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 (including inter/multidisciplinary aspects)

A critical concern in the modern world is security. Effectively protecting the ports, airports, trains and other transportation systems from malicious attacks, fighting the trafficking of drugs, firearms and even people, and securing proprietary and sensitive information over the ever-growing cyber-networks, comprise some of the principal axes of this critical task. The main challenge in all of these problems is that maximum security must be obtained with a limited number of available resources. For instance, the total number of security agents available to simultaneously protect a multitude of designated targets may not be sufficient to provide full security coverage at an airport. This calls for the design of appropriate resource allocation techniques which, given the available resources, would result in maximum security.

Security resource allocation and scheduling problems comprise one of the many application areas that have recently been shown to greatly benefit from game-theoretic approaches. Indeed, as a solid mathematical framework to model strategic decision making, game theory has proved useful in many real-world applications from economics and political science to logic, computer science and psychology. In this paradigm, the problem is cast as a "game" and the objective is to find a solution whereby each "player" makes choices to maximise her own *utilities*, which may often be in conflict with those of her opponent. 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 player an outcome (value) is assigned, which depends on the pair of actions taken by both players. In cases where these outcomes are known, game-theoretic approaches have provided impressive results. Since 2007, the so-called ARMOR software[1] is used at the Los Angeles International Airport (LAX) to effectively determine checkpoints on the roadways leading to the airport, and to canine patrol routes within terminals. Similarly, such programs as IRIS,[2] PROTECT,[3] and TRUSTS[4] are respectively being deployed at the US Federal Air Marshals, the US coast guard patrolling, and the Los Angeles Metro system's fare inspection strategy.

A severe limitation of these models is that the true utility functions are usually unknown and must be estimated by experts or obtained from historical data. As a result, the potentially high estimation errors or the lack of historical data from which is suffering any newly established security system, will often render the security game solver useless. Therefore it is of importance to use methods that can quickly collect relevant data in order to estimate the parameters of the game and quickly be operational.

Machine learning lies at the crossroad between statistics and computer science. The common goal is to design programs able to actively and intelligently gather data and extract information from the collected data, autonomously using them to make strategic decisions. Based on theoretically sound statistical methods, machine learning techniques are ubiquitously being deployed in a variety of modern applications ranging from robotics to personalised product recommendation.

Learning is a very well fitted approach to tackle security games as they are often of repeated form, played daily between a defender and possible attackers. Repeated security games allow for the continuous collection of data, which, through data-driven approaches can in turn be used to estimate the parameters of the game. Another possibility is in fact that the defender can in turn test the game and collect (in the most efficient and less costly manner) more information about the game. The key objective forming the basis of this grant proposal, is thus to design efficient and theoretically sound, data-driven methods that can actively interact with the environment to *learn* a fair model through repeated games. Using machine learning, the purpose of this proposal is to create *practical*, *scalable* and *robust* methods for security games.

[1] James Pita et al. "Deployed ARMOR protection: the application of a game theoretic model for security at the Los Angeles International Airport". In: *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems: industrial track*. International Foundation for Autonomous Agents and Multiagent Systems. 2008, pp. 125–132.

[2] Jason Tsai et al. "IRIS-a tool for strategic security allocation in transportation networks". In: (2009).

[3] Eric Shieh et al. "Protect: A deployed game theoretic system to protect the ports of the united states". In: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems. 2012, pp. 13–20.

[4] Zhengyu Yin et al. "TRUSTS: Scheduling randomized patrols for fare inspection in transit systems using game theory". In: *AI Magazine* 33.4 (2012), p. 59.

State-of-the-art

i. 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 an attacker (e.g., terrorists/drug 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 corresponding pay-offs. From the defender's perspective, this corresponds to efficiently allocating a limited number of resources to secure some predefined targets. These utilities for both players can be stored in two matrices, $\boldsymbol{A}$ for the defender, $\boldsymbol{B}$ for the attacker, where $\boldsymbol{A}_{i,j}$ and $\boldsymbol{B}_{i,j}$ are the utilities for the defender, respectively the attacker, when the defender plays her strategy i and the attacker responds with her j^{th} strategy. Solutions to such games rely on randomised strategies, making the defender's scheme highly unpredictable for the attacker, thus giving rise to a significant advantage over the original mechanisms that are based on deterministic human schedulers. Another important feature of this paradigm is that it allows to obtain theoretical guarantees. Specifically, in the case of games that are fully competitive between the two players (i.e. the so-called zero-sum games), the solution is provably robust in that it provides guaranteed performance against *any* possible attacker. Moreover, such guarantees hold, even if the defender's strategy is completely revealed to the attacker. An extension of this guarantee to a more general (non zero-sum) game can be provided by Stackelberg equilibrium, a notion that generalises the famous Nash equilibrium.[5] Indeed, this is well-suited to the scenario where the defender's strategies may be at risk of revelation, leak or discovery through repetitive interactions.

ii. Uncertainty in Security Games

Uncertainty is endemic in most real-world applications. Unlike standard game-theoretical approaches, in some scenarios, the players is uncertain about their utilities (the $A_{i,j}$'s), or those of the other players'. For instance, in the context of security games, the random selection of passengers for security checks at an airport is a source of uncertainty in this game, where the outcome is random and the probability of successful security enforcement is unknown. As also confirmed by several empirical studies in fraud and cybercrime detection, this phenomenon can significantly decrease the defender's performance.[6] Extensive studies have been dedicated to the design of security games that are robust with respect to uncertainty about the environment.[7] However, an important observation is that much more can be done in the case of *repeated* security games. Indeed, this repetition allows the defender to further reduce her uncertainty about the model and intelligently *learn* how to improve her performance over time. Specifically, in this case, a security game solver can autonomously take intelligent decisions at repeated instances of the game, by carefully collecting, extracting and acting upon information from historical data. As discussed further in the subsequent section, this is precisely the scope of machine learning tools.

iii. Indispensable Tools from Machine Learning

Machine learning consists of data-driven techniques with strong ties to the fields of statistics (allowing them to account for uncertainty about the data) and optimisation (so as to quickly converge to the desired solution, minimising for instance, the number of actions required to achieve a specific level of performance).

[5] Dmytro Korzhyk et al. "Stackelberg vs. Nash in Security Games: An Extended Investigation of Interchangeability, Equivalence, and Uniqueness." In: *J. Artif. Intell. Res.(JAIR)* 41 (2011), pp. 297–327.

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

[7] Michele Aghassi and Dimitris Bertsimas. "Robust game theory". In: *Mathematical Programming* 107.1-2 (2006), pp. 231–273; Thanh H Nguyen et al. "Regret-based optimization and preference elicitation for stackelberg security games with uncertainty". In: *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI).* 2014, pp. 756–762; Christopher Kiekintveld, Towhidul Islam, and Vladik Kreinovich. "Security Games with Interval Uncertainty". In: *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems.* AAMAS '13. St. Paul, MN, USA: International Foundation for Autonomous Agents and Multiagent Systems, 2013, pp. 231–238. ISBN: 978-1-4503-1993-5.

One of the fundamental problems in machine learning, relevant to our research objectives in this proposal, is the *multi-armed bandit problem.* It corresponds to a scenario where the learner is required to actively collect data from an environment in order to solve a given task. The solutions built for this problem have found many practical applications from adaptive routing in a network to medical trials of new medicines.[8] The bandit problem is a repeated game where, at each time step the same (fixed) set of actions is available to the learner. In its simplest form there are K actions (arms). At each round t, the environment allocates rewards to each arm (described as a vector $l^t \in \mathbb{R}^k$) while the learner simultaneously chooses an arm $I(t)$ to pull in order to obtain a reward $l^t_{I(t)}$. An important constraint in this setting is that the player is not allowed to observe the hypothetical reward that would have been collected had another arm been selected instead. This therefore corresponds to a simplified security games where the strategies of the other players are fixed and only an external environment allocates the rewards. One can think of a game where the K arms/actions correspond to the K possible security strategies whose values are determined according to some unknown underlying probability distribution. The average per action reward can be estimated through sampling, i.e. via pulling the corresponding arm.

While the multi-arm bandit games are extremely indispensable to active learning, an important question is how they can be used in general security games where one has to also take into account the actions of the other players.

iv. Existing results

Some interesting advancements in security games have recently been made through links with optimisation and machine learning methods. These methods mostly focus on the case where the attacker's preferences are not fully known and are thus to be learned; the learning objective is achieved through a repeated a game. Some recent work analyse the number of required queries to learn the optimal defender's strategy.[9] Marecki et. al. and Qian et. al.[10] take an empirical Bayesian approach where, given a prior distribution, planning techniques based on Partially Observable Markov Decision Processes (POMDPs) are used to update the posterior over the adversary's preferences. The main theoretical drawback of this planning method is in that the algorithm is based on Upper Confidence Trees (UCT), which, are provably sub-optimal.[11] Recently, an extended analysis has been given[12] 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. The latter work shows strong connections with adversarial bandit theory.

Main Goal

The purpose of this proposal is to create *practical*, *scalable* and *robust* methods for security games. First, we target *practicality* in the sense that our algorithms would be autonomous in handling the uncertainty in the model and would actively be working at reducing it by interacting with the environment in which the game takes place. Specifically, we aim to broaden the scope of repeated security game problems where the initial uncertainty about the players utilities can be overcome through using learning techniques in conjunction with repetitive plays of the game. These techniques are from the extremely active field of machine learning research. Second, we target *scalability* so that the solvers could handle a potentially extremely large number of possible actions. To that end, simplifying structure assumptions such as combinatorial structure will be

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

[9] **blum2014learning**; **letchford2009learning**.

[10] Janusz Marecki, Gerry Tesauro, and Richard Segal. "Playing Repeated Stackelberg Games with Unknown Opponents". In: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*. AAMAS '12. Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 821–828. ISBN: 0-9817381-2-5, 978-0-9817381-2-3; Yundi Qian et al. "Online planning for optimal protector strategies in resource conservation games". In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems.* International Foundation for Autonomous Agents and Multiagent Systems. 2014, pp. 733–740.

[11] **munos2014bandits**.

[12] Maria-Florina Balcan et al. "Commitment without regrets: Online learning in Stackelberg security games". In: (2015).

considered as well as the simplifying submodular property of the objective function. Finally the *robustness* is a key issue in security games. We propose to insure robustness in three different ways. First, one can be very conservative and assume that an adversary actually chooses (in the most adversarial way) the data that we do not know. Second, we introduce offline learning where the test and the learning happen before the strategies is actually use in the real world. Third, we might be required to learn defence strategies that do not possess large variances in their performance.

Our goal is to have a theoretically sound approach by designing efficient algorithms for which we can provide finite sample analysis.

.

Objective 1: Scalability in Pure exploration Bandits via Submodularity. As discussed previously bandit problems come very handy to model repeated games under uncertainty. Therefore addressing the question of scalability in repeated security games must be first addressed in the context of bandits. In security games, it is natural to make use of the combinatorial nature of the actions, placing a limited number of checkpoints in a finite number of locations. We believe that combinational bandit techniques, which also form part of the fellow's area of expertise, will be key in solving this problem. However, naively enumerating all possible actions as in the standard formulation of security games makes the computations rapidly intractable. A key observation is that in many cases the players' performance evaluation function is submodular. This particularly arises in the context of maximal coverage problem for sensor (or checkpoint) placement.[13] This submodularity property can in turn be used to provide tractable and almost optimal algorithms. The fellow provided work on addressing submodular maximization techniques in a bandit setting[14] under a regret analysis. As discussed further in the objective 2, we are also interested the bandit formulation called pure exploration that will provide a safer scenario to tests the security strategies. As this bandit setting is part of the fellow expertise a very natural first step will be to consider combinatorial pure exploration under submodularity.

Objective 2: Pure Exploration in Security Games. As discussed previously, the main challenge in applying security games to real-world scenarios is that the players' utility matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are unknown and must be learned. The learning strategy depends on the particular setting of the problem. In the first objective we consider the case where the defender is in fact able to safely examine her defensive strategies before applying them online. A familiar real-world example is the security system at an airport, which can be tested many times before being deployed as the principal defence scheme.

We observe that in this formulation of the problem, we can capture the learning module of the security game by a recent bandit setting called *pure exploration*, where the learner is only evaluated at the end of an exploration phase comprised of a limited number of interactions with the environment. Pure exploration has generally proved useful in many practical scenarios. Specifically, its application to parallel action selection in robotic planning has been extensively studied by the fellow.[15] However, applying this approach to security games is an open research path with a strong potential to bring new and interesting angles to the domain. In this formulation, we assume both utility matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ completely unknown. During the test phase, at every mock repetition of the game, the defender is able to probe an entry $\boldsymbol{A}_{i,j}$ of its utility matrix $\boldsymbol{A}$, corresponding to the outcome obtained when the attacker plays her strategy j and the defender responds with her i^{th} strategy. The value obtained is *a noisy version* of the true entry $\boldsymbol{A}_{i,j}$. What makes this setting particularly well-suited to real-world security applications is in that, 1. unlike existing work, no assumption is required to be made about the defender's knowledge of its utility matrix $\boldsymbol{A}$. Instead, the utilities can be learned offline through probing the individual entries of $\boldsymbol{A}$, and 2. the stochastic framework gives a natural model for the players' uncertainty about the environment.

We aim to 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

[13] **krause2011randomized**.

[14] **gabillon2013adaptive**.

[15] V. Gabillon et al. "Multi-Bandit Best Arm Identification". In: *Proceedings of the Advances in Neural Information Processing Systems 25.* 2011, pp. 2222–2230.

strategy given a fixed number of tests. In the classical multi-armed bandit problem, these optimisation objectives respectively correspond to a *fixed confidence* setting[16] and a *fixed budget* constraint.[17] In order to extend these classical results to the setting proposed above, we will first carefully characterise the data-dependent complexity (hardnesss) of the problem and then move on to designing algorithms to best capture the obtained hardness. In standard bandit, the complexity of an arm is inversely proportional to the gap $\Delta_i = \mu^* - \mu_i$ between the value of the best option μ^* and that of option i. Extensions of this notion have been designed for combinatorial bandits.[18] The fellow is currently working on a improved version of the state-of-the-art result. The complexity of an arm defined in these combinatorial games is more complex than in the simple multi-armed bandit problem as it involves combinatorial quantities. Similarly in security games, we expect the complexities of each entry of the matrix for the defender to depend also on the actions available to the attacker. Therefore we will study this problem by gradually increasing its difficulty as we will explore different partial feedback structures. First we note that a recent work[19] on query complexity, corresponds to the simpler deterministic version of this problem where it is assumed that the probing outcome corresponds to the *true value* of $\boldsymbol{A}_{i,j}$. Thus they do not analyse the number of queries required to correctly estimate each entry of the utility matrix, which in turn depends on the performance of the strategies that would use this entry. Therefore our first approach will be to combine ideas from pure exploration and the deterministic 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 more adversarial learning problem where the defender chooses a strategy and only observes the value of the game when the attacker best-responds 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 of the utilities must be performed online while actually playing the security game. In this context it is of high importance for the agent to learn the utilities as fast as possible. This means that only providing an analysis to demonstrate asymptotic convergence is definitely not enough. To address repetitive learning in security games we will here assume that the utility matrix $\boldsymbol{A}$ is unknown to defender and that, at each repetition of the game the defender will best respond to his current strategy. Therefore we are considering a setting that is related to the analysis of Stackelberg equilibrium. A natural quantity of interest is the *cumulative regret*, which is standard in multi-armed bandit analysis. In our context, the objective is or the learner-defender build a series of defence strategies π_t for $t = 1, \ldots, n$ to minimize the expected cumulative regret after n pulls $R(n)$ defined as

$$R(n) = nv(G) - \sum_{t=1}^{n} \pi_t A b(\pi_t).$$

where the v(G) is the value of the game when the two players play according to the Stackelberg equilibrium of the game. The cumulative regret would be therefore defined the difference between the sum of rewards collected by always using the *best security strategy* in hindsight and the sum of rewards actually collected by the forecaster. This is an online game where, at each time set t of the game, a fundamental trade-off arises for the forecaster between the simultaneous need to select the security strategy solution he currently thinks is the best in order to maximise the immediate security given his current knowledge (exploitation) while also wanting to test possible other strategies that might or might not be better (exploration).

[16] O. Maron and A. Moore. “Hoeffding races: Accelerating model selection search for classification and function approximation”. In: *Proceedings of the Advances in Neural Information Processing Systems 7.* 1993; E. Even-Dar, S. Mannor, and Y. Mansour. “Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems”. In: *Journal of Machine Learning Research* 7 (2006), pp. 1079–1105.

[17] S. Bubeck, R. Munos, and G. Stoltz. “Pure Exploration in Multi-Armed Bandit Problems”. In: *Proceedings of the Twentieth International Conference on Algorithmic Learning Theory.* 2009, pp. 23–37; J.-Y. Audibert, S. Bubeck, and R. Munos. “Best Arm Identification in Multi-Armed Bandits”. In: *Proceedings of the Twenty-Third Conference on Learning Theory.* 2010, pp. 41–53.

[18] Shouyuan Chen et al. “Combinatorial pure exploration of multi-armed bandits”. In: *Advances in Neural Information Processing Systems.* 2014, pp. 379–387.

[19] **goldberg2014query**.

Our goal here will be to provide a regret analysis for the online stochastic repeated security games. The first step will be to study how to design a new algorithms borrowing ideas from the UCRL algorithm[20] that was extending the bandits ideas of the classical UCB analysis to general reinforcement learning problems. As most of the approach in online learning, UCB implements a strategy that is optimistic in face of uncertainty, playing any strategy that could be the best given the level of uncertainty. It would be therefore very interesting to see if this optimism principle is still optimal in a case of an adversarial game especially in a case where the adversary knows the true utilities of the games and moreover knows your levels of uncertainty. For this project collaboration with Bruno Scherrer, a fellow's co-author, working at Inria Nancy, would prove fruitful as he has recently analysis how reinforcement learning asymptotic analysis could be applied to zero sum games.[21] Approaches bases on Thompson sampling[22] will be also considered as they have proved very efficient in practice and correspond to the area of expertise of Prof. Leslie.

Finally an important other possible additional requirement is that we 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 requirement has been well-studied in the statistical community has a *risk-averse* requirement. Recently this requirement has been considered in a multi-armed bandit framework.[23] An implementation in the security games specific context is therefore natural.

Objective 4: Learning Stackelberg Equilibrium against Combinatorial Adversaries. Objective 4 will be devoted to solving security games with more complex action structures. Real-world security problems often involve large, complex networks. This includes, for instance, complex routes, or computer/communication networks. Indeed, taking advantage of the inherent structure of the problem is of great importance which can significantly help create efficient and computationally tractable algorithms. We propose to extend the recent work of Balcan[24] that proposes to learn Stackelberg equilibrium strategies against k unknown adversaries, to the case where the set of adversaries is of combinatorial nature. To this end a link with the classical work of combinatorial adversarial bandit is relevant.[25] Note that the fundamental problem of combinatorial bandits has never been considered in the the context of a Stakelberg games. 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 stucture is present in numerous real-word applications. In light of the ever-growing, modern, social and communication networks, an extremely important application area is that of (mobile) smuggler arrest in a network.[26] 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 been studied in its repeated form, where a pursuit-evasion game is played multiple times against a series of different smugglers. Therefore, the strategy of the defender is *not adaptive* to the previous observations collected about the attackers' historical strategy and is therefore suboptimal. Since this simple problem possesses a graph structure, it can be effectively used as a preliminary step to test security games on a graph for which learning methods are necessary. Moreover, this setting can be viewed as an instance of the work on general adversarial combinatorial bandits.[27] Note however

[20] Peter Auer, Thomas Jaksch, and Ronald Ortner. "Near-optimal regret bounds for reinforcement learning". In: *Advances in neural information processing systems.* 2009, pp. 89–96.

[21] Bruno Scherrer. "Approximate Dynamic Programming for Two-Player Zero-Sum Markov Games". In: (2015).

[22] **russo2014information**.

[23] Amir Sani, Alessandro Lazaric, and Rémi Munos. "Risk-Aversion in Multi-armed Bandits". In: *Advances in Neural Information Processing Systems 25.* Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 3275–3283.

[24] Balcan et al., "Commitment without regrets: Online learning in Stackelberg security games".

[25] Nicolo Cesa-Bianchi and Gábor Lugosi. "Combinatorial bandits". In: *Journal of Computer and System Sciences* 78.5 (2012), pp. 1404–1422.

[26] Manish Jain et al. "A double oracle algorithm for zero-sum security games on graphs". In: *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1.* International Foundation for Autonomous Agents and Multiagent Systems. 2011, pp. 327–334.

[27] Cesa-Bianchi and Lugosi, "Combinatorial bandits".

that the latter setting does not capture all the structure specific to the graph pursuit-evasion problem. More specifically, it does not restrict the attacker to play according to a *path* on the graph. Therefore the objective is to design specific algorithms in cases where the problem it is possible to take advantage of specific graphical structure of the problem. We 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. The fellow is currently working to establish a collaboration with Dr Michal Valko, a world-famous expert in active learning on graphs and part of Inria Lille in France. This collaboration will be greatly beneficial not only in achieving the second objective of this proposal, but also to strengthen international links between Lancaster University in the UK and Inria research organisation in France.

Originality and innovative aspects of the research programme:
This grant proposal is original has it proposes to develop a bridge between two areas of research (Security games and Machine learning) that have both proved their practical capacities but have only rarely been used jointly. By addressing some of the fundamental challenges of this connection we expect to bring more attention to the potential of this connection and bring the two communities to collaborate more in order to design software that solve important security problems. Our approach is to design algorithms are both theoretically grounded and of practical use in real world problems. Moreover by collaborating with with the Security Lancaster Departement in the University of Lancaster, we hope to apply and test our methods to real world problems.

Timeliness and relevance:
Security games has found numerous applications in the recent 10 years and has proved that it could bring a significant improvement over systems designed by humans. Making this systems even more autonomous so that they can autonomously improve by constantly learning is a key issue to bring low maintenance security systems. We believe machine learning is providing most of the necessary tool to reach this goal. Machine Learning is one of the most rapidly growing community in computer science research as it is currently making a difference in key industrial sectors already building strong recommendation systems, and being used by the most important companies of our time to create self-driving cars or new pattern recognition algorithms. The connection between security games and machine learning is evident, work that have connects those two worlds are still rare and it will be the source of many new research challenges. Finally bringing more robust security systems is of high importance for Europe has recent event have shown how ensuring the security of network communications from spy or securing public transportation is crucial.

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.

The proposed project is primarily a research project, and the main training objectives are to enhance the fellow's scientific skills. Dr Gabillon is already an expert in the modern theory of bandits, including best arm identification, and reinforcement learning. Therefore this project's main training objective for Dr Gabillon will be to develop his skills and knowledge in statistical learning methods and game theory. The supervisor is an expert in both areas, and will of course assist the development of the Researcher. Further expertise in Lancaster from whom the Researcher will learn includes the Statistical Learning group, in which the Researcher will be based, and the broader Statistics Research Group.

Considering also the research group in operations research within the Management School, Lancaster 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 expertise in this area, and indeed brings expertise from a complementary aspect of online learning and decision-making in the design and analysis of algorithmic approaches to learning, especially with combinatorial bandit problems. Dr Gabillon's expertise in best-arm identification will be of great interest 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. Furthermore his expertise in combinatorial bandits complements current industrially-funded research of the Supervisor.

In addition to his research skills, Dr Gabillon will learn from the host's world-leading expertise in developing 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 preparing funding applications by co-authoring proposals for UK and EU funding agencies with Prof. Leslie and others.

5. Attend staff training workshops designed specifically for early-career researchers, and specifically the Research Development Programme, a structured development route for researchers, designed to promote impactful research and to support development beyond a disciplinary area.

6. Participate in teaching and research supervision at undergraduate and graduate level. This will not be obligatory, but the fellow will be given the opportunity to benefit from peer observation, mentoring, and constructive criticism.

The UK Concordat to Support the Career Development of Researchers is an agreement between funders and employers of research staff to improve the employment and support for researchers and research careers in UK higher education. Lancaster University is fully committed to the Concordat to Support the Career Development of Researchers and has put in place an Action Plan to support the full implementation of the Concordat at Lancaster. Furthermore, throughout the fellowship, 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. In addition, Dr Gabillon will have regular meetings with the host supervisor to discuss his research and to receive advice.

1.3 Quality of the supervision and the hosting arrangements

Qualifications and experience of the supervisor(s)

Information regarding the supervisor(s) must include the level of experience on the research topic proposed and document its track record of work, including the main international collaborations. Information provided should include participation in projects, publications, patents and any other relevant results. To avoid duplication, the role and profile of the supervisor(s) should only be listed in the "Capacity of the Participating Organisations" tables (see section 6 below).

Hosting arrangements[28]

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

[28] The hosting arrangements refer to the integration of the Researcher to his new environment in the premises of the Host. It does not refer to the infrastructure of the Host as described in Criterion Implementation.

game theory, with 19 refereed articles in top journals of several different research fields, and collaborators from France, Singapore, USA and Australia. His research on contextual bandit algorithms[29] 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,[30] stochastic approximation,[31] and the mathematics of statistically-inspired reinforcement learning.[32] Prof. Leslie is the holder of a Google Faculty Award which funds a student to investigate multiple-action selection in bandits. Prior to his relocation to Lancaster, he was a senior lecturer in the statistics group of the School of Mathematics, University of Bristol. He continues to be 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.

Prof. Leslie's mentoring approach is one of 'guided freedom' in which the mentee takes responsibility for their own research, while regular discussions ensure that dead ends are avoided and promising openings are exploited. In the 10 years since taking up a Faculty position, he has supervised 17 PhD students, 2 post-doctoral fellows, numerous MSc and undergraduate dissertations, and an undergraduate secondment from ENS Lyon.

Hosting arrangements

Dr Gabillon will be embedded within the statistical learning group which is lead by Prof. Leslie. This is a team of 5 academic staff and around 5 PhD students within the Department of Mathematics and Statistics. The Researcher will participate in weekly group meetings and benefit from advice from the senior scientists in the group, including the Supervisor, on 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 of which have approximately weekly seminars. These exciting initiatives will provide multiple further opportunities to develop informal mentoring relationships in addition to the formal process which takes place for all staff at Lancaster University; to ensure integration within these networks the Researcher will be introduced to the groupings of researchers, invited to deliver a seminar on his research, and will participate in away days in which strong relationships are developed.

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

Professional maturity in academia would be ideally reached by leading a dynamic research group actively working on fundamental problems at the interface of game theory and online learning, with strong impact in real-world applications. Dr Gabillon has shown an extremely high potential to achieve this goal. As evident from his solid publication record, he has strong expertise in the domain, always giving equal importance to theory and applications. The fellow has also demonstrated strong ability to acquire new knowledge and become highly productive in a short period of time. Indeed, a significant result of his PhD thesis is in bringing classical reinforcement learning algorithms closer to daily life. Moreover, aside from providing theoretical guarantees for his proposed methods, this entailed spending a significant amount of time single-handedly managing extensive parallel-computing experiments over a grid of computers, a task for which he had no prior knowledge. During the course of his PhD, through a 6-months internship at a major US R&D lab (Technicolor Research Laboratory, Palo Alto), he had the opportunity to collaborate with a new team of R&D researchers. He quickly became productive and his efforts in this short period of time have resulted in the publication of two peer-reviewed papers at prestigious international conferences in machine learning. Through this experience he has also obtained valuable knowledge about industrial research and

[29] **MayEtAl2012**.

[30] **LeslieCollins03**; **LeslieCollins05**; **LeslieCollins06**; **ChapmanEtAl2013**; **PerkinsLeslie2014**.

[31] **LeslieCollins03**; **PerkinsLeslie2012**; **PerkinsLeslie2014**.

[32] **LeslieCollins05**; **LarsenEtAl2010**.

its interaction with academia. This has given him the ability to better understand the research pathways to produce high-impact results and establish significant collaborations with industry.

At the start of the fellowship, Dr Gabillon will be closely mentored by Professor Leslie at Lancaster University. He will also have access to the university's research resources, and will be able to further develop his research and supervision skills, which will greatly contribute to achieving professional maturity. At Lancaster University, the fellow will also have the unique opportunity to establish inter-disciplinary collaborations through the recently established STOR-i program, a quality research training interface between statistics and industry.

2 Impact

[**TODO:** Demonstrate: worthwhile outreach, good communication strategy (are there existing connections that can be exploited?), adequate discussion of impact on researcher's career, indication of how outreach activities will be assessed, strategies for exploitation of outcomes.]

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 a training oppotunity in two key additional research competences. Firstly, the Researcher will develop an in depth knowledge of cutting edge statistical theory, and bring that to bear within bandit algorithms. Training will be received from leading scientists in statistics and operations research at Lancaster University, and the many international visiting researchers who visit the department. Secondly, the Supervisor is a leading expert on learning in games, as well as bandit algorithms, and will mentor the Researcher to bring ideas from bandits into the game theoretical scenarios of this research proposal. 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 pure research opportunities, Dr Gabillon will work within Lancaster University's extremely effective framework for industrial collaboration. He will develop skills in how to manage the industry/academia relationship to ensure mutually beneficial outcomes. This relationship-management will be a key skill for academics in the future; Lancaster University, and particularly the Department of Mathematics and Statistics, is currently a world-leading institution in developing such relationships. The Researcher 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

Public engagement *Researchers should ensure that their research activities are made known to society at large in such a way that they can be understood by non-specialists, thereby improving the public's understanding of science. Direct engagement with the public will help researchers to better understand public interest in priorities for science and technology and also the public's concerns.*

Dissemination, exploitation of results *All researchers should ensure, in compliance with their contractual arrangements, that the results of their research are disseminated and exploited, e.g. communicated, transferred into other research settings or, if appropriate, commercialised. Senior researchers, in particular, are expected to take a lead in ensuring that research is fruitful and that results are either exploited commercially or made accessible to the public (or both) whenever the opportunity arises.*

With the launch of the Data Science Institute, Lancaster University will be inaugurating a "Data Science Network", in conjunction with Lancaster University's Knowledge Business Centre, 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. This network will bring together academic data scientists with local companies in regular show and tell sessions. The Researcher 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 articles directly from researchers to the public; the Researcher 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 "The Engaging Researcher Course", a one-day experiential training course to explore public engagement activities that researchers can get involved in.

The excellent and innovative research generated in this project will of course be published Open Access in the world's leading academic journals and conferences. 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. We will discuss results directly with companies in Lancaster University's Knowledge Business Centre, 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. 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; the Researcher 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. The Research Support Office of Lancaster University has extensive experience of industrial engagement and will assist in the management of IP and any patents that may arise from the research.

3 Implementation

[**TODO:** Show them: specific tasks and clearly-defined outputs/deliverables; host institution has capacity to support researcher; coherent workplan (including justification for the scheduling); metrics to assess progress; clear management structure (ie what is done beyond regular supervisor meetings); risk management and contingency plans; quality management procedures]

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

Describe the different work packages. The proposal should be designed in such a way to achieve the desired impact. A Gantt Chart should be included in the text listing the following:

- *Work Packages titles (for EF there should be at least 1 WP);*
- *List of major deliverables;*[33][34]
- *List of major milestones;*[35]
- *Secondments if applicable.*

The schedule should be in terms of number of months elapsed from the start of the project.

Work packages

WP1: Combinatorial bandits The Fellow will develop new approaches to combinatorial bandits, investigating the pure exploration problem and the online regret problem. This WP builds upon current research of the Fellow 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.

[33] A deliverable is a distinct output of the action, meaningful in terms of the action?s overall objectives and may be a report, a document, a technical diagram, a software, etc.

[34] Deliverable numbers ordered according to delivery dates. Please use the numbering convention <WP number>.<number of deliverable within that WP>. For example, deliverable 4.2 would be the second deliverable from work package 4.

[35] Milestones are control points in the action that help to chart progress. Milestones may correspond to the completion of a key deliverable, allowing the next phase of the work to begin. They may also be needed at intermediary points so that, if problems have arisen, corrective measures can be taken. A milestone may be a critical decision point in the action where, for example, the researcher must decide which of several technologies to adopt for further development.

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 the Researcher, 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 The Fellow 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. After completion of WP1, the Researcher will switch attention to working in combinatorial games in parallel with WP2, and this WP will continue 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. The Fellow 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 be started after WP2 has been completed in month 12, and run until the end of the project. **Deliverable 4.1** is a paper describing the 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 an assumption that the attacker always plays a best response to the current mixed strategy will be relatively straightforward, and greater emphasis can be placed on WP3 straight away. If the results here are not so strong, more effort will need to be given to WP2 in order to obtain suitable building blocks for WP3.

Milestone 2 is at the end of 1 year of the project, when WP1 and WP2 will both be completed, and WP3 is in progress. This will give an opportunity to take stock and decide the problems to be addressed in WP4. 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. However weaker game-theoretical results may necessitate initial focus on bandit approaches in WP4.

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

Develop your proposal according to the following lines:

- *Project organisation and management structure, including the financial management strategy, as well as the progress monitoring mechanisms put in place;*
- *Risks that might endanger reaching project objectives and the contingency plans to be put in place should risk occur.*

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 of the research. The Fellow will be a member of the Department of Mathematics and Statistics, and more specifically the Statistical Learning group lead by Prof. Leslie. The Fellow will also be assigned a formal mentor under standard Lancaster University human resources procedures, who will be a second point of contact for the Researcher. During the project, Dr Gabillon will be responsible for the research work, and will meet weekly with Prof. Leslie to discuss results, challenges and research strategies. The Fellow will also be responsible for the management of the project; he will be supervised in this task through monthly management and mentoring meetings with the Supervisor, in which progress against the workplan and career development plan will be discussed.

Clearly there are risks at each stage of an ambitious research project such as this. The two-pronged approach mitigates some of this risk: if developments in combinatorial approaches prove to be difficult then greater focus will be placed on the game theory, and vice versa. That being said, both of WP1 and WP2 contain elements that are clearly addressable based on current work. A solid foundation can thus be laid while the Fellow and Supervisor develop a working relationship, in preparation for the more ambitious objectives in the latter part of the project.

3.3 Appropriateness of the institutional environment (infrastructure)

- *Give a description of the main tasks and commitments of the beneficiary and partners (if applicable).*
- *Describe the infrastructure, logistics, facilities offered in as far they are necessary for the good implementation of the action.*

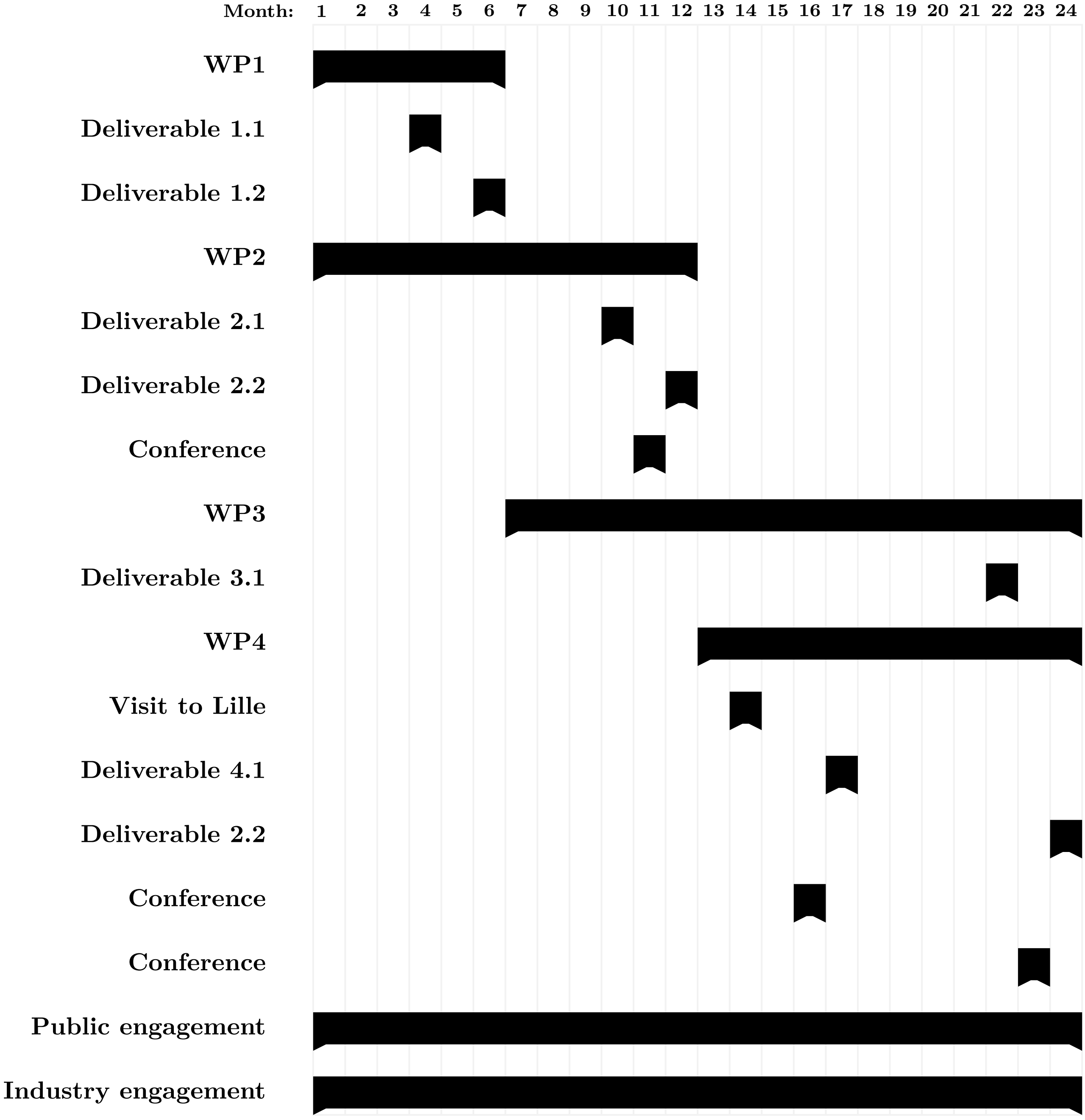
The Researcher will be hosted in the Department of Mathematics and Statistics, Lancaster University. Prof. Leslie will provide the main mentorship and research supervision. The Statistical Learning group, and the Statistics Research Group beyond that, will provide further immediate support to the Researcher. The Department has extremely strong links with research groups in Operations Research in Lancaster University Management School, through the STOR-i Centre for Doctoral Training, and with Computer Science, through the Data Science Institute. Therefore multiple researchers in cognate areas will contribute to the project with informal mentorship and research leadership, 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 to allow the researcher to carry out the project.

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

The active contribution of the beneficiary to the research and training activities should be described. For GF also the role of partner organisations in Third Countries for the outgoing phase should appear. Additionally a letter of commitment shall also be provided in Section 7 (included within the PDF file of part B, but outside the page limit) for the partner organisations in Third Countries. NB: Each participant is described in Section 5. This specific information should not be repeated here.

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 the Researcher 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 the Researcher 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.

Gantt chart Reflecting work package, secondments, training events and dissemination / public engagement activities



4 CV of the Experienced Researcher

During the course of my studies several invaluable experiences have greatly contributed to my desire to pursue a research-based career in Computer Science. I have had the opportunity to participate in stimulating research projects, in such areas as Machine Learning or Signal Processing. From my early years as an undergraduate student I have tried to keep the balance between theory and application. After three years of intensive Mathematics and Physics studies I entered TELECOM SudParis, a Telecommunication engineering school. There, on the one hand my engineering education made me comfortable with programming (C/C++, Java) and Network issues (LANs, WANs) and on the other and I personally got involved in a research project on PCA algorithms which has lead to a publication at ICASSP 2009. In 2008, I continued with my graduate studies in Applied Mathematics as a master student with focus on Statistical Learning where I developed solid background Machine Learning theory (including a course on Graphical Models by Francis Bach and one on Reinforcement Learning by Rémi Munos). Still I completed my master with an internship at INRIA research lab where I applied statistical learning techniques to help design a realistic automatic ad-server for Orange Inc affiliated websites. This work has launched a collaboration which is still in progress.

My current research involves 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 in our algorithms development is to ensure our algorithms capture the real complexity of a problem and testing in practice their performances in real world problems. During my PhD, I investigated Reinforcement Learning (RL) which is a field where one tries to solve complex systems where an agent has to learn from its environment. 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 define a policy through an associated value function as this value function is often poorly approximated. Therefore, this class of algorithms is expected to perform better than its value-based counterparts whenever the policies are easier to represent than their value functions. However, CBPI algorithms can require large number of samples from the environment. To improve the CBPI efficiency, I proposed new hybrid approaches using value function 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 core problems to model any problem involving adaptiveness. We designed a sampling strategy to solve several bandit problems in parallel (which led to two publications in NIPS 2011 & 2012).

During the course of my Ph.D. I worked as an research intern for 6 months at Technicolor Labs in Palo Alto California under the supervision of Branislav Kveton. Our primary goal was to improve the questionnaire asked to elicit movie preferences of users for a recommendation website. The problem was cast as an adaptive submodular maximization problem. The novelty was that we consider this problem in the case where the preferences of the users are not supposed to be known to build the questionnaire but need to be learned (which led to a publication in NIPS 2013 & AAAI 2014).

As a post-doctorate in the Queensland University of Technology, under the supervision of Peter Bartlett, I am conducting research in online learning. My first project deals with combinatorial pure exploration bandits, is set in a stochastic setting and could model network routing problem (online shortest-path problem). The second one is set in the non-stochastic setting (adversarial) bandit setting where the goal is to give a simple setting of this bandit game that admits an exact minimax solution. This therefore is a more theoretical question that draws connection with game theory.

Through the experiences already described I developed my ability to work in a team environment. The international conferences, internships and summer schools I have been attending gave me the opportunity to learn and exchange with researchers from diverse horizons. In addition, teaching computer science (Algorithmic with Python & Databases) for Master and Licence students keeps enriching my communication skills. I build up my programming skills through my curriculum in a telecommunication engineering school and later through the lectures and practical sections I gave. Moreover most of my projects have involved programming part which have made me comfortable with coding in Python and C++.

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**
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 article

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 the "Classification-based Policy Iteration" algorithms. In addition to proposing theoretical analysis of this methods (C3,C5), we implemented a complex and extensive experimental studies of the performance of this algorithms in the famous benchmark of the game of Tetris. Our result show that for the first time a reinforcement learning methods performs well in Tetris even improving on the state-of-the-art techniques. Moreover, while these state-of-the-art techniques were based on black-box optimisation techniques that requires a lots of samples from the environment, our methods require 10 times less samples to learn Tetris strategies with same performance (C8).

- *Industry:* ***Constrained Learning for Orange Ad Server:*** Orange, the french leading company in telecommunications had made a contract with the research team SequeL in order to turn their online web-advertising services automatic. My initial goal was to make a survey of the machine learning literature and find an appropriate solution optimizing their clic-per-rate revenues. This solution had to take into account specific new constraints on the limited and known number of display per ads. Finally, a new approach was proposed combining linear programming and bandits algorithms with experiments on synthetic data. The results was published and awarded in a french speaking conference (C2) and started a collaboration between SequeL and Orange which is still running.

- *Industry:* ***Adaptive Questionnaire Design at Technicolor Inc:*** During research and developpement internship at Technicolor, the primary goal was to improve the questionnaire asked to elicit movie preferences of users for a recommendation website. The problem was cast as an adaptive submodular maximization problem. The novelty was that we considered this problem in the case where the preferences of the users are unknown but need to be learned in order to build an adaptive questionnaire (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.
  3^{rd} 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.
  3^{rd} 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 25 academic staff, 10 research associates and around 50 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 — from infrastructure and fundamentals through to globally relevant problem domains and the social, legal and ethical issues raised by the use of Data Science.
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. [**TODO:** Computing facilities?]
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 [**TODO:** Other big grants?]. 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]