START PAGE

MARIE SKLODOWSKA-CURIE ACTIONS

Individual Fellowships (IF)
Call: H2020-MSCA-IF-2015

PART B

"OSEGA"

This proposal is to be evaluated as:

[Standard EF]

Contents

0 List of Participants

Participants	**Legal Entity Short Name**	**Academic**	**Non-academic**	**Country**	**Dept. / Division / Laboratory**	**Supervisor**	**Role of Partner Organisation**
Beneficiary							
Lancaster University	ULANC	X		United Kingdom	Department of Mathematics and Statistics	Prof David Leslie	

1 Excellence

1.1 Quality, innovative aspects and credibility of the research

A critical concern in the modern world is security. Effectively protecting transportation systems from malicious attacks, combating traffickers of drugs, firearms and even people, and securing sensitive information on ever-growing cyber-networks, comprise some of the principal axes of this critical task. In all of these problems, maximum security must be obtained with a limited number of available resources. For instance, the total number of security agents is typically less than the number of targets that need to be protected. This calls for the design of appropriate resource allocation techniques which optimise security under the constrained resources available. A critical challenge, yet to be addressed, is to design solutions capable of handling uncertainty about adversaries' interests and the dynamics of the environment.

Security resource allocation and scheduling problems comprise one of many application areas recently been shown to benefit from game-theoretic approaches. As a solid mathematical framework to model strategic decision making, game theory has proved useful in diverse real-world applications including economics, political science and computer science. A problem is cast as a "game" and the objective is to find a solution whereby each "player" makes choices to maximise her own *utility.*A "security game" corresponds to a competition between a defender and an attacker. To solve a security game, all possible actions (attacks and defences) of the two players are enumerated, and for each pair of actions a utility is assigned to each player. In cases where the outcomes are known, game-theoretic approaches have provided impressive results. Since 2007, the so-called ARMOR software[1] has been used at Los Angeles International Airport to determine checkpoints on roadways leading to the airport, and to determine canine patrol routes within terminals. Similar deployments have been made by the US Federal Air Marshals, the US coast guard, and to design the Los Angeles Metro system's fare inspection strategy.[2]

A severe limitation of these models is that they assume the utility functions to be known. In actual fact they must be estimated by experts or obtained from historical data. As a result, estimation errors or lack of historical data (both of which are likely in a quickly-evolving security scenario) may render the game-theoretical solution useless. Therefore it is of importance to use methods that can quickly collect the most relevant data in order to estimate the parameters of the game and quickly reach excellent performance.

This project will deploy approaches of statistics and machine learning to security games played repeatedly between a defender and attackers. Repeated security games allow for collection of data, which can in turn be used to estimate the game and influence future behaviour. Our key objective is to design efficient and theoretically sound, data-driven methods that can actively interact with the environment to *learn* and *act* in security games of realistic scales, which must therefore be *practical*, *scalable* and *robust.*

State-of-the-art

Security meets Game Theory From a game-theoretic perspective, a security problem is viewed as a two-player game that captures the interaction between a defender (e.g., border patrols, metro inspectors, network administrators) and attackers (e.g., terrorists/smugglers, illegal metro users, malicious cyber attackers). The action of the defender (attacker) is defined as selecting a subset of targets to protect (attack). For each defender/attacker action pair, *utilities* are defined as the players' gain or loss, and the players' objectives are to maximise their own utility. The expected utilities for both players can be stored in two matrices, $\boldsymbol{A}$ for the defender, $\boldsymbol{B}$ for the attacker. Entries $\boldsymbol{A}_{i,j}$ and $\boldsymbol{B}_{i,j}$ are the expected utilities for the defender and attacker, respectively, when the defender plays strategy i and the attacker responds with strategy j. Solutions to such games rely on randomised (mixed) strategies, making each player's behaviour unpredictable to the other. If a fully competitive setting, in which the attacker's gain is the defender's loss (i.e. zero-sum games) a fully robust strategy can be calculated for the defender, in that it provides guaranteed performance against *any* possible attacker, even if the defender's strategy is completely revealed to the

[1] J. Pita et al. "Deployed ARMOR protection: the application of a game theoretic model for security at the Los Angeles International Airport". In: *Conference on Autonomous agents and multiagent systems: industrial track.* 2008, pp. 125–132.

[2] J. Tsai et al. "IRIS-a tool for strategic security allocation in transportation networks". In: *Conference on Autonomous agents and multiagent systems.* 2009; E. Shieh et al. "Protect: A deployed game theoretic system to protect the ports of the united states". In: *Conference on Autonomous Agents and Multiagent Systems.* 2012, pp. 13–20; Z. Yin et al. "TRUSTS: Scheduling randomized patrols for fare inspection in transit systems using game theory". In: *AI Magazine* 33.4 (2012), p. 59.

attacker. A generalisation to more general (non zero-sum) games, which is particularly relevant to security games, is the Stackelberg equilibrium[3] in which the defender's mixed strategy is first publicised and the attacker plays a best response to this mixed strategy. It is this Stackelberg security game framework that will be considered in the proposed project.

Uncertainty in Security Games Most standard game-theoretical analyses assume that the payoff matrices are known in advance. However uncertainty is endemic in most real-world applications. For instance,the random selection of passengers for security checks at an airport is a source of uncertainty in this game, where probability of successful security enforcement is unknown in advance. As also confirmed by several empirical studies in fraud and cybercrime detection, this phenomenon can significantly decrease the defender's performance.[4] Extensive studies have been dedicated to the design of security games that are robust with respect to uncertainty about the environment.[5] However, much more can be done in the case of *repeated* security games. The repetition allows the defender to reduce her uncertainty about the world and intelligently *learn* how to improve performance over time. Specifically, a security game solver could autonomously take intelligent decisions at repeated instances of the game, to carefully collect and act upon the most relevant available data. As discussed further below, this is precisely the area where mathematical machine learning has the strongest results.

Bandit problems Mathematical machine learning is a modern amalgamation of statistics and optimisation. A fundamental problem in machine learning is the *multi-armed bandit.* It is a one-player repeated game in which, on iteration t, the player selects action $i(t)$ and received reward l_t, a random variable with expectation $a_{i(t)}$. An important constraint is that the player does not observe the reward that would have been collected had another action been selected instead. Solutions to this problem have found many practical applications from adaptive routing in a network, to web advertising, to clinical trials.[6]

In the bandit problem two related tasks are commonly considered. One is the *online decision-making* task, in which the reward for each decision must be taken into account; this task is relevant to the immediate deployment of the system which learns as it acts. The standard performance metric here is the regret, $T \max_i a_i - \sum_{t=1}^{T} a_{i(t)}$, giving the difference between the maximal expected reward that could have been achieved if full information were available in advance, and the (expected) reward for the selected actions. The other task is that of *pure exploration*, in which a learning phase is permitted during which received rewards do not matter, allowing a training phase prior to system deployment. The performance metric in pure exploration is the quality of the arm selected immediately after the training phase. Application of pure-exploration to action selection in robotic planning has been extensively studied by Dr Gabillon.[7]

In general security games, one can conceive of a non-adaptive attacker who uses a fixed distribution over actions through time. This results in a bandit problem for the defender. Bandit strategies are therefore important for active learning in security games. However it is as yet an important open question how to take into account adaptive and strategic behaviour of the attacker and retain effective performance guarantees. It is this aspect of learning in security games which will be developed in the proposed project.

Existing results Some recent advances have been made for learning in security games. Most approaches focus on the case where the attacker's preferences are not fully known and are learned through repeated plays of the game: some analyse the number of queries required to learn the optimal defender's strategy,[8] others take a Bayesian approach, with techniques based on Partially Observable Markov Decision

[3] D. Korzhyk et al. "Stackelberg vs. Nash in Security Games: An Extended Investigation of Interchangeability, Equivalence, and Uniqueness." In: *Journal Artificial Intelligence Reseqrch* 41 (2011), pp. 297–327.

[4] J. S. Granick. "Faking It: Calculating Loss in Computer Crime Sentencing". In: *ISJLP* 2 (2005), p. 207; P. Swire. "No cop on the beat: Underenforcement in e-commerce and cybercrime". In: *J. on Telecomm. & High Tech. L.* 7 (2009), p. 107.

[5] M. Aghassi and D. Bertsimas. "Robust game theory". In: *Mathematical Programming* 107.1-2 (2006), pp. 231–273; T. H. Nguyen et al. "Regret-based optimization and preference elicitation for stackelberg security games with uncertainty". In: *AAAI Conference on Artificial Intelligence.* 2014, pp. 756–762; C. Kiekintveld, T. Islam, and V. Kreinovich. "Security Games with Interval Uncertainty". In: *Conference on Autonomous Agents and Multi-agent Systems.* 2013, pp. 231–238.

[6] S. Bubeck and N. Cesa-Bianchi. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems". In: *arXiv preprint arXiv:1204.5721* (2012).

[7] V. Gabillon et al. "Multi-Bandit Best Arm Identification". In: *Neural Information Processing Systems.* 2011, pp. 2222–2230.

[8] A. Blum, N. Haghtalab, and A. D. Procaccia. "Learning optimal commitment to overcome insecurity". In: *Advances in Neural*

Processes used to update a posterior over the adversary's preferences.[9] Recently, a more relevant analysis has been given[10] for the case of multiple attackers, where at each round of the game, a single attacker is chosen adversarially from a fixed, finite, set of known attackers. This corresponds to a case where the utility matrix $\boldsymbol{B}$ is chosen adversarially from a set of k known matrices, and shows strong connections with adversarial bandit theory. However all of these security games results rely on restrictive assumptions about prior knowledge and observability, and also on the number of available actions to each player being reasonably small. In this proposal we consider realistic feedback settings, and combinatorial actions (choosing which k of the n potential targets to protect) so the total number of actions available is enormous.

Vision The purpose of this project is to create *practical*, *scalable* and *robust* methods for security games. First, we target *practicality* in the sense that our methods will be autonomous in handling uncertainty in the model and will actively reduce this uncertainty by interacting with the environment in which the game takes place during repeated plays of the game. We will do so by combining existing security game research with the extremely active field of multi-armed bandit research. Second, we target *scalability* so that the methods will apply with an extremely large number of possible actions. This will be achieved by making simplifying structure assumptions, such as a combinatorial structure in which an action consists of selecting k objects from n. Finally *robustness* is a key issue in security games in several senses. Methods should not break down in the face of adversarial and adaptive play by the attacker. Furthermore methods should not perform well only in average — worst case performance is extremely important. We will therefore devise a theoretically sound approach in which *practical*, *scalable* and *robust* algorithms are developed and finite time performance guarantees are provided.

Objective 1: Scalability in Pure Exploration Bandits via Submodularity. Given the combinatorial nature of many security games, it is necessary to use this structure intelligently. In particular, naively enumerating all possible actions makes the computational constraints of standard algorithms intractable. We will address this problem by first studying combinatorial bandit techniques,a central part of Dr Gabillon's area of expertise. A key observation is that in many cases the players' performance utility function is submodular. One example is in the maximal coverage problem for sensor (or checkpoint) placement.[11] This submodularity property can in turn be used to provide tractable and almost optimal algorithms for bandits.[12] Objective 1 is therefore to complete the analysis of combinatorial bandits by addressing the (as yet unsolved) pure exploration problem under submodularity assumptions.

Objective 2: Pure Exploration in Security Games. As discussed previously, a major barrier to applying security games to real-world scenarios is that the players' utility matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are unknown and must be learned. In the first of two objectives on standard security games we consider the case where the defender is in fact able to safely examine her defensive strategies before applying them online. A real-world example is an airport security system, which can be tested many times before being deployed as the principal defence scheme. During the test phase, the defender is able to probe an entry of its utility matrix at every repetition of the game. The value observed is *a noisy version* of the true entry $\boldsymbol{A}_{i,j}$.

We will design a strategy for the defender to either minimise the number of tests needed to identify an excellent strategy with a given level of confidence or to maximise her probability of identifying the best strategy given a fixed number of tests.[13] In order to extend these classical results to the setting proposed

Information Processing Systems. 2014, pp. 1826–1834; J. Letchford, V. Conitzer, and K. Munagala. "Learning and approximating the optimal strategy to commit to". In: *Algorithmic Game Theory.* Springer, 2009, pp. 250–262.

[9] J. Marecki, G. Tesauro, and R. Segal. "Playing Repeated Stackelberg Games with Unknown Opponents". In: *Conference on Autonomous Agents and Multiagent Systems.* 2012, pp. 821–828; Y. Qian et al. "Online planning for optimal protector strategies in resource conservation games". In: *Conference on Autonomous agents and multi-agent systems.* 2014, pp. 733–740.

[10] M.-F. Balcan et al. "Commitment without regrets: Online learning in Stackelberg security games". In: *Proceedings of the Sixteenth ACM Conference on Economics and Computation.* 2015.

[11] A. Krause, A. Roper, and D. Golovin. "Randomized sensing in adversarial environments". In: *IJCAI Proceedings-International Joint Conference on Artificial Intelligence.* Vol. 22. 3. 2011, p. 2133.

[12] V. Gabillon et al. "Adaptive Submodular Maximization in Bandit Setting". In: *Neural Information Processing Systems.* 2013, 2697–2705.

[13] O. Maron and A. Moore. "Hoeffding races: Accelerating model selection search for classification and function approximation". In: *Neural Information Processing Systems.* 1993; J.-Y. Audibert, S. Bubeck, and R. Munos. "Best Arm Identification in Multi-Armed

above, we will first carefully characterise the data-dependent hardness of the problem, extending recent relevant results for combinatorial bandits[14] and similar results currently in preparation by Dr Gabillon. Of course, in games these complexity results are more challenging than in the bandit problem since the complexities depend also on the actions available to the attacker. Therefore we will study this problem by gradually increasing its difficulty with different partial feedback structures. First we note that recent work[15] on query complexity, corresponds to the deterministic version of this problem where it is assumed that the probing outcome corresponds to the *true value* of $\boldsymbol{A}_{i,j}$. Therefore our first approach will be to combine ideas from pure exploration and the query complexity setting in a context where the defender can individually sample from any entry of the matrix. A second more challenging setting will be to consider the adversarial learning problem where the defender chooses strategy i but Attacker chooses j; the attacker might be either oblivious to the defender, or playing a Stackelberg best-response to the defender strategy.

Objective 3: Regret Analysis of Repeated Security Games. In some applications a newly created security system is not provided with any historical data and cannot be tested before being used in production. Here, the learning must be performed online while playing the security game. In this online decision-making context it is of high importance for the agent to learn the utilities as fast as possible. This means that only demonstrating asymptotic convergence[16] is inadequate. To address online decision-making in security games we will here assume that the utility matrices are unknown to the defender and that, at each repetition of the game, the attacker will best respond to the defender's strategy (if π is a mixed strategy of the defender, then denote $b(\pi)$ the best response of the attacker). Therefore we are considering a setting that is related to the analysis of Stackelberg equilibrium. A natural quantity of interest is the *cumulative regret*, defined as

$$R(n) = T \max_{\pi} \left[\pi A b(\pi)\right] - \sum_{t=1}^{T} \pi_t A b(\pi_t).$$

This compares the actual reward received (assuming the attacker always performs as well as they can) to the reward achieved at Stackelberg equilibrium (when the defender chooses the best possible mixed strategy under the knowledge that the attacker will best respond to it). The objective is for the defender to build a series of defence strategies π_t for $t = 1, \ldots, T$ to minimize the cumulative regret $R(T)$.

This game-theoretic scenario, is actually strongly related to the bandit problem, in that the best-responding assumption on the attacker leads to a situation where the reward to the defender depends only on her ow (mixed) strategy. One solution is thus to treat it as a bandit problem with continuous action space consisting of the set of all probability distributions on the original discrete action space. However a more efficient solution is likely to be obtained by explicitly considering the game-theoretical nature. In particular, most current approaches for online decision-making in bandits, such as upper confidence bound methods,[17] implement a strategy that is optimistic in face of uncertainty. It will be extremely interesting to discover if this optimism principle still holds in an adversarial game, or whether a more cautious approach is needed.

Finally an interesting additional requirement is to learn security strategies that are not only of good quality in average but also whose performance is not subject to large variance when used on a daily basis. This *risk-averse* requirement has been well-studied in the statistical community, and has recently been considered in a multi-armed bandit framework.[18] An implementation in the security games specific context is a very natural extension to the main body of work in this objective.

Objective 4: Learning in combinatorial games. Objective 4 will be devoted to solving security games with more complex action structures. Real-world security problems often involve large, complex

Bandits". In: *Conference on Learning Theory*. 2010, pp. 41–53.

[14] S. Chen et al. "Combinatorial pure exploration of multi-armed bandits". In: *Neural Information Processing Systems*. 2014, 379–387.

[15] P. W. Goldberg and S. Turchetta. "Query Complexity of Approximate Equilibria in Anonymous Games". In: *arXiv:1412.6455* (2014).

[16] D. S. Leslie and E. Collins. "Generalised weakened fictitious play". In: *Games and Economic Behavior* 56.2 (2006), pp. 285–298; A. C. Chapman et al. "Convergent Learning Algorithms for Unknown Reward Games". en. In: *SIAM Journal on Control and Optimization* 51.4 (Jan. 2013), pp. 3154–3180.

[17] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-Time Analysis of the Multi-Armed Bandit Problem". In: *Machine Learning* (2002).

[18] A. Sani, A. Lazaric, and R. Munos. "Risk-Aversion in Multi-armed Bandits". In: *Neural Information Processing Systems*. 2012.

networks. This includes, for instance, complex routes, or computer/communication networks. The size of the action spaces for both attacker and defender is often combinatorially large. Standard results with convergence times that increase with the size of the action spaces become extremely weak in such settings. We will therefore take advantage of the inherent combinatorial structure of the problem to create efficient and computationally tractable algorithms in these large games. In particular, we will develop the approaches of both Balcan in simple Stackelberg games and in adversarial combinatorial bandits.[19] The issue of scalability will be addressed in light of the results found in Objective 1.

Objective 5: Repeated Network-Security Games. As a more concrete application of Objective 4, this objective will focus on the particular combinatorial structure that is a graph as this structure is present in numerous real-word applications. In light of the ever-growing, modern, social and communication networks, a canonical example is that of smuggler arrest in a network.[20] This has received significant attention in the community, especially in response to the Mumbai attacks of 2008, after which Mumbai Police started to schedule a limited number of inspection checkpoints on the road throughout the city. This problem has not yet been studied in its repeated form, where a pursuit-evasion game is played multiple times against a population of smugglers. Therefore, the currently deployed strategy of the defender is *not adaptive* to observations collected about the attackers' historical strategy and is therefore sub-optimal. We will therefore develop adaptive strategies, using the approaches developed in Objectives 1–4. However, the graph structure provides additional constraints on the action spaces, and provides additional information, when compared with the general combinatorial problem. In particular, the set of actions available to the attacker is restricted to a set of paths through the network, and the the graph structure provides strong information about sensible choices of checkpoints (for example, aligning them all along one route through the network is a particularly unfortunate choice, but is not ruled out by a generic combinatorial structure). Furthermore, absence or presence of a smuggler on one link of the graph will likely provide information about which other other links were utilised on that iteration. Therefore the objective is to design specific algorithms in in situations where it is possible to take advantage of specific graphical structure of the problem. We will start by defining a notion that captures the hardness of the task depending on characteristics of the graph that we would discover. Note that, although the goal is to generate algorithms for security games on graphs, the results to be obtained will be expected to lay grounds for research in a more general setting of active learning with graph structure. Dr Gabillon has held initial discussions on this topic with Dr Michal Valko, a world-famous expert in active learning on graphs and part of INRIA Lille in France. This project will allow the formation of a productive and lasting collaboration with Dr Valko, which will be greatly beneficial not only in achieving the this objective, but also to strengthen international links between Lancaster University in the UK and INRIA in France.

1.2 Clarity and quality of transfer of knowledge/training for the development of the researcher in light of the research objectives

The overall training objective is to significantly develop Dr Gabillon's scientific, organisational, communication and technology transfer skills. This will enable him to continue building his portfolio of outstanding research to attain a position of independence and gain recognition in the international research community.

Dr Gabillon is already an expert in the modern theory of bandits, including best arm identification, and reinforcement learning. Therefore this project's main scientific training objective will be to develop his skills and knowledge in statistical learning methods and game theory. Prof. Leslie is an expert in both areas. Further expertise in Lancaster from which Dr Gabillon will learn includes the Statistical Learning group, in which he will be based, and the broader Statistics Research Group.

Lancaster University is the leading UK institution in bandit theory, with expertise in index policies (Glazebrook, Kirkbride, Jacko), Thompson sampling and contextual bandits (Grunewalder, Leslie) and application in medical trials (Vilar). Dr Gabillon will have ample opportunity to further develop his

[19] Balcan et al., "Commitment without regrets: Online learning in Stackelberg security games"; N. Cesa-Bianchi and G. Lugosi. "Combinatorial bandits". In: *Journal of Computer and System Sciences* 78.5 (2012), pp. 1404–1422.

[20] M. Jain et al. "A double oracle algorithm for zero-sum security games on graphs". In: *Conference on Autonomous Agents and Multiagent Systems.* 2011, pp. 327–334.

expertise in this area, and indeed brings expertise from a complementary aspect of online learning and decision-making, especially in the area of combinatorial bandit problems. A specific side benefit of Dr Gabillon's fellowship at Lancaster will be transfer of his expertise in best-arm identification to the Medical and Pharmaceutical Statistics research group. He will present his research in this area to the research group and discuss possible applications in clinical trial design.

In addition to scientific skills, Dr Gabillon will learn from Lancaster's world-leading expertise in industrially-inspired statistics. Statistical researchers in Lancaster have constant exposure to external companies, through the STOR-i Centre for Doctoral Training, and the Data Science Institute. While embedded in this culture, Dr Gabillon will be given the opportunity to:

1. Gain further experience of developing industry/academic partnerships by working with Profs. Leslie and Eckley and other staff in STOR-i and the Data Science Institute in technology transfer activities.
2. Develop public communication skills by presenting research results to varied audiences.
3. Participate in the organisation of workshops in Lancaster and at the Royal Statistical Society.
4. Receive training on applying for funding by co-authoring proposals for UK and EU funding agencies.
5. Attend training designed specifically for early-career researchers, including the Research Development Programme, designed to promote impactful research and support extra-disciplinary development.
6. Participate in teaching and research supervision (undergraduate and graduate). This will not be obligatory, but Dr Gabillon will have the opportunity to benefit from peer observation and mentoring.

Lancaster University is fully committed to the UK Concordat to Support the Career Development of Researchers, an agreement between funders and employers of research staff to improve the employment and support for researchers and research careers in UK higher education. Furthermore, Dr Gabillon will adhere to the European Charter for Researchers, and the training objectives will be managed through a Personal Career Development Plan that Prof. Leslie and Dr Gabillon will write together. This plan will be revised regularly throughout the fellowship to ensure that all objectives are met.

1.3 Quality of the supervision and the hosting arrangements

Qualifications and experience of the supervisor(s)

Prof. Leslie leads the Statistical Learning research group in the Department of Mathematics and Statistics, Lancaster University, and is Theme Lead for Foundations in Lancaster University's new Data Science Institute. He is a world-leading researcher in statistical learning, Bayesian inference, decision-making and game theory, with collaborators from France, Singapore, USA and Australia. His research on contextual bandit algorithms[21] is used by many of the world's largest companies to balance exploration and exploitation in real-time website optimisation. He is expert in the mathematics of learning in games, stochastic approximation, and the mathematics of statistically-inspired reinforcement learning.[22] Prof. Leslie is the holder of a Google Faculty Award which funds a student to investigate multiple-action selection in bandits. He is co-director of the £1.5m EPSRC-funded cross-disciplinary decision-making research group at the University of Bristol, and was on the management team of the £5.5m ALADDIN project, a large strategic partnership between BAE Systems and EPSRC, involving researchers from Imperial College, Southampton, Oxford, Bristol and BAE Systems. In the 10 years since taking up a Faculty position, Prof. Leslie has supervised 17 PhD students (5 now in Academic positions), 2 post-doctoral fellows, numerous MSc and undergraduate dissertations, and an undergraduate secondment from ENS Lyon.

[21] B. May et al. "Optimistic Bayesian sampling in contextual-bandit problems". In: *Journal of Machine Learning Research* (2012), 2069–2106.

[22] D. S. Leslie and E. J. Collins. "Convergent Multiple-timescales Reinforcement Learning Algorithms in Normal Form Games". In: *Annals of Applied Probability* 13.4 (2003), pp. 1231–1251; S. Perkins and D. Leslie. "Asynchronous stochastic approximation with differential inclusions". en. In: *Stochastic Systems* 2 (2012), pp. 409–446; Leslie and Collins, "Generalised weakened fictitious play"; T. Larsen et al. "Posterior Weighted Reinforcement Learning with State Uncertainty". In: *Neural Computation* 22 (2010), pp. 1149–1179; S. Perkins and D. Leslie. "Asynchronous stochastic approximation with differential inclusions". en. In: *Stochastic Systems* 2 (2012), pp. 409–446; Chapman et al., "Convergent Learning Algorithms for Unknown Reward Games"; S. Perkins and D. S. Leslie. "Stochastic Fictitious Play with Continuous Action Sets". In: *Journal of Economic Theory* 152 (2014), pp. 179–213.

Hosting arrangements

Dr Gaabillon will be embedded within the statistical learning group, lead by Prof. Leslie. This is a team of 5 academic staff and around 5 PhD students within the Department of Mathematics and Statistics. Dr Gabillon will participate in weekly group meetings discussing research direction and management, personal development, workshop organisation, teaching, and other aspects of academic life. The group also has extremely strong links with both the Data Science Institute (www.lancaster.ac.uk/dsi/) and the STOR-i Centre for Doctoral Training (www.stor-i.lancs.ac.uk/); each provides a weekly seminar series. These exciting initiatives provide numerous further opportunities for informal mentoring in addition to that provided by the official mentor scheme. To ensure integration Dr Gabillon will be invited to deliver a research seminar to each network, and will participate in the respective away days.

1.4 Capacity of the researcher to reach and re-enforce a position of professional maturity in research

Professional maturity for Dr Gabillon consists of leading a dynamic research group actively working on mathematical problems at the interface of game theory and online learning, with strong impact in real-world applications. In a short period of time, Dr Gabillon has already developed broad expertise in the domain, with a strong publication record in both theory and applications. Indeed, a significant result of his PhD thesis is in bringing classical reinforcement learning algorithms closer to daily life.

During a 6-months internship at a major US R&D lab (Technicolor Research Laboratory, Palo Alto), Dr Gabillon developed collaborations with a team of industrial researchers. He quickly became productive and the internship produced two peer-reviewed papers at prestigious international conferences. He also obtained valuable knowledge about industrial research and its interaction with academia. This has given him a crucial understanding of the pathways to impact and how to establish industrial collaborations.

At the start of the fellowship, Dr Gabillon will be closely mentored by Prof. Leslie at Lancaster University, and will have access to the university's research resources. He will thus further develop his research and supervision skills, which will greatly contribute to achieving professional maturity.

2 Impact

2.1 Enhancing research- and innovation-related skills and working conditions to realise the potential of individuals and to provide new career perspectives

Dr Gabillon is already a leading researcher in the mathematics of bandit algorithms and reinforcement learning. This fellowship provides training in two key additional competences. Firstly, he will develop an in depth knowledge of relevant cutting edge statistical theory, through working with leading scientists in statistics and operations research at Lancaster University, and Lancaster's many international visiting researchers. Secondly, Prof. Leslie is a leading expert on learning in games, and will mentor Dr Gabillon to bring ideas from bandits into the game theoretical framework. This significant broadening of the researcher's skill set will give him an extremely solid foundation on which to build a future research career.

In addition to research development, Dr Gabillon will work within Lancaster University's extremely effective framework for industrial collaboration. He will develop skills to ensure mutually beneficial outcomes from industry/academia relationships, which is fast-becoming a key skill for academics. The Department of Mathematics and Statistics, is a world-leader in developing such relationships; Dr Gabillon will both be introduced to prospective industrial partners, and receive mentoring as he develops his own relationships.

2.2 Effectiveness of the proposed measures for communication and results dissemination

Lancaster University's Data Science Institute and Knowledge Business Centre[23] will be inaugurating a "Data Science Network' which will bring together academic data scientists with local companies in regular show and tell sessions. Dr Gabillon will be a regular participant at these events, enabling bi-directional communication of opportunities and requirements, and the building of a network of industry contacts. In addition, Lancaster University supports researchers to write for the Conversation, a news service delivering

[23] The Knowledge Business Centre is an innovation hub providing a gateway for business/academic interaction which allows the transfer of expertise between Lancaster's academics, regional businesses and community partnerships through training and technology transfer activities

articles directly from researchers to the public; Dr Gabilon will make use of this support to produce expository articles explaining the benefits that adaptive data science approaches can deliver to society. Finally, to ensure successful public engagement, Dr Gabillon will attend Lancaster University's "Engaging Researcher Course" which explores public engagement activities for researchers.

The research generated in this project will of course be published Open Access in the world's leading academic journals and conferences, and all code generated will be released under standard Open frameworks. Prof. Leslie currently works with several companies, both large and small, including the Defence Science and Technology Laboratory who have a current interest in security games. Dr Gabillon will be mentored to develop similar relationships. He will also work with Security Lancaster (www.lancs.ac.uk/security-lancaster) to ensure the results of the current project are shared with relevant industrial and government partners. A particularly successful mechanism deployed extensively at Lancaster is the industrially-sponsored MSc or PhD project, which allows the supervisor's research to be both developed and deployed directly within a company; Dr Gabillon will be encouraged to join appropriate supervisory teams to help both disseminate the project's research and develop an industrial research network to enhance his future career.

3 Implementation

3.1 Overall coherence and effectiveness of the work plan, including appropriateness of the allocation of tasks and resources

Work packages

WP1: Combinatorial bandits Dr Gabillon will develop new approaches to combinatorial bandits (Objective 1), investigating the pure exploration problem and the online regret problem. This WP builds upon current research of Dr Gabillon and can be completed in months 1–6. This will result in **Deliverable 1.1**, a paper on pure exploration in combinatorial bandits, and **Deliverable 1.2**, a paper on regret in combinatorial bandits with submodular reward structures.

WP2: Security games Objectives 2 and 3 will be considered in this Work Package, which will develop algorithms for both pure exploration and online performance guarantees in security games. This WP brings together the knowledge of the Researcher and the Supervisor, and will therefore also be started in Month 1. However it will take longer to complete due to the greater level of novelty for Dr Gabillon, so will continue for 12 months. **Deliverable 2.1** is a paper on pure exploration in security games; **Deliverable 2.2** is a paper on online regret bounds in security games.

WP3: Combinatorial security games Dr Gabillon will address Objective 4 with the development of methods for security games with combinatorial action spaces. This package combines the results of WP1 and WP2, and will initially be carried out in parallel with WP2, running through until the end of the project. **Deliverable 3.1** is a paper presenting the results of this research.

WP4: Network defence Objective 5 will be addressed, with the application of combinatorial work (WP1 and WP3) to network problems. Dr Gabillon will visit Dr Michal Valko learn about techniques in networks and how to integrate them with the previously-developed combinatorial results. He will also collaborate with researchers from Security Lancaster to develop applications in supply chain protection. This package will commence after WP2 has been completed in month 12, and run until the end of the project. **Deliverable 4.1** is a paper describing a bandit approach to network defence, and **Deliverable 4.2** is a paper describing the game-theoretical results.

Major milestones

Milestone 1 is the completion of WP1. If strong results are obtained in this first phase of research, then extending to combinatorial games under a Stackelberg response assumption will be relatively straightforward, and greater emphasis can be placed on WP3 straight away. If the WP1 results are not so strong, more effort will need to be given to WP2 in order to obtain foundations for WP3.

Milestone 2 is at the end of month 12 of the project. Results from WP1 and WP2 will be assessed and the problems to be addressed in WP4 decided. If the game-theoretical results in WP2 are strong, and WP3 is progressing well, then the full game-theoretical approach can be addressed directly in WP4. Weaker game-theoretical results may necessitate an initial focus on bandit approaches in WP4.

3.2 Appropriateness of the management structure and procedures, including quality management and risk management

The Research Support Office at Lancaster University has extensive experience of managing European project grants, and will be responsible for administering the project budget, legal aspects and potential commercial exploitation. Dr Gabillon will be a member of the Dept. of Mathematics and Statistics, and more specifically the Statistical Learning group lead by Prof. Leslie. He will also be assigned a formal mentor under standard Lancaster University procedures, who will be a second point of contact. Dr Gabillon will be responsible for the research work, and will meet weekly with Prof. Leslie to discuss results, challenges and research strategies. Dr Gabillon will also be responsible for the project management; he will be assisted in this through monthly management and mentoring meetings with Prof. Leslie, in which progress against the workplan and career development plan will be discussed.

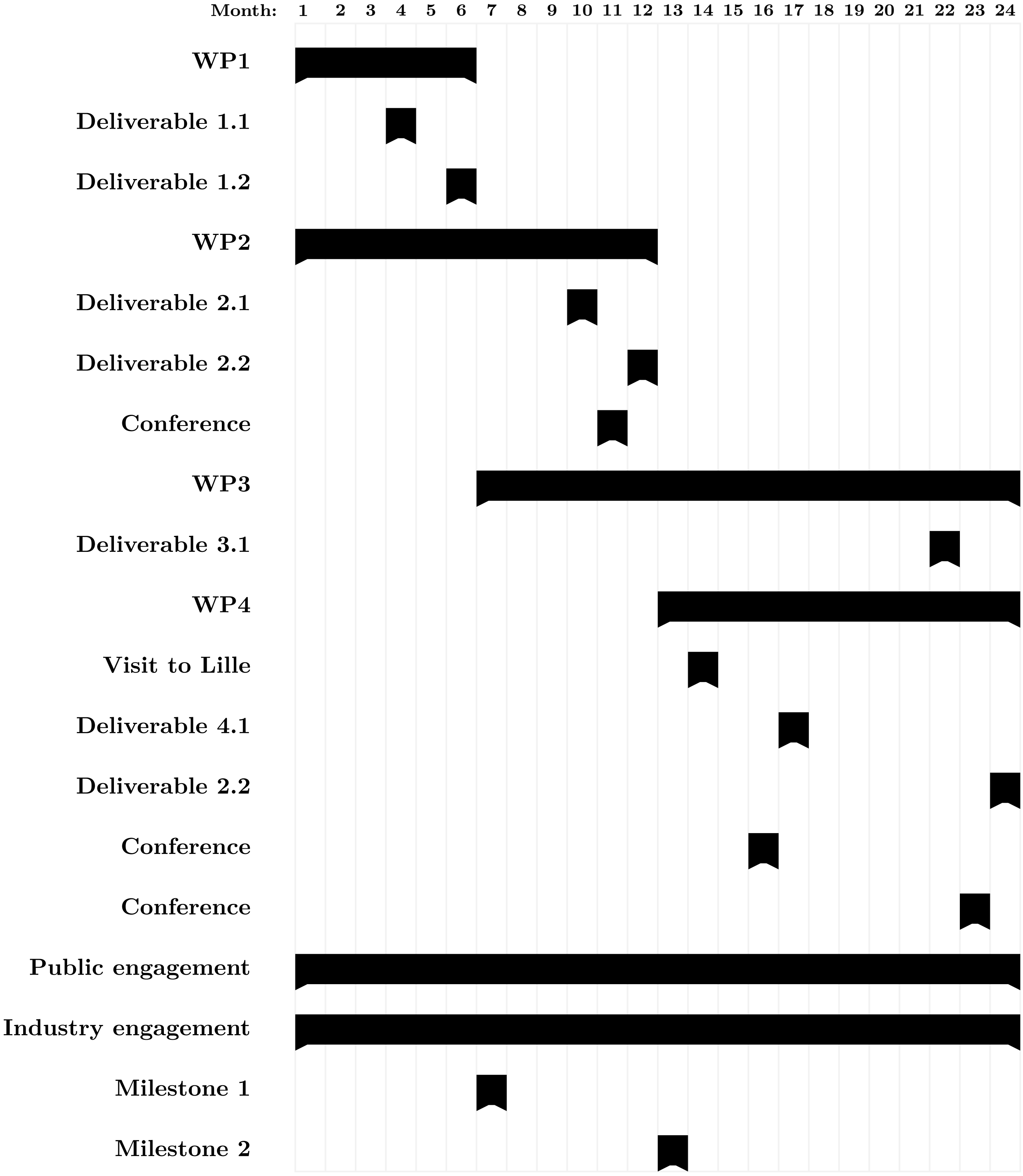
There are risks at each stage of this ambitious research project. The two-pronged approach mitigates some of this risk: if combinatorial cause difficulty then greater focus will be placed on the game theory, and vice versa. Both of WP1 and WP2 contain elements which are lower-risk while still likely to yield high-quality research outputs. A solid foundation can thus be laid while Researcher and Supervisor develop a relationship, in preparation for more ambitious objectives in the latter part of the project.

3.3 Appropriateness of the institutional environment (infrastructure)

Dr Gabillon will be hosted in the Department of Mathematics and Statistics, Lancaster University. Prof. Leslie will provide the main mentorship and research supervision. The Statistical Learning group, and the Statistics Research Group beyond that, will provide further immediate support to Dr Gabillon. These research groups work closey with the departments of Operations Research (through the STOR-i Centre for Doctoral Training) and Computer Science (through the Data Science Institute). Many researchers in cognate areas will thus contribute informal research leadership and mentoring, as well as providing an environment with multiple relevant research seminars. In terms of physical resources, the Department will provide high quality office space and standard IT facilities, including high performance computing.

3.4 Competences, experience and complementarity of the participating organisations and institutional commitment

The Department of Mathematics and Statistics at Lancaster University was ranked fifth equal in the United Kingdom in the most recent Research Excellence Framework assessment. The Department has a thriving research environment, with 50 faculty, 11 post-doctoral fellows, and 72 PhD students. The Department has numerous government- and industry-funded research projects, many of which relate to industrially-motivated statistics and operations research and are related to the currently-proposed project. The skill set of Dr Gabillon complements that of the Beneficiary by providing expertise in current algorithmic approaches to bandit algorithms and reinforcement learning. The host institution in return provides expertise in statistical methodology appropriate to online inference, and game theoretical learning, and a strong track-record of working with industry to ensure the fundamental research is relevant and generates impact. In addition Dr Gabillon will develop links with Security Lancaster (www.lancaster.ac.uk/security-lancaster/), in which researchers are currently addressing the security of supply chains using game-theoretical approaches, to both develop test cases for the current research project and build links with their network of industry and government collaborators.


Month: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
WP1
Deliverable 1.1
Deliverable 1.2
WP2
Deliverable 2.1
Deliverable 2.2
Conference
WP3
Deliverable 3.1
WP4
Visit to Lille
Deliverable 4.1
Deliverable 2.2
Conference
Conference
Public engagement
Industry engagement
Milestone 1
Milestone 2

4 CV of the Experienced Researcher

My primary research interest involve the investigation of machine learning techniques to create algorithms that, in some way, adapts to its users, or more generally learns from its environment. The approach is both theoretical and application oriented. A major objective is to ensure the proposed algorithms capture the real complexity of a problem and to study their scalability towards real-world applications.

During my PhD, I investigated *reinforcement learning* which is a field where an agent has to learn from its environment in order to maximise some measure of long-term performance. More precisely, the focus was on a class of algorithms called "Classification-based Policy Iteration" (CBPI) which are algorithms that learn directly the policies as output of a classifier. Thus they avoid, as in the standard RL techniques, to directly rely on value function as the approximation of these is sometime poorer than the policy approximation. To improve the CBPI sample-efficiency, I proposed new hybrid approaches using both value function and policy approximations in the CBPI framework that leverage the benefits of both approaches (which led to two publications in ICML 2011 & 2012 while a journal paper has been published in JMLR). Moreover, we applied our techniques in the game of Tetris, a domain where RL techniques had obtained poor results, and learned a controller removing on average 50.000.000 lines (the best in the literature, to the best of our knowledge which is reported in a paper in NIPS 2013).

I also investigated *bandit problems*. Bandit problems are the core mathematical formulation for modelling of adaptive and sequential decision-making. My interest was to design algorithms that solve challenging combinatorial problems. A first step has been to tackle the case of parallel bandit problems (which led to two publications in NIPS 2011 & 2012). As a post-doctorate, I am currently studying the extension to more general and challenging combinatorial settings. Moreover, I also made use of the bandit approach to prospose new solutions to industry problems during two of my internships. During my research internship at Technicolor Labs, while trying to improve the questionnaire asked to elicit movie preferences of users for a recommendation website, a combinatorial learning problem arose. We designed a scalable bandit algorithms by using the submodular property that naturally arises in this problem (this led to a publication in NIPS 2013 & AAAI 2014). During my master internship I proposed a combination of bandits techniques with linear programming to design an ad server software for the Orange Telecommunications company that takes into account limited budget constraint other various ad categories (the subsequent publication was awarded a best paper award in a French conference). My research also involves more theoretical objectives. As a postodoctoral fellow, I currently investigate the fundamental learning limits of the non-stochastic (adversarial) combinatorial bandit problem. The goal is to give a simple formulation of this bandit game that admits an exact minimax solution. Seen as a two-player game this problem is very closely connected to game theory that we are investigating (a subject also very relevant to this proposal).

Curriculum Vitae of the Applicant, Victor Gabillon

Education

PhD in Computer Science (Accessit Award of the AI French Association, AFIA) **June 2014**
Team SequeL, INRIA Lille - Nord Europe, France
Title: "Budgeted Classification-based Policy Iteration"
Domains: Reinforcement learning & Bandits games
Supervisors: Mohammad Ghavamzadeh & Philippe Preux
Examiners: Peter Auer (Leoben University), Olivier Cappé (Télécom ParisTech), Shie Mannor (Technion) and Csaba Szepesvári (Alberta University)

M.Sc. Image Processing & Statistical Learning with honours **Sep 2009**
École Normale Supérieure, Cachan, France

Engineering Degree in Information Technology **Sep. 2009**
TELECOM SudParis, Évry, France

Professional Activities

Postdoctoral Research Fellow in Statistics *full time* **Nov 2015 – ongoing**
Supervisor: Peter Bartlett
School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia
PhD Researcher *full time* **Oct 2009 – June 2014**
Team SequeL, INRIA Lille - Nord Europe, France
Research Engineer *full time* **Mar 2013 – Sep 2013**
Technicolor Research Group, Palo Alto, USA.
External Lecturer *part time* **Fall 2012**
Lille 1 University, France
External Lecturer *part time* **2010 – 2011**
Lille 3 University, France
Research Engineer *full time* **June 2008 – Sep 2008**
Chinese Academy of Science, Beijing, China.

Awards & Grants

Postdoctoral Research Fellowship in Statistics **Nov 2015**
Two-year fellowship funded by the Queensland University of Technology
Second place award for the best French PhD in Artificial Intelligence **June 2015**
Award from AFIA, the French Association for Artificial Intelligence.
Best applied paper award **Jan 2010**
Award from the EGC conference, French speaking conference on knowledge mining and management.
PhD Grant **Oct 2009**
Three-year grant funded by the French Ministry of Research

Research Expeditions

3 months at Berkeley Statistic Departement, USA **Mar – June 2015**
Hosted by Peter Bartlett
One week at Inria Nancy-Grand Est, France **June 2012**
Hosted by Bruno Scherrer of the team Maia

Peer Reviewer

I have been an official reviewer for the Neural Information Processing Systems (NIPS) international conference in 2014 and 2015 and I have reviewed papers for the Machine Learning Jounal and the Journal of Machine Leaning Research (JMLR).

Invited Presentations

Talks other than Conference presentations
Talk Oxford Robotics Research Group Seminar, Oxford, UK, May 2014
"Classification-Based Policy Iteration perform well in the game of Tetris".
Talk Gatsby Reinforcement Learning Research Group, London, UK, May 2014
"Classification-Based Policy Iteration perform well in the game of Tetris".
Talk Team Maia Seminar, Nancy, France, June 2012
"Pure Exploration Bandits".
Talk Co-Adapt Seminars, Marseille, France, May 2012
"Pure Exploration Bandits for Brain-Computer Interface?".

Publications

Peer-reviewed journal article

J1. Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon & Matthieu Geist, ***Approximate Modified Policy Iteration***, to appear in Journal of Machine Learning Research (JMLR).

Peer-reviewed conference articles

C9. Victor Gabillon, Branislav Kveton, Zheng Wen, Brian Eriksson & S. Muthukrishnan, ***Large Scale Optimistic Adaptive Submodularity***. AAAI 2014, 28th Conference of the Association for the Advancement of Artificial Intelligence. Oral presentation at Quebec City, Canada, July 2014.

C8. Victor Gabillon, Mohammad Ghavamzadeh & Bruno Scherrer, ***Approximate Dynamic Programming Finally Performs Well in the Game of Tetris***. NIPS 2013, 27th Conference on Neural Information Processing Systems. Poster presentation at South Lake Tahoe, Nevada, December 2013.

C7. Victor Gabillon, Branislav Kveton, Zheng Wen, Brian Eriksson & S. Muthukrishnan, ***Adaptive Submodular Maximization in Bandit Setting***. NIPS 2013, 27th Conference on Neural Information Processing Systems. Poster presentation at South Lake Tahoe, Nevada, December 2013.

C6. Victor Gabillon, Mohammad Ghavamzadeh & Alessandro Lazaric, ***Best Arm Identification: A unified approch to fixed budget and fixed confidence***. NIPS 2012, 26th Conference on Neural Information Processing Systems. Poster presentation at South Lake Tahoe, Nevada, December 2012.

C5. Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon & Matthieu Geist, ***Approximate Modified Policy Iteration***. ICML 2012, 29th International Conference on Machine Learning. Long lecture presentation at Edinburgh, Scotland, June 2012.

C4. Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric & Sébastien Bubeck, ***Multi-Bandit Best Arm Identification***. NIPS 2011, 25th Conference on Neural Information Processing Systems. Poster presentation at Granada, Spain, December 2011.

C3. Victor Gabillon, Alessandro Lazaric, Mohammad Ghavamzadeh & Bruno Scherrer, ***Classification-based Policy Iteration with a Critic***. ICML 2011, 28th International Conference on Machine Learning. Lecture presentation at Bellevue, USA, June 2011.

C2. Victor Gabillon, Jérémie Mary & Philippe Preux, ***Affichage de publicités sur des portails web***. EGC 2010, 10th French-speaking International Conference on Knowledge Extraction and Management. Lecture presentation of long article at Hammamet, Tunisia, January 2010. Best applied paper award.

C1. Jean-Pierre Delmas & Victor Gabillon, ***Asymptotic performance analysis of PCA algorithms based on the weighted subspace criterion***. ICASSP 2009, International Conference on Acoustics, Speech and Signal Processing. Poster presentation at Taipei, Taiwan, April 2009.

Peer-reviewed workshop article

W1. Victor Gabillon, Alessandro Lazaric & Mohammad Ghavamzadeh, ***Rollout Allocation Strategies for Classification-based Policy Iteration***. Workshop on Reinforcement Learning and Search in Very Large Spaces International Conference on Machine Learning, Lecture presentation at Haifa, Israel, June 2010.

Major research achievements & industrial innovations

- *Research:* ***Reinforcement Learning is Finally Competitive:*** We proposed a new family of reinforcement learning methods based on “Classification-based Policy Iteration” algorithms. In addition to providing thorough theoretical analysis of these methods (C3,C5), we implemented the proposed algorithms and ran extensive experimental studies to analyse their performance using the game of Tetris as a benchmark (the experiments involved computing on a grid of computers). Our results show an unprecedented performance of reinforcement learning methods in Tetris, improving upon the state-of-the-art techniques. Moreover, while these state-of-the-art techniques were based on black-box optimisation methods that require a large number of samples from the environment, our methods are able to learn Tetris strategies and achieve the same performance with 10 times less number of samples (C8).

- *Industry:* ***Constrained Learning for Orange Ad Server:*** Orange, the leading company to provide telecommunication solutions in France, had made a contract with the research team SequeL in order to automate their online web-advertising services. My initial goal was to make a survey of the machine learning literature and find an appropriate solution to optimise their click-through rate revenues. This solution had to take into account specific new constraints on the limited and known number of display per ads. I proposed a new approach combining linear programming and bandit algorithms and validated its performance on synthetic data. The results received the best paper award at the French conference on Data Extraction and Knowledge Management (C2). Moreover, the project initiated an ongoing collaboration between SequeL and Orange.

- *Industry:* ***Adaptive Questionnaire Design at Technicolor Inc:*** During the course of my PhD, I participated in a 6-months R&D internship program at Technicolor Inc.'s research lab. The primary goal of my project was to improve upon an online questionnaire which aimed to elicit users' movie preferences and generate a recommender system. I cast the program as an adaptive submodular maximisation problem. The novelty of my approach was to consider a completely realistic formulation of the problem where the users' preferences were unknown and to be actively learned in order to build the adaptive questionnaire. My efforts in this short period of time resulted in the publication of two peer-reviewed papers at prestigious international conferences in machine learning (C7, C9).

Teaching

In the past 5 years I taught 216 hours of undergraduate and master's courses in France.
Instructor:

- *Introduction to algorithmic and programming with Python.*
  48 hours (lectures and practical sessions). Winter 2010, Fall 2011 & Fall 2012
  1^{rst} year of Master *Computer science and document* at Lille 3 University and 1^{rst} year of Licence *Physics-Chemistry* at Lille 1 University.

Teaching assistant:

- *SQL and Python.* 36 hours (practical sessions). Fall 2010.
  3rd year of Licence *Mathematics and computer science applied to social sciences* at Lille 3 University.

- *Designing databases and object-oriented programming.* 36 hours (practical sessions). Winter 2011.
  3rd year of Licence *Mathematics and computer science applied to social sciences* at Lille 3 University.

5 Capacities of the Participating Organisations

Beneficiary: Lancaster University

General Description	Lancaster University is a top ten UK university. The Department of Mathematics and Statistics, within the Faculty of Science and Technology, hosts one of the largest and strongest statistics research groups in the UK comprising 50 academic staff, 11 post-doctoral fellows and around 70 FTE research students. In the 2014 Research Excellence Framework assessment, the Mathematical Sciences at Lancaster were ranked fifth overall and third in terms of the impact of research. Research is supported by grants from the UK Research Councils, the European Commission, and industrial sponsors. The statistics research group is also a fundamental partner in Lancaster's new Data Science Institute, which aims to act as a catalyst for Data Science, providing an end-to-end interdisciplinary research capability.
Role and Commitment of key persons (supervisor)	Prof. David Leslie, PhD in Mathematics (University of Bristol, 2003). 17 PhD students and 2 post-doctoral fellows supervised. 5% FTE time commitment to the project throughout the 24 month duration.
Key Research Facilities, Infrastructure and Equipment	The Department of Mathematics and Statistics is housed in dedicated space at Lancaster University. The Researcher will be provided with office space and basic equipment within the Department. Researchers in have access to the Department's own computer support (2.6FTE computer technicians) and computer cluster (nearly 500 computer cores, 800GB of memory). These computing facilities are supplemented by access to Lancaster University's High-End Computing cluster (1700 computer cores, 8TB of memory, 32TB of high performance filestore).
Independent research premises?	Yes
Previous Involvement in Research and Training Programmes	Between 2001 and 2005 the department held the Marie Curie Training Site status for its PhD programme. The Postgraduate Statistics Center (PSC) was founded in 2005 as the only Centre for Excellence in Teaching and Learning focussing on postgraduate statistics in the UK. The PSC is still operative and runs three Masters degrees (Statistics, Quantitative Methods, and Quantitative Finance) and coordinates the PhD programme in statistics.
Current involvement in Research and Training Programmes	Together with the Management School, the Department hosts and runs STOR-i, a multi-million pound EPSRC-funded Centre for Doctoral Training in Statistics and Operational Research in partnership with industry. The Centre was established in 2010 and funds 12 PhD students per year. The department is also a key player in the Academy for Phd Training in Statistics, a collaboration between major UK statistics research groups to organise courses for first-year PhD students in statistics and applied probability nationally. The group hosts one node of a multi-institution Programme Grant on Intractable Likelihood, and received industrial funding from companies including Shell, BT, Google and Unilever. The Department's Medical and Pharmaceutical Statistics Research Unit works closely with the pharmaceutical industry and public sector research institutes to develop novel statistical methods for the design and analysis of clinical trials. It leads the EU-funded research training network IDEAS (www.ideas-itn.eu) and is an integral part of the Medical Research Council funded North-West Hub for Trials Methodology Research.
Relevant Publications and/or research/innovation products	Perkins, S. and Leslie, D.S. (2014) Stochastic fictitious play with continuous action sets. *Journal of Economic Theory* **152**, 179–213. Chapman, A.C., Leslie, D.S., Rogers, A. and Jennings, N.R. (2013) Convergent learning algorithms for unknown reward games. *SIAM Journal on Control and Optimization* **51**, 3154-3180. May, B.C., Korda, N., Lee, A. and Leslie, D.S. (2012) Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research* **13**, 2069–2106. Larsen, T., Leslie, D.S., Collins, E.J. and Bogacz, R. (2010) Posterior weighted reinforcement learning with state uncertainty. *Neural Computation* **22**, 1149–1179. Leslie, D.S. and Collins, E.J. (2003) Convergent multiple-timescales reinforcement learning algorithms in normal form games. *Annals of Applied Probability* **13**, 1231–1251.

ENDPAGE

MARIE SKLODOWSKA-CURIE ACTIONS

Individual Fellowships (IF)
Call: H2020-MSCA-IF-2014

PART B

“OSEGA”

This proposal is to be evaluated as:

[Standard EF]