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]

TABLE OF CONTENTS

0 List of Participants

Participants	Legal Entity Short Name	Academic	Non-academic	Country	Dept. / Division / Laboratory	Supervisor	Role of Partner Organisation
Beneficiary							
- [TODO: NAME]							
Partner Organisation							
- NAME							

1 Excellence

[**TODO:** Convince the reader that this is state of the art, important, timely and relevant. It needs to demonstrate quality and innovation. There needs to be some detail on the methodological approach. We need to give some specifics on complementary skills development (in particular leadership), and give clarity and specifics on how VG will reach independence.]

Please note that the principles of the European Charter for Researchers and Code of Conduct for the Recruitment of Researchers promoting open recruitment and attractive working conditions are expected to be endorsed and applied by all beneficiaries in the Marie Sklodowska-Curie actions.

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, and firearms, and securing proprietary and sensitive information over the ever-growing, modern 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 coverage.

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. As discussed further in the sequel, 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 proved remarkably useful in providing maximum security.

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.

In general, in most real-world scenarios the actual outcomes of the game are not available to be fed to the solver, and must be estimated via expert educated guess or data-driven approaches. However, the data from which these values can be obtained are usually scarce and the error in expert data may be high, rendering the security game solver useless. An example of this situation can occur when, one of the players cannot completely foresee the outcome of oneâĂŹs action, or is bound to ignore what utilities the opponent is aiming to maximise.

Machine learning is a field of artificial intelligence where the goal is to design software able to **1)** actively (and smartly) collect data in order to **2)** extract information from this data so that **3)** the machine itself can make use of this information to take autonomous decisions. Based on solid statistical theoretical background, machine learning has been applied in a very large variety of modern applications ranging from robotics to personalised product recommendation.

[1] **pita2008deployed**.
[2] **tsai2009iris**.
[3] **shieh2012protect**.
[4] **yin2012trusts**.

Learning is possible in security games as this game is often played and repeated on a daily basis between the security services and the possible attackers. Therefore each day the security agency can collect more data about the parameters of the game. Another possibility is in fact that the security agency can itself 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 goal will be either **1)** to limit the costs imposed by this extra need to learn the model or **2)** to actually design an efficient data mining procedure before the game starts that guarantee a good defence strategy.

State-of-the-art

Security meets Game Theory

[**TODO:** better distinction between minimax nash stacjkelberg] 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 from the attacker. 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. In the case of games that are fully competitive between the two players (i.e. the so-called zero-sum games), these methods are provably robust in that they provide guaranteed performance against *any* possible attacker. In this case, such guarantees hold, even if the defender's strategy is completely revealed to the attacker. The extension of this guarantee to a more general (non zero-sum) game is provided by Stackelberg equilibrium, a notion that generalises the famous Nash equilibrium.[5]

Sequential decision-making under uncertainty

Machine learning is a field of artificial intelligence where the goal is to design software able to extract information from data so that the machine itself can make use of this information to take autonomous decisions. The problem of sequential decision-making under uncertainty arises in everyday life, when we try to find an answer to questions like how to navigate from home to work, how to play and win a game (e.g., backgammon, poker, or the game of Tetris that has been used as an experimental testbed in Victor's thesis), how to retrieve our information of interest from the Internet, how to optimize the performance of a factory, etc. Such applications naturally involves uncertainty about the consequence of the player action (in the game of Tetris, the next falling piece is sampled randomly, in a security game the success of randomly selecting passengers to be searched is inerrantly unpredictable). To solve this problems machine learning has proposes methods and techniques that borrow and extend a lot from the fields of statistics (in order to take into account the uncertainty around the data) and optimisation (to create fast converging methods, minimising for instance the number of actions to get to specific level of performance).

A core framework for learning is the *multi-armed bandit problem*. It is a game where at each time step the same fixed set of actions is available to the player. It can be formalized as a game between an environment and a forecaster. 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 simultaneously the forecaster chooses to pull an arm $I(t)$ and observes a reward $l^t_{I(t)}$. The fundamental property is that the player does not get to observe the reward he would have collected if he had selected another arm instead. In a security context you could imagine that the K arms/actions are K possible security strategies whose value are following some random law and that can only estimated by testing them /sampling them /pulling

[5] **korzhyk2011stackelberg**.

the corresponding arm. For this proposal we focus on two different ways to measure the performance of interest of the methods. The first formulation corresponds to the classical cumulative regret setting where the forecaster tries to constantly choose the security strategy with the highest value on average. The second one is the "pure exploration" setting, where the forecaster can uses the exploration phase, a limited phase during which he can freely pull the arms he chooses to, in order to identify the best security strategy among the K.

[**TODO:** Maybe add success of bandits.]

- The cumulative regret setting is the standard formulation for multi-armed bandits. In this formulation, the objective for the forecaster is to minimize the expected cumulative regret after n pulls $R(n)$ defined as

$$R(n) = \max_i \sum l_{i,t} - \mathbb{E}\left[\sum_{t=1}^{n} l_{I(t),t}\right].$$

Here a fundamental trade-off arises between the simultaneous need to select the solution we currently think is the best in order to maximise the immediate reward (exploitation) while also wanting to test possible other choice/actions that might or might not be better (exploration).

The problem has been studied under two main assumptions about the power of the environment. In a first setting the environment chooses the rewards in a stochastic manner, following some predetermined (however unkwnon to the player) distributions to allocate the rewards to the arms). This setting permits is very handy to model noise in the data and will be useful in security games when the defender is uncertain about the stochastic outcome of one of its action. A popular efficient algorithm for it is Thompson Sampling as it has both proved very efficient,[6] handy and as recently started to be theoretically good [7]. In the second setting, no stochastic assumptions are made and it is actually as if an adversary could arbitrarily choose and design all the rewards.[8] This latter setting is therefore a harder one and will be useful when modelling a competition between our two players, the defender and the attacker.

Note that a very numerous of variation of the initial games ranging from considering infinite number of arms to continuous actions[9] permits these methods to adapt to a very large mount of challenges

- The pure exploration is a relatively new setting, where the forecaster is only evaluated at the end of an exploration phase comprising a limited number of pulls. Contrary to the cumulative regret setting, the rewards collected before the end of the game are not taken into account. We see a very interesting application of this setting to security games when the defender can test his defensive strategy before putting it in application in the real world. The objective is a play between the probability of error and the samples (pulls) required. In the *fixed confidence* setting (see e.g.,[10]), the defender would typically try to minimize the number of test (and hence the cost) needed to achieve a fixed confidence on the quality of the returned best estimated defence strategy while in the *fixed budget* setting (see e.g.,[11]), the number of tests of the exploration phase is fixed and is known by the defender, and the objective is to maximize the probability of returning the best strategy at the end of the phase.

Frontiers of Security Games: From handling uncertainty towards self-learning algorithms
Some of the main issues forming the primary focus of research in security games have been scalability, the ability for the designed methods to handle problems with very large number of actions for the players, or devising strategies that take advantage of the attacker's potentially limited rationality or bounded memory.[12] Closer to our current interest, another important research goal that has been extensively addressed is to devise methods that are robust with respect to uncertainty about the environment.[13] Granick, for example, argues that weaknesses in our understanding of the measurability of losses serve as

[6] **Chapelle11EE**.
[8] **Auer03NS**.
[9] **Wang08AI**; **Abbasi-Yadkori11IA**; **Dani07TP**.
[10] **Maron93HR**; **Even-Dar06AE**.
[11] **Bubeck09PE**; **Audibert10BA**.
[12] **tambe2012game**.
[13] **aghassi2006robust**; **Nguyen14RO**; **Kiekintveld:2013**.

an impediment in sentencing cybercrime offenders.[14] Swire adds that deterring fraudsters and criminals online is hampered if we cannot correctly aggregate their offences across different jurisdictions.[15] These interest in dealing with uncertainty in the data shows how important and crucial and costly it could be and that it is a real issue in real problems.

However, still little has been done to give a realistic active learning solution to the uncertainty about the unknown environment. Achieving this goal is indeed crucial, since algorithms that make use of environmental knowledge are arguably more reliable than those merely designed to be robust against this lack of information. With this motivation, some interesting advancements have very recently been made through links with optimisation and machine learning methods. These methods focus mostly 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.[16] propose analyses in terms of the number of required queries to learn the optimal defender's strategy. Marecki et. al. and Qian et. al.[17] 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, as shown by,[18] are provably sub-optimal. Recently an extended analysis is given by[19] 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. The latter work shows strong connections with adversarial bandit theory.

Main Goal

The purpose of this proposal is to design new methods for security games that are even more practical in the sense that they 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. We desire 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 booming field of machine learning. First we will look at new instances of these problem that are related to our area of bandit expertise and that serve real purpose.

To solve real world problems our higher priority fields of investigations will be:

- *Explore stochastic assumption:* Aside from contributing to adversarial setting, One of our interest is to make Stochastic assumptions when dealing with noise in the model and adversarial assumption when dealing with the adversary to make our approach both realistic and robust. So far stochastic noise in the model in conjunction with learning has been untouched while it can happen when phenomenon are not inherently adversarial (sensor that work stochastically action that have stochastic outcomes, checkpoint might not stop deterministically the attacks and the probability of success needs to be determined, here learned).

- *Efficient learning algorithms:* Our goal is to have a theoretically sound approach by designing efficient algorithms for which we can provide finite sample analysis. Mention my previous work?

- *Scalable learning algorithms:* Extending the previous approaches to complex problem that involves some combinatorial structure is also important. Submodularity emtion work?

- *Different feedback structures:* The complexity of learning highly depends on the quality of the feedback that the learned can collect. From the most informative full information feedback to the many variations of partial feedback setting it is of importance, as discussed in Objective 1, to quantify how the nature of the feedback affects the performance and to focus on scenario that correspond to real world example.

[14] **granick2005faking**.
[15] **swire2009no**.
[16] **blum2014learning**; **letchford2009learning**.
[17] **Marecki12PR**; **qian2014online**.
[18] **munos2014bandits**.
[19] **Balcan15CR**.

- *Robust learning:* When dealing with security game, robustness is a key issue. While the need for learning inherantly add some ecurity cost we want to look at instance of the learning problem that preserve some notion of robustness. The first way to be robust is to be very conservative and assume that an adversary actually chooses (in the most adversarial way) the data that we do not know A first way to reply to that is to use setting where no assumption use of *adversarial* algorithm. Another possible concern that might happen in some problems pure best arm identification Another possibility is that in some situation we do not want the learning process to happen during the use of the program but before hand. Then we can assume that we use of a pre launch exploration phase where we try to learned as precisely as possible the model given some budget constraint or some targeted performance guarantees.

  An this need for robustness needs to be mitigated depending on the application. some real world problem might need an extra care on robustness like terrorist attack while others are not that sensitive to it. In the latter case we should not be to conservative in the learning to be able to learn faster.

  we might be required to learn defence strategies that are not necessarily the best in expectation but instead also guarantee not to possess large variances in their performance. Here we plan to make connection with risk averse learning algorithm.

- *Adaptivity ?*

Objective 1 Pure exploration in Stackelberg games Bringing the pure exploration in bandits to Stacklerberg games is the natural first phase of our project. As it is first one of the main domain of expertise of Victor and two it brings a first conservative way to address learning without bringing to much risks for the learner. Indeed in this setting the learning of the unknown model happens during an exploration phase before putting on the market. This for instance means that he can run tests of the security in a variety of predetermined attack scenario and therefore probe his own probability of defence.

We would make the stochastic assumption as here it accounts for noise in our model and not adversary actions. The objective of this approach is to determine the best strategy during a given exploration phase and is therefore closely related to the general theory of optimisation and has been study in the discrete context of multi arm bandit as pure exploration problems.[20] This initial work has been extended in a flurry variant setting where one tries to find the best(s) arms. Victor has a nice expertise in that and has participated to the extension and application of such a framework in more and more complex problem (cite my work?) and is working on extension to combinatorial bandits that would improve upon the seminal work by Chen.

Taking into account the particular structure of the problem will be necessary when dealing with Stackelberg equilibrium in security games. There the function to optimise is even more complex. What is the complexity here?

• **Complexity:** The hardness of the best arm identification problem in the stochastic setting can be interpreted as the total number of pulls required to discriminate the best arm(s) from the others. In simple multi-arm bandit setting it is defined as the sum of the complexity of each suboptimal options, where the complexity of a suboptimal option i is invertionally proportional to the gap $\Delta_i = \mu^* - \mu_i$ the difference between the value of the the best option μ^* and the value of option i

More precisely the complexity H is defined as

$$H = \sum_k \frac{1}{\Delta_k^2}, \tag{1}$$

Extensions of this complexity notion have been designed in more complex setting like combinatorial bandits citeChen. Note that Victor is currently working on a improved version of this result. In a combinatorial

[20] **Audibert10BA**.

setting, the forecaster must make a recommendation that is of combinatorial structure The (combinatorial) decision set is $\mathcal{C} \subseteq 2^K$ is such that any decision $U \in \mathcal{C}$ is a set of arms $U \subseteq \mathcal{K}$ and its value is the sum of their values, $\mu_U = \sum_{i \in U} \mu_i$. The value gap between two decisions is denoted by $\Delta_{U,V} = \mu_U - \mu_V$ and $U^* = \arg\max_{U \in \mathcal{C}} \mu_U$ is the best decision

$$\Delta_k^{\odot} = \begin{cases} \mu^* - \max\limits_{V \in \mathcal{C}: k \in V} \mu_V & \text{if } k \notin V^* \\ \mu^* - \max\limits_{V \in \mathcal{C}: k \notin V} \mu_V & \text{if } k \in V^* \end{cases}$$

In case of of games the picture would be even more complex as the complexity would depend on the actions that adverasry have available.

One first step is to relax the problem as shown in Krause et al finding the best response to a given adversary. This is known to be is NP hard problem but can be solve almost optimally be a greedy algorithm thank to a sub modularity property of the problem. This gives rise to a first objective which would be to learning optimise stochastic submodular function under a pure exploration setting. Note that I worked on similar subject with learning in submodular functions.

connections with risk averse (Cite the work of Amir Sani) maybe a separate section for this. talk about the classical cumulative regret setting also!

this also ask the long term question of convergence when both player are learning at the same time their utilities.

Objective 2 Learning more complex adversarially chosen attacker in Stakleberg The idea would be to extend the work of Balcan using more complex bandit algorithms. They use a version with k known attackers. We can assume that k is extremely large but there is some structure that permits us to use for instance combinatorial bandits.

Objective 3 Repeated Network Security Games The security issue naturally has application in graph problem that model the network of roads/ connection between computers that agents might need to secure. Therefore there has been study that apply game theory to this problems. For instance it has been used to monitor road barrage in mumbai (connection) The goal is there to put some check point on a road to stop some terrorist. Its a one shot game where you try to minimise the probability of the player to pass. Utilities are not really defined and complex here You just want to maximise the probability of catching the attacker. We are interested in a version of this game that is repeated . Everyday the same problem arises. We would minimise the cumulative regret. Therefore the defender can be adaptive and if the attacker is not smart and repeat always the same plan we will catch him often (not totally a worse case scenario). This can be seen actually has a specific problem of adversarial combinatorial bandits where the attacker is limited to a very specific structure of losses which are path in a graph. We can expect to use the specificity of the graph by using some result from spectral graph theory. Maybe also we can use this theory to solve some issue with the scalability of the algorithm.

Originality and innovative aspects of the research programme:

timeliness and relevance:

Security is booming since 5 years mahcine laerining also the conjuntion of the two is definietly relevant but still largely unexplored Europe wants security migrants/ spy

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

Outline how a two way transfer of knowledge will occur between the researcher and the host institution, in view of their future development and past experience: (please see Section 5.2 of this Guide):

- *Explain how the Experienced Researcher will gain new knowledge during the fellowship at the hosting organisation(s)*

- *Outline the previously acquired knowledge and skills that the researcher will transfer to the host organisation*

The overall trianing 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

Required sub-heading:

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[21]

The text must show that the Experienced Researcher should be well integrated within the hosting organisation(s) in order that all parties gain the maximum knowledge and skills from the fellowship. The nature and the quality of the research group/environment as a whole should be outlined, together with the measures taken to integrate the researcher in the different areas of expertise, disciplines, and international networking opportunities that the host could offer.

For GF both phases should be described - for the outgoing phase, specify the practical arrangements in place to host a researcher coming from another country, and for the incoming phase specify the measures planned for the successful (re-)integration of the researcher.

Describe briefly how the host will contribute to the advancement of their career. In that context the following section of the European Charter for Researchers refers specifically to career development:

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 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 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[22] 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,[23] stochastic approximation,[24] and the mathematics of statistically-inspired reinforcement learning.[25] 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

[21] 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.

[22] **MayEtAl2012**.

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

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

[25] **LeslieCollins05**; **LarsenEtAl2010**.

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 on project 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

Applicants should demonstrate how their proposed research and personal experience can contribute to their professional development as an independent/mature researcher.

Please keep in mind that the fellowships will be awarded to the most talented researchers as shown by the proposed research and their track record (Curriculum Vitae, section 4), in relation to their level of experience.

Dr Gabillon's intention is to continue developing until he is able to build and lead a multi-disciplinary research group, developing innovative and impactful research in machine learning for decision-making, in collaboration with industry partners. The initial stages of the researcher's career indicate that this is a realistic goal, with multiple high-quality publications in a short time frame, and already significant international research experience (see Section 4). He has worked both as an academic researcher, and within a company [**TODO:** Which company, doing what?], and the proposed research will broaden his research foundations significantly to allow further development as an independent researcher. Working closely with Lancaster University's Data Science Institute, and STOR-i Centre for Doctoral Training, will not only develop a professional network of impact-aware academics, it will foster skills in the development of academia/industry partnerships.

Furthermore, Dr Gabillon will establish a 2-year Personal Career Development Plan which he will update regularly under the mentorship of Prof. Leslie, the project Supervisor. The Fellow will have access to Lancaster's central training resources, which will be used to further develop skills in teaching, public engagement, research supervision and management, which will contribute greatly to a position of professional maturity. Moreover, by participating in weekly seminars, group meetings, and informal gatherings, the Fellow will have the opportunity to receive valuable feedback about his career development from the experience scientists in the Statistics Research Group. This research group is visited by dozens of internationally-leading scientists each year, which will provide further opportunities to develop a research network and reach a position of professional maturity.

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

Explain the expected impact of the planned research and training, and new competences acquired during the fellowship on the capacity to increase career prospects for the Experienced Researcher after this fellowship finishes.

Demonstrate also to what extent competences acquired during the fellowship, including any secondments will increase the impact of the researcher?s future activity on European society, including the science base and/or the economy

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 with Lancaster University's extremely effective mechanisms 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 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

The new knowledge generated by the action should be used wherever possible to advance research, to foster innovation, and to promote the research profession to the public. Therefore develop following three points.

- *Communication and public engagement strategy of the action*
- *Dissemination of the research results*
- *Exploitation of results and intellectual property rights*

Concrete plans for the above must be included in the Gantt Chart (see point 3.1). The following sections of the European Charter for Researchers refer specifically to public engagement and 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", bringing together academic data scientists with local companies in regular show and tell sessions. The Researcher will be a regular participant at these events, enabling him to present aspects of his research to the local business community, develop an understanding of the business requirements for this kind of user, and build 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, and Dr Gabillon will be mentored to develop similar relationships. 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;*[26][27]
- *List of major milestones;*[28]
- *Secondments if applicable.*

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

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 and the legal aspects. 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

[26] 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.

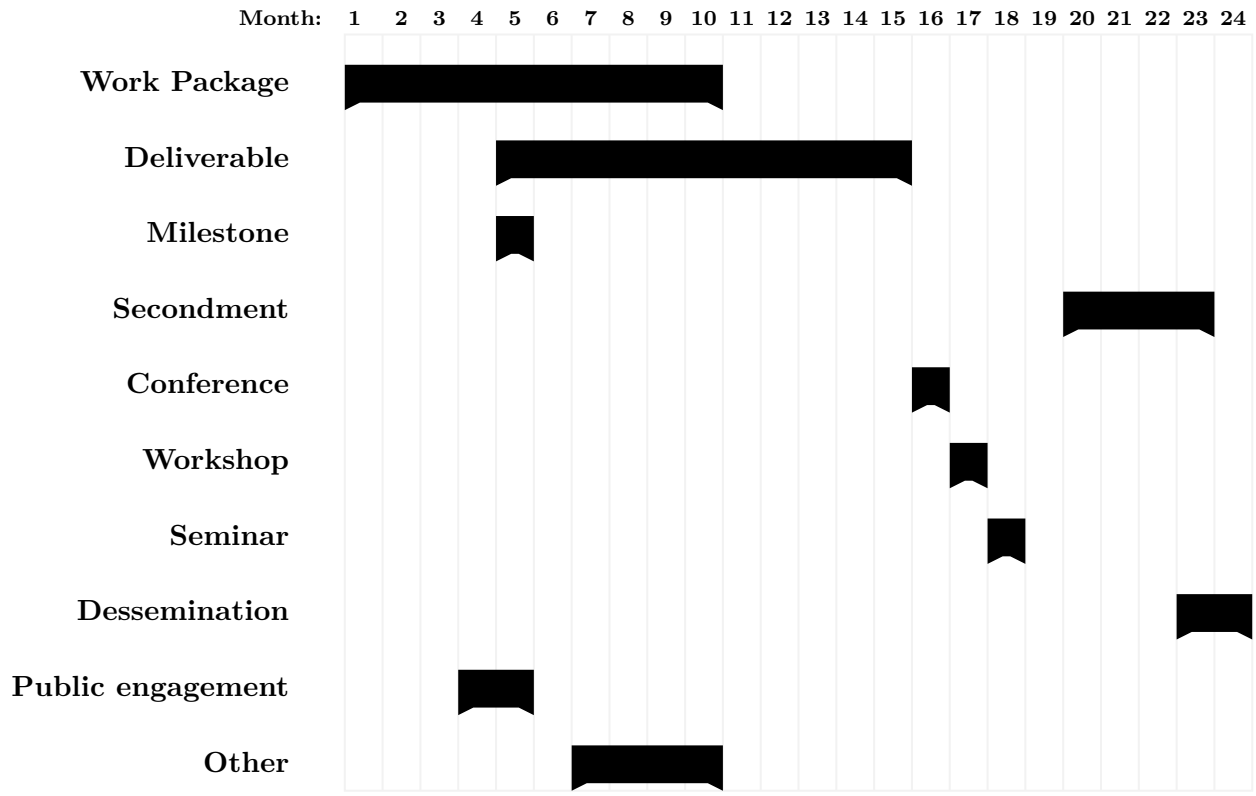
[27] Deliverable numbers ordered according to delivery dates. Please use the numbering convention <WP number>.<number of deliverable within that WP>. For example, deliverable 4.2 would be the second deliverable from work package 4.

[28] 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.

mentoring meetings with the Supervisor, in which progress against the workplan and career development plan will be discussed.

[**TODO:** Discuss risk management.]

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



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.

4 CV of the Experienced Researcher

This section should be limited to maximum 5 pages and should include the standard academic and research record. Any research career gaps and/or unconventional paths should be clearly explained so that this can be fairly assessed by the independent evaluators. The Experienced Researchers must provide a list of achievements reflecting their track, and this may include, if applicable:

1. Publications in major international peer-reviewed multi-disciplinary scientific journals and/or in the leading international peer-reviewed journals, peer-reviewed conference proceedings and/or monographs of their respective research fields, indicating also the number of citations (excluding self-citations) they have attracted.
2. Granted patent(s).
3. Research monographs, chapters in collective volumes and any translations thereof.
4. Invited presentations to peer-reviewed, internationally established conferences and/or international advanced schools.
5. Research expeditions that the Experienced Researcher has led.
6. Organisation of International conferences in the field of the applicant (membership in the steering and/or programme committee).
7. Examples of participation in industrial innovation.
8. Prizes and Awards.
9. Funding received so far
10. Supervising, mentoring activities

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 RÃľmi 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).

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 a combinatorial set of possible choices, 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) where the goal is to give a simple setting of this bandit game that admits an exact mimimax 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++.

My long term career goal is to become a researcher. I wish to gain professional experience at an environment that will allow me to expand my knowledge and capabilities through collaborations with researchers who can mentor and inspire me. I am confident that GRASP will provide me with such an environment and much more. It fits my willingness to lead research that can find real applications, particularly in artificial intelligence for games or robotic purpose (as I already worked on the Tetris game). I believe that my background in Machine Learning will permit me to take on GRASP challenge on designing powerful lifelong learning algorithms. I also believe that my diverse research background, and my prior exposure to similar research environments make me a unique candidate for the internship program at GRASP. I look forward to conducting research at GRASP world-class research environment, while nurturing my innovative and practical abilities.

Curriculum Vitae of the Applicant, Dr Victor Gabillon

Education

- **June 2014 : PhD in Computer Science**

  in Team SequeL, INRIA Lille - Nord Europe, France. *Title:* "Budgeted Classification-based Policy Iteration"
  Domains: Reinforcement learning & Bandits games
  Supervisors: Mohammad Ghavamzadeh & Philippe Preux

- **2008-09: M.Sc. in applied mathematics, ÃĽcole Normale SupÃŕrieure, Cachan, France.**

Cursus MVA (image processing & statistical learning) with honours.
Relevant courses: Reinforcement Learning, Graphical Models, Statistical learning (SVM, Boosting...).

- **2006-08: Engineering degree, TELECOM SudParis, ÃĽvry, France.**

Graduate school of engineering committed to the development of information technology.
Relevant courses: Programming, Statistics, Information Theory, Image Processing, LANs & WANs.

4.1 Publications

Victor Gabillon, Branislav Kveton, Zheng Wen, Brian Eriksson & S. Muthukrishnan, ***Large Scale Optimistic Adaptive Submodularity***. AAAI 2014, 28^{th} Conference on Artificial Intelligence. Oral presentation at Quebec City, Canada, July 2014.

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

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

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

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

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

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

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.

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

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.

Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon & Matthieu Geist, ***Approximate Modified Policy Iteration***, JMLR.

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]