# Paper Summary: Towards A Rigourus Science of Intepretable Machine Learning

Daniel Braithwaite

September 15, 2017

The growing number of situations in which Machine Learning (ML) systems are used has lead to an increasing interest in systems which not only achieve a high accuracy but also provide explanations of there output. This situation leads to a difficult questions such as, is the system safe? Is the system non-discriminatory? Can the system provide an explanation? Quantifying the answer to these questions is often not possible. If a ML system can provide an explanation of its answers then this can be used to verify whether a system is safe or non-discriminatory. If the goal is to create intepretable ML systems what does this mean? There is little to no agreement on what intepretability is and now it can be measured.

The need for interpretable ML systems is real, in 2018 a European Union regulation will come into effect which requires any algorithm that makes decisions based on user level predictors which "significantly effect" users must provide explanation.

In the context of ML systems intepretability is defined as "Ability to explain or to present in understandable terms to a human". There exists specific information which one might want to extract from an ML system, such as the idea of fairness, the system should not be biased against some protected group.

Intepretability in ML systems is not always necessary, prehapse the consequence of an incorrect result does not have any significant effect.