2025

Banc Ceannais na hÉireann
Central Bank of Ireland
Eurosystem

Insight
RESEARCH IRELAND CENTRE FOR DATA ANALYTICS

Call Note: Central Bank PhD Programme in Artificial Intelligence & Data Science

2025 Call

Deadline: 11th April 2025

Introduction

As part of its efforts to broaden the Central Bank's research agenda, the Bank is launching a **Call for Proposals** to address key research topics for 2025. The PhD programme is intentionally designed to be broad in scope, encompassing various aspects of AI and Data Science applied to areas of strategic interest to the Bank. The programme also aims to foster wider public interest in the transformative potential of AI and Data Science. The purpose of this Call Note is to outline the broad topics of interest and, within those, some of the specific areas for PhD projects in 2025.

This programme seeks to explore fundamental research alongside its application to critical public policy issues, such as understanding algorithms, their reasoning, and decision justification. A transdisciplinary perspective can provide valuable insights into these complex challenges. The proposed PhD supervision team could include at least one Insight FI/PI and may also incorporate co-supervisors from outside the Insight investigator network.

With the AI Act coming into force in 2025, there is an increasing focus on AI supervision and the fundamental questions that will shape policy in the years ahead. This is particularly important for the Bank, as the use of AI in high-stakes settings—such as financial services—could reshape the risk landscape for consumers and impact fundamental rights. AI-related risks can be assessed through various lenses, including interpretability, fairness, ethical data use, and the broader context of AI deployment, as well as the recourse available to users. Under the EU AI Act, high-level requirements are set for explainability, bias mitigation, and accountability. To effectively supervise AI risks, the Bank aims to foster research within the Irish research community, enhancing understanding of these fundamental AI challenges and their practical applications.

Important Dates

Some important dates for the 2025 call are as follows

	Call Item	Date
1	Programme Briefing to prospective supervisors	Friday March 7th 2025
2	Full Call Document available and on websites	Friday March 21st 2025
3	Submission deadline for applications	Friday April 11th 2025
4	Decision on successful applications	Wednesday April 30th 2025

Proposed Topic Areas for 2025

1. Interpretable machine learning models for high stakes settings

In the context of the types of data typically used in financial services and of relevance to the Bank (credit, underwriting, fraud/AML), current machine learning practice for modelling a particular problem is based on finding a single good model given a dataset. Typically, this model used is a black-box model – a model too complex for people to understand directly. Their use is driven by perceived better performance and incentives/information asymmetry. By their nature, black box models' mistakes are hidden. Interpretable machine learning models are – in some sense - understandable and therefore make mistakes or fail in understandable ways. In high stakes settings (i.e., where the consequences of the decision have profound impacts on people) like financial services as well as other essential services like access to social services, medical care, interpretability is very closely related to the justification for a decision.

Recent research has shown that interpretable models are as good as black box models regarding their performance (accuracy) on tabular-type data. This research has focused on developing methodologies to identify from the set of all possible models, how to find many approximately equally good interpretable models of particular types can exist given a dataset and specific types of models.[1]

Some extensions to this approach and its applications to high stakes uses including in financial services that could be explored as PhD topics are characterising and measuring the set or space of many good models, extending this to other types of models/algorithms and loss functions. This could also include characterising important variables for prediction across this space of models, starting to explore how unbiased or fair models can be found in this many models paradigm, and potentially extending this to intersectional fairness/unbiasedness where more than one protected characteristic is of interest.

There are some other important open questions that could also be explored as part of this topic area. First, how interpretable models can be edited to include important domain knowledge or constraints and how that affects the set of good models – as constraints or edits may be applied across many models in these sets. An interesting application using sparse Generalised Additive Models is presented here. Second, how to present and choose among these potentially large numbers of 'good' models once they have been identified remains an open question, requiring consideration and development of novel approaches based on context, users visualisation or methods. These new algorithms/approaches can be applied to tabular (i.e., spreadsheet like) and other data typically used in financial services like insurance risk assessment and pricing, as well and other domains where there are high stakes in terms of decisions.

Many of these approaches focus on particular types of learners or hypothesis space. A third approach could focus on both extending the type of learners and finding ways to

[1] As outlined by Rudin et al. (2024), real-world datasets can give rise to many approximately-equally-good models. Leo Breiman called this phenomenon the Rashomon Effect (Breiman, 2001)

search the space of good models. In terms of extending learners, optimal/sparse approaches with classes of learners like Optimal Decision Trees, using various approaches like satisfiability (SAT) formulations or dynamic programming with branch and bound pruning. Searching over rule-based classifiers could be extended, including methods that do not rely on enumeration. This could also potentially involve combing statistical and machine learning approaches to have flexible interpretable models like Explainable Boosted Machines. Neural Additive Models and extensions to distributional regression (NAMLSS) where this flexibility is needed in a domain specific context. Different model types may depend on different variables to represent a given dataset. In turn, this may matter to how predictions are interpreted and explained to users. Therefore, leveraging this many model approach to produce stable variable importance measures may be conducive to coherent interpretability.

2. Explanation of AI

Interpretability of an AI system is the ability for a person to understand how and why a model performed the way it did in a specific context. This includes the ability to understand the rationale behind its decision or behaviour. Interpretable models are useful compared to black boxes as they are able to be de-bugged and their explanations verified or refuted.

Explanations are the degree to which a system, and a set of governance practices and tools support a person's ability to understand the rationale underlying the behaviour of the system. An explainable model is a predictive model where some methods are applied to provide post-hoc explanations. The explanation is based on querying the 'black box' to provide an account of what the algorithm may have done. Typically, this is done through looking at the inputs and outputs and/or using another model to 'explain' the black box. In this case, they are approximations not explanations. Approximations of explanations methods for black boxes can be problematic, misleading and potentially contradictory. Interpretability and explainability are different concepts applying to differing sets of models.

However, until recently creating interpretable models for tabular data could sometimes be much more difficult than creating black box models for a variety of reasons related to computational complexity and domain specific nature of interpretability. Simpler models are thought to be generally more interpretable: it is straightforward to understand why and how the AI arrives at its decisions. Creating simpler interpretable models is such one approach (explored under Topic 1).

However, there are other approaches. These include developing representations translating low-level representations into ones that are interpretable by humans. How these representations can be constructed and how they could be used in high stakes contexts like financial services and other public interest settings remain open questions.

Another approach is to use sparse explanations. How sparsity of explanations can be implemented in interpretable and other types of models depend on the type of modelling approach used. Interesting open questions are how to control sparsity in certain types or models or algorithms, develop other approaches like case-based reasoning, disentangling neural networks, and distillation.

When warranted - such as for complex data like images, text, or a combination of both- creating sparse explanations for complex types of models including black-box models is an emerging research topic. These types of sparse explanations may matter for explaining recommendation, pricing, or lending decisions, or investment advice in financial services as well as other high-stakes applications.

There is evidence that interpretable models are valued by the public based on a multi-country study Nussberger et al (2023) find this is when there are high stakes applications of AI or where it performs a gatekeeping role allocating scarce resources. While algorithmic transparency through interpretable and grounded explanations is necessary for responsible high stakes AI and it is valued, it does not provide any guarantees that it is understood or used by people who receive them. This last step is critical for transparency, accountability, and recourse.

The use of an explanation depends on the context for the explanation and who is giving/receiving the explanation, as well as the type of explanation they receive (see financial services and healthcare examples). Understanding both the use of AI in context and drawing on Human Computer Interaction (HCI) research behind how explanations are constructed, presented, and understood/received is underexplored in several domains including financial services. This is despite this being a key component of responsible use of AI.

For example, human factor research has shown that explanatory variable based methods for insurance investment product recommendations from algorithms did not perform well in a series of experiments, and essentially the same as no explanations at all in their acceptance by differing groups of people (customers, supervisors). Research on supervisors review of AI-AML/CTF models in France concluded that current explainable-AI techniques fall short of regulators' expectations to provide accurate and faithful information about AI system's inner workings. Therefore, several open questions could be explored in this area explain AI outputs in contexts relevant for financial services. These include both development of methods and their use or acceptance in practice for various groups of stakeholders. Overall, there are several aspects related to interpretability, explanations and real-world high stakes contexts that could be combined.

In the area of Large Language Models (LLMs), interpretability and explainability is in its infancy and the implications are only starting to be understood more clearly. LLMs exhibit capabilities across a wide array of tasks. This means a possibly broader notion of interpretable AI or at the least conceptual and empirical approaches for interpreting LLMs are needed.

On one hand, the ability to explain in natural language allows LLMs to expand the scale and complexity of explanation and the ability to provide an interactive and tailored explanation to the user and their context. On the other, the question of the fidelity or faithfulness of the explanation provided by the LLM (provably true versus the reason stated by the LLM), their inscrutability, and their currently enormous training computational costs could limit their use as part of an interpretable AI system.

This leads to some fundamental research questions: what constitutes an explanation when an LLM can give a response to almost any question and where this answer (output) is variable? Aside from this conceptual challenge, a key concern is faithfulness -

the extent to which an explanation accurately reflects a model's reasoning process, particularly when it is difficult for humans to comprehend.

There are some approaches emerging in the literature. A general approach involves framing this evaluation as a measurement challenge. This could draw on social sciences to handle the evaluation of abstract and sometimes unsettled concepts. Among the specific approaches that could be adopted is using a rubric-basis for explanations. An interesting research direction could be to use smaller LLMs or distilled large open source LLMs to help users customise the algorithms explanations to their needs, including designing rubrics customised to explanations. This too has challenges including developing strategies for eliciting the necessary feedback, how to explore the range of alternative explanations in the relevant socio-technical context and dealing with adversarial explanation contexts including how trustworthy is the answer.

Recent work has hinted at what may be useful in developing appropriate reliance on LLM explanations This may be particularly important when one black-box AI is used to explain another or evaluate the answers of another. One approach here is to score the trustworthiness of the answer. A shortcoming of this approach is it depends on evaluating inputs and outputs of a black box LLM model using another black box model.

A second approach could be concept-based – using Concept Bottleneck models applied to NLP tasks that overcome some of the limitations of previous approaches. This may allow human understanding of errors, adverse, or unacceptable outputs. Some avenues for exploration may be concept generation and its implications for the rest of this framework, how concepts are corrected, potentially unlearned, and their faithfulness. In addition to these approaches, there are others including neuro-symbolic based methods.

Faithfulness matters for users and recourse. The ability of LLMs and more complex systems comprised of several LLMs to provide plausible explanations (self-explanations) does not mean they are faithful. It has already been shown how a black-box explanation system developed to defend a black-box decision system can manipulate decision recipients into accepting an intentionally discriminatory decision model. Open research questions could be how to do this safely and detect malicious or manipulative explanation, and trustworthiness of LLMs or some variant of them intermediating the explanation in high stakes settings.

Application and Evaluation

The application and evaluation process consists of multiple stages:

- Call for Expressions of Interest (EOI): This document serves as the official Call for Expressions of Interest from potential PhD supervisors. The Call will be posted on the CBI, Insight, and other relevant websites.
- Frequency of the Call: The decision on the Call's frequency will be made jointly by the CBI and Insight, with an expected cycle of once per year.
- Submission Guidelines: Detailed submission instructions, including deadlines and the submission portal, will be provided at the time of the Call.
- Proposal Submission: Potential supervisors must submit a PhD research proposal relevant to one of the specified research areas.

- Review and Shortlisting: An Insight Review Board will assess, and shortlist proposals based on the evaluation criteria outlined below.
- CBI Review and Selection: Shortlisted proposals will be submitted to the CBI for review and funding recommendations.
- Notification and Recruitment: Successful supervisors will be notified and invited to recruit PhD scholars following the host university's PhD recruitment process, ensuring alignment with the Irish National Framework of Qualifications (Level 10)

Eligibility Criteria

To be eligible for funding under the CBI PhD Programme, supervisors must meet their institution's criteria for PhD supervision. The following key principles apply:

- Institutional Affiliation: Applicants must be affiliated with an eligible Irish research institution, such as universities, institutes of technology, or other research-performing organizations recognized by Research Ireland.
- Researcher Status: Applicants must be Principal Investigators, meaning they hold a position that allows them to supervise postgraduate students. This could be a permanent academic position or a contract extending beyond the proposed PhD duration. Early-career Principal Investigators are encouraged to apply as main or co-supervisors.
- Research Area: Proposed research must align with one of the Call's specified topics.
- Ethical and Legal Compliance: Applicants must ensure their research adheres to ethical and legal requirements, including necessary approvals for studies involving personal data or human participants.
- Residency Requirement: While applicants do not need to be Irish citizens, they must reside in the Republic of Ireland to supervise the student.
- Co-Supervision: Each proposal must include at least one approved Insight Investigator as a supervisor or part of the supervision team. Cross-institutional and interdisciplinary co-supervision is encouraged to provide broader perspectives on these complex challenges.

How to apply

Before submitting a proposal, all applicants are encouraged to read this **Call Document** in its entirety.

All proposals must be submitted via the online application form, available **here** [https://forms.gle/uCsbzrJq1WRtf8JS9]. This submission link is also accessible from the news section of the **Insight website** [https://www.insight-centre.org/news/] and through the Central Bank's website.

Application Form Structure

1. Research Topic:
   - Specify which of the Central Bank's research areas this proposal addresses.

2. Lay Abstract*(Max: 300 words)*:
   - Provide a summary of the proposal suitable for a non-specialist audience.

3. Research Proposal *(Max: 500 words)*
   - Aims, objectives, and central research questions.
   - How existing literature has informed the proposal.
   - Contribution to advancing the state-of-the-art in the field.

4. Research Design & Methodology *(Max: 500 words)*
   - Outline the methodologies and approach to be employed.

5. Project Timeline & Risk Management *(Max: 500 words)*
   - Key milestones and deliverables.
   - Potential risks and mitigation strategies.

6. Dissemination & Impact Strategy: *(Max: 500 words)*
   - Plans for publications, conferences, Education & Public Engagement (EPE), and knowledge exchange.
   - Strategies for measuring the research's impact.

Only fully completed applications[2] received prior to the application deadline will be considered for evaluation. The evaluation consists of a three-stage process:

Evaluation Process

The evaluation follows a three-stage process:

1. Eligibility Assessment: Administrative compliance check.
2. Technical Evaluation: Assessment based on the evaluation criteria below.
3. Final Decision & Notification: Selection of top-ranked applications for funding

Evaluation Rubric

Applications will be assessed based on the following criteria, with the allocated weighting for each section. This list serves as a guideline rather than a strict framework, outlining key aspects assessors will consider.

Excellence and Innovation (35%)

- Novelty and ambition in relation to the state-of-the-art in Ireland and beyond.
- Demonstrated understanding of the relevant research landscape.
- Validity and reliability of the proposed concept and approach, including interdisciplinary elements.
- Research track record of the supervisor.

[2] Fully complete means containing all the relevant information to enable eligibility and technical assessment and as described herein.

- Eligibility of the supervisor and their institution, ensuring compliance with the programme's requirements. Early-stage supervisors are encouraged to apply as lead or co-supervisors

Relevance and Impact (35%)

- Alignment with the needs of the CBI and broader research community.
- Relevance to national and EU policies (e.g., National AI Strategy, EU AI Act).
- Contribution to responsible, ethical, and trustworthy AI regulation.
- Potential for influencing policymaking and industry practices.
- Strategies for effective dissemination and stakeholder engagement.
- Ethical data management practices.
- Evidence of value-added from trans- trans-disciplinary collaboration.
- Gender balance among co-applicants where applicable

Quality and Efficiency of Implementation (30%)

- Coherence and effectiveness of the research work plan.
- Quality of the research framework, including clear deliverables, milestones, and a credible breakdown of activities.
- Robust research management and risk mitigation strategies, including scheduling, dependency identification, and monitoring.
- Strength of supervisory management and oversight, including risk management strategies.

Conclusions

This call note has outlined the topic areas, eligibility and how to apply, and evaluation process for the 2025 Insight – CBI PhD programme.

All proposals must be submitted via the online application form, available **here** [https://forms.gle/uCsbzrJq1WRtf8JS9]. This submission link is also accessible from the news section of the **Insight website** [https://www.insight-centre.org/news/] and through the Central Bank's website.

Please note the deadline for proposal submission is **Friday April 11th 2025**.

Any queries should be addressed to the programme email: cbifellowship@insight-centre.org