National Cyber Security Centre

The Alan Turing Institute

Security for Artificial Intelligence: Call for Proposals

Closing date: 30 September 2022

Contents

Summary

The Alan Turing Institute's Defence and Security programme in partnership with the National Cyber Security Centre (NCSC) is inviting proposals from academic researchers for research to improve our understanding of Artificial Intelligence (AI) security, particularly with a view to practical risks and implications.

Available funding

Funding will be available for 2 to 3 short projects over the duration of 3 months. Each application can be for a maximum of £40,000 (not including VAT). The proposal should be defined such that the research can be completed by mid-March 2023.

The NCSC believes that this topic is of significance across all applications. We will be campaigning for more attention to be given to the topic at a national scale and seeking additional funding for research from both government and industry partners. We therefore expect future funding to become available and we will run another call for proposals.

Terms and conditions

The funding will be made available under The Alan Turing Institute's Defence & Security programme Research Service Agreement terms and conditions. For copy of the terms please contact Alaric Williams by emailing dsprogramme@turing.ac.uk.

You will be required to confirm your university's acceptance of these terms as part of this application process.

The research will be funded at Full Economic Cost and VAT will apply. Budgets for attendance to publicise and disseminate the work should be included within the proposal. Funding should be allocated to allow for travel to the NCSC office in London for a workshop in March 2023.

Eligible costs include:

- Salary of personnel working directly on the project – this could include, for example, PIs, postdoctoral research associates, research assistants, data managers, data scientists or software engineers.
- travel and subsistence for project researchers (e.g., attending conferences, travelling to/from the Turing/other collaborators).
- conference or event attendance fees (where conference/event is directly applicable to the research project).
- Cloud computing or other high performance computing costs.
- Other costs which are specifically justified for the project e.g., books, meeting room or catering costs, specific laptops.
- Open access publications.

Background

The NCSC exists to help make the UK the safest place to live and work online. The NCSC support the most critical organisations in the UK, the wider public sector, industry, SMEs as well as the general public. As part of this the NCSC understands cyber security and distils this knowledge into practical guidance that we make available to all and uses industry and academic expertise to nurture the UK's cyber security capability.

This call focuses on the security of AI; as AI is being used to make a wider range of decisions, including those of high significance, it becomes likely that these systems will become the direct target of attackers. Being able to extract important information or manipulate the processes and outputs of these intelligent tools could have serious impacts, not just in the field of security, but much wider.

The NCSC is working towards contributing to advice, guidance, assurance and standards including effective consideration of Machine Learning (ML) security. This includes the adoption of AI/ML across the UK enabled with an understanding of the security aspects. As part of this work, the NCSC is developing Security of AI Principles for publication, aimed guiding at all stakeholders involved in the process of developing, deploying and operating a system with a Machine Learning (ML) component. The aim is to help them make educated decisions by aiding them in assessing specific threats to their system. They are designed to be wide ranging and agnostic to data type, model algorithm and deployment environment, working towards the idea of best practice in this area. The NCSC is interested in the tooling and deployment of techniques to understand and detect these security concerns.

The research from this call will help the NCSC improve understanding of security vulnerabilities in AI systems. Knowing the limits of what can be extracted or manipulated will guide the focus of the NCSC security of AI research.

Context for the research

For this call we are focussing on the advancement in securing artificially intelligent systems. While there is significant interest in the usage of artificial intelligence, the end-to-end understanding of what is needed to ensure that AI-based tools are secure is still limited.

Secure algorithmic design provides only one component for providing security for artificial intelligence. There are security risks throughout the ML development lifecycle, from requirements scoping and data collection to implementation, maintenance, and decommissioning. There is a need to understand the prevalence of these risks in a real-world context and evaluate the methods and tools available. This needs to extend to future solutions which are needed to provide a solution to security of AI. A part of this is understanding where current ML systems are being exploited and what data exists or needs to be created within systems that could send an alert to make system owners and users aware that an ML system has been targeted. The long term aim of this research is to enable further AI security research and tool development to assist anyone involved in ML deployment mitigate the risks. Evidence showing real-world use cases and limitations in detecting compromises in current ML algorithms and security tooling is a key output of this research.

Research challenges

The research proposals should clearly address at least one aspect of these principal challenge areas:

- Assessing the **transferability** of attacks. This could be from one machine learning algorithm to another or from a larger model compacted to be deployed on the edge or in resource-constrained environments.
- Detecting security **vulnerabilities** in AI models and ML systems. What approaches and measures should be used to detect when our models/systems have been compromised? This also includes automated detection of vulnerabilities in AI models (perhaps using AI itself for the automation!).

- Assessing the behaviour of intelligent systems and understanding how to detect **degradation** in their behaviour due to malicious activity. This includes identifying effective mechanisms for correcting or compensating for malicious behaviour.
- Best practices for mitigating **model inversion**, including model training and deployment for minimising the amount of insight an attacker might be able to gain from the model using outputs and statistics returned or test data classified.
- Best practices for **building secure intelligent systems**. Understanding which algorithms and decision models are most suitable for a range of problem classes and use cases and providing guidance into selection criteria. This additionally includes requirements for the design of user interfaces and system architecture for deploying secure AI.

The importance and impact of secure artificial intelligence is wider than just those of the security application. Any context where there is a viable need to secure the intelligent system would be considered a valid use case for this research.

Eligibility

To be eligible to apply you must:

- Be part of a UK university or research institute. Commercial organisations or overseas universities or institutes are not eligible.
- Have permission from your organisation to apply, i.e., ensure your organisation agrees to the Terms and Conditions provided and that you submit an approval of submission letter from your research/finance office stating this. An example of a letter is available on request.

How to apply

Application submissions should be no more than six sides of A4 and should include a breakdown of all costs involved, including equipment, travel & expenses etc.

Applications must be submitted via the online portal at https://ati.flexigrant.com/. If you have not already done so, all applicants must first register on the system and provide basic details to create a profile. If you have any questions regarding the application form or using the online system, please contact the programme inbox dsprogramme@turing.ac.uk.

Please use the budget template provided in the Flexigrant application form**.** Please note, applicants will also need to upload on Flexigrant an 'approval of submission' letter from your research/finance office to confirm costs are correct. The Principal Investigator must ensure the same is received for all collaborators / universities on multi party applications.

We must receive your application by **1600 on Friday 30 September 2022**.

What should be in the proposal?

Each proposal must make it very clear how it addresses the challenge areas described above. Proposals should also include details of any planned engagement with 'real world' security.

The proposal should specifically address each of the following items:

- Background: An outline of the context of the research.

- Aim: A description of what understanding of the topic space the research is progressing and what potential impact it will have in practice.

- Relevance to the call: A description of which challenges the research addresses, and how it addresses them.
- Data: Whether the research is planning to create or make use of any specific datasets, how they will be generated and handled, and their availability prior to the start of the research.
- Resources: An overview of the timescales, resources and structure of the research. A workplan should illustrate how these aspects combine to progress the research. The resources being used should be detailed, and CVs for named and visiting researchers included where these are known. Where researchers aren't known, a clear recruitment strategy should be identified to ensure completion within the time frame. Any external collaborators and advisors should be identified.
- Method: An outline of how the research will be carried out, detailing techniques and approaches that intend to be used. An indication of the level of previous experience of these approaches should be included.
- Potential impact: A description of how the research will push forward the understanding of the challenges and potential remediation for securing artificial intelligence.

The projects will be expected to provide a summary report on the activities at the end of the project. There would also be expectation that a representative from successful projects would be able to attend a working with the NCSC in March 2023 to disseminate the findings of the research. This should be accounted for in the project planning.

If you are employed by one of the Institute's 13 university partners, please contact your University Liaison Manager (a list of University liaison managers is available on the Turing website) to make them aware of your application. They can provide support, answer questions and involve you as part of the Turing community at your university from now on.

If you are employed at a university that received a Turing Network Development Award, please contact your Award lead (a list of Turing Network Development Award Leads - scroll to the bottom of the page) – to make them aware of your application.

Assessment and review

The assessment and review will follow the following stages:

1) Stage 1.0: Eligibility and triage
2) Stage 2.0: Expert review
3) Stage 3.0: Expert review panel

Following eligibility checks, proposals will be reviewed by an expert assessment panel comprising representatives from academia, industry and HMG. The panel will rank the proposals based on scores and panel consensus following review.

The assessment panel will consider three key criteria:

- Quality: This will consider the method and concepts for the proposed research. This will assess if the methods are suitable for delivering the desired outputs and pushing forward fundamental understanding in the field.
- Viability: This will assess how feasible it is to practically carry out the proposed research, and if it can be delivered in the time frame. This will account for the

difficulty of the tasks, logistical factors surrounding delivery, and the track record of the proposed research team.
- Significance: This will consider the relevance to the call and the themes that are represented.
- Justification of resources: This will consider whether the proposal is appropriately resourced and suitable expenditure has been included in the budget.

Each of the criteria will be scored by the panel from 0 - 10. While all four criteria will have equal weighting in evaluation, there will be a minimum requirement on significance to be considered for approval.

Key Dates

Deadlines are as follows

Activity	**Date**
Proposals to be Submitted*	Friday 30 September 2022
Announcement of Results	Monday 31 October 2022
Research Starts*	Mid-November
Research Completed and deliverables submitted	Mid-March

*Proposals must be submitted via Flexigrant by 16:00 Friday 30 September 2022.

**Any project agreements not signed by Monday 14 November 2022 may result in funding offer being withdrawn and going to an application on the reserve list.

Post-award information

Project meetings

Successful applicants will be expected to attend a kick-off meeting and a project close meeting, with a Technical Partner from the D&S programme Partner/s. These may take place online, at the Turing, at the NCSC, or at the project lead's university.

Screening of researchers

This research is not at a classified level so formal security clearance (see https://www.gov.uk/guidance/security-vetting-and-clearance) is not required.

Outputs required

We require that all projects will produce the following:

- If applicable, the application to and approval from the relevant research ethics committee.
- Progress summaries (up to one page) and meetings.
- Summary report to cover activities and outcomes at the end of the project.
- Attendance to a working at the NCSC in March 2023 to disseminate the findings of the research.
- If applicable, any source code, compilation, use documentation and material associated with the outputs delivered.

Outputs acceptance criteria

The summary report shall describe the entire project in sufficient detail to explain comprehensively the work undertaken and results achieved - including all relevant technical

details of any hardware, software, process or system developed there under. The technical detail shall be sufficient to permit independent reproduction of any such process or system.

For advice and guidance on reproducibility, please visit The Turing Way project resources; the online book is available here:https://the-turing-way.netlify.com/introduction/introduction Contributions and discussion are also welcome here:https://github.com/alan-turing-institute/the-turing-way#about-the-project

If outputs do not meet the acceptance criteria, re-work will be requested before final acceptance.

Publications

Please note, approval from the D&S programme is sometimes required prior to publication; in such cases, approval will not be unreasonably withheld.

The funders are committed to full and open publication of the research outputs in line with academic practices.

We encourage researchers to submit their findings to a high-quality peer-reviewed journal or conference, on an open-access basis (funding for open-access fees will be available on a case-by-case basis).

We expect a 'green' open access version of any papers to be published (if allowed by journal/conference - please check http://www.sherpa.ac.uk/romeo/index.php) either as a pre-print on (e.g.) the ArXiv (https://arxiv.org/) or in an institutional repository.

We also encourage datasets and research code to be openly shared too where possible - for example on the Turing's Github repository. All publications, reports and code should reference the support of the Turing Defence & Security programme.

Reporting and dissemination

Extracts from reports may be collated into update papers for the D&S Programme Board, Strategic Partners Board, Turing Innovations Ltd Board, and the Turing's Trustee Board.

Awardees may also be required to present their work to members of the D&S programme, the D&S Programme Board and/or other invited audience during the award period.

Reporting allows further identification and signposting of potential additional opportunities for the benefit of the awardees and the Turing; for example, opportunities from across the Turing's network such as new collaborations, external/public engagement, media/press, other funding availability, speaking slots at or invitations to events/conferences/seminars.

Queries

Please contact Alaric Williams, Programme Manager, The Alan Turing Institute, dsprogramme@turing.ac.uk.