The Alan Turing Institute

Defence & Security Grand Challenge:

AI Security – Call for Proposals

Closing date: 02 September 2024, 15:00

Contents

Summary

The Alan Turing Institute's Defence & Security Grand Challenge is seeking to commission research projects in AI Security. The research call is part of a wider multi-year programme in AI security, working closely with UK Government agencies.

The programme aims to have real-world applied impact through deployable machine learning solutions realised over the coming years. This research call is seeking to support novel research ideas, testing of real-world solutions with datasets where possible and, if appropriate, literature reviews. Projects will need to be delivered between the period 01 September 2024 – 31 March 2025. Projects may be initiated after 01 September 2024 and applicants may determine the duration but the project should be defined as such that it can be completed by 31 March 2025.

Should your project benefit from an extended duration (to take place between April 2025 – March 2026) please provide details of an optional extension as part of your application. Please note follow-on funding is not guaranteed beyond March 2025.

Given the multiple interpretations of the term "AI", and the various instantiations of what may be termed "AI models and systems", our collective definition of AI includes systems based on current and future generative AI, e.g. Large Language Models (LLMs) and image, video and audio generation; as well as now-standard deep learning systems using ML for classification and regression, e.g. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), and "classic" ML systems in everyday use, e.g. those using control systems and statistical learning methods.

Available Funding

The total funding available per project is £50,000 - £200,000 ex VAT. The research will be funded at 100% Full Economic Cost and VAT will apply.

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.
- Overheads, estates and indirects
- 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.

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 a copy of the terms please contact Alaric Williams by emailing dsprogramme@turing.ac.uk.

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

The Requirement

The Alan Turing Institute is seeking to commission a range of external subject matter experts/teams for 3-6 months, to deliver projects in Model Security, Data Security, Learning Security, AI and Cyber Security and AI model Cybersecurity Evaluations and Safeguards Analysis. Below are descriptions of technical areas of interest. We would expect projects or related research activities to be in the defence and/or national security context.

Each project should be led by a theorist drawn from a relevant academic domain. They will seek analytical solutions that carry more strategic worth than experimental results. We encourage projects to reveal 'dead-ends' in the solution endeavour that are normally supressed in an academic industry, in which only successes are traditionally reported.

2. Model Security

Theoretical and practical guarantees of security and robustness for AI/ML systems require deep understanding of model structure, e.g. neural network layers, activation functions, etc., and the effects of this structure on model behaviour. The choice of structure – often chosen in pursuit of optimal performance – can have a range of security implications, from attack susceptibility to risks of data exposure.

2.1 Characterisation, assessment and assurance

Can we test model structure for potential security issues, and can we do this in a privacy preserving (non-invasive) manner? We will investigate whether the presence (or absence) of certain components and design patterns indicate vulnerability to attack; and whether such analyses could be used for AI/ML model security assurance.

2.2 Training Evolution

Recent breakthroughs in deep learning theory have enabled some understanding of complex model behaviour, from first initialisation and throughout the optimisation process. Our question is whether these tools, e.g. the Neural Tangent Kernel (NTK), a model function approximator, can be used to make statements about security during the dynamics of training. For example, are there regions of higher or lower security in a gradient flow? Can we detect them, or even target regions of high security?

2.3 Model Inversion

This is the ability to acquire representations or samples from the pre-image of a trained model function. What can we say mathematically about (approximately) solving inverse problems for AI/ML? Can we guarantee inversion robustness, and what structural traits might allow such guarantees?

2.4 Geometry of sensitivity and attack detection

Can the tools of geometry (specifically tropical geometry) be used to characterise models' sensitivity to new data points and the presence of nefarious data such as poisoned samples or adversarial examples?

2.5 Stability and trustworthiness

Is model stability a good measure of trust? Are unstable models more vulnerable to adversaries? For example, could a stability statement be defined, e.g. Lipschitz (using a distance metric such as Gromov-Hausdorff), and could this be used to make claims about security?

2.6 Security implications of efficiency

To make very large models tractable, we often turn to methods that reduce complexity, e.g. sparsity, surrogate models, pruning, quantisation, etc. This often reduces task performance, but can we analyse mathematically the security implications, e.g. inadvertently revealing hidden structure in the original model or private training data?

2.7 Detecting interference

If an attacker interferes with a model, this could be very subtle, e.g. manipulating the weights of a neural network. Can we develop effective methods for detecting interference from the model behaviour alone? Similar to out-of-distribution detection for when data has been 'manipulated' naturally, can we detect nefarious manipulation?

2.8 Firmware and hardware security

When AI/ML models are deployed, their operation depends inherently on their representation in a device, e.g. their binary representation. Does compilation or conversion to a deployment-ready format introduce its own security issues? Certain patterns that betray information about parameters θ or training data x may be more apparent in these formats; can we defend against such revelation?

3. Data Security

Many of the privacy and security concerns in AI/ML systems lie in the data, primarily its representation both inside and outside the model. Several of our research questions involve the protection of data during training, inference and deployment. In addition to the safeguarding of training data (previously seen) and test data (previously unseen), this problem area includes the identification of synthetic data arising from generative modelling.

3.1 Data security guarantees

What data transformations and (differential) privacy guarantees can we make, such that task performance and computational complexity remain reasonable, but there is resilience to follow-on inference attacks that target the training data?

3.2 Statistical tests for synthetic data

Developing a range of rigorous, model-agnostic statistical tests for the presence of synthetic data, where there is little to no knowledge of the generative model type. Ideally a suite of tests can be produced for a mixture-of-experts decision to flag a sample of data as 'suspicious' (in the sense that it was produced by a model rather than nature).

3.3 Generative manipulation detection

Identification of generative manipulation or synthesis in regions of input data, particularly when the data are high-dimensional and evolve over time in a complex manner e.g. parts of video frames, segments of audio or passages of text.

3.4 Symmetries and security

Does the presence of symmetries in data imply security vulnerability? The detection of symmetries in data is an active research topic in mathematics, and symmetries often arise naturally in ML models, e.g. translation invariance in convolutional neural networks (CNNs). Can we use the new methods of symmetry identification to highlight potential data security issues?

3.5 Safeguarding data manipulation

Methods for prohibiting post hoc generative manipulation of data, e.g. mathematical "watermarking", or signing/certifying provenance.

3.6 Interaction effects and security

Understanding of the relationship between data x, the model parameters θ and structure f, and how security issues might arise from such analysis. For example, can we connect geometric structures in the parameters θ to knowledge and features in the data x?

3.7 Manifold hypothesis and security

The manifold hypothesis states that data x often lie in a low-dimensional space (manifold) inside the so-called ambient dimension. This implies that there is natural redundancy in many datasets, and we are interested in whether this affects data security directly and assumptions about data security in AI/ML systems. The tools of manifold learning and Riemannian optimisation could be used to formally explore this.

4. Learning Security

The methods used to train and optimise AI/ML models lie at the heart of their resulting behaviour, therefore the learning methods themselves are an important area for study in AI/ML security. Models can behave very differently depending on the training method used, and the resulting parameter choice(s). The parameters can be a target for attackers given their role in processing raw data x through the model f.

4.1 Loss surface information and security

Model loss surface characterisation and its role in ML security, e.g. what does knowledge of the loss surface reveal about model vulnerability? Can the loss surface be used to assess vulnerability for formal verification and assurance? Can the characteristics be used for detecting malicious activity, e.g. data poisoning.

4.2 Training data reconstruction

Large AI/ML models tend to memorise large sets of training data in pursuit of performance, effectively encoding private information in the more easily accessible θ. Training data reconstruction attacks exploit this by revealing elements of x given only θ. These attacks currently work on simpler model architectures, but progress is rapid. Can we defend against them effectively? What properties of model structure, data and learning can be designed to mitigate against these attacks?

4.3 Locking and guardrails

Building safeguards against downstream generation, so that models cannot be used for nefarious purposes. This is an area that presently commands a lot of attention, but we are interested in whether safeguards can be built into the learning algorithms themselves, rather than post-training.

4.4 Implicit regularisation and security

Research into implicit regularisation properties of deep neural networks is becoming increasingly popular, where the actual model complexity is naturally controlled within a highly complex setup, e.g. over-parameterisation. Does implicit regularisation also lead to security issues?

4.5 Topology of training

Model parameters can vary by intention, i.e. during training or fine-tuning, or by an attacker, e.g. manipulating weights in situ. Can we take advantage of recent mathematical approaches that topologically characterise trained models from untrained ones, and bona fide models from manipulated ones?

4.6 Scalable homomorphic encryption

Homomorphic encryption provides a gold standard for secure AI/ML learning, but it is currently not scalable to the levels of modern systems with high-dimensional θ, i.e. those used in deep learning. Can we make compromises for scaling? If so, how might these affect the security; both theoretically and practically.

5. AI and Cyber Security

These are problems associated with the cybersecurity, deployment, provenance, social and sociotechnical aspects of AI/ML.

5.1 Security of compressed and low resource ML

Modern ML models are often large, containing billions of parameters and requiring significant resources to train, store and run. A smaller but cheaper version of the model may provide adequate performance in many applications. A range of compression techniques can be applied, including quantisation, distillation and pruning. Adversarial attacks and defences can be brittle, and transferability across models can't be guaranteed. The vulnerability to attack of a compressed model may therefore be different to that of its parent model, and this effect may vary depending on the domain or type of attack.

ML in low-power, constrained resource environments, where "resource" refers to power, memory, compute, etc., rather than volumes of training data (as in "low resource language"). We need to understand the current and future applications for ML on edge/embedded devices such as phones, Internet of Things (IoT) devices and Industrial Control Systems (ICS) sensors due to their potential use in systems critical to the UK. By understanding how compressed models are secured, we can understand any vulnerabilities that may arise, and understand how best to tailor existing guidance to edge/embedded applications.

Questions that we would like to explore further include:

- We have mixed empirical evidence as to whether quantised models are inherently more or less robust to direct attack than unquantised models. Can a relationship be derived formally, and under what conditions?
- What compression characteristics make a difference to fundamental robustness of a model?
- What are the characteristics of an attack that make it easier or hard to transfer between standard and compressed models?
- How can we protect against model reverse engineering and tampering, particular in edge and/or resource-constrained environments where encryption may be challenging?

5.2 AI model monitoring

NCSC Guidelines [6] refer to logging and monitoring in several places. For example, they state that developers should "measure the outputs and performance of [a] model and system such that [they] can observe sudden and gradual changes in behaviour affecting security" and "monitor and log inputs to [a] system [...]to enable compliance obligations, audit, investigation and remediation in the case of compromise or misuse."

We have undertaken some work to understand how to capture and analyse model inputs and outputs to be able to detect adversarial attacks on ML systems. This work has focussed purely on inputs and outputs, treating the model itself as a closed box. We would like to know what additional information can and should be captured, represented and analysed when you have full access to the model, in order to detect adversarial attack. This could include neuron activations in a neural network context.

5.3 AI model provenance

Suppose we have a model that we initially don't trust, for example because we can't validate its supply chain; can we verify where that model comes from, how it has been trained and modified, and information about its supply chain? Questions include:

- Are there patterns of changes in model files that indicate particular types of difference between model versions, for example fine-tuning vs backdooring via modifying weights?
- To what extent can we validate that a model has been trained on the dataset it claims to have been trained on?
- (How) can we build or validate the "family tree" of a model?
- How might we go about detecting and mitigating backdoors caused by manipulation of any part of the MLOps pipeline, including training data, fine-tuning/Reinforcement Learning from Human Feedback (RLHF), trained model weights, serialisation and compilation [1]?

5.4 Socio-technical aspects of AI usage and how that overlaps with AI security

All real-world AI deployments currently (and for the foreseeable future) exist in human-centric end-to-end systems. As identified in a recent Turing whitepaper [8], there are a number of significant high-level gaps in the research landscape at the intersection of AI and Cybersecurity.

A more recent academic study [9] also identified AI Security research gaps from a technical perspective, but still with a socio-technical lens. We would like to better understand how each of these gaps can be addressed in the context of securing real-world end-to-end AI systems, and how we can eventually deliver impact from our research through standards [2] and possible regulation.

5.5 The relationship between EDI and AI security

Tackling the challenge of AI Security requires both the strongest possible talent pool and mitigating any possible harms from inequality of access to security. Some of the problem areas that must be addressed include:

- Understanding the trade-offs between security mitigations and bias, including how to implement EDI-aware auditing of AI systems.
- Recruiting a diverse talent pool for AI security research, — meta-research studies will be required on making sure a diverse range of voices are represented.
- Other sociotechnical considerations, including:
    - Intersection of diversity characteristics and trust in AI systems
    - Equality of access to benefits of AI security

5.6 Research into specific AI vulnerabilities and mitigations

We would like to better understand and map the adversarial landscape of possible vulnerabilities in the ML supply chain, and in the development and deployment pipelines [1]. We would also like to dig deep into researching specific vulnerabilities and optimal mitigations for them.

5.7 Trade-offs between security and performance, fairness, bias and other factors

Implementing security mitigations in practice usually requires trade-offs between performance, security, fairness, bias and other aspects of robust and reliable ML. We need to have at least some idea of what these trade-offs are in order to understand the risks associated with implementing a specific mitigation.

In a recent draft paper [5], the factors considered were: Task Performance, Robustness, Engineering Overhead, Training Overhead, Inference Overhead, Interpretability and Explainability, Fairness and Bias, Longevity and Reuse. All research into security mitigations should aim to take into account as many of these factors as possible. However, our research found that reporting in the academic literature was ad hoc and patchy at best.

Some work has been done in ML privacy to develop frameworks to assess and quantify the effect of mitigations such as Differential Privacy (DP); however, this focusses only on certain aspects of a specific domain (performance vs privacy when implementing DP). We would like to extend this to other aspects of ML security, eventually deriving a (set of) framework(s) that allow the rigorous, practical assessment of all the relevant risks in a specific use case.

5.8 Research into mitigating the cybersecurity threat from AI

The potential for both autonomous attacker agents, and for deep and novel AI capabilities across the full range of Mitre Attack [4] Tactics, Techniques and Procedures (TTP) raises the potential for significant cybersecurity threat from AI in the near to medium-term future (as well as to AI models themselves [3]). However, as much of this work is speculative and could be sensitive, technical research in this space will not be our primary focus for the first year. However, we do want to prepare for a range of scenarios, perhaps through strategic threat modelling exercises. This may mean bringing in lessons from many previous cyber incidents over the years, drawing on other fields (such as epidemiology) for inspiration, and leveraging new methods such as game-theoretic analysis. We also need to be able to effectively share knowledge on AI-driven ‘cyber incidents’ and track their prevalence.

6. AI model Cybersecurity Evaluations and Safeguards Analysis

Below we describe several research directions that are currently of interest. The focus is the security threat to and the threat from AI systems, respectively.

6.1 Effectiveness/robustness of Frontier AI cybersecurity evaluations

Modern ML models are often large. Evaluations are taking place of Frontier AI models, both for understanding their potential for cyber misuse (giving attackers additional or scaled-up capabilities), and for understanding the vulnerability of existing model safeguards (the security protections implemented in and around models to try to prevent them outputting dangerous content).

To make this more concrete – imagine we have a model, to which we have applied a set of mitigations and on which we have performed some set of (potentially expensive) evaluations – what guarantees do we have that the results of our evaluation will hold given a range of types of modification to the model? Such modifications could include fine tuning, quantisation, and distillation. How do we help make evaluations more robust in the face of ever-changing frontier models with many variants?

6.2 Attempting to quantify the coverage of cybersecurity evaluations

Current evaluation playbooks rely on a set of red-teaming activities established from human best practice. However, particularly for AI systems, we don’t have a good sense of how comprehensive these evaluations are likely to be, nor of general principles to ensure maximal coverage of likely issues.

Having performed a set of tests/evaluations, to what extent can we make conclusions about the general behaviour and likely vulnerability of a model? How do we develop the optimum evaluation approach for a given model and task, given a model that we cannot formally verify/validate?

6.3 Cyber offence/defence balance

As cyber capabilities of models increase these capabilities could be used for both offensive and defensive purposes. To understand the potential impact of these capability increases, we are interested in understanding the cyber offense/defence balance better to:

- Identify and research which cyber capabilities of models might tip the balance significantly in favour of offense or defence.
- Research on the time lag or adoption rate of new advances in cyber from a defensive perspective, across various industries.
- Research on what kinds of evaluations might be effective to assess if there is a significant shift in the offense/defence balance, with a focus on evaluations of frontier AI models that might indicate when this might be shifted in favour of the offenders.

6.4 How to define “red line” and “yellow line” capability threshold evaluations

Even once a cyber risk or threat may have been identified as worth evaluating, identifying the specific capabilities and methods of evaluating these risks is hard. Questions we are interested in exploring further are:

- How to define a capability/capabilities threshold of a frontier AI systems that might indicate a significant increased risk level for a cyber risk, a “red line” or “trigger” capability.
- How to build an evaluation or series of evaluations that might assess this risk. Are we interested in comprehensiveness in assessing a range of capabilities? How much do we care about the balance between false negatives and false positives?
- How to define a “yellow line” or “safety margin” threshold based on a red line capability threshold and evaluation.
  - What should inform how we define “yellow lines”? E.g. lead time in implementing mitigations, buffer to account for possible error bars in evaluation results or mitigating against a shortfall in capability elicitation.
  - Does how we think about false negatives and false positives differ for these evaluations?

6.5 Investigate whether there are any narrow AI cyber tools that could pose significant harm

Currently, a large proportion of resource building evaluations in industry are from frontier AI labs. As such there is a focus on building cyber evaluations on state-of-the-art models produced by those labs, evaluations focused on general purpose LLM or multi modal systems. We would like to explore further whether there are narrow AI systems currently available or likely to be available in the future that might pose a significant risk (>$100bn per year in harms).

6.6 Investigate the impact of fine-tuning open-source models on the performance of cyber skills

Opensource AI models are likely to pose different risks to closed source models given that they can be more easily finetuned and have their safeguards bypassed to be used for malicious purposes. We are interested in understanding whether the facility to finetune and easily bypass safeguards of opensource models, can result in significant increases in performance of cyber capabilities. And if so, in which capabilities we might see this performance increase and using what kinds of datasets to improve cyber capabilities.

6.7 Survey on safeguards

We are interested in developing a better understanding of the range of measures that frontier model developers can take to limit the risks from model vulnerabilities. We've benefitted from prior taxonomising work (for example, NIST's adversarial ML taxonomy [7] and MITRE ATLAS [3]), and we think developing a shared understanding of safeguards and their limitations will make it easier to analyse and improve security going forward.

We are especially excited about work relevant to near-future LLM-based systems, and work that focuses on ML-related vulnerabilities as opposed to traditional security issues. Safeguards may include techniques to:

- improve robustness to adversarial prompts,
- prevent models from possessing certain explicitly harmful capabilities,
- monitor system abuse and take action against malicious users,
- and reduce hallucinations and other unintended behaviour.

Alongside discussion of existing safeguards, a safeguards literature review could discuss the known limitations of such safeguards.

6.8 Safeguard deep dives

We would like to better understand the limitations of certain safeguards, including through both empirical demonstrations of failures and conceptual limitations. Our projects in this area focus on a particular safeguard, investigating both conceptual limitations of safeguard classes, and/or limitations related to specific deployments of model safeguards (e.g. use of low-quality data or poorly designed safeguard instantiations). For example, we would be excited about projects to better understand:

- the theoretical and empirical limitations of using LLM-based classifiers to filter user input and model output, for example by setting up realistic filtering schemes and exploring failures
- the cost and efficacy of different forms of pre-training data filtering, for example by applying a range of techniques to remove sensitive information from a sample dataset and exploring their effectiveness
- the considerations and efficacy of abuse monitoring solutions, which can use multi-interaction user data to flag malicious users—for example, exploring using withheld classifiers with different failure modes as compared to classifiers used to filter user input or model output
- the techniques and feasibility of removing poisoned data samples prior to training, or removing backdoors or other poisoning artifacts after model training
- the types of identity verification when using APIs or chat interfaces, and the difficulty in evading such schemes to create new accounts after banning
- the techniques and limitations of patching vulnerabilities and broader adversarial training

6.9 Understanding future vulnerabilities

Though current vulnerabilities are concerning, we are especially concerned with vulnerabilities persisting in future systems. Accordingly, we are interested in collecting evidence as to whether particular vulnerabilities will persist in future model generations, with more capable, larger models trained on more data.

In addition to exploring the vulnerability of future models to current attacks, we are also interested in understanding attacks that may become more desirable if attackers are willing to invest more resources into implementing attacks, for example by using small-scale experiments to explore efficacy of higher-resource attacks. Example projects include:

- Exploring the scaling behaviour of a particular attack or class of attacks. For example, the Gemini Team found more capable Gemini models are more robust to random search and GCG attacks, but less robust to semantic attacks; Mazeika et al. (2024) found that bigger models do not appear to become more robust to attacks like GCG-T and AutoDAN; and Anil et al. (2024) found bigger models were more vulnerable to many-shot jailbreaks.
- Estimating the efficacy and costs of data poisoning attacks against future models, for example by collecting evidence as to the quantity of attacker-controlled data necessary for various attacker objectives.
- Exploring jailbreak attacks which exploit possible capabilities of future models, for example by exploiting models which are more capable than human-in-the-loop monitoring systems.

6.10 Safeguard evaluations

In addition to better understanding safeguards and future vulnerabilities, we are also interested in performing and reporting evaluations on deployed or soon-to-be-deployed systems. We are interested in iterative improvements to our evaluation methodology, for example:

- Creating better datasets to measure efficacy of jailbreaks, like harmful tasks which require multi-turn interactions or longer-horizon agentic behaviour.
- Exploring better proxies for measuring the correctness of sensitive harmful information.
- Developing more detailed risk models of attacker patterns of concern. Beyond such iterative improvements, we are interested in:
    - Improved red teaming methodology, such as attempts to advantage red teamers by providing them with extra information or access as compared to real-world threat models.
    - Methods to affirmatively argue for model safety which do not rely on failure of a resource- and time-bound red team.
    - Evaluating safeguards beyond refusal behaviour, for example methods to adversarially evaluate unlearning, pretraining filtering, harmful content classifiers, identity verification, or abuse monitoring

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 can agree to the Terms and Conditions and that you include as part of the application an approval of submission letter from your research/contracts/finance office stating this. An example of a letter is available on request.
- We expect applicants to seek the relevant ethics approvals from their institute prior to submitting their application.

How to apply

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. Your approval of submission letter should also confirm that your research/finance office have reviewed that the costs provided are correct.

The submission approval letter must confirm that:

- if not already covering the entire period of the project, then the contract of employment for project researchers will be amended and/or extended as necessary to enable the successful completion of the project.
- that the researcher is already employed by the university and no recruitment is needed to fulfil this project
- the project will be given full access to the facilities, equipment and personnel as required by the application.
- the costs included in the application have been correctly calculated with the support of the Research Office / Finance Department (or equivalent).
- the terms and conditions of the agreement have been reviewed by the Research Office / Legal Department (or equivalent)
- the letter signatory is authorised to approve the submission of applications for funding and the application has met all internal approval procedures

The Principal Investigator must ensure the same is received for all collaborators / universities on multi-party applications.

We must receive your complete application by **1500 on 02 September 2024**.

What should be in the proposal?

The proposal should:

- Be in OpenDocument, MS Word or PDF
- Describe the scope and technical approach of the proposed work –
    - This should be a narrative description of the principals/solutions the project would aim to achieve and how those solutions may relate to the numbered topics (2 – 6) in the requirement section.
- Describe how the approach would lead to the desired results (and any real-world impact).
- Include a description of how the task is decomposed, thematically i.e. by work package, include optional extension at the end.
    - For each work package, what activities will be undertaken to produce the results.
- Answer questions such as:
    - What is innovative about the proposed research?
    - Why are you uniquely placed to undertake this task?
    - What is the expected scientific impact?
    - How do you intend to demonstrate this impact?
- Identify any key risks and mitigations.
- Reference related work or/and relevant experience.
- Include FEC cost.

If you are employed by a University that is a member of the Turing University Network, please contact your Turing Liaison (a list of Turing Liaisons 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.

Deliverables

Deliverables will be confirmed prior to contracts being signed. We anticipate projects will generally achieve the following deliverables (based on 6-month project). Please include additional or alternative deliverables in your proposal for consideration.

Ref.	Deliverable	Due by Start Date = T0	Format	Description
D1	Project Kick off Meeting	T0+1M	Meeting	Meeting to initiate the project
D2	Technical update meetings	Monthly	Meeting	technical update meetings
D3	Mid-project update report	T0+2M,	Technical Report	An update on technical progress against the outputs identified in the proposal/contract.
D4	Final report	T0+6M	Technical report	Final technical report describing the methodology and technologies used, along with a discussion of recommendations for next steps.
D5	Any source code / tools / proof-of-concept tool	T0+6M	Software	Any source code including code for initial proof-of-concept tools, along with documentation, and instructions for install and use
D6	End-of-phase presentation / workshop	T0+6M	Presentation + demonstration	Technical presentation / workshop summarising key findings from the report with key stakeholders
D7	Project Closure Meeting	T0+6M	Meeting	Meeting to close the project or agree follow on steps

Assessment and review

Following eligibility checks, proposals will be reviewed by an assessment panel who will rank the proposals based on score.

The assessment panel will consider the following 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. Consider whether the proposed research is likely to deliver real-world and/or research impact
- Justification of resources: This will consider whether the proposal is appropriately resourced and suitable expenditure has been included in the budget.
- Knowledge and expertise: Candidates should demonstrate they fulfil the bullet points in the requirement section.

Each of the criteria will be scored and while all 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*	Monday 02 September 2024
Announcement of Results	Monday 16 September 2024
Earliest Project Start Date by**	Tuesday 01 October 2024
Six-month research project starts by**	Tuesday 01 October 2024
Five-month research project starts by**	Monday 04 November 2024
Four-month research project starts by**	Monday 02 December 2024
Three-month research project starts by**	Tuesday 07 January 2025
Research Completed and deliverables submitted for all projects	Monday 31 March 2025

*Proposals must be submitted via Flexigrant by **15:00 Monday 02 September 2024**.

**Any project agreements not signed 10 working days prior to deadline start date above 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. Applicants will also be expected to join regular project progress meetings. These may take place online, at the Turing, at UK Government partner site, 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. Successful applicants may however be required to complete a Personal Particulars - Research Workers form for a security screening in accordance with UK Governments baseline personnel security standard (see https://www.gov.uk/government/publications/government-baseline-personnel-security-standard).

Publications

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

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, The Alan Turing Institute, Programme Manager dsprogramme@turing.ac.uk

References:

[1] E. Clifford, I. Shumailov, Y. Zhao, R. Anderson, and R. Mullins. ImpNet: Imperceptible and blackbox-undetectable backdoors in compiled neural networks. In 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), pages 344–357. IEEE, 2024.

[2] Centre for Emerging Technology and Security. (CETaS). Towards Secure AI. https://cetas.turing.ac.uk/publications/towards-secure-ai.

[3] MITRE. MITRE Adversarial Threat Landscape for AI Systems (Atlas). https://atlas.mitre.org/.

[4] MITRE. MITRE Attack Enterprise Matrix. https://attack.mitre.org/matrices/ enterprise/.

[5] NCSC. At What Cost? Towards More Practical Reporting of Machine Learning Security Mitigations. 2024.

[6] NCSC and CISA. NCSC Guidelines for Secure AI System Development (co-developed with CISA). https://www.ncsc.gov.uk/collection/ guidelines-secure-ai-system-development.

[7] Oprea and A. Vassilev. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, National Institute of Standards and Technology, 2023.

[8] K. M. Roba Abbas, J. Pitt, K. M. Vogel, and M. Zafeirakopoulos. Artificial Intelligence (AI) in Cybersecurity: a socio-technical research roadmap. 2022.

[9] S. Soare. Security of AI: A Literature Review. 2024 - ‘Available on request’