Contents

Introduction 1

1 Preliminaries on Interpretablity and Complexity 3

1.1 Cat Recognition, a bottom-up example to get started . . . 3

1.2 But what is interpretability after all? A brief literature discussion . . . 6

1.2.1 Transparency . . . 7

1.2.2 Explanations or Post-hoc interpretability . . . 7

1.2.3 Interpretability vs. Explainability . . . 8

1.2.4 Desiderata . . . 8

1.3 Background in Complexity Theory . . . 9

2 A framework to measure and compare model interpretability 13

2.1 Models and instances . . . 13

2.2 Explainability Queries . . . 14

2.3 Interpretability in terms of complexity . . . 16

3 The computational complexity of interpreting different models 18

3.1 Specific models . . . 18

3.2 Main interpretability theorems . . . 19

3.3 The complexity of interpretability queries . . . 21

3.3.1 The complexity of MinimumChangeRequired . . . 22

3.3.2 The complexity of CheckSufficientReason . . . 27

3.3.3 The complexity of MinimalSufficientReason . . . 28

3.3.4 The complexity of MinimumSufficientReason . . . 30

3.3.5 The complexity of CountCompletions . . . 36

4 The parameterized complexity of interpreting Neural Networks: deeper is harder than shallow 41

4.1 Required background in parameterized complexity . . . 42

4.2 Statement of a parameterized result and sketch of proof . . . 43

4.2.1 The graph interpretation of (r)MLPs . . . 44

4.3 Constructions and transformations . . . 44

4.4 A proof for Theorem 4.1 . . . 48

4.5 Application of the parameterized result . . . 56

5 Applications to Fairness and Bias Detection 57

5.1 About the notion of (un)fairness used . . . 57

5.2 Definitions of bias . . . 58
5.3 On the complexity of bias detection . . . 59

6 Discussion **67**
A About our results . . . 67
B Related work . . . 68
C Open problems and future research directions . . . 69

Bibliography **70**