# Research Proposal

Remi Desmartin

May 11, 2022

## 1 Background

The increased complexity of critical software systems has made traditional testing tools insufficient for detecting edge failure cases. One solution, which has gained momentum in the recent years, is the use of formal methods, that use rigorous mathematical tools to model the system, formulate specifications and verify that they are met [**?**].

The introduction of artificial intelligence (AI) components, often relying on machine learning (ML) algorithms is a contributing factor in the complexification of safety-critical systems. Neural networks (NN), a group of ML algorithm, have shown exceptional performance at dealing with noisy data and have been deployed in safety-critical applications like autonomous cars [] or automated trading agents [**?**]. Because ML components' failure can lead to the of the whole system's failure [**?**], formally verifying the system necessarily involves verifying the ML component.

Imandra [**?**] is a verification tool that has been successfully applied to verifying complex systems in the FinTech domain [**?**], like trading algorithms and blockchain smart contract infrastructure. It combines features of both functional languages and interactive and automated theorem provers. Imandra's logic is based on a pure, higher-order subset of OCaml, and functions written in Imandra are at the same time valid OCaml code that can be executed. Imandra's mode of interactive proof development is based on a typed, higher-order lifting of the *Boyer-Moore waterfall* [**?**] for automated induction, tightly integrated with novel techniques for SMT modulo recursive functions. In short, Imandra is both a functional programming language in which programs can be implemented and executed, and a reasoning engine which can apply formal verification techniques to these programs.

Verifying ML components (especially NNs) has been the focus of recent research, essentially divided into SMT-based tools [**?**, **?**] and those based on abstract interpretation [**?**, **?**], for instance studying the propagation of geometric shapes between the layers. The biggest shortcoming of NNs is their sensitivity to small variations in the input [**?**]. To counter that, a large part of verification efforts focus on proving that NNs classify consistently examples that are close within the input space. This property is called robustness. Multiple formal

definitions are used [**?**], and even though it is the primary focus of NN verification its scope is limited. Other properties like fairness or domain-specific property are also desirable and investigated.

## 2 Research Question

Currently, verification of ML algorithms such as neural networks requires dedicated tools. These tools are often limited to narrow use cases: a single family of NNs (e.g. MLP with Relu activation, RNN) or a single type of properties (e.g. a single definition of robustness). In addition, the integration into more generic and verification frameworks that are able to certify the larger systems containing ML.

There is a need for a flexible and generic NN verification framework that allows building systems with ML components with strong safety and security guarantees.

In particular, we will investigate the following aspects:

1. What is the role of ML-components in complex systems ? (e.g. simple inputs to controllers, or controllers themselves?);
2. How are NN components verified ?
3. the integration of ML components' verification in larger verification frameworks

More questions will arise from the exploration of these initial questions.

## 3 Aims

Our aim is to contribute to the available verification tools for NNs in order to make it easier to deploy reliable and trustworthy software. To that end, our objectives are the following:

1. Investigating the questions above
2. integrate third-party NN verifiers into a generic verification tool like Imandra
3. to produce a fully functional ML library in Imandra, which will include domain-specific proof heuristics.

## 4 Methodology

First, the initial investigation will help us define our library's structure, such as the data types used and exposed API. In order to evaluate our library, it will be evaluated against standard NN Verification benchmarks such as the ones proposed by the NNVC [].

# 5 Research Environment and Supervision

1. Imandra's intellectual and infrastructural support (financial ofcourse too! :-)); 2. LAIV's environemnt on AI verification: seminars, schools etc 3. The Edinburgh Center for Robotics (NB: Imandra pays extra for your training within the ECR): engagement with broader range of AI academics and industries, extra training

Imandra will provide infrastructural support for this research, providing funding and guidance for the duration of the project. Prof. Ekaterina Komendentskaya, Dr. Grant Passmore and ?? will co-supervise this research. Passmore is the creator of Imandra and the co-founder and co-CEO of the company that comercialises it.

# 6 Feasibility

The projective timeline is that roughly one year is dedicated to each objective of the project. I started learning about automated reasoning and formal verification of NNs during my Master's dissertation. This dissertation allowed me to consolidate my knowledge in functional programming, a central part of the Imandra theorem prover, and allowed me to get acquainted with automated reasoning and verification.

As Imandra Inc is a partner of the Lab for AI Verification (and a sponsor for this research), we will have support in terms of infrastructure and technical support. An earlier collaboration for the Master's dissertation was fruitful.

# 7 Potential Impact

Doing this research within the Lab for AI Verification would allow to contribute to the discussion and research on NN Verification, in particular by presenting early results to seminars organised by the lab and to specialised conferences like FoMlAs.

The partnership with Imandra Inc. means that successful results would impact directly their users by allowing them to formally verify their ML-based systems.