy environment [73], and released a preliminary set of agronomy-games [74] that feature intrinsically stochastic and coupled dynamics to foster open and reproducible RL research on agronomy challenges. However, *I argue that open-source RL simulators are not enough to foster applicability, and should be complemented with an open-source application platform.* We note that large recommender systems from the industry are private, black boxes, which hinders the development of novel strategies and compliance from users, while providing explainability [75, 76] is critical in experimental sciences especially in a risk-averse context. On the other hand, there is no open-science recommender platform dedicated to assist experimentation in large dynamical systems, while this could stimulate collaborative efforts on applicability of RL methods, similarly to efforts on reproducibility.

a.3 Methodology and organization

StaRLux aims to advance statistical RL to assist experimental sciences by contributing to core RL, RS and statistics theory, developing novel sample efficient and scalable algorithmic strategies, and by providing open-source softwares for applicable RL research in agroecology. The scientific work packages (WP) are organized logically, from most fundamental and generic to most applied and specific.

WP1 **Unlocking RL efficiency for experimental environments**. To achieve sample efficiency in stochastic environments (C1) with complex interactions (C2), I intend to invent novel strategies.

•*Stochastic environments.* Pursuing the efforts that have been successful to improve bandit strategies [77, 78, 79] and permeate to model-based MDP learning [80, 81, 82],[83, 84, 85], I intend to adapt more bandit theory to stochastic MDPs. Encouraged by the recent excellent [86] first to provably reach optimality for some (small) MDPs, I intend to uncover novel strategies for large systems with the help of parametric [87] and distributional [88, 89] models. Then, rather than opposing bandits and deep-learning algorithms, I will also investigate a complementary mix suitable to large stochastic systems.

•*Structured interactions and auxiliary information.* I will study how to both leverage and discover structure of influential variables in a learning context. This requires innovative ideas that may considerably increase the sample-efficiency of learning strategies. I will also adapt auxiliary information theory [90] to incorporate expert knowledge such as prediction of certain variables, which is unusual in MDPs.

To optimally decide when to act (C3), and adapt to shifting conditions (C4), I will deviate from MDPs.

•*When to intervene, when to observe.* I will address the exploration-exploitation trade-off in control models alternative to (rested) MDPs, better suited at handling autonomous (restless) dynamical sys-

tems and decision over time. PD-MPs share similarities with MDPs such as the availability of dynamic programming but have not been studied from a learning viewpoint.

•*Shifting environments and progressive learning.* Since considering all micro-actions yields computational limits, apparent e.g. in the lower performance bounds of MDPs [91], I will study alternative models to refine and incorporate novel actions and objectives continuously in the learning process.

WP2 **Scaling-up experimentation with group-RL** To achieve scalability the second work-package focuses on core challenges of contextual learning. I here focus on enabling simultaneous learning on a vast array of environments, to handle (C5) diverse practitioner objectives, (C6) charting best policies and performing group-sequential exploration (C7).We here mix group-sequential and contextual theory.

•*Context-goal exploration-exploitation trade-off.* I intend to revisit the exploration-exploitation trade-off under a contextual, personalized objective. This is unusual as usually a similar objective is assumed for all contexts. I will build on the idea that similar contexts still share similar dynamics, which can be exploited to speed-up learning. To be successful, strategies should carefully balance exploration and exploitation accross several environments and leverage novel linear structure of objectives.

•*Charting good policies across contexts.* I will extend the best policy identification [92, 93, 94] from sequential hypothesis testing literature to the contextual setup, that is when dealing not with a single, but a collection of environments. The challenge is here to derive contextual hypothesis testing strategies able to stop with minimum number of trials and output in which contexts a policy is most beneficial.

•*Batch coordinated exploration and joint-risk.* Encouraged by excellent preliminary results **[95, 96]** for group-sequential risk-aware RL in agriculture, my goal is to extend these strategies to dynamical systems. I will especially focus on addressing group-risk aversion that penalizes recommending too similar actions for the same contexts, adapting ideas from random and diversity enforcing strategies [97, 98, 99].

WP3 **Making RL companions applicable in practice** After two work packages on core theory we now move to foster applicability of RL in a real-world application. I choose agroecology due to the many collaborations I developed with researchers in agronomy over the past few years (see appendix) that led to identify many challenges **[22]**, and as agroecology data is easier to share than the highly-sensitive medical data. I focus on reproducibility in comparing strategie (C8), explainability of companions (C9) and providing open-source platforms for agronomists (C10), to stimulate interdisciplinary research.

•*Reliable rankings and RL coopetition.* Sequential design of experiments may automatically adapt the number of independent executions of each strategy to produce statistically reliable rankings of RL strategies. However, optimal methods often assume a parametric score model, an assumption rarely met by the distribution of performance of RL strategies. I will thus extend finite-time guarantees to non-parametric distributions, following the recent efforts in **[79]**, [100] based on sub-sampling, and combining with the rich (though often asymptotic) literature on non-parametric tests [101, 102, 103, 104]. I will apply such tools for reproducible research to an open-source RL coopetition platform, based on the stochastic agronomy environments **[73, 74]** that we will enrich with increasingly challenging novel features.

•*Explainability of RL companions and event prediction.* To provide explainability [105] and in particular prediction of effects or counter-effects of decision [106], suitable to agroecology environments, I will derive refined theoretical study of trajectory evolution and risk occurrence in a complex, dynamical system. This also includes non-linear dynamics, which requires innovative ideas, especially regarding risk.

•*Agro-recommendation platform and practitioner engagement.* I intend to develop a recommender platform integrating most of our research results on experimentation companions into an open-science platform ready to be used by actual practitioners. This will contrast with private recommender systems from the industry and has the potential to stimulate interdisciplinary research on the basis of a solid reproducible, reliable system. Developing this platform requires engineering efforts, on top of research efforts. We will build on the proof of concept platform I initiated in the ending SR4SG project.

While apparently disconnected, I argue that these three work-packages collectively form the coherent back-bone of **StaRLux** *to effectively develop the next generation of decision companions in RL. More specifically, it is intended to yield sample efficient, scalable and applicable strategies that may really assist experimental sciences, on the basis of sound, theoretically-grounded principles.*

a.4 Expected results and impact

Scientific impact. StaRLux will make core contributions in reinforcement learning, sequential statistics and control theory at large, paving the way towards general A.I. All the challenges being treated from a generic standpoint, with the exception of the RL coopetition and the agro-recommendation platform, specific to agriculture, results of **StaRLux** have the potential to impact (all?) experimental sciences at large. We expect applied researchers will start applying our methodology in their experiments.

StaRLux will also foster multidisiplinary collaborations with experimental sciences. In particular, I expect collaborative projects with research centers specialized in agroecology research, such as the CIRAD, with unit AIDA or LIANE, the CIAT, INRAE or U. Bihar with which I already developed recent collaborations. The goal would be to organize trials with practitioners to experimentally validate the platform, compare RL companions to classical experimental design on real fields and prepare larger involvement of practitioners. For obvious ethical, safety and work reasons, such trials will not be done in **StaRLux** but in an external collaborative project with identified practitioners, and after **StaRLux** has unlocked some key challenges to provide sufficient safety guarantees. We expect such collaboration can occur in year 3 or 4, and will continually benefit from the progress of the project. Methodological results may also benefit side projects such as B4H with Hospital of Lille.

Industrial impact. The impact of **StaRLux** will not only be substantial from the theoretical side. Success in the **StaRLux** project is expected to significantly reduce the amount of interactions required by learning agents, and unveil sound techniques to move away from purely simulated worlds and games, to experimental sciences addressing real-world systems. 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.

Societal impact. By addressing challenges of experimental sciences with little data and real-life constraints, providing open-source codes and algorithmic platform targeting high reproducibility standards for real-life experimentation companions, and organizing international RL challenges to bring visibility to such questions, **StaRLux** is expected to get significant echos in applied sciences, possibly assisting thousands of researchers and practitioners on the long run. 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.

a.5 Risk assessment and chances of success

This is a high-risk/high-gain project. Risks are mitigated by the established expertise of the PI in the field of reinforcement learning and statistics, his network of collaborators and supervision experience, and the various research programs he conducted at national (e.g. ANR JCJC) or international level (e.g. with Univ. Kyoto, IISc Bangalore). The objectives are backed-up with promising preliminary results, and collaborations on complementary projects. The project, being organized in rather independent entities, is flexible and robust to potential changes. I will also allocate complementary skills and human effort on each task., combining Ph.D. students, Postdocs and Engineers. The external risks regarding competition from other academic and large industrial research groups is limited: they do not have the unique expertise of the PI, nor the same research focus or interest e.g. in agronomy companions, for which I already initiated strong collaborations (Cirad, Inrae, Univ. Bihar). Last but not least, I am committed to dedicate the next 15-20 years of my life to effectively transform RL to assist experimental sciences, and especially agroecology, in addressing urgent societal hazards.

a.6 Resources and commitment of the PI

I will commit 70% of my time on the project. The other human resources allocated to the project include four Ph.D. students, two 2-year postdocs and two 36-month engineer. I also request traveling money to invite external collaborators (3 weeks/y), present works in international venues, organize a challenge (workshop) and fund internships and equipment (computers, books, etc).

We have indicated internal references **[this way]**, and external references [this way] in this document.

References

[1] J. Bruinsma, *World agriculture: towards 2015/2030: an FAO perspective*. Routledge, 2017.

[2] W. W. A. P. U. Nations) and UN-Water, *Water in a changing world*. Earthscan, 2009.

[3] K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, and P. Daszak, "Global trends in emerging infectious diseases," *Nature*, vol. 451, no. 7181, pp. 990–993, 2008.

[4] 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.

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

[6] M. A. Altieri, "Agroecology: the science of sustainable agriculture. boulder," *Westview Press. Part Three Dev. Clim. Rights*, vol. 238, pp. 12 052–12 057, 1995.

[7] S. R. Gliessman, *Agroecology: the ecology of sustainable food systems*. CRC press, 2014.

[8] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, p. 2674, 2018.

[9] L. Benos, A. C. Tagarakis, G. Dolias, R. Berruto, D. Kateris, and D. Bochtis, "Machine learning in agriculture: A comprehensive updated review," *Sensors*, vol. 21, no. 11, p. 3758, 2021.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] J. Pineau, "Reproducible, reusable, and robust reinforcement learning," *Advances in Neural Information Processing Systems*, 2018.

[15] 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.

[16] J. Neyman and E. S. Pearson, "Ix. on the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, 1933.

[17] E. L. Lehmann, "The fisher, neyman-pearson theories of testing hypotheses: one theory or two?" *Journal of the American statistical Association*, vol. 88, no. 424, pp. 1242–1249, 1993.

[18] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[19] M. Greenwood, "The statistician and medical research," *British medical journal*, vol. 2, no. 4574, p. 467, 1948.

[20] S. M. Stigler, "The history of statistics in 1933," *Statistical Science*, vol. 11, no. 3, pp. 244–252, 1996.

[21] A. Kolmogorov, "Fundamental concepts of probability," 1933.

[22] 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.

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

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

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

[26] V. D. Milman, "A new proof of a. dvoretzky's theorem on cross-sections of convex bodies," *Funkcional. Anal. i Prilozen*, vol. 5, pp. 28–37, 1971.

[27] M. Talagrand, "Concentration of measure and isoperimetric inequalities in product spaces," *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, vol. 81, no. 1, pp. 73–205, 1995.

[28] M. Ledoux, *The Concentration of Measure Phenomenon*, ser. Mathematical surveys and monographs. American Mathematical Society, 2001.

[29] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[30] M. Raginsky, I. Sason *et al.*, "Concentration of measure inequalities in information theory, communications, and coding," *Foundations and Trends® in Communications and Information Theory*, vol. 10, no. 1-2, pp. 1–246, 2013.

[31] G. Shafer and V. Vovk, *Game-Theoretic Foundations for Probability and Finance*. John Wiley & Sons, 2019, vol. 455.

[32] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon, "Time-uniform, nonparametric, nonasymptotic confidence sequences," *The Annals of Statistics*, vol. 49, no. 2, pp. 1055–1080, 2021.

[33] O.-A. Maillard, "Mathematics of statistical sequential decision making," Ph.D. dissertation, Université de Lille, Sciences et Technologies, 2019.

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

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

[36] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

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

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

[39] 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.

[40] 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.

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

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

[43] 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.

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

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

[46] 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.

[47] 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.

[48] 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.

[49] 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.

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

[51] 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.

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

[53] 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.

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

[55] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: introduction and challenges," in *Recommender systems handbook*. Springer, 2015, pp. 1–34.

[56] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: Techniques, applications, and challenges," *Recommender Systems Handbook*, pp. 1–35, 2022.

[57] 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.

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

[59] 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.

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

[61] 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.

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

[63] 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.

[64] 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.

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

[66] 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.

[67] 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.

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

[69] 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.

[70] 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.

[71] 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.

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

[73] 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

[74] 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

[75] 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.

[76] 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.

[77] 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.

[78] O.-A. Maillard, "Boundary crossing probabilities for general exponential families," *Mathematical Methods of Statistics*, vol. 27, no. 1, pp. 1–31, 2018.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] S. R. Chowdhury, A. Gopalan, and O.-A. Maillard, "Reinforcement learning in parametric mdps with exponential families," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1855–1863.

[88] 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.

[89] D. W. Nam, Y. Kim, and C. Y. Park, "Gmac: A distributional perspective on actor-critic framework," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7927–7936.

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

[91] 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.

[92] 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

[93] 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.

[94] 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.

[95] 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.

[96] 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.

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

[98] 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.

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

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

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

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

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

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

[105] 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.

[106] 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.

[107] O.-A. Maillard, R. Munos, and G. Stoltz, "A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences," in *Proceedings of the 24th annual Conference On Learning Theory*. JMLR Workshop and Conference Proceedings, 2011, pp. 497–514.

[108] A. Garivier and O. Cappé, "The kl-ucb algorithm for bounded stochastic bandits and beyond," in *Proceedings of the 24th annual conference on learning theory*. JMLR Workshop and Conference Proceedings, 2011, pp. 359–376.

[109] 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.

[110] O.-A. Maillard, T. A. Mann, and S. Mannor, ""how hard is my MDP?" the distribution-norm to the rescue," in *Advances in Neural Information Processing Systems 27 (NIPS)*, 2014, pp. 1835–1843.

B1.b. Curriculum vitae

Personal information MAILLARD Odalric-Ambrym, French, October 12th 1984, divorced
ORCID 0000-0001-7935-7026 Personal website Gitlab Google sholar

Education

2019 Habilitation, Sciences Mathématiques, Université de Lille, France.

2011 Ph.D. "*Sequential Learning: Bandits, Statistics and Reinforcement.*", Université Lille 1, France, Advisors: Rémi Munos (Inria, Lille) & Philippe Berthet (IMT, Université de Toulouse)

2008 Diploma from École Normale Supérieure de Cachan, selectively given for outstanding cursus.

2008 M.Sc research in Comput. Sci.: "Master Vision et Apprentissage", ENS Cachan, France (rank 1).

2007 M.Sc research in Mathematics: "Théorie Statistique", Université Rennes 1, France (rank 1).

Current position(s)

2019- Chargé de Recherche Habilité (CRCN HDR), Scool Team, Inria Lille - Nord Europe, France.

Previous position(s)

2008-2011 PhD student, ED-SPI, Université Lille 1 and Université Paul-Sabatier, France.

2011-2013 Postdoctoral fellow, Information-Technology, Montanuniversität Leoben, Austria.

2013-2014 Postdoctoral fellow, EECS, Technion, Israël.

2014-2015 Senior Researcher, EECS, Technion, Israel.

2015-2016 Chargé de Recherche (CR2) Inria, Tao team, Inria Saclay - Île de France, France.

2016-2019 Chargé de Recherche (CRCN) Inria, SequeL team, Inria Lille - Nord Europe, France.

Fellowships and grants

I have been PI of several national and international projects:
Reliant "Real-life bandits", Inria-Japan Associate Team, CoPI J.Honda (U. Kyoto), 2022-2024, 30k.
RLRL"Real-Life Reinforcement Learning", CoPI B. De Saporta (U. Montpellier), 2021, 20k.
AppRenf "Apprentissage par Renforcement", Chaire of IA, CoPI P.Preux (U. Lille), 2021-2024, 500k.
SR4SG "Sequential Recommendation for Sustainable Gardening", Inria A-Ex., 2019-2023, 200k.
BADASS "Bandits against non-Stationarity and Structure", ANR JCJC, 2016-2020, 180k.
"Contextual MABs with hidden structure", IFCAM, CoPI A. Gopalan (IISc Bangalore), 2016, 7k.
PROMO , PEPS JCJC - INS2I, CoPI R. Bardenet (CNRS, Lille), 2015, 7k.
"Outil générique de prévision/contrôle de flux hydrauliques", Inria-Carnot, Prolog, 2015, 50k.
Programme Invité Digiteo (Invitation of A. Gopalan for 3 weeks), 2015, 4k.
I secured the following fellowships:
2016-2020. Recipient of Inria "Prime d'encadrement doctoral et de recherche" (PEDR).
2013-2014. Postdoctoral fellowship FP7 European Project SUpReL, Israel.
2011-2013. Postdoctoral fellowship FP7 European Project CompLacs, Austria.
2008-2011. Ph.D. fellowship from ENS Cachan (selective application, rank 1), France.

Supervision of graduate students and postdoctoral fellows

I have supervised the following smart people:
Tuan Dam (Postdoc, 2022-2023, Chaire IA), Timothée Mathieu (Postdoc, 2021-2023, Inria),
Mohammad Sadegh Talebi (Postdoc, 2018-2020, ANR), Hernán Carvajal (Engineer, 2021-2023, Inria).
I continue this list with the Ph.D. students I have advised since my habilitation (2019):
S. Vashishtha (100%, 2022-2025, Chaire IA fellow) F. Fabre (33%, 2021-2024, U. Reunion & CIRAD)
F. Pesquerel (100%, 2020-2023, ENS fellow) P. Saux (50%, 2020-2023, ENS fellow)
D. Baudry (50%, 2019-2022, Inria fellow) R. Ouhamma (50%, 2019-2023, Ecole Polytechnique fellow)
H. Saber (100%, 2019-2022, Inria fellow) E. Leurent (50%, 2017-2020, CIFRE Renault; 2 PhD awards)
Before my habilitation, I also co-supervised[1] S. Akhtar (2017, ANR) R. Alami (2016-2020, CIFRE Orange Labs) R. Allesiardo (2015-2017, CIFRE Orange Labs) N. Galichet (2016, Université Orsay).

[1] in French, doctoral school officially recognizes supervision only after habilitation.

I have advised 18 Master internships since 2015. Also, foreign Ph.D. Students regularly get in touch to visit and work with me for a few months during their Ph.D. Below is a list of recent visiting students:

- Chuan-Zheng Lee, Stanford University, Stanford fellow, May-August 2019.
- Arun Verma, IIT Bombay, Raman-Charpak fellow, June-Nov 2019.
- Ramakrishnan K, IISc Bangalore, Raman-Charpak fellow, May-July 2022.
- Shubhada Agrawal, TIFR Bombay, funded by Inria Scool, March-May 2022.

Likewise, young researchers get in touch to visit me, e.g. recent visits of J. Komiyama (2019, Research Associate, University of Tokyo), A. Mehrabian (2019, Postdoctoral Researcher, McGill University).

Teaching activities
I have been teaching **Statistical Reinforcement Learning** in several M.Sc over the last few years: **École Polytechnique**, 2019-2023 (5y) Master 2 "Artificial Intelligence & Advanced Visual Computing (Ai-ViC)" (48h, ∼20 students) and 2020-2022 (3y) Executive Master Data-Science (3h, ∼50 students) **École Centrale Supélec**, 2021/2022, Master 2 Mention IA (24h, ∼70 students) and **École Centrale de Lille**, 2017/2018, Master 2 "Données, Apprentissage, Décision (DAD)" (18h, ∼40 students)
I also taught **Bandits** in the **Reinforcement Learning Summer School (RLSS)**, 01-12 July 2019, `https://rlss.inria.fr/`, (5h, ∼300 students)

Organisation of scientific meetings
2020-2022: all planned workshops and meetings were canceled due to Covid.
2019 - Co-organization of Reinforcement Learning Summer School `https://rlss.inria.fr`.
2018 - PI of "ANR BADASS Lecture Series", hosting P. Grünwald, at Inria Lille – Nord europe.
2017 - PI of Workshop "Sequential Structured Statistical Learning" at Institut des Hautes Études Scientifiques (IHES).
2014 - PI of NIPS Workshop "From Bad models to Good policies"

Institutional activities
Inria CR hiring commitee: 2018/2019/2020. Inrae CR MathNum hiring commitee: 2021.
Inria Comission Emploi Recherche (CER): 2020/2021/2022.

Reviewing activities
International conferences (55 PC) and journals (+50 reviews):

Algorithmic Learning Theory, Bernoulli Journal, Artificial Intelligence and Statistics, The Annals of Statistics, Conference On Learning Theory, European Conference on Machine Learning, European Workshop on Reinforcement Learning, International Conference on Machine Learning, Journal of Machine Learning Research, Machine Learning Journal, Neural Information Processing Systems, IEEE Transactions on Automatic Controls, IEEE Transactions on Information Theory, Theoretical Computer Science. I am **editor** of the Journal of Machine Learning Research. I also reviewed for the French ANR (Agence Nationale de la Recherche) and Hungarian Scientific Research Fund (OTKA). Since 2019 (Habilitation), I regularly participate in Ph.D. juries.

Memberships of scientific societies
SSFAM

Major collaborations
P.Auer & R.Ortner (U. Leoben, Austria), S.Mannor (Technion, Israël), A.Gopalan (IISc Bangalore, India), A.Durand (U. Laval, Canada), J.Honda (U. Kyoto, Japan), R. Gautron (CIRAD and CIAT, Colombia).

Other 08/2015, part of the 200 "promising young researchers" internationally selected to attend the 3rd Heidelberg Laureate Forum and meet Abel, Fields and Turing laureates.

Career breaks None. The year before COVID was especially complicated at a personal level. COVID slowed down my publications and collaborations, but, I believe, like almost anyone on earth.

Appendix. Current research grants and any on-going applications related to the proposal of the PI (Funding ID)

Current grants (*or indicate No funding*)

Project title	Funding source	Amount (€)	Period	Role of the PI	Relation to **StaRLux**
SR4SG (A)	Inria A.Ex.	200k€	2019/2023	PI	Preliminary study
RELIANT (H)	Inria ass.team	6k€/year	2022/2024	PI	RL Foundation
3RSDM (A)	Inria-Inrae	4k€	2022/2025	PI	Real-world risk
Apprenf (T)	AI Chair	500k€	2020/2023	PI	RL Foundation
B4H (H)	I-SITE ULNE	110k€	2019/2023	I	Practitioners contact
BIP-UP (H)	ANR PRC	520k€	2023/2026	I	Practitioners contact
DC4SCM (A)	Inria ass.team	6k€/year	2022/2024	I	Practitioners contact
Pl@ntAgroEco (A)	PEPR	180k€	2023/2027	collaborator	Software
EpiRL (T)	ANR	6k€	2023/2026	collaborator	None

The letters indicate the focus of each project: (T)heory, (A)groecology, (H)ealthcare
Gray-colored lines indicate projects having a time overlap with **StaRLux**.

Apprenf is about advancing reinforcement learning bottlenecks from a theoretical standpoint. The project does not overlap with **StaRLux**.
EpiRL is about epistemic reinforcement learning theory, this a project of theoretical nature, with little investment from the PI.

SR4SG is about crowdsourcing and sharing of good practices in agroecology. The project does not overlap with **StaRLux**. Within the project, we have developed the RL platform FarmGym that we expect to extend to host a community challenge on experimental sciences in **StaRLux**. We also developed a POC crowdsourcing and recommendation software. Along the way, we developed links with agronomy researchers from CIRAD, Inrae and U. Bihar.
DC4SCM (with U. Bihar, India) is an Inria associate team about performing data-collection for smart crop management using environment-friendly practices, in university farms in india. A prototype tool has been introduced and several time series of observations and interventions have been collected on various aspects of climate, soil, insect, plant and interventions on the same fields. The next step is to test classical design of experiment against novel sequential approaches in simplified experiments.
3RSDM is about decision making under environment risks, applied to forest management. This is a joint project with Inrae team Mistis in Toulouse, with potential application in forest management.
Pl@ntAgroEco is about extending Pl@ntNet identification tools to agriculture. I participate to supervise a postdoctoral fellow on fundamental tools for some active recommendation tasks.

RELIANT (with U. Kyoto, Japan) is an Inria associate team about overcoming fundamental bottlenecks that prevent bringing modern bandit theory to real applications. The funding is limited to scientific visits.
B4H (with Inserm, Lille Hospital) is about putting bandits in real clinical environment, with physicians.
BIP-UP (with Inserm, Lille Hospital) is about planning personalized visits for patients follow-up. It is an extension of the project B4H, involving actual clinical follow-up in hospital, in collaboration with surgery specialists and researchers in healthcare.

On-going and submitted grant applications (*or indicate None*) None

B1.c. Early achievement track record

Most important publications I have authored 58 articles in top-tier machine learning conferences and statistics journals. I introduce below various font styles to clarify my role as a contributor (my co-authors may have similar roles, I only use this convention for myself): **Main contributor** with small-caps and bold font, **Key contributor** with bold font. This role can be combined with a Key supervisor role, using underlined font. Standard contributor and supervisor are indicated with standard font. This terminology is further detailed in my extended publication list. See also my 🎓Google sholar profile.

1. Olivier Cappé, Aurélien Garivier, **Odalric-Ambrym Maillard**, Rémi Munos and Gilles Stoltz. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

This paper comes from the fusion of [107] that I did during my PhD with R. Munos (PhD advisor) and G. Stoltz, and [108]. It was decided to use alphabetical order.

2. **Odalric-Ambrym Maillard**, Timothy A Mann, and Shie Mannor. How hard is my mdp?" the distribution-norm to the rescue". *Advances in Neural Information Processing Systems* (NIPS), volume 27, pages 1835–1843, 2014. *Oral, 3% acceptance rate*
This work was done during my postdoc at Technion, Haifa.

3. **Odalric-Ambrym Maillard**. Boundary crossing probabilities for general exponential families. *Mathematical Methods of Statistics*, 27(1):1–31, 2018.
Following this publication, that solves a 30-year old problem as a tribute to the work of T.L. Lai, I was invited as plenary opening speaker at the 20th IBIS (Information-Based Induction Sciences) conference, Sapporo and also at the RIKEN institute (Tokyo). This also sowed seeds for the Inria-Japan associate team RELIANT, which I am now co-heading with J. Honda from Kyoto University.

4. Mahsa Asadi, Mohammad Sadegh Talebi, Hippolyte Bourel, and **Odalric-Ambrym Maillard**. Model-based reinforcement learning exploiting state-action equivalence. In *Asian Conference on Machine Learning* (ACML), pages 204–219. PMLR, 2019. *Best student paper award*

5. Edouard Leurent, Denis Efimov, and **Odalric-Ambrym Maillard**. Robust-adaptive control of linear systems: beyond quadratic costs. in *34th Conference on Neural Information Processing Systems* (NeurIPS), 2020. *Oral, 1% acceptance rate*
This is a collaboration with the Inria Valse team specialized in Control, and Renault. E. Leurent, whom I co-supervised with D. Efimov, received two national prizes for his PhD thesis about Safe Reinforcement Learning for Autonomous Driving.

6. Dorian Baudry, Patrick Saux, and Odalric-Ambrym Maillard. From Optimality to Robustness: Dirichlet Sampling Strategies in Stochastic Bandits. In *35th International Conference on Neural Information Processing Systems* (NeurIPS), Sydney, Australia, December 2021.
This work done with my PhD students introduces provably optimal non-parametric bandit strategies.

7. Hippolyte Bourel, **Odalric-Ambrym Maillard**, and Mohammad Sadegh Talebi. Tightening exploration in upper confidence reinforcement learning. In *International Conference on Machine Learning* (ICML), pages 1056–1066. PMLR, 2020.
Mohammad is my former postdoctoral fellow. This work introduces UCRL3, significantly improving the performances of optimistic RL strategies.

8. Sayak Ray Chowdhury, Aditya Gopalan, and **Odalric-Ambrym Maillard**. Reinforcement learning in parametric mdps with exponential families. In *International Conference on Artificial Intelligence and Statistics* (AI&STATS), pages 1855–1863. PMLR, 2021.
This is a collaboration with A. Gopalan from IISc Bangalore and his student, after I visited IISc Bangalore for a few weeks. I have known Aditya since my postdoc in 2014, and we regularly discuss.

9. Romain Gautron, Odalric-Ambrym Maillard, Philippe Preux, Marc Corbeels, and Régis Sabbadin. Reinforcement Learning for crop management. *Computers and Electronics in Agriculture*, 200:107182, July 2022.

This is a publication in agriculture (not machine learning), done with R. Gautron, in Ph.D. at CIRAD and CGIAR in agronomy. His Ph.D. enabled to initiate solid links with the CIRAD and to step in research with world-experts in agroecology. This opens an exciting avenue of research at the crossing of machine learning, statistic and agronomy, giving motivation to **StaRLux**.

10. Fabien Pesquerel and **Odalric-Ambrym Maillard**. IMED-RL: Regret optimal learning of ergodic Markov decision processes. In *NeurIPS 2022 - Thirty-sixth Conference on Neural Information Processing Systems*, Thirty-sixth Conference on Neural Information Processing Systems, New-Orleans, United States, November 2022.
In this work, done with my student, we obtained the first provably exact optimal (as opposed to order optimal) learning algorithm for regret minimization in MDPs. Though restricted to ergodic MDPs, this strategy, inspired from multi-armed bandits, displays striking performance outperforming the state-of-the-art beyond ergodic MDPs, while being computationally cheap.

Research monographs and any translations thereof.

1. **Odalric-Ambrym Maillard**. *Mathematics of statistical sequential decision making.* Habilitation thesis, Université de Lille, Sciences et Technologies, 2019.

Selected invited presentations (since 2018)
Invited speaker at SAVI workshop, Eindhoven, 2022.
Invited speaker at Institut Montpelliérain Alexander Grothendieck, Montpelllier, 2021.
Invited speaker at JFPDA, Invited chair at CNIA, online, 2021.
Invited speaker at Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 2019.
Invited speaker at Workshop on Empirical Processes and Applications to Statistics, Besancon, 2019.
Invited speaker at 3rd non-stationary day, Institut Henry Poincaré (IHP), Paris, 2019.
Invited speaker at RIKEN institute, Tokyo, 2018.
Opening-conference plenary speaker at 20th Information-Based Induction Sciences (IBIS) conference, Sapporo, 2018.

Prizes/Awards/Academy memberships etc. I received in 2012 the **AFIA prize** (Association Française pour l'Intelligence Artificielle) for the best French Ph.D. thesis in Artificial Intelligence.
Several works that I led received distinctions: NeurIPS oral presentation [109], NIPS oral presentation [110], ACML best student paper [84]. My Ph.D. student E. Leurent received two national Ph.D. prizes.

In 2022, researchers named an algorithm (initially described in my PhD manuscript) after me, see Bian, J., & Jun, K. S. Maillard Sampling: Boltzmann Exploration Done Optimally, In *International Conference on Artificial Intelligence and Statistics* (AISTATS) 2022. It is considered by the authors a replacement for the Thompson sampling algorithm introduced in [18], an extremely popular 90-year old strategy in multi-armed bandits enjoying strong optimality guarantees.

Ability of the PI I am an established researcher in the field of Bandit and Reinforcement Learning theory, with a strong appeal to theoretical foundations to which I contributed significantly over the last ten years. My team and I are highly experienced with the regret minimization proof techniques that we have mastered and developed. I am aware of the main other projects in RL and have direct connections with other world experts in various aspects of the field. I already participated to two FP7 European projects, conducted various projects at national (e.g. ANR JCJC) or international level (e.g. Univ. Kyoto, IISc Bangalore) and nurtured several young researchers in their career. Additionally, I developed, in the recent years a tight connection with researchers in agronomy (CIRAD, CIAT, INRAE) and healthcare (Lille Hospital) hence can access fresh problems and a realistic experimental standpoint. As a result, I am regularly contacted by international students and researchers from various places to organize collaborations. Overall, my expertise and connections make me uniquely qualified to conduct this project to success. Last but not least, I am ready to dedicate the next 15-20 years of my life ensuring that RL companions for assisting agroecology at large becomes a reality.