Chirag Arora

FinalThesisBind-17-101

Research Paper

Document Details

Submission ID

trn:oid:::27535:146366995

Submission Date

Jul 24, 2026, 1:24 PM GMT+5:30

Download Date

Jul 24, 2026, 1:32 PM GMT+5:30

File Name

FinalThesisBind-17-101.pdf

File Size

1.6 MB

85 Pages

37,077 Words

221,747 Characters

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Match Groups

289 Not Cited or Quoted 7%
Matches with neither in-text citation nor quotation marks

26 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
6% Publications
5% Submitted works (Student Papers)

Match Groups

- **289** Not Cited or Quoted 7%
  Matches with neither in-text citation nor quotation marks
- **26** Missing Quotations 1%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
6% Publications
5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Student papers	University of Nicosia on 2024-12-08	<1%
2	Internet	arxiv.org	<1%
3	Internet	github.com	<1%
4	Internet	raw.githubusercontent.com	<1%
5	Publication	José R. Zubizarreta, Elizabeth A. Stuart, Dylan S. Small, Paul R. Rosenbaum. "Hand...	<1%
6	Internet	www.bradyneal.com	<1%
7	Internet	mikenguyen13.github.io	<1%
8	Publication	Gourav Yadav, Piyush Moghe. "<b>Literature Review on Fake News Detection Usi...	<1%
9	Publication	Lanners, Quinn. "Domain-Guided Machine Learning: Trustworthy Methods for Ca...	<1%
10	Internet	www.ncbi.nlm.nih.gov	<1%

11 Publication

Guanglei Hong. "Causality in a Social World", Wiley, 2015 <1%

12 Internet

edoc.ub.uni-muenchen.de <1%

13 Publication

Durai Rajamanickam. "Causal Inference for Machine Learning Engineers", Spring... <1%

14 Publication

Peng Ding. "A First Course in Causal Inference", CRC Press, 2024 <1%

15 Internet

migariane.github.io <1%

16 Internet

cdn.bookey.app <1%

17 Internet

repository.essex.ac.uk <1%

18 Publication

Thomas P.A. Debray, Tri-Long Nguyen, Robert W. Platt. "Comparative Effectivenes... <1%

19 Internet

dokumen.pub <1%

20 Student papers

NSS COLLEGE OF ENGINEERING, PALAKKAD on 2025-07-24 <1%

21 Internet

grokipedia.com <1%

22 Internet

rucore.libraries.rutgers.edu <1%

23 Publication

David S Lee, Thomas Lemieux. "Regression Discontinuity Designs in Economics", J... <1%

24 Publication

Imai, Kosuke, and Marc Ratkovic. "Covariate balancing propensity score", Journal ... <1%

25 Internet

www.drugchemsci.com <1%

26 Student papers

North West University on 2026-07-20 <1%

27 Student papers

Victorian Institute of Technology on 2026-06-10 <1%

28 Internet

metricgate.com <1%

29 Internet

www.mdpi.com <1%

30 Publication

Mingchao Qi, JunQiang Zhao, Yan Feng. "An optimized public opinion communica... <1%

31 Publication

Ter-Minassian, Lucile. "Towards Trustworthy AI: From Local Explanations to Caus... <1%

32 Internet

coinbro.pro <1%

33 Internet

gpapadogeorgou.netlify.app <1%

34 Internet

pmc.ncbi.nlm.nih.gov <1%

35 Internet

www.proceedings.com <1%

36 Student papers

Erasmus University of Rotterdam on 2024-10-22 <1%

37 Publication

Habib Nurfaizal, Afrizal Zein. "Design and Development of a Vulnerability Simulati... <1%

38 Publication

Kret, Meghan. "Probabilistic Temporal Segmentation for Audio Moment Retrieval... <1%

39 Publication

Litschko, Robert Manfred. "Resource-Lean Transfer Methods for Cross-Lingual Inf... <1%

40 Publication

Rebecca R. Andridge, Ruoqi Song, Brady T. West. "A Sensitivity Analysis Framewor... <1%

41 Publication

Verme, Paolo. "Perceived Freedom", Encyclopedia of Quality of Life and Well-Bein... <1%

42 Internet

www.stata.com <1%

43 Internet

"Another look at the regression discontinuity design", 'Institute for Fiscal Studies' <1%

44 Student papers

Anna University on 2025-02-26 <1%

45 Publication

Freimut Bodendorf, Mathias Kraus. "Dimensions of Intelligent Analytics for Smart... <1%

46 Student papers

Harrisburg University of Science and Technology on 2026-04-17 <1%

47 Internet

datafield.dev <1%

48 Internet

deepai.org <1%

49 Internet

edoc.hu-berlin.de <1%

50 Internet

fastercapital.com <1%

51 Internet

link.springer.com <1%

52 Internet

opus.campus02.at <1%

53 Publication

"ICT Systems and Sustainability", Springer Science and Business Media LLC, 2026 <1%

54 Internet

athens.src.uchicago.edu <1%

55 Student papers

Capella University on 2025-05-13 <1%

56 Student papers

Columbia University on 2026-05-07 <1%

57 Publication

Guido W Imbens, Jeffrey M Wooldridge. "Recent Developments in the Econometri... <1%

58 Student papers

University of Hong Kong on 2024-03-30 <1%

59 Internet

academic.oup.com <1%

60 Publication

Richard Blundell, Monica Costa Dias. "Alternative approaches to evaluation in em... <1%

61 Student papers

University of Malaya on 2026-07-14 <1%

62 Student papers

Zambia Centre for Accountancy Studies on 2025-11-30 <1%

63 Internet

aclanthology.org <1%

64 Internet

jasonyanglu.github.io <1%

65 Internet

researchrepository.universityofgalway.ie <1%

66 Internet

www.ijcaonline.org <1%

67 Internet

www.repository.cam.ac.uk <1%

68 Publication

Dong, Juncheng. "Efficient Learning From Weak Supervision: From Data and Mod... <1%

69 Publication

Doostmohammadi, Ehsan. "Toward Understanding and Enhancing the Training a... <1%

70 Student papers

University of St. Gallen on 2018-05-28 <1%

71 Publication

Vinod Kumar Chauhan, Devendra Singh Dhami, Boyan Gao, Xin Wang, Lei Clifton, ... <1%

72 Publication

"Computational Linguistics and Intelligent Text Processing", Springer Science an... <1%

73 Publication

A.T. Panter, Sonya K. Sterba. "Handbook of Ethics in Quantitative Methodology", ... <1%

74 Publication

Abejide Ade-Ibijola, Solomon Oyelere, Jude Osakwe. "Innovation in Africa for Indu... <1%

75 Publication

Deshpande, Shachi. "Probabilistic and Generative Models for Causality and Decisi... <1%

76 Publication

Marco Caliendo. "SOME PRACTICAL GUIDANCE FOR THE IMPLEMENTATION OF PR... <1%

77 Publication

Navjeevan, Manvendu. "Meditations on Econometric Modeling", University of Cali... <1%

78 Student papers

University of Stellenbosch, South Africa on 2010-09-12 <1%

79 Internet

open.uct.ac.za <1%

80 Internet

therealistjuggernaut.com <1%

81 Internet
www.kcl.ac.uk <1%

82 Internet
www.manageengine.com <1%

83 Publication
Ghosheh, Ghadeer. "Individualized Data Generation for Electronic Health Records... <1%

84 Publication
Lavanya S. K., Subalalitha Chinnaudayar Navaneethakrishnan. "Building Tamil Te... <1%

85 Student papers
New Zealand School of Education on 2026-04-01 <1%

86 Publication
Peter C. Austin. "The performance of different propensity-score methods for esti... <1%

87 Student papers
Southern New Hampshire University - Continuing Education on 2026-02-24 <1%

88 Student papers
William Jessup University on 2026-06-19 <1%

89 Internet
carlos-mendez.org <1%

90 Internet
escholarship.org <1%

91 Internet
mmcommunications.vn <1%

92 Internet
www.nlsinfo.org <1%

93 Publication
Brand, J.E.. "The effects of job displacement on job quality: Findings from the W... <1%

94 Publication
Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini et al. " : A Suite... <1%

95 Student papers

George Washington University on 2026-07-15 <1%

96 Publication

James J. Heckman, Robert J. Lalonde, Jeffrey A. Smith. "The Economics and Econo... <1%

97 Publication

Jau-er Chen, Chien-Hsun Huang, Jia-Jyun Tien. "Debiased/Double Machine Learnin... <1%

98 Publication

Kamran, Fahad. "Aligning Machine Learning Solutions With Clinical Needs", Unive... <1%

99 Publication

Lyn Corno, Eric M. Anderman. "Handbook of Educational Psychology", Routledge, ... <1%

100 Publication

Ngo, Nghia Trung. "Multidimensional Transfer Learning in Natural Language Pro... <1%

101 Student papers

Rutgers University, New Brunswick on 2023-11-18 <1%

102 Publication

Shah, Aashna P.. "Redefining Normal in Clinical Medicine.", Harvard University <1%

103 Publication

Sicheng Lin, Meng Xu, Xi Zhang, Shih-Kang Chao, Ying-Kai Huang, Xiaolin Shi. "Bal... <1%

104 Student papers

University of Cape Town on 2020-12-02 <1%

105 Student papers

University of Maryland, Global Campus on 2026-02-25 <1%

106 Student papers

University of Melbourne on 2023-05-11 <1%

107 Student papers

Victorian Institute of Technology on 2026-06-03 <1%

108 Student papers

Westcliff University on 2024-10-14 <1%

109 Publication

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu et al. "Pre-Trained Models: Past, ... <1%

110 Publication

Zahra Dasht Bozorgi, Irene Teinemaa, Marlon Dumas, Marcello La Rosa, Artem Po... <1%

111 Internet

causalml-book.org <1%

112 Internet

ebin.pub <1%

113 Internet

hal.univ-lorraine.fr <1%

114 Internet

jobs.anitab.org <1%

115 Internet

scholarshare.temple.edu <1%

116 Internet

slds-lmu.github.io <1%

117 Internet

vdoc.pub <1%

118 Internet

www.micromindercs.com <1%

119 Internet

www.vive.dk <1%

120 Publication

Adapa, Pavan Dharma. "Beyond English: Auditing and Mitigating Cross-Lingual D... <1%

121 Publication

Alshammari, Hamed. "AI-Generated Text Detector for Arabic Language", Universi... <1%

122 Publication

Booil Jo. "Causal inference in randomized experiments with mediational processe... <1%

123 Student papers

College of Public Health on 2026-03-12 <1%

124 Publication

Daryl Pregibon. "More bang for their bucks", Proceedings of the 1st international ... <1%

125 Student papers

Erasmus University of Rotterdam on 2019-07-07 <1%

126 Student papers

Erasmus University of Rotterdam on 2019-10-07 <1%

127 Publication

Germano Antônio Zani Jorge. "An investigation of extractive summarization meth... <1%

128 Publication

Kutub Thakur, Al-Sakib Khan Pathan. "Transforming Cybersecurity with Machine ... <1%

129 Publication

Park, Sangbaek. "Estimating Individual Treatment Effects Using Emerging Metho... <1%

130 Publication

Qian, Eric Yibin. "Essays in Macroeconometrics: Heterogeneity, Spillovers, and Dy... <1%

131 Publication

Teppei Yamamoto. "Understanding the Past: Statistical Analysis of Causal Attribu... <1%

132 Student papers

Thesis 2025 on 2025-10-26 <1%

133 Student papers

Thesis 2026 on 2026-04-13 <1%

134 Student papers

Universidad Politécnica de Madrid on 2022-03-26 <1%

135 Student papers

Universidad Politécnica de Madrid on 2022-05-10 <1%

136 Student papers

Universiteit van Amsterdam on 2019-07-03 <1%

137 Student papers

University Of Tasmania on 2018-04-21 <1%

138 Student papers

University of St. Gallen on 2010-05-26 <1%

139 Student papers

University of Sydney on 2022-09-05 <1%

140 Publication

Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu, Lars Lyberg. "Handbook of Comp... <1%

141 Publication

Yarlagadda, Dedeepya. "A RAG-Augmented Graph Neural Network for Evidence-B... <1%

142 Internet

bonndoc.ulb.uni-bonn.de <1%

143 Internet

discovery.ucl.ac.uk <1%

144 Internet

docshare.tips <1%

145 Internet

edoc.pub <1%

146 Internet

etda.libraries.psu.edu <1%

147 Internet

files.eric.ed.gov <1%

148 Internet

mdpi-res.com <1%

149 Internet

nvlpubs.nist.gov <1%

150 Internet

ojs.aaai.org <1%

151 Internet
orca.cf.ac.uk <1%

152 Internet
repository.rice.edu <1%

153 Internet
www.daneshnamehicsa.ir <1%

154 Internet
www.eurecom.fr <1%

155 Internet
www.frontiersin.org <1%

156 Internet
www.mep.ki.se <1%

157 Internet
www.mit.edu <1%

158 Internet
www.researchsquare.com <1%

159 Internet
www.uyik.org <1%

160 Publication
Bing Xue, Ahmed Sameh Said, Ziqi Xu, Hanyang Liu, Neel Shah, Hanqing Yang, Phi... <1%

161 Student papers
College of Public Health on 2026-03-16 <1%

162 Publication
Ebrat, Danial. "Toward Intelligent Personalization: Integrating Large Language M... <1%

163 Student papers
Instituto Politecnico Nacional on 2020-02-25 <1%

164 Publication
Lecture Notes in Computer Science, 2015. <1%

165 Student papers

National Tsing Hua University on 2026-05-15 <1%

166 Publication

Parikh, Harsh J.. "Causal Inference for High-Stakes Decisions", Duke University, 20... <1%

167 Publication

Stefan Wager, Susan Athey. "Estimation and Inference of Heterogeneous Treatm... <1%

168 Student papers

Trident University International on 2025-03-12 <1%

169 Student papers

University of St. Gallen on 2021-09-27 <1%

170 Student papers

Vrije Universiteit Amsterdam on 2026-06-29 <1%

171 Publication

Aurore Delaigle, Paul Gustafson. "Handbook of Measurement Error Models", CRC ... <1%

172 Publication

Barack W. Wanjawa, Lilian D. A. Wanzare, Florence Indede, Owen Mconyango, La... <1%

173 Student papers

Bocconi University on 2024-09-18 <1%

174 Student papers

CSU, San Jose State University on 2026-05-13 <1%

175 Publication

Debasis Chaudhuri, Jan Harm C Pretorius, Debashis Das, Sauvik Bal. "Internationa... <1%

176 Student papers

Erasmus University of Rotterdam on 2019-07-07 <1%

177 Publication

Joseph L. Schafer, Joseph Kang. "Average causal effects from nonrandomized stu... <1%

178 Publication

Liu, Xiaole Alyssa. "Essays in Healthcare Operations.", New York University <1%

179 Publication

Liu, Zihan. "Effective Transfer Learning for Low-Resource Natural Language Unde... **<1%**

CHAPTER 1
INTRODUCTION

1.1 Background and Motivation

AI and data-driven systems are often built and tested under ideal conditions, such as clean datasets, predictable distributions, and controlled threat environments. However, they are deployed in situations where these conditions are rarely met. The difference between how a method performs in theory and in practice is not just a small adjustment issue. It is a significant problem across various fields, and this thesis addresses it directly. Three areas highlight this issue clearly. In causal inference, the growing number of machine learning estimators has made choosing the right method genuinely challenging. This thesis reveals a surprising finding: on datasets with poor covariate overlap, a simple difference-in-means estimator outperforms complex methods like Causal Forests by more than ten times in root mean squared error. In this case, more complexity can actually harm performance. In cybersecurity, a breach in June 2025 exposed around sixteen billion login credentials from major platforms. These credentials were structured in the typical URL-username-password format used by infostealer malware. While tools to prevent such breaches are available and well-documented, organizations still rely on less effective alternatives. The gap between available security technology and what is actually used remains the main unsolved problem in the field. In natural language processing, sentiment analysis has expanded from single-language text classification to include multilingual and multimodal systems. Yet, improvements seen in benchmark datasets do not consistently translate to real-world deployment, where low-resource languages are often overlooked, code-switching occurs frequently, and evaluation standards vary. All three areas share a common structural issue: systems can fail, sometimes dramatically, when real-world conditions deviate from the design assumptions. Understanding when and why this occurs—and how to address it—is the primary focus of this thesis.

1.2 Thesis Theme and Unifying Argument

The three studies in this thesis all ask the question: why do systems that use artificial intelligence fail when things do not happen as planned in the real world and what can people who work with these systems do about it? When we look at why these systems fail we see that it is because of the way things are distributed. If the people who get a treatment and the people who do not get the treatment are different then even the best ways of guessing what will happen do not work. This is not because the two groups are too different. Because the treatment affects people in different ways. This changes the way people should figure out what method to use. In the area of cybersecurity systems fail because people do not use the methods. There are ways to check if someone is who they say they are that are very good at stopping messages and people can buy these methods. Instead people use weaker methods like sending a secret code to someones phone. This study looks at why this happens and it is because of usability and organisational reasons. When we try to figure out what people mean by what they say systems

also fail. If a computer program is trained on texts that're mostly in English it does not work well with texts in other languages or texts that mix languages or texts that have pictures and words. There is a difference between how well these programs work in tests and how well they work in real life. The main point of all these studies is that people who design systems need to understand the situation they are designing for before they pick a method. This means that the context is more important, than the method and this is the idea of this thesis.All three studies show that this is true and it is the idea of this thesis.

1.3 Research Objectives

We will focus on three goals in this thesis. The first goal is to see how well causal estimators work when things are not perfect. We will test seven estimators with different datasets to find out which one works best in real life situations. We want to know which estimator to use when the data's not exactly what we want. The second goal is to look at cybersecurity. We will study how malware steals information and how we can prevent this. We also want to know why companies do not use all the security tools they have. We will look at ways to authenticate users and see which ones are the best. The third goal is to look at problems with sentiment analysis. We want to know how well computers can understand what people mean in languages and with different types of data. We will look at how language differences and lack of data can make it hard for computers to understand what people mean. We will also look at how using types of data can make things more complicated for computers. We want to know how we can make computers better, at understanding what people mean in situations.

1.4 Scope and Limitations

Each study is designed to be thorough in its method than covering a lot of ground. The test for cause-and-effect uses three sets of data from the RealCause framework. LaLonde PSID, LaLonde CPS and Twins. It looks at how average treatment effect estimation works. It does not look at treatment effects that depend on conditions or how confident we can be in the results or areas outside of work economics and health studies. The review of cybersecurity looks at how info-stealers work how to stop account takeovers and how to prevent session hijacking. It does not cover ransomware attacks that flood systems or threats to the supply chain.This review uses 166 sources from 2015 to 2025. Most of the studies it looks at are, from North America and Western Europe.The sentiment analysis review is limited to 18 studies published between 2020 and 2023 addressing multilingual or multimodal approaches. It is synthetic rather than experimental — it does not run original models or produce new benchmark results. All three studies acknowledge directions for future work that fall outside their current scope. The cybersecurity findings in particular may require reassessment as the threat landscape evolves.

1.5 Thesis Organisation

Chapter 2 provides a unified literature review spanning all three domains. It surveys the potential outcomes framework and the landscape of classical and machine learning-based

causal estimators; the evolution of credential theft from early keyloggers to modern Malware-as-a-Service platforms and the authentication technologies designed to counter them; and the progression of sentiment analysis from rule-based text classifiers to multilingual pre-trained transformers and multimodal fusion architectures. The chapter concludes by drawing the three literatures together under the unifying theme of robustness under real-world violations.

Chapter 3 presents the first contributing paper: a systematic benchmark of seven causal estimators across three RealCause datasets evaluated over one hundred independent samples each, supplemented by a controlled knob-sweep experiment varying overlap and treatment effect heterogeneity across a five-by-three grid. The chapter reports the key findings — the sophistication paradox, the T-Learner's unique robustness, and the primacy of heterogeneity over overlap as a driver of estimator ranking — and presents a practical two-stage decision framework grounded in these results.

Chapter 4 presents a PRISMA-compliant systematic review of infostealer malware and credential theft prevention, drawing on 166 sources. It covers how modern infostealers work, how effective current detection and authentication methods are in practice, and why the security-implementation gap persists despite the availability of stronger alternatives. Emerging technologies — behavioural biometrics, visual cryptography, and blockchain-based identity — are assessed against a structured multi-dimensional framework.

Chapter 5 presents a systematic review of multilingual and multimodal sentiment analysis, guided by seven research questions. It examines the challenges of linguistic diversity and code-switching, benchmarks state-of-the-art models including mBERT, XLM-R, and multimodal transformers, surveys strategies for handling low-resource language data scarcity, and identifies the limitations that prevent current systems from generalising reliably across languages and modalities.

Chapter 6 provides a cross-domain synthesis and discussion, placing the three studies in dialogue with one another. It argues that the sophistication paradox, the security-implementation gap, and the model generalisation gap in multilingual NLP are manifestations of the same structural challenge, and draws implications for the design of AI-driven systems that are intended for deployment under realistic, rather than idealised, conditions.

Chapter 7 concludes the thesis with a summary of contributions, explicit answers to the three research objectives stated in Section 1.3, and directions for future research in each domain and at their intersection.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction to the Review

This chapter situates the three contributing studies of this thesis within their respective bodies of literature, tracing the intellectual lineage from which each study emerges and identifying the gaps that each is designed to address. The review is organised into five substantive sections. The first two sections cover the domain of causal inference: the theoretical foundations of the potential outcomes framework and the landscape of estimators that implement it. The third and fourth sections are about cybersecurity. They talk about how credential theft through infostealer malware has changed over time. These sections also discuss the technologies used to protect authentication and sessions from threats. The fifth section looks at how sentiment analysis has evolved. It started with text classifiers that could only understand one language. Now it uses transformers and multimodal fusion architectures. The final section ties everything together. It explains that real-world conditions often do not match what AI systems are designed for. This mismatch causes AI performance to degrade. This issue affects all three areas discussed. This review is not meant to cover everything. It aims to provide background information. This information helps readers understand Chapters 3 4 and 5. It also explains why these chapters are important given the state of each field. The cybersecurity sections focus on theft and cybersecurity measures. The sentiment analysis section focuses on sentiment analysis. The concluding section ties the literatures, on cybersecurity and sentiment analysis together.

2.2 Causal Inference in Observational Studies: Theoretical Foundations

2.2.1 The Potential Outcomes Framework

The potential outcomes framework, which was made official by Rubin in 1974 and he built on the work of Neyman from 1923 is the basis of causal inference. The main idea of the outcomes framework is simple: for every single unit there are two possible results. One if the unit gets treatment and one if it does not get treatment. For example the potential outcomes for a unit can be Y_i(1) if the unit gets the treatment and Y_i(0) if the unit does not get the treatment. The effect of the treatment on an unit is found by looking at the difference between Y_i(1) and Y_i(0). The big problem with figuring out inference as Holland said in 1986 is that we can only see one of the potential outcomes for each unit. The potential outcomes framework is important because it helps us understand this problem. The potential outcomes framework is used to think about the outcomes, like Y_i(1) and Y_i(0) and how they are related to the treatment. The other is a counterfactual that must be estimated. The estimand of primary interest in most applied work is the Average Treatment Effect (ATE), defined as $E[Y(1) - Y(0)]$, which aggregates individual treatment effects across the entire population of interest. A related quantity is the Average Treatment Effect on the Treated (ATT), $E[Y(1) - Y(0) \mid T = 1]$, which restricts attention to the subgroup of units that actually received treatment. When the study population is divided into subgroups, the Conditional Average Treatment Effect (CATE), $E[Y(1) - Y(0) \mid$

X = x], captures how the treatment effect varies with covariates X. In a randomised controlled trial (RCT), treatment assignment T is independent of the potential outcomes by design, so the ATE can be consistently estimated by the simple difference in observed means between treated and control groups — what the benchmark study in Chapter 3 calls the "naive" estimator. The challenge of causal inference in observational studies is that treatment assignment is not random: units select into or are assigned to treatment in ways that correlate with their potential outcomes, creating confounding bias that the naive estimator does not correct.

2.2.2 Identifying Assumptions for Observational Data

Identification of the ATE from observational data requires three assumptions, each of which imposes a restriction on the data-generating process that cannot be verified from the data alone. The first is unconfoundedness, also called conditional ignorability or selection on observables (Rosenbaum and Rubin, 1983). It states that conditional on the observed covariates W, the treatment assignment T is independent of the potential outcomes: $\{Y(0), Y(1)\} \perp T \mid W$. In plain terms, this assumption requires that all variables that jointly influence both treatment assignment and outcomes have been observed and included in W. When this assumption is plausible, the conditional average treatment effect is identified from observational data by adjusting for W. The second is the positivity assumption, also called overlap. It requires that every unit in the study has a positive probability of receiving each treatment value, regardless of its covariate values: $0 < P(T = 1 \mid W = w) < 1$ for all w in the support of W. Overlap ensures that for any covariate profile, there exist both treated and control units from which to estimate counterfactual outcomes. When overlap is violated — that is, when some covariate profiles are exclusively observed in the treated group or exclusively in the control group — counterfactual predictions must rely on extrapolation, which becomes increasingly unreliable as the degree of mismatch grows. The LaLonde datasets used in Chapter 3 are canonical examples of an overlap violation: treated units are drawn from participants in the National Supported Work randomised experiment, while control units are drawn from the Panel Study of Income Dynamics (PSID) or the Current Population Survey (CPS), populations with substantially different pre-treatment characteristics. The third assumption is the Stable Unit Treatment Value Assumption (SUTVA), which has two components: no hidden versions of treatment (the treatment is well-defined and the same for all units) and no interference between units (the potential outcomes of unit i are unaffected by the treatment assignments of other units $j \neq i$). SUTVA is often violated in settings with social interactions, contagion, or network effects.

2.2.3 The Propensity Score and Its Role

Rosenbaum and Rubin (1983) introduced the propensity score $e(W) = P(T = 1 \mid W)$ as a scalar summary of the covariate vector W for the purpose of controlling confounding. Their key result is that if unconfoundedness holds given W, it also holds given e(W) alone — a dimension-reduction result that enables balancing through the propensity score rather than the full covariate vector. This result underlies two major classes of estimators: propensity score matching, which constructs matched treated-control pairs with similar propensity scores, and inverse probability weighting (IPW), which reweights units by the inverse of their probability

of receiving the treatment they actually received. The propensity score also provides a natural diagnostic for overlap. Examining the distribution of estimated propensity scores in the treated and control groups reveals whether the two groups occupy the same region of covariate space. Standardised mean differences (SMD) per covariate and the overlap coefficient of the propensity score distributions are commonly used summary diagnostics. When propensity scores approach zero or one for many units, the IPW weights become extreme, making the weighted estimator highly sensitive to individual observations and prone to large variance. The decision framework developed in Chapter 3 is grounded in these propensity score diagnostics.

2.3 Classical and Machine Learning-Based Causal Estimators

2.3.1 Classical Estimators: IPW, OLS, and Doubly Robust Methods

Inverse Probability Weighting (IPW), in its simplest form, reweights the observed outcomes by the inverse propensity score to create a pseudo-population in which treatment is independent of covariates. The Horvitz-Thompson estimator and the Hájek estimator are two well-known implementations. IPW is consistent if the propensity score model is correctly specified, but its variance can be extremely large when propensity scores are near the boundary of zero or one. Clipping propensity scores (for example, to the interval [0.01, 0.99]) mitigates the worst variance amplification at the cost of introducing some bias. Ordinary Least Squares (OLS), used as an outcome regression or ANCOVA estimator, fits a linear model of the outcome on treatment and covariates, then uses the fitted model to impute counterfactual outcomes for each unit. OLS is consistent if the outcome model is correctly specified — that is, if the true relationship between covariates and outcomes is approximately linear. Its key vulnerability under overlap violations is that it must extrapolate its fitted linear surface into covariate regions populated only by treated or only by control units. If the true effect surface is nonlinear or if the two groups occupy substantially different covariate regions, this extrapolation introduces systematic bias. As the results in Chapter 3 show, this vulnerability is severe on the LaLonde datasets, where OLS achieves RMSE of $14,941 on LaLonde PSID. Doubly robust estimators, most prominently the Augmented Inverse Probability Weighting (AIPW) estimator introduced by Robins, Rotnitzky, and Zhao (1994) and formalised by Bang and Robins (2005), combine an outcome regression with a propensity score model. Their defining property is that the estimator remains consistent if either the outcome model or the propensity score model is correctly specified — not necessarily both. This double robustness is theoretically appealing because it provides protection against misspecification of one of the two components. However, when both models are misspecified, AIPW inherits the bias of both components, and as Chapter 3 demonstrates, on datasets with severe overlap violations AIPW's performance (RMSE $12,889 on LaLonde PSID) is not substantially better than OLS.

2.3.2 Meta-Learner Frameworks: T-Learner and X-Learner

The meta-learner framework, formalised by Künzel, Sekhon, Bickel, and Yu (2019), treats causal effect estimation as a function estimation problem and uses any regression method as a plug-in "base learner." Rather than defining a single estimator, meta-learners define a strategy

for using regression models to estimate counterfactual outcomes, with CATE or ATE estimated as a derived quantity.

The T-Learner (Two-model Learner) fits two separate outcome models: one trained exclusively on treated units, $\hat{\mu}_1(W) = E[Y \mid T = 1, W]$, and one trained exclusively on control units, $\hat{\mu}_0(W) = E[Y \mid T = 0, W]$. The ATE is then estimated as the mean difference between the two models' predictions across all units: $ATE = (1/n) \Sigma [\hat{\mu}_1(W_i) - \hat{\mu}_0(W_i)]$. This stops the problem of using information from one group to make guesses about another group, which's what causes problems for things like OLS and AIPW and even the X-Learner when the groups are really different. The T-Learner is very good at handling this problem. Chapter 3 has world examples that show this is true and it shows that the T-Learner is the best, at handling different situations especially when the groups are not the same size or do not overlap much. The X-Learner, which was also introduced by Künzel et al. In 2019 is a kind of estimator that is used when the treated group and the control group are not the same size. In its first stage, it imputes individual treatment effects for treated units using the control outcome model, and for control units using the treated outcome model. In its second stage, it combines the two sets of imputed effects using a propensity score-weighted average. The X-Learner's motivation is that the second stage exploits the more abundant group to improve estimates for the smaller group. Under good overlap, this design provides efficiency gains. Under poor overlap, however, the second-stage cross-group predictions are precisely the extrapolations that the T-Learner avoids, explaining why the X-Learner performs substantially worse than the T-Learner on the LaLonde datasets (RMSE $10,422 vs $6,995 on LaLonde PSID).

2.3.3 Causal Forests and Double Machine Learning

Causal Forests, introduced by Wager and Athey (2018), extend the random forest framework to directly target heterogeneous treatment effects. Each tree in the forest is built using a splitting criterion designed to maximise treatment effect heterogeneity within leaves rather than outcome prediction accuracy. "Honesty" — a key design principle of causal forests — requires that the data used to determine the tree structure be separate from the data used to estimate leaf-level treatment effects, reducing overfitting. Causal Forests are consistent for the CATE under regularity conditions and are designed to produce valid confidence intervals for heterogeneous effects. Double Machine Learning (DML), developed by Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), addresses the problem of estimating low-dimensional treatment effect parameters in the presence of high-dimensional nuisance functions (outcome and treatment propensity models). DML uses cross-fitting — a sample-splitting procedure — to estimate the nuisance functions on a held-out fold and then constructs a Neyman-orthogonal moment function whose score is insensitive to first-order perturbations of the nuisance functions. This Neyman orthogonality property ensures that errors in estimating the nuisance functions do not translate into first-order bias in the treatment effect estimate, enabling root-n consistent and asymptotically normal estimation even when the nuisance functions are estimated with flexible machine learning methods. The EconML implementation used in Chapter 3 combines DML with a causal forest — using random forests for the nuisance functions and a causal forest for the CATE — in what is called CausalForestDML. While this

combination provides strong theoretical guarantees under well-specified conditions, it requires cross-group prediction of residuals and is therefore exposed to extrapolation error under poor overlap. The finding in Chapter 3 that Causal Forest achieves the worst RMSE among all seven estimators on both LaLonde datasets ($15,850 and $20,421) is particularly striking given its theoretical sophistication, and forms the core of what Chapter 3 terms the "sophistication paradox."

2.3.4 Benchmarking Causal Estimators: Existing Work and Its Gaps

Several prior benchmarking studies have compared causal estimators. The Atlantic Causal Inference Conference (ACIC) Data Challenges (Dorie et al., 2019) have provided a recurring benchmarking platform using fully synthetic data-generating processes, enabling controlled evaluation of many estimators. The advantage of fully synthetic benchmarks is complete control over the DGP; the disadvantage is that the synthetic distributions may not reflect the confounding structures, covariate distributions, or outcome relationships observed in real applications. At the other extreme, fully observational datasets such as the original LaLonde (1986) dataset have known ground-truth ATE estimates from the embedded RCT component, but only for the treated group — making rigorous evaluation of ATE estimation across the full sample distribution impossible. Curth and van der Schaar (2021) benchmarked meta-learners on the Infant Health and Development Program (IHDP) dataset and found strong dataset-dependence in estimator rankings, but their comparison did not include Causal Forest, DML, or any classical estimators. The RealCause framework (Neal, Huang, and Raghupathi, 2021) bridges the gap between fully synthetic and fully observational benchmarks by fitting deep generative models — specifically TARNet architectures — to experimental datasets with known ground truth and using them to generate realistic synthetic samples with verifiable causal estimands. Gentzel, Pruthi, and Jensen (2021) used RealCause to benchmark 1,568 estimator configurations but focused on classical statistical methods and did not include modern ML estimators such as Causal Forests, meta-learners, or DML. The benchmark study in Chapter 3 directly extends Gentzel et al. by adding four ML estimators, conducting the benchmark across all three RealCause datasets at 100 samples each, and providing a systematic analysis of how overlap and treatment effect heterogeneity interact to determine estimator performance.

2.4 Credential Theft and Infostealer Malware

2.4.1 The Credential Theft Ecosystem

Credential theft is still one of the commonly exploited attack vectors in cybersecurity. It targets usernames, passwords, session tokens and cryptographic keys. The scale of this problem has grown a lot. For example RockYou2021 collected 8.4 billion passwords from breaches. Also in June 2025 Cybernews researchers found about sixteen billion credentials in thirty databases. These databases included accounts from Google, Apple, Facebook, GitHub, Telegram and government services. The way the exposed records are structured, with a URL, username and password is typical of what infostealer malware produces. This suggests that the dataset was collected from campaigns. Stolen credentials can be used for two types of attacks. One is

account hijacking, where attackers use the stolen credentials to log in as the victim. The other is session hijacking, which's more advanced. It involves stealing session tokens to take over an already authenticated session without needing the password. Modern infostealers can steal both types of information in an infection. We will look at how each attack type works and what countermeasures are available against them, in Chapter 4. Credential theft and stolen credentials are issues. They can lead to account hijacking and session hijacking. So understanding theft and how to prevent it is crucial.

2.4.2 Detection Approaches: What Infostealer Malware Can and Cannot Do

To catch infostealer malware you need to have security measures in place at the endpoint, network and behavioural levels. No single method can cover all the ways these malwares attack. The traditional antivirus software works well against known infostealer malware types it can stop about 85 to 95 percent of them. But it is not very good at stopping ones it can only stop about 40 to 60 percent of those. The people who make infostealer malware usually release versions every two to four weeks. This means that there is a time period when the malware can infect computers without being detected. The infostealer malware can even change its code, which makes it hard for the antivirus software to detect it. There is another way to detect infostealer malware it is called detection. This method uses Endpoint Detection and Response systems to monitor what the malware is doing on the computer. It looks for things like the malware trying to access the browsers password database or trying to inject code into programs. This method can detect 70 to 90 percent of infostealer malware types. The problem with detection is that it can sometimes think that a legitimate program is malware this is called a false positive. This can happen because some legitimate programs, like password managers can look like infostealer malware to the Endpoint Detection and Response system. Some Endpoint Detection and Response systems use machine learning to detect infostealer malware. These systems can detect about 85 to 92 percent of infostealer malware types. The people who make infostealer malware can test their malware against these systems and make it so that it is not detected. Another way to detect infostealer malware is to use a sandbox. A sandbox is an environment where the malware can run without hurting the computer. This method can detect 90 to 95 percent of simple infostealer malware types. The people who make infostealer malware know about sandboxes and they can make their malware so that it is not detected. Infostealer malware can also avoid being detected at the network level. This is because the malware can use different addresses to communicate with its creators and it can make its traffic look like legitimate traffic. The main problem is that the people who make infostealer malware have an advantage. They can test their malware against all the detection systems before they release it.. The people who are trying to stop the malware have to detect every single type of it. Usually the detection of infostealer malware happens months after the infection, not days. This means that authentication systems cannot rely on detecting infostealer malware in time. Infostealer malware is a problem and it is hard to detect.They must remain secure even after credentials have already been stolen — which is precisely why Chapter 4's prevention analysis focuses on credential-agnostic approaches such as FIDO2/WebAuthn.

2.5 Authentication and Session Security

2.5.1 Multi-Factor Authentication Methods

Multi-factor authentication (MFA) reduces account takeover risk by requiring a second verification factor beyond the password. The major implementations differ substantially in their security and usability profiles, and a direct quantitative comparison is central to Chapter 4's contribution. SMS-based one-time passcodes (OTP) are the most widely deployed MFA method, used by approximately 60–70% of organisations that mandate MFA, but they provide no resistance to phishing or man-in-the-middle attacks (0% phishing resistance) and are vulnerable to SIM-swapping, in which an attacker socially engineers a mobile carrier into transferring the victim's phone number to an attacker-controlled SIM. Lee et al. (2020) tested five major US carriers and found SIM-swapping success rates between 39% and 77% using basic impersonation techniques. NIST SP 800-63B has deprecated SMS-based MFA for high-security applications. Despite these vulnerabilities, SMS-based MFA remains dominant because of its low deployment cost and high user familiarity. Time-based one-time passcodes (TOTP), generated by authenticator applications (Google Authenticator, Authy, Microsoft Authenticator), provide substantially better security than SMS: they are immune to SIM-swapping and provide approximately 50% resistance to phishing (real-time phishing kits can intercept and replay TOTP codes, though the narrow validity window limits the attack). TOTP authentication time is approximately 6.3 seconds with a 6% user error rate. Hardware security keys implementing the FIDO2/WebAuthn standard provide the strongest available authentication. Lang et al. (2017) documented Google's deployment of hardware security keys to over 85,000 employees: over the course of a one-year study period, the deployment eliminated all successful phishing attacks against enrolled accounts. FIDO2 achieves this through origin-bound public key cryptography: the cryptographic signature produced by the key is bound to the specific domain of the service being authenticated, making it impossible to use credentials obtained through a phishing site (which would have a different origin) against the legitimate site. Platform authenticators — fingerprint and face recognition integrated into smartphones and laptops, implementing the same FIDO2 protocol — achieve equivalent security with zero marginal hardware cost and an authentication time of approximately 2.1 seconds. The cost of authentication methods is often assumed to favour simpler approaches, but Forrester Research analysis demonstrates the opposite over a three-year horizon. For a 10,000-user organisation, password-only systems have the lowest initial cost ($25,000) but incur high ongoing operational costs — primarily password resets and account recovery, which account for 30–50% of help desk ticket volume — resulting in a three-year total cost of ownership of $580,000 ($58 per user).

2.5.2 Session Hijacking and Its Countermeasures

Session hijacking is when someone steals your session to get into your account without needing your password. This happens by getting the special tokens that keep you logged in. There are four ways this happens:

- Hackers use scripts to steal your cookies through XSS attacks.
- They force you to use a session ID that they control, known as session fixation.
- They intercept your tokens from internet traffic called MITM attacks.
- They use malware on your device to read your browser cookies which is hard to defend against.

To prevent this at the network level using HTTPS is key. Between 2015 and 2017 more websites started using HTTPS from 40% to 80% making it harder for hackers to intercept your data. Adding HSTS, which stops browsers from connecting unencrypted reduces MITM attacks by a lot, around 95%.

At the application level there are three cookie settings that help:

- HttpOnly stops JavaScript from reading your cookies, which reduces cookie theft by 83%.
- The Secure flag makes sure cookies are only sent over encrypted connections reducing hijacking by 91%.
- SameSite=Lax, which became Chromes default in 2020 reduces CSRF attacks by 94%.

However just using one of these is not enough. A study by Calzavara et al. In 2018 found that 23% of sites using HttpOnly were still vulnerable. So it's best to use all three settings along, with Content Security Policy headers and checking user input.

Tokens that prevent CSRF attacks work well stopping 99.7% of CSRF-based attacks when used correctly. Session binding — tying tokens to client IP address and user-agent — detected 92% of hijacking attempts in controlled evaluations.

2.6 Sentiment Analysis: From Text to Multilingual and Multimodal Systems

2.6.1 Foundations of Sentiment Analysis

Sentiment analysis, also called opinion mining, is the computational study of subjective information expressed in natural language. The field addresses two related but distinct tasks: emotion extraction, which identifies the emotional state expressed by a text (happy, sad, angry, etc.), and opinion extraction, which identifies the opinion holder, the entity being evaluated, and the polarity of the expressed sentiment. An example of the latter: in "Alex enjoys the battery life of the laptop," Alex is the opinion holder, the battery life is the aspect of the laptop entity being evaluated, and the sentiment is positive. Early approaches to sentiment analysis were rule-based or lexicon-based: sentiment lexicons such as SentiWordNet and AFINN assigned positive or negative scores to individual words, and sentiment scores were computed by aggregating lexical scores with simple heuristics for negation and intensification. These approaches were fast and interpretable but brittle: they depended heavily on the completeness of the lexicon, failed on idiomatic expressions, and could not handle context-dependent polarity reversal (for example, "the battery doesn't last long" is negative, but so is "the battery doesn't

drain quickly" — a simple negation rule would get the second wrong). Machine learning approaches replaced lexicon-based methods as labelled datasets became available. Supervised models using Naive Bayes, Support Vector Machines, and eventually deep learning on bag-of-words or TF-IDF representations substantially outperformed rule-based systems.

2.6.2 Multilingual Pre-Trained Language Models

The introduction of the Transformer architecture (Vaswani et al., 2017) and the subsequent development of large pre-trained language models represented a paradigm shift in natural language processing. Models pre-trained on massive corpora using self-supervised objectives such as masked language modelling (BERT; Devlin et al., 2019) or causal language modelling (GPT; Radford et al., 2018) learn rich contextual representations of language that can be fine-tuned on small labelled datasets with strong results. The key insight is that pre-training on large unlabelled corpora transfers linguistic knowledge that would be impractical to learn from small task-specific datasets. Multilingual BERT (mBERT), released by Devlin et al. (2019), extended this approach to 104 languages by pre-training a single transformer model on Wikipedia text from all supported languages simultaneously. The shared architecture and joint pre-training produce representations that support zero-shot cross-lingual transfer: a model fine-tuned on English sentiment data can be applied to French or German texts without French or German training examples. However, mBERT was trained without explicit cross-lingual alignment objectives, and its performance on low-resource languages is constrained by the amount of text from those languages present in its training data. XLM (Conneau and Lample, 2019) and XLM-R (Conneau et al., 2020) improved on mBERT by incorporating Translation Language Modelling (TLM) as an additional pre-training objective, which uses parallel corpora (aligned translations of the same text in multiple languages) to explicitly align representations across languages. XLM-R, trained on 2.5 terabytes of filtered Common Crawl text across 100 languages, substantially outperforms mBERT on multilingual benchmarks, particularly for low-resource languages. It offers robust zero-shot learning capabilities and more consistent cross-lingual transfer. mT5 (Xue et al., 2021), a multilingual variant of the T5 model, treats all NLP tasks as text-to-text generation problems, enabling a single model to handle classification, translation, and generation in multiple languages through a unified interface.

2.6.3 Multimodal Sentiment Analysis

Multimodal sentiment analysis moves beyond text to jointly analyse speech, facial expressions, body language, and visual context — recognising that emotion is expressed across multiple channels simultaneously, each carrying information the others do not.

The core technical challenge is fusion — combining heterogeneous modality representations effectively. Early approaches used simple concatenation or decision-level combination. Tensor fusion networks (Zadeh et al., 2017) improved on this by modelling cross-modal interactions through outer products of unimodal representations, capturing pairwise and three-way feature interactions that flat concatenation misses.

Attention-based fusion mechanisms, developed with the rise of transformers, allow the model to learn which modality to weight more heavily depending on the input context. Tensor and attention-based fusion approaches achieve over 80% accuracy on CMU-MOSI and CMU-MOSEI benchmarks. VisualBERT (Li et al., 2019) and LXMERT (Tan and Bansal, 2019) extend the BERT architecture to joint vision-language tasks by aligning image regions with text tokens through a shared transformer. These models achieve strong results on tasks that combine textual and visual cues. CLIP (Radford et al., 2021) aligns text and image embeddings using a contrastive loss, offering zero-shot visual-text matching capabilities. Multimodal fusion models (MMBT, UNITER) combine features from different modalities through learned fusion layers, achieving competitive results on social media datasets where images and text reinforce each other. Audio-specific models combining Mel-Frequency Cepstral Coefficients (MFCC) features with CNNs achieve up to 92.7% accuracy on audio sentiment benchmarks — the highest accuracy of any single-modality approach reviewed in Chapter 5 — because acoustic features such as pitch, tempo, and energy encode affective information that text representations do not capture. The Self-MM model, which uses self-supervised learning to align multimodal representations with textual anchors, achieved an 84.30% F1 score on binary sentiment classification tasks. MTSA on CMU-MOSEI achieved the lowest Mean Absolute Error of 0.541 among reviewed models. The limitation shared by all current multimodal models is the dependence on large, aligned multimodal training datasets, which are expensive to collect and label, and which are predominantly available only in English. Extending multimodal sentiment analysis to low-resource languages requires both the multilingual model development described above and the creation of multimodal datasets in those languages — a doubly demanding requirement that has constrained progress.

2.6.4 Benchmark Datasets and Evaluation

The choice of benchmark dataset significantly affects the relative performance of sentiment analysis models. Standard evaluation metrics include accuracy, F1 score (particularly macro-F1 for class-imbalanced settings), and Mean Absolute Error (MAE) for regression-style sentiment intensity prediction. The fragmentation of evaluation across different datasets, metrics, and task formulations makes it difficult to make direct comparisons of model performance across papers. Among the frequently used datasets in the systematic review underlying Chapter 5, CMU-MOSI and CMU-MOSEI are the dominant multimodal benchmarks. For multilingual text tasks, XLM-T demonstrated F1 score improvements from 40.29% to 56.39% when fine-tuned for cross-lingual sentiment classification on Twitter data, highlighting the potential of domain-specific fine-tuning even for models with strong multilingual pre-training.

2.6.5 Strategies for Low-Resource Language Development

Cross-lingual transfer learning exploits the shared representations learned by models like mBERT and XLM-R to apply models fine-tuned on high-resource languages to low-resource target languages with zero or few labelled examples. Zero-shot cross-lingual transfer works surprisingly well for structurally similar language pairs but degrades significantly when source

and target languages are typologically distant. Crowdsourcing through platforms like Amazon Mechanical Turk, and through community initiatives such as Masakhane (which focuses on African languages) and AI4Bharat (which focuses on Indian languages), engages native speakers to create linguistically authentic annotated datasets. Mozilla's Common Voice project has collected speech data in multiple languages through community contribution. The IndicNLP dataset covers multiple Indian languages. These community-driven efforts address the problem that commercially motivated data collection tends to prioritise high-resource languages with large markets. Weak supervision and self-supervised approaches use noisily labelled or unlabelled data — social media posts with emojis, for instance, where the emoji provides a proxy sentiment label — to train models without expensive manual annotation. The abundance of unstructured multilingual text on the web makes these approaches attractive for low-resource settings where labelled data is scarce but raw text is available.

2.7 Cross-Cutting Theme: Robustness Under Real-World Violations

The three literatures reviewed in this chapter, though occupying different domains, converge on a shared structural challenge that forms the thesis theme. In causal inference, the challenge is that practitioners face distributional violations — departures from the overlap and homogeneity assumptions under which many estimators are designed — and the existing literature has not clearly established which estimators degrade gracefully under these conditions and which fail catastrophically. The conventional diagnostic workflow prioritises overlap checks, but the evidence reviewed here and the results of Chapter 3 indicate that treatment effect heterogeneity is the more important driver of estimator performance degradation, requiring a revision of standard practice. In cybersecurity, the challenge is that a substantial security-implementation gap exists between the theoretical capability of available authentication methods and their actual deployment. Hardware security keys and FIDO2 platform authenticators offer documented near-complete phishing resistance, yet SMS-based OTP — with zero phishing resistance — remains dominant. The technical, usability, and organisational factors that sustain this gap are well-documented in the literature reviewed here, and the emergence of increasingly capable detection-evading infostealers makes closing the gap more urgent. In causal inference, diagnosing heterogeneity before overlap determines the appropriate estimator. In authentication, diagnosing the actual threat vectors and deployment constraints of an organisation's environment determines which authentication methods will close the security gap in practice rather than merely on paper. In sentiment analysis, understanding the linguistic and modal characteristics of the target deployment environment — its languages, modalities, and cultural contexts — is prerequisite to selecting or building a model that will generalise. This principle, which emerges independently from the literatures reviewed here, constitutes the intellectual foundation on which Chapters 3, 4, and 5 build.

CHAPTER 3

BENCHMARKING CAUSAL ESTIMATORS UNDER OVERLAP AND HETEROGENEITY VIOLATIONS

3.1 Introduction and Research Questions

Causal inference from observational data is one of the most consequential and contested problems in applied statistics. Across economics, epidemiology, political science, and industry analytics, practitioners routinely need to estimate the effect of a treatment, intervention, or policy using data that was not generated by a randomised experiment. The reasons randomisation is often unavailable are varied: it may be ethically impermissible to withhold a potentially beneficial treatment from a control group; it may be retrospectively impossible to randomise over decisions that have already been made; or it may simply be too expensive or logistically infeasible in the context at hand. In these settings, observational estimators — methods that attempt to recover causal effects from non-randomised data by statistically adjusting for the differences between treated and control groups — are the only available tool. The practical challenge for any analyst working with observational data is that the landscape of available estimators has expanded dramatically over the past decade. Classical methods — inverse probability weighting, outcome regression via ordinary least squares, and doubly robust combinations of the two — have been joined by a growing family of machine learning-based causal estimators including meta-learners, causal forests, and double machine learning. The theoretical properties of these newer methods under ideal conditions are well-understood: they can approximate arbitrary treatment effect surfaces without parametric assumptions, they achieve minimax-optimal rates of convergence in large samples, and they can be combined with cross-fitting procedures that provide valid uncertainty quantification. The question that their theoretical guarantees do not directly answer, however, is how they behave when the idealized conditions under which those guarantees are derived — particularly the assumption of adequate covariate overlap between treated and control groups — are violated, as they routinely are in practice. This chapter presents a systematic benchmark designed to answer that question directly. It makes four principal contributions. First, it establishes empirically that a sophistication paradox exists: on datasets with poor covariate overlap, the naive difference-in-means estimator substantially outperforms state-of-the-art machine learning methods, with Causal Forest achieving root mean squared error (RMSE) between $15,850 and $20,421 compared to $1,132–$2,925 for the naive estimator on the same datasets. Second, it identifies the T-Learner as uniquely robust among machine learning estimators under distributional mismatch, achieving statistically significant improvements over all six competitors on LaLonde CPS and over five of six on LaLonde PSID. Third, it isolates the mechanism driving the performance differences: a controlled knob-sweep experiment demonstrates that it is treatment effect heterogeneity, not overlap degradation, that drives estimator ranking reversal — a finding that directly inverts the conventional practitioner diagnostic workflow. Fourth, it translates these empirical findings into an actionable two-stage decision framework that practitioners can apply before committing to an estimator. The chapter is organised as follows. Sections 3.5

through 3.8 present the five principal findings in detail. Section 3.9 presents the practical decision framework derived from those findings. Section 3.10 discusses the mechanistic explanations for the observed patterns, the limitations of the study, and directions for future work.

The research questions motivating this chapter are:

RQ1. Across a benchmark spanning classical and machine learning causal estimators, how does performance — measured by RMSE of the ATE estimate — vary as a function of the degree of covariate overlap between treated and control groups?

RQ2. Among machine learning estimators, which is most robust to distributional mismatch, and what structural property of that estimator explains its robustness?

RQ3. What is the relative importance of covariate overlap and treatment effect heterogeneity in determining estimator performance rankings? Which dimension should practitioners diagnose first?

RQ4. Can the empirical findings of the benchmark be translated into a reproducible, diagnostics-driven decision procedure that a practitioner can apply before selecting an estimator?

3.2 Datasets and the RealCause Framework

3.2.1 The Problem with Existing Benchmarking Approaches

Any credible benchmark of causal estimators must resolve a fundamental tension. Evaluating an estimator requires knowledge of the true causal effect — the ground truth against which estimates can be compared. In a randomised controlled trial, this ground truth is accessible because treatment assignment is independent of potential outcomes by design. In a purely observational dataset, it is not, because the potential outcome under the counterfactual treatment condition is by definition unobserved. Previous benchmarking approaches have resolved this tension in two ways, each with serious limitations. Fully synthetic benchmarks, such as those used in the Atlantic Causal Inference Conference (ACIC) Data Challenges, generate data entirely from known parametric distributions, ensuring that ground-truth causal effects are available. The limitation of this approach is that the synthetic data-generating processes may not faithfully represent the confounding structures, covariate distributions, or outcome relationships encountered in real applications. A method that performs well on a particular synthetic DGP may do so by implicitly exploiting that DGP's specific structure rather than by being generally robust. Specifically, the generative model is a TARNet architecture, a treatment-aware representation network that learns separate outcome heads for treated and control units while sharing a covariate representation layer. Once fitted, the generative model can be sampled repeatedly to produce new datasets that have the same joint distributional structure as the original experimental data — including the same covariate distributions,

treatment propensity surface, and outcome functions — but with known ground-truth potential outcomes generated by the model itself. The key advantage of this approach over fully synthetic benchmarks is that the data-generating process is grounded in real observational structure, including the specific confounding patterns, covariate correlations, and outcome heterogeneity present in the original data. The key advantage over using real observational datasets directly is that ground-truth potential outcomes are available for every unit in every generated sample, enabling exact RMSE evaluation of ATE estimates.

3.2.2 The Three Benchmark Datasets

This chapter uses all three datasets available in the RealCause benchmark library: LaLonde PSID, LaLonde CPS, and Twins. Each dataset is available as 100 independently generated samples from the fitted generative model, where each sample contains covariates W, binary treatment T, observed outcome Y, and ground-truth potential outcomes Y(0) and Y(1) for every unit. This structure allows exact computation of the true ATE for each sample and therefore exact evaluation of estimation error.

LaLonde PSID is derived from Dehejia and Wahba's (1999) re-analysis of LaLonde's (1986) seminal study evaluating the econometric evaluation of job training programmes. The treated group in this dataset consists of participants in the National Supported Work (NSW) Demonstration, a randomised job training programme that ran in the mid-1970s in the United States. The NSW programme enrolled economically disadvantaged adults — long-term welfare recipients, ex-drug addicts, ex-criminal offenders, and youth school dropouts — and randomly assigned them to either a temporary subsidised employment programme or a control condition. The outcome of interest is post-treatment earnings in 1978 (in 1978 US dollars). In the LaLonde PSID variant, the control group is not drawn from the NSW randomised trial itself but from the Panel Study of Income Dynamics (PSID), a representative survey of the US labour force population. The PSID control units are general labour market participants and differ substantially from NSW participants on pre-treatment characteristics: they have higher pre-treatment earnings, higher rates of employment, higher education levels, and different demographic compositions. This deliberate substitution creates a severe overlap violation — the treated and control covariate distributions are substantially non-overlapping — making LaLonde PSID the canonical stress test for observational causal estimators.

LaLonde CPS uses the same NSW treated group but substitutes a control group drawn from the Current Population Survey (CPS), another representative survey of the US labour force population. The CPS control group is larger than the PSID control group and somewhat different in its distributional properties, but the fundamental structure is the same: a treated group of NSW programme participants and a control group of general labour market participants. The overlap violation is present but somewhat less extreme than in LaLonde PSID, making LaLonde CPS a useful complement that represents an intermediate level of distributional mismatch. The Twins dataset provides the positive control necessary to distinguish performance differences attributable to distributional mismatch from those attributable to other factors. Together, the three datasets span a meaningful range of overlap

conditions from severe mismatch (LaLonde PSID) through moderate mismatch (LaLonde CPS) to near-perfect overlap (Twins).

3.3 Estimator Descriptions and Identifying Assumptions

3.3.1 Identifying Assumptions Common to All Estimators

All seven estimators benchmarked in this chapter are designed to estimate the population Average Treatment Effect, defined as ATE = E[Y(1) − Y(0)], from observational data. They differ in the statistical strategy they employ to do so, but they all rely on the same three identifying assumptions from the potential outcomes framework. The first assumption is **unconfoundedness**, also called conditional ignorability or selection on observables. Formally: {Y(0), Y(1)} ⊥ T | W, where W denotes the observed covariate vector. This assumption states that conditional on the observed pre-treatment covariates, the treatment assignment T carries no additional information about the potential outcomes — in other words, all variables that jointly influence both the probability of treatment and the potential outcomes have been measured and are included in W. Unconfoundedness is a strong assumption that cannot be verified from the data; its plausibility depends entirely on the subject-matter knowledge of the analyst about what determines treatment selection. Within the RealCause framework, unconfoundedness is satisfied by construction because the generative model specifies the complete relationship between covariates, treatment assignment, and potential outcomes. The LaLonde datasets represent a deliberate, canonical violation of positivity: because treated units are drawn from the NSW programme population and control units are drawn from general labour market surveys, there exist covariate profiles that are observed exclusively or predominantly in one group, making propensity scores approach zero or one for many units. The RealCause overlap knob, described in Section 3.4, allows systematic variation of the degree to which this positivity violation is present. The third assumption is **SUTVA** — the Stable Unit Treatment Value Assumption — which has two components. The first component is the no-interference requirement: the potential outcomes of unit i depend only on unit i's own treatment assignment, not on the treatment assignments of other units. The second component is the no-hidden-versions-of-treatment requirement: the treatment is well-defined and takes the same form for all treated units. SUTVA is satisfied by construction in the RealCause framework.

3.3.2 The Seven Estimators

Naive Difference-in-Means. The naive estimator computes the ATE as the simple difference between the mean observed outcome in the treated group and the mean observed outcome in the control group: ATE_naive = Ȳ_1 − Ȳ_0, where Ȳ_1 is the sample mean outcome among treated units and Ȳ_0 is the sample mean outcome among control units. This estimator performs no covariate adjustment whatsoever. It is consistent for the true ATE only under conditions of random treatment assignment — that is, when treatment is independent of potential outcomes unconditionally, not merely conditionally. Under confounding, the naive estimator will be biased by the difference in potential outcomes between groups that would have existed even in the absence of treatment. Despite this theoretical limitation, the naive estimator serves a crucial

role in the benchmark as the performance baseline and, paradoxically, as the best-performing estimator on the LaLonde datasets, as Section 3.5 demonstrates.

Inverse Probability Weighting (IPW). IPW reweights each observation by the inverse of its estimated probability of receiving the treatment it actually received, creating a pseudo-population in which treatment assignment is independent of covariates. For a treated unit i, the weight is $1/\hat{e}(W_i)$, where $\hat{e}(W_i)$ is the estimated propensity score; for a control unit, the weight is $1/(1 - \hat{e}(W_i))$. The propensity scores are estimated using logistic regression of treatment on all covariates. To prevent extreme weights when propensity scores are near zero or one — which occurs frequently under poor overlap — propensity scores are clipped to the interval [0.01, 0.99] before computing weights. IPW is consistent if the propensity score model is correctly specified. Its primary weakness under poor overlap is high variance: when many units have propensity scores near the boundary, even with clipping the weights can be large, making the estimator sensitive to individual observations.

OLS/ANCOVA. The OLS estimator fits an ordinary least squares regression of the outcome Y on the treatment indicator T and all covariates W: $Y = \alpha + \tau T + \beta^T W + \varepsilon$. The estimated ATE is the coefficient $\hat{\tau}$ on the treatment indicator. This estimator is consistent if the outcome model is correctly specified — that is, if the true relationship between covariates and outcome is approximately linear and does not vary with treatment status. Under poor overlap, the OLS estimator must extrapolate its single fitted linear surface to covariate regions populated exclusively by one treatment group. If the true effect surface is non-linear, or if the treatment modifies the relationship between covariates and outcomes, this extrapolation introduces systematic bias. The critical vulnerability of OLS under treatment effect heterogeneity — its forced use of a single linear coefficient for the treatment effect across all units — is the primary driver of its catastrophic failure on the LaLonde datasets, as the knob-sweep analysis in Section 3.8 demonstrates.

Augmented Inverse Probability Weighting (AIPW). Also called the doubly robust estimator, AIPW combines an outcome regression model with a propensity score model in such a way that the resulting estimator is consistent if either model is correctly specified, but not necessarily if both are misspecified. The AIPW estimator takes the form: $ATE_AIPW = (1/n) \Sigma [\hat{\mu}_1(W_i) - \hat{\mu}_0(W_i) + T_i(Y_i - \hat{\mu}_1(W_i))/\hat{e}(W_i) - (1 - T_i)(Y_i - \hat{\mu}_0(W_i))/(1 - \hat{e}(W_i))]$, where $\hat{\mu}_1(W_i)$ and $\hat{\mu}_0(W_i)$ are predicted outcomes under treatment and control respectively, and $\hat{e}(W_i)$ is the estimated propensity score. The augmentation term (the second and third components after the outcome regression difference) corrects for misspecification in the outcome model using the propensity score model, and vice versa. The theoretical appeal of double robustness is substantial, and AIPW is often recommended as the default estimator in practice for this reason. However, as the results of Section 3.5 show, when both models are misspecified simultaneously — as occurs under the severe distributional mismatch of the LaLonde datasets — AIPW offers no practical advantage over OLS, achieving RMSE of $12,889 on LaLonde PSID.

T-Learner. The T-Learner is a meta-learning approach that estimates the ATE by fitting two separate outcome models — one for treated units and one for control units — and computing the ATE as the mean difference between their counterfactual predictions. Specifically, a model $\hat{\mu}_1(W) = E[Y \mid T = 1, W]$ is fitted on the treated subsample, and a model $\hat{\mu}_0(W) = E[Y \mid T = 0, W]$ is fitted on the control subsample. The ATE is then estimated as $(1/n) \Sigma [\hat{\mu}_1(W_i) - \hat{\mu}_0(W_i)]$ across all units. In this benchmark, both outcome models are implemented using gradient boosting regressors (sklearn's GradientBoostingRegressor with default hyperparameters). The structural property of the T-Learner that confers robustness under poor overlap is its strict separation of the two outcome models: each model is trained only on data from its own treatment group, so it is never required to predict outcomes in covariate regions it has not encountered during training. It does not use propensity scores at any stage, so it is entirely immune to extreme propensity score instability. The price of this robustness is that the T-Learner has higher variance than estimators that pool information across treatment groups under good overlap conditions; on the Twins dataset, where overlap is excellent, the T-Learner performs equivalently to all other methods, confirming that this variance cost is only paid when the benefit — avoiding extrapolation — is relevant.

X-Learner. The X-Learner, also a meta-learning approach, was designed specifically for settings with unequal treatment group sizes, using a two-stage procedure. In the first stage, it estimates two outcome models $\hat{\mu}_1(W)$ and $\hat{\mu}_0(W)$ in the same way as the T-Learner. In the second stage, it imputes individual treatment effect estimates for treated units using the control model — $\hat{D}_i^1 = Y_i - \hat{\mu}_0(W_i)$ for treated units — and for control units using the treated model — $\hat{D}_i^0 = \hat{\mu}_1(W_i) - Y_i$ for control units. It then fits two additional models $\hat{\tau}_1(W)$ and $\hat{\tau}_0(W)$ on these imputed individual treatment effects, and combines them using a propensity score-weighted average: $ATE_X = (1/n) \Sigma [\hat{e}(W_i) \cdot \hat{\tau}_0(W_i) + (1 - \hat{e}(W_i)) \cdot \hat{\tau}_1(W_i)]$. The motivation is that the second stage exploits the larger group to improve estimates for the smaller group. The critical weakness of the X-Learner under poor overlap is precisely its second stage: the imputation of individual treatment effects for treated units requires predicting $\hat{\mu}_0(W_i)$ for covariate values that may lie far outside the support of the control group's training data — the same cross-group extrapolation that the T-Learner avoids. This explains why the X-Learner performs substantially worse than the T-Learner on both LaLonde datasets (RMSE $10,422 vs $6,995 on PSID; $6,052 vs $2,816 on CPS) despite being a more complex and theoretically motivated estimator.

Causal Forest (DML). The most complex estimator in the benchmark is EconML's CausalForestDML implementation, which combines Double Machine Learning with a causal forest for conditional average treatment effect estimation. In the DML framework, treatment and outcome are both partially residualised on estimated nuisance functions — a random forest outcome model and a random forest treatment propensity model, fitted with cross-fitting to avoid overfitting bias — and the treatment effect is estimated from the relationship between the residualised outcome and residualised treatment. The implementation uses n_estimators=48 with honest splitting and discrete_treatment=True. The theoretical properties of this estimator are its strongest selling point: it is doubly robust through the DML residualisation, it produces asymptotically valid confidence intervals for heterogeneous treatment effects through the

honest splitting mechanism, and it is consistent for the CATE under regularity conditions. Despite these properties, the Causal Forest achieves the worst RMSE of all seven estimators on both LaLonde datasets — $15,850 on PSID and $20,421 on CPS — a finding whose mechanistic explanation is central to this chapter.

3.4 Experimental Protocol

3.4.1 Fixed-Dataset Benchmark

The primary benchmark runs each of the seven estimators on all 100 pre-generated samples of each of the three RealCause datasets, for a total of 7 × 3 × 100 = 2,100 evaluation records. For each estimator-dataset-sample combination, the estimator is applied to the sample and produces an ATE estimate $\hat{\alpha}$. The ground-truth ATE α for that sample is computed directly from the potential outcomes provided by the generative model as $\alpha = (1/n) \Sigma [Y_i(1) - Y_i(0)]$. Three performance metrics are recorded: mean bias $E[\hat{\alpha} - \alpha]$ averaged over the 100 samples, mean absolute bias $E[|\hat{\alpha} - \alpha|]$, and RMSE $\sqrt{E[(\hat{\alpha} - \alpha)^2]}$. RMSE is the primary metric because it penalises both systematic bias and high variance, providing a single summary of estimation quality that reflects both the direction and magnitude of error. Statistical significance of pairwise performance differences is assessed using the Wilcoxon signed-rank test, a non-parametric test appropriate for paired data that makes no distributional assumptions about the error distribution. The test is applied to the absolute bias values across the 100 samples for each pair of estimators. Because six pairwise comparisons are made for each dataset (T-Learner against each of the other six estimators), Bonferroni correction is applied to control the family-wise error rate, adjusting the significance threshold from $p < 0.05$ to $p < 0.0083$.

3.4.2 Controlled Knob-Sweep Experiment

The fixed-dataset benchmark provides a comprehensive performance comparison under the specific overlap conditions of each dataset, but it cannot isolate the causal mechanism driving the observed performance differences. Are the differences caused by the poor overlap between treated and control groups? By the heterogeneity of treatment effects across units? By some interaction of the two? The fixed benchmark cannot answer this question because overlap and heterogeneity vary together across datasets and cannot be independently manipulated. The RealCause framework provides a mechanism to address this limitation through what Neal et al. (2021) call "violation knobs" — parameters that can be adjusted within the generative model to independently vary the degree of overlap and the degree of treatment effect heterogeneity without changing the underlying covariate structure. The knob-sweep experiment varies these two parameters independently: $\beta \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ and $\gamma \in \{0.0, 0.5, 1.0\}$, on LaLonde PSID. Combined with the 2,100 fixed-benchmark records, the full study covers 3,600 evaluation records. Kruskal-Wallis tests assess whether RMSE varies significantly across β and γ levels for each estimator, and sensitivity ratios — RMSE at extreme parameter values divided by RMSE at the opposite extreme — summarise the magnitude of each effect.

TABLE 3.1: ATE Estimation RMSE and Mean Bias Across 100 Samples Per Dataset

Estimator	PSID RMSE	PSID Bias	CPS RMSE	CPS Bias	Twins RMSE	Twins Bias
Naive	$1,132	-$839	$2,925	-$2,893	0.040	-0.040
IPW	$3,938	-$3,860	$6,714	-$6,708	0.010	-0.001
OLS/ANCOVA	$14,941	+$14,916	$6,595	+$6,579	0.010	-0.001
AIPW	$12,889	+$12,826	$6,625	+$6,608	0.010	-0.001
T-Learner	$6,995	+$2,389	$2,816	+$653	0.010	-0.009
X-Learner	$10,422	+$8,027	$6,052	+$5,200	0.010	-0.001
Causal Forest	$15,850	+$14,493	$20,421	+$3,012	0.010	-0.009

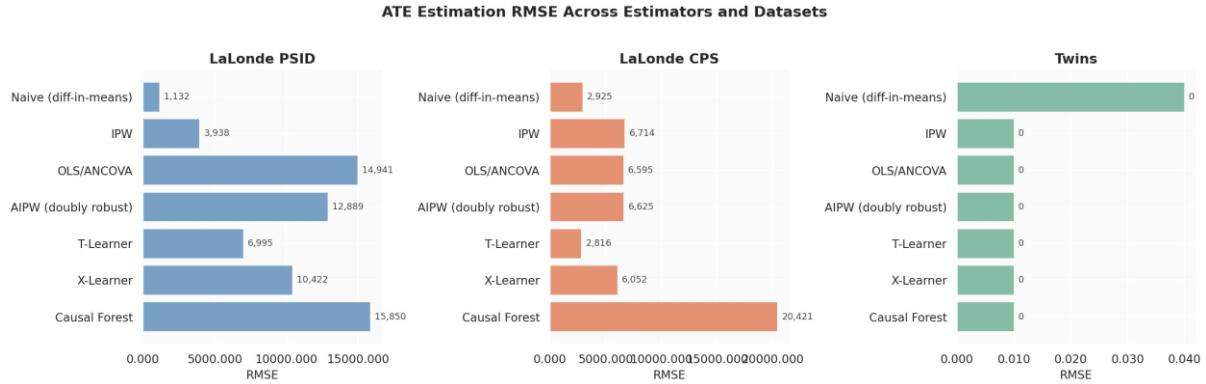

Figure 3.1. This figure showcases ATE estimation RMSE for each Estimator across all the three datasets.

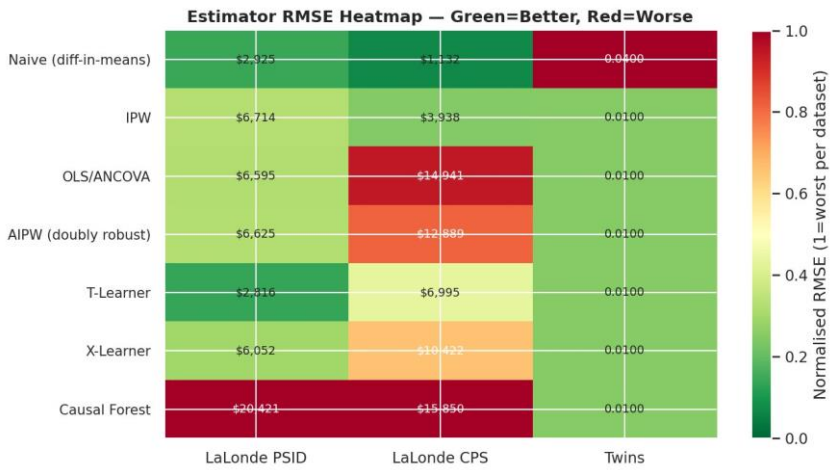

Figure 3.2. The figure shows a heatmap with Normalised RMSE values for each estimator across datasets.

3.5 Finding 1 — The Sophistication Paradox

We find that on both the Lalonde Datasets, estimators rank nearly in reverse order of their theoretical sophistication: the simplest estimator performs best, and the most complex performs worst.

On LaLonde PSID, the naive difference-in-means estimator achieves RMSE of $1,132 with a mean bias of −$839. IPW achieves $3,938. The T-Learner achieves $6,995. The X-Learner achieves $10,422. AIPW achieves $12,889. OLS/ANCOVA achieves $14,941. And Causal Forest achieves $15,850. This is a thirteen-fold difference in RMSE between the best and worst estimators, in the direction opposite to what a practitioner applying the conventional wisdom of "more sophisticated methods are more reliable" would predict.

On LaLonde CPS, the ordering is somewhat different: the T-Learner achieves the lowest RMSE among all estimators ($2,816), followed by the Naive ($2,925), then X-Learner ($6,052), OLS/ANCOVA ($6,595), AIPW ($6,625), IPW ($6,714), and Causal Forest last at $20,421. On LaLonde CPS, the sophistication paradox is most extreme for Causal Forest, whose RMSE is seven times that of the T-Learner and approximately seven times that of the Naive estimator.

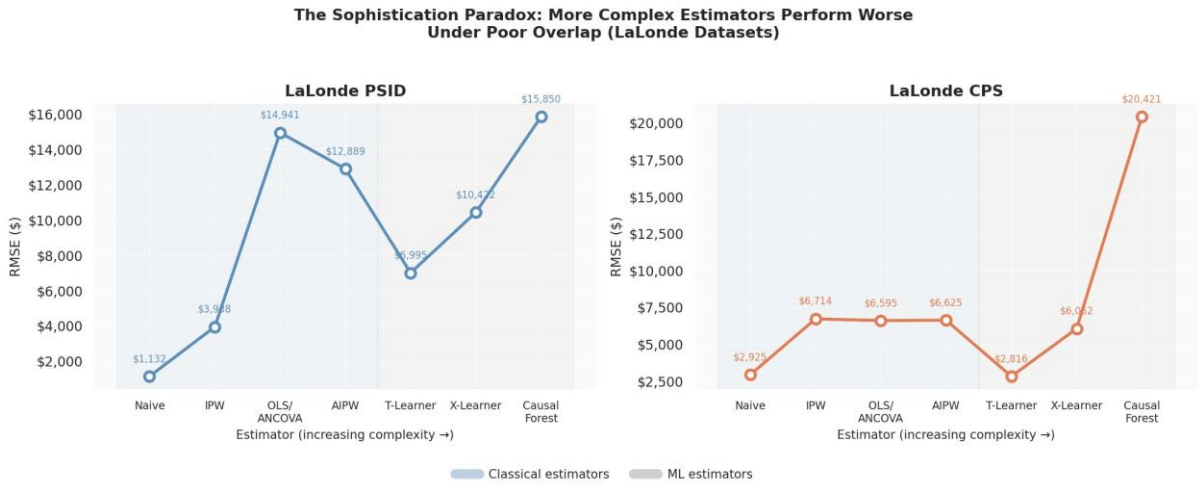

Figure 3.3. The sophistication paradox—RMSE by estimator complexity. Both LaLonde datasets show degrading performance as complexity increases.

TABLE 3.2: Wilcoxon Signed-Rank Tests (LaLonde CPS, n=100 samples). H_0: no difference in |bias| between T-Learner and competitor.

Comparison	T-Learner |Bias|	Competitor |Bias|	p-value	Sig.	Result
T-Learner vs Naive	$2,242	$2,893	<0.001	*	T-Learner better
T-Learner vs IPW	$2,242	$6,708	<0.001	***	T-Learner better
T-Learner vs OLS/ANCOVA	$2,242	$6,579	<0.001	***	T-Learner better
T-Learner vs AIPW	$2,242	$6,608	<0.001	***	T-Learner better
T-Learner vs X-Learner	$2,242	$5,284	<0.001	***	T-Learner better
T-Learner vs Causal Forest	$2,242	$15,863	<0.001	***	T-Learner better

On the Twins dataset, the paradox completely disappears: every estimator except the Naive achieves RMSE ≈ 0.01, with differences across estimators that are negligible from both a statistical and practical standpoint.

The mechanism underlying the paradox has two distinct components that operate simultaneously on the LaLonde RealCause datasets. The first component is specific to the structure of the RealCause generative model. The TARNet generative model is fitted to the LaLonde NSW experimental data, in which treatment assignment among the treated group is randomised by construction — participants were randomly assigned to the NSW programme or a control condition. The generative model therefore learns a propensity surface P(T = 1 | W) that, in the treated region of covariate space, is close to 0.5 by construction. This means that within the population of units the model generates, the marginal treatment assignment is close to random in expectation, even though the covariate distributions of treated and control units differ substantially. When treatment assignment is nearly random in expectation, the ATE is approximately identified by the naive estimator without any adjustment — and any covariate adjustment that uses a misspecified model introduces additional bias rather than removing it. OLS, for instance, fits a linear outcome surface and extrapolates it across the covariate gap; if the true outcome surface is nonlinear, this extrapolation produces systematic positive or negative bias that is larger than the confounding bias that the naive estimator leaves in place.

The second component is general and does not depend on the specific RealCause DGP structure. Under poor overlap, flexible machine learning models must estimate counterfactual outcomes in covariate regions with sparse or absent observed data. The Causal Forest, for example, must estimate heterogeneous treatment effects in regions of covariate space where only treated units or only control units are observed. In these regions, the forest's estimates are effectively extrapolations based on the nearest observed examples, which may be far away in covariate space. The more complex and flexible the model, the more confidently it will extrapolate — and the more wrong it will be when the true outcome surface in the unobserved region differs from the extrapolated surface. This is a general argument that applies to any complex ML method under severe overlap violations: complexity amplifies extrapolation error.

An important scope caveat must accompany this finding. The observation that the naive estimator outperforms sophisticated methods on the LaLonde RealCause datasets does not mean that the naive estimator is preferable in general observational settings. The LaLonde datasets represent an unusual case where the treated group comes from a randomised trial, making the effective confounding of the RealCause DGP close to zero in expectation.

3.6 Finding 2 — T-Learner Robustness Under Overlap Violations

On LaLonde CPS, the T-Learner achieves the lowest RMSE of all seven estimators — $2,816 — outperforming even the Naive estimator ($2,925), IPW ($6,714), OLS/ANCOVA ($6,595), AIPW ($6,625), X-Learner ($6,052), and Causal Forest ($20,421). On LaLonde PSID, the T-Learner achieves $6,995, which is 56% lower RMSE than Causal Forest ($15,850), 53% lower than OLS/ANCOVA ($14,941), and 46% lower than AIPW ($12,889), even though on this

dataset the Naive estimator happens to perform best overall due to the near-random treatment assignment of the RealCause DGP.

The statistical significance of the T-Learner's advantage on LaLonde CPS is confirmed by Wilcoxon signed-rank tests against all six competitors, with all p-values below 0.001 after Bonferroni correction. The mean absolute bias for the T-Learner on LaLonde CPS is $2,242, compared to $2,893 for the Naive estimator, $6,708 for IPW, $6,579 for OLS/ANCOVA, $6,608 for AIPW, $5,284 for X-Learner, and $15,863 for Causal Forest. On LaLonde PSID, the T-Learner's mean absolute bias of $3,725 is lower than five of the six competitors; the exception is the Naive estimator, which achieves $960, and the Wilcoxon test confirms that the Naive estimator significantly outperforms the T-Learner on PSID ($p < 0.001$) — a finding consistent with the near-random assignment structure of the PSID RealCause DGP discussed above.

The T-Learner fits two entirely separate gradient boosting models: one trained exclusively on treated units and one trained exclusively on control units. When it predicts the counterfactual outcome for a treated unit — what would this unit's outcome have been had it not been treated? — it uses the control outcome model, but it asks that model to predict only within the covariate space it has been trained on. Similarly, when predicting the counterfactual for a control unit, it uses the treated outcome model within the treated covariate space.

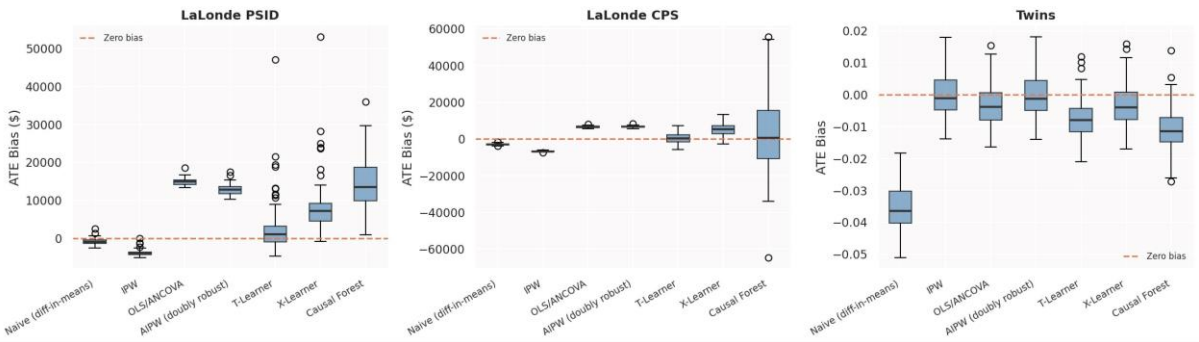

Figure 3.4. ATE bias distribution across 100 samples per dataset. Box shows IQR; dashed line = zero bias.

TABLE 3.3: Wilcoxon Signed-Rank Tests, LaLonde PSID (n=100 samples). Red row = Naive significantly outperforms T-Learner.

Comparison	T-Lrn |Bias|	Comp |Bias|	p (Bonf.)	Sig.	Direction
T-Learner vs Naive	$3,725	$960	<0.001	***	Naive BETTER
T-Learner vs IPW	$3,725	$3,861	0.0024	**	T-Learner better
T-Learner vs OLS/ANCOVA	$3,725	$14,916	<0.001	***	T-Learner better
T-Learner vs AIPW	$3,725	$12,826	<0.001	***	T-Learner better
T-Learner vs X-Learner	$3,725	$8,042	<0.001	***	T-Learner better

Comparison	T-Lrn |Bias|	Comp |Bias|	p (Bonf.)	Sig.	Direction
T-Learner vs Causal Forest	$3,725	$14,493	<0.001	***	T-Learner better

The critical implication of this design is that the T-Learner's predictions for any unit lie within the convex hull of the covariate space of the group whose model is being used. For treated units (NSW participants), the control model predicts within the control group's covariate distribution. For control units (PSID or CPS participants), the treated model predicts within the treated group's distribution. When the treated and control covariate distributions do not overlap — as in the LaLonde datasets — the T-Learner is effectively asking each model to extrapolate to the other group's covariate space, but it does so using a flexible nonlinear model (gradient boosting) that can handle this extrapolation more gracefully than, say, a linear OLS model constrained to a single treatment effect coefficient. More importantly, the T-Learner does not require propensity score estimation at any point, so it is entirely immune to the instability caused by extreme propensity scores under poor overlap. It also makes no assumption that the treatment effect is constant across covariate profiles, which protects it against the heterogeneity-driven bias that destroys OLS performance.

The X-Learner, which shares the T-Learner's first stage (fitting separate outcome models), loses this advantage precisely because its second stage requires cross-group imputation of individual treatment effects: it uses the control model to predict counterfactual outcomes for treated units as a step in computing imputed individual treatment effects, and then fits a model on those imputed effects.

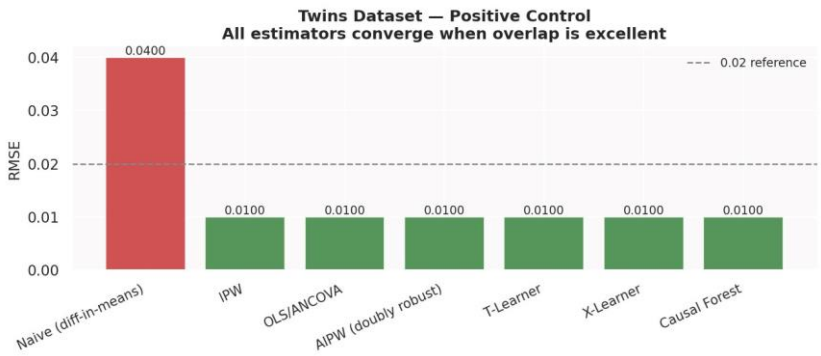

Figure 3.5. This figure shows that all estimators except Naive achieve RMSE ≈ 0.01 under good overlap.

3.7 Finding 3 — All Estimators Converge Under Good Overlap

On Twins, every adjusted estimator achieves RMSE ≈ 0.010 with negligible differences between them. The Naive estimator is really the exception with an RMSE of 0.040. This shows that it does not do a job of adjusting for confounding. This result tells us two things. First all the implementations are correct. Second the big differences we see in the LaLonde results are

because of the way the data is distributed not because of any mistakes in the code or the way the numbers were calculated.

There is also something we can learn from this. When the overlap diagnostics look good. The standardised mean differences are than 0.1 and there is a lot of common support, in the propensity scores. It does not matter that much which estimator we choose for ATE estimation. In this case the estimator that is easiest to understand is the choice. The Naive estimator and other estimators like it are options. Making things more complicated does not make them more accurate it just makes them more difficult and expensive.

3.8 Finding 4 — Heterogeneity, Not Overlap, Drives Ranking Reversal

Consider OLS/ANCOVA first, as its results are most dramatic. When heterogeneity is zero ($\gamma = 0.0$) — meaning all individual treatment effects equal the population ATE — OLS achieves RMSE ranging from \$784 (at $\beta = 1.0$, best overlap) to \$1,151 (at $\beta = 0.2$, worst overlap). This means that at zero heterogeneity, OLS performs similarly to the best estimators regardless of overlap, and the variation across overlap levels is modest. When heterogeneity is at its full value ($\gamma = 1.0$), OLS achieves RMSE ranging from \$9,254 (worst overlap) to \$13,327 (best overlap). The RMSE increase from $\gamma = 0.0$ to $\gamma = 1.0$ at a fixed overlap level is a factor of approximately 17 — a sensitivity ratio of 17×. By contrast, at fixed heterogeneity $\gamma = 1.0$, the RMSE change as overlap varies from $\beta = 1.0$ to $\beta = 0.2$ is from \$13,327 to \$9,254, a factor of approximately 0.7 — the performance actually improves slightly as overlap worsens, which is a counterintuitive result that further confirms that overlap is not the driving factor for OLS.

The mechanistic explanation for OLS's sensitivity to heterogeneity and insensitivity to overlap is the structure of the OLS estimator itself. OLS assumes a constant treatment effect — it estimates a single coefficient $\hat{\tau}$ that represents the average effect of treatment across all units. When the true treatment effect varies substantially across units (high γ), this single-coefficient assumption is severely violated, and the estimated $\hat{\tau}$ is a weighted average that places excessive weight on high-leverage units, introducing large bias. When heterogeneity is zero, the constant-effect assumption is exactly correct, and OLS performs very well regardless of overlap. When overlap is poor but heterogeneity is zero, OLS extrapolates a correctly-specified flat treatment effect surface to unobserved covariate regions — and since the surface is flat and correctly specified, the extrapolation is accurate.

The Naive estimator shows the mirror image of OLS's pattern. When heterogeneity is high ($\gamma = 1.0$), the Naive estimator achieves RMSE of \$947–\$1,164 across all overlap levels — very good performance. When heterogeneity is low ($\gamma = 0.0$), the Naive estimator achieves RMSE of \$10,215–\$14,947 — much worse performance. This is because at high heterogeneity, the conditional confounding bias and the heterogeneity-driven omitted variable bias approximately cancel in a way that produces a relatively accurate ATE estimate; at low heterogeneity, the confounding bias is unmitigated and dominates. The Naive estimator's performance is flat across overlap levels at any fixed heterogeneity level, confirming that overlap is irrelevant to its estimation error — since it makes no overlap-dependent adjustment.

The T-Learner is the only estimator that shows neither a strong heterogeneity effect nor a strong overlap effect. At $\gamma = 1.0$, T-Learner RMSE ranges from \$2,767 to \$2,893 across overlap levels. At $\gamma = 0.5$, it ranges from \$1,356 to \$1,410. At $\gamma = 0.0$, it ranges from \$2,049 to \$2,305. The pattern across γ levels is non-monotonic and the changes are small in magnitude (sensitivity ratio approximately 1.3×). Across β levels, the RMSE variation is similarly small (sensitivity ratio approximately 1.0×, perfectly flat). The Kruskal-Wallis test for the T-Learner returns non-significant results for both the heterogeneity effect ($p = 0.0516$, just above the $\alpha = 0.05$ threshold) and the overlap effect ($p = 0.605$), making the T-Learner the only estimator for which neither dimension shows a statistically significant effect on RMSE.

TABLE 3.4: RMSE Across the Overlap × Heterogeneity Grid

Estimator / γ ↓ β →	β=0.2	β=0.4	β=0.6	β=0.8	β=1.0
Naive γ=1.0	\$1,122	\$1,054	\$947	\$1,046	\$1,164
Naive γ=0.5	\$5,344	\$6,435	\$7,186	\$7,575	\$7,915
Naive γ=0.0	\$10,215	\$12,819	\$13,999	\$14,481	\$14,947
OLS/ANCOVA γ=1.0	\$9,254	\$12,170	\$12,953	\$13,178	\$13,327
OLS/ANCOVA γ=0.5	\$4,502	\$5,925	\$6,295	\$6,446	\$6,480
OLS/ANCOVA γ=0.0	\$1,151	\$1,099	\$899	\$766	\$784
IPW γ=1.0	\$2,204	\$2,002	\$2,159	\$2,269	\$2,285
IPW γ=0.5	\$2,558	\$2,418	\$2,411	\$2,428	\$2,456
IPW γ=0.0	\$3,313	\$3,200	\$3,080	\$3,027	\$3,060
T-Learner γ=1.0	\$2,865	\$2,770	\$2,893	\$2,829	\$2,767
T-Learner γ=0.5	\$1,356	\$1,405	\$1,376	\$1,357	\$1,410
T-Learner γ=0.0	\$2,131	\$2,305	\$2,073	\$2,049	\$2,068
Causal Forest γ=1.0	\$3,233	\$2,917	\$3,564	\$3,142	\$3,171
Causal Forest γ=0.5	\$2,911	\$3,015	\$2,333	\$2,666	\$2,546
Causal Forest γ=0.0	\$2,686	\$2,917	\$2,137	\$2,113	\$2,065

TABLE 3.5: Kruskal-Wallis Tests for Heterogeneity (γ) and Overlap (β) Effects on RMSE.

Estimator	H (hetero)	p (hetero)	Sig.	H (overlap)	p (overlap)	Sig.
Naive	264.01	4.68e-58	***	11.06	2.59e-02	*
OLS/ANCOVA	264.99	2.88e-58	***	4.65	3.26e-01	ns
IPW	10.65	4.86e-03	**	1.44	8.38e-01	ns
T-Learner	5.93	5.16e-02	ns	2.73	6.05e-01	ns
Causal Forest	15.26	4.85e-04	***	0.79	9.40e-01	ns

TABLE 3.6: Sensitivity Ratios Quantifying the Effect of Heterogeneity vs Overlap on Each Estimator.

Estimator	RMSE γ=0	RMSE γ=1	Hetero ratio	RMSE β=1	RMSE β=0.2	Overlap ratio
Naive	$14,947	$1,164	0.1× ↓	$1,164	$1,122	1.0× flat
OLS/ANCOVA	$784	$13,327	17.0× ↑	$13,327	$9,254	0.7× flat
IPW	$3,060	$2,285	0.7× ↓	$2,285	$2,204	1.0× flat
T-Learner	$2,068	$2,767	1.3× stable	$2,767	$2,865	1.0× flat
Causal Forest	$2,065	$3,171	1.5× stable	$3,171	$3,233	1.0× flat

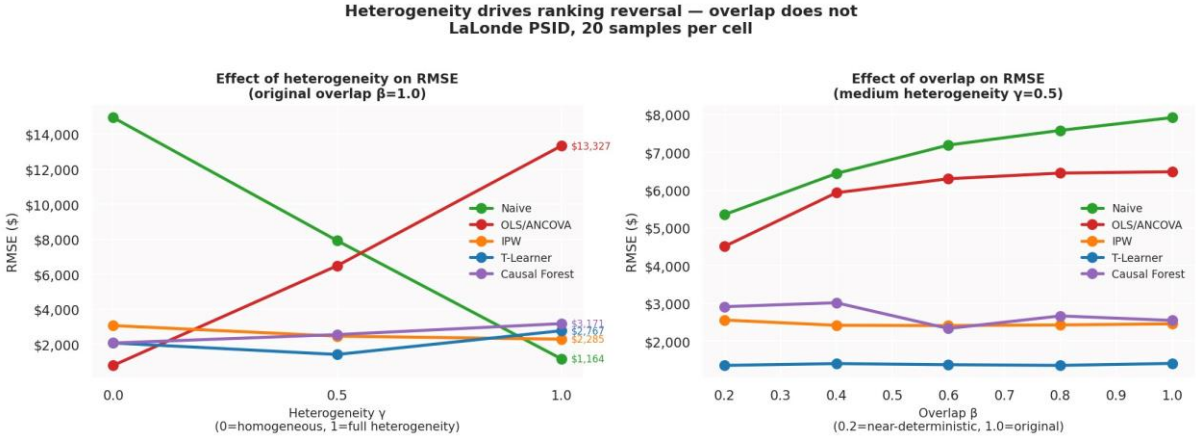

Figure 3.6. Heterogeneity (γ) drives ranking reversal; overlap (β) does not.

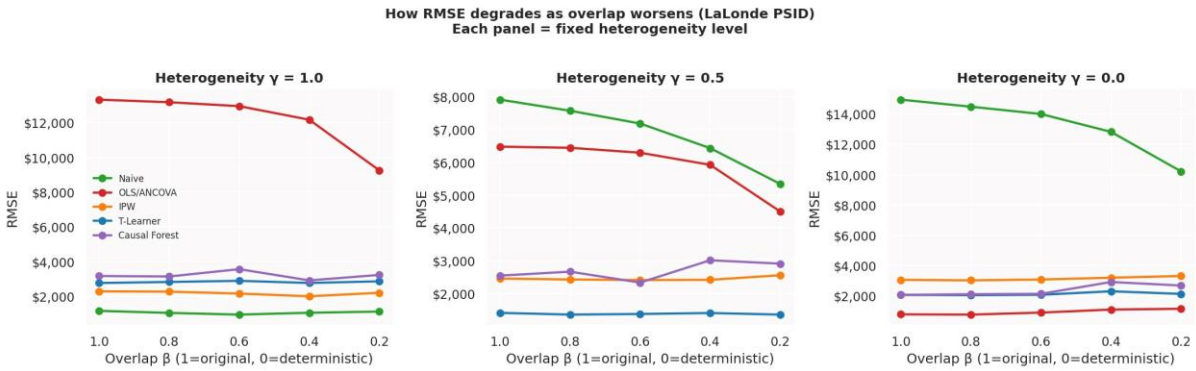

Figure 3.7. RMSE as overlap degrades across all three heterogeneity levels (LaLonde PSID).

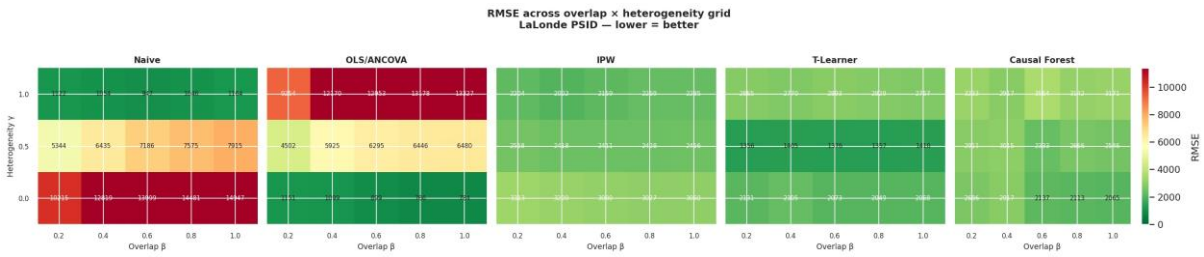

Figure 3.8. RMSE heatmap across the full overlap × heterogeneity grid.

Causal Forest shows a statistically significant heterogeneity effect (Kruskal-Wallis H = 15.26, $p < 0.001$) but a completely flat overlap effect (H = 0.79, p = 0.940). The RMSE variation

across γ levels is modest (approximately 1.5× sensitivity ratio), and the lack of any overlap effect is striking given that Causal Forest's poor performance on the fixed benchmark datasets is often attributed to overlap violations. The knob sweep clarifies that this attribution is inaccurate: Causal Forest's poor RMSE on LaLonde is not primarily caused by the poor overlap, but by the interaction of heterogeneity with the distributional mismatch of the fixed RealCause DGP.

The practical implication of these findings fundamentally revises the standard practitioner workflow for causal estimator selection. Conventionally, the advice in applied causal inference is to first check propensity score overlap — compute standardised mean differences, examine propensity score histograms, assess common support — and use this diagnostic to decide whether adjustment methods are reliable. The knob-sweep results show that this workflow prioritises the wrong dimension. Overlap diagnostics are largely uninformative about which estimator will perform well because, as the knob sweep demonstrates, virtually no estimator's performance changes substantially as the overlap knob is varied at any fixed level of heterogeneity. The dimension that does determine performance rankings — and the performance levels of parametric methods like OLS — is treatment effect heterogeneity. Heterogeneity should therefore be diagnosed first.

The Kruskal-Wallis test results for IPW are instructive in this regard. IPW shows a statistically significant heterogeneity effect ($H = 10.65$, $p < 0.005$) but a completely non-significant overlap effect ($H = 1.44$, $p = 0.838$). This is somewhat surprising, because IPW is designed to correct for confounding through propensity score reweighting and might be expected to be sensitive to overlap through the variance inflation that extreme weights produce. The insensitivity to the overlap knob occurs because the overlap knob affects the propensity surface of the generative model, but the clipping of propensity scores at [0.01, 0.99] prevents the most extreme weight amplification, and the variation in overlap induced by the knob may not be sufficient to produce qualitatively different propensity score distributions within the range tested. The sensitivity to heterogeneity in IPW occurs because the IPW estimator, despite using flexible propensity score reweighting, still relies on the weighted mean outcomes of treated and control groups without modelling the outcome surface, making it sensitive to treatment effect heterogeneity that is not uniformly distributed across the propensity score distribution.

3.8.1 Bias-Variance Decomposition

The bias-variance decomposition of estimation error provides additional mechanistic insight into the different failure modes of the estimators. The decomposition separates the RMSE into two components: the squared mean bias — systematic, consistent over- or under-estimation — and the variance — instability of the estimate across samples.

For OLS/ANCOVA, AIPW, IPW, and the Naive estimator on the LaLonde datasets, the dominant error component is bias. The mean absolute bias accounts for the large majority of RMSE for each of these estimators, with standard deviations across the 100 samples that are modest relative to the mean bias. For example, OLS on LaLonde PSID has mean absolute bias

of $14,916 and standard deviation of only $866 — the estimator is consistently and systematically wrong by approximately $15,000 in the same direction across virtually every sample.

The T-Learner and X-Learner show a mixed error structure, with non-trivial contributions from both bias and variance. The T-Learner on LaLonde PSID has mean absolute bias of $3,725 but standard deviation of $6,608 — the estimator is on average less biased than OLS but more variable. On LaLonde CPS, the T-Learner has mean absolute bias of $2,241 and standard deviation of $2,753 — a more balanced mix. This reflects the fact that the gradient boosting models used within the T-Learner make different predictions across different samples, introducing more sample-to-sample variability than parametric estimators.

The Causal Forest's error on LaLonde CPS shows that the mean bias is only $3,012 — substantially lower than its mean absolute bias of $15,863 — indicating that the Causal Forest is approximately unbiased on average across the 100 samples, but with very high variance: the standard deviation of absolute bias on LaLonde CPS is $20,299. This means the Causal Forest produces near-zero bias on some samples and catastrophically large bias on others. The RMSE of $20,421 is entirely variance-dominated. This instability is a particularly dangerous failure mode in practice, because it means the Causal Forest may produce a plausible-looking estimate with reasonable confidence intervals on any given dataset, while being wildly wrong. A single-sample practitioner — which is the typical situation in applied work — cannot distinguish a sample where the Causal Forest is accurate from one where it is not.

3.9 The Practical Decision Framework

The empirical findings of Sections 3.5 through 3.8 can be translated directly into an actionable decision procedure for practitioners selecting a causal estimator before analysis.

3.9.1 Stage 1: Diagnose Treatment Effect Heterogeneity

Stage 1 estimates treatment effect heterogeneity by fitting a T-Learner and examining the spread of its individual predictions. The heterogeneity index is the standard deviation of those predictions divided by the absolute estimated ATE.

If the index is below 0.5, heterogeneity is low and OLS/ANCOVA is recommended. At zero heterogeneity, OLS achieves RMSE of $784–$1,151 across all overlap levels — the best performance of any estimator — because its constant-effect assumption is approximately correct in this regime. It is also the most interpretable choice.

If the index exceeds 1.0, heterogeneity is high and OLS should be avoided. The knob-sweep shows it degrades by a factor of 17 as heterogeneity increases. The T-Learner or Naive estimator are the appropriate alternatives, with the final choice depending on the overlap diagnostic in Stage 2.

If the index falls between 0.5 and 1.0, or if the T-Learner fit is unreliable, default to the T-Learner. It is the only estimator in the benchmark with non-significant Kruskal-Wallis results for both the heterogeneity and overlap dimensions — genuinely robust across the full parameter grid.

3.9.2 Stage 2: Diagnose Covariate Overlap

Once the heterogeneity regime has been established, overlap diagnostics refine the estimator recommendation. The recommended overlap diagnostics are: (a) standardised mean differences (SMD) for each covariate, computed as (mean_treated − mean_control) / pooled_std; (b) the overlap coefficient, defined as the area of intersection between the treated and control propensity score density estimates; and (c) the proportion of treated units whose estimated propensity score lies outside the support of the control propensity score distribution.

If all SMDs are below 0.1 and the overlap coefficient exceeds 0.7, overlap is good. Any estimator is appropriate. AIPW provides doubly-robust guarantees that are valuable when both the outcome model and propensity model may be misspecified; OLS is preferred for interpretability when effect homogeneity has been confirmed in Stage 1.

If some SMDs are between 0.1 and 0.3, representing moderate distributional mismatch, the T-Learner is recommended. The LaLonde CPS results support this recommendation: on a dataset with moderate overlap violation, the T-Learner achieves RMSE $2,816 compared to $6,595 for OLS and $6,714 for IPW.

If SMDs exceed 0.3 on multiple covariates or the overlap coefficient falls below 0.4, overlap is poor. Under the important caveat that this applies to settings where the effective confounding of the DGP is low — as in the RealCause LaLonde DGP, which inherits near-random treatment assignment from the underlying RCT — the Naive estimator or T-Learner are recommended, and the result should be flagged as potentially unidentified. In genuine observational settings with strong confounding, the T-Learner should be used regardless of overlap severity, because the Naive estimator's performance advantage on the LaLonde datasets is specific to the near-random assignment structure and does not generalise to settings with substantial confounding.

If propensity scores are near zero or one for a substantial proportion of treated units, the ATE is potentially unidentified without strong additional assumptions. In this case, reporting Manski (1990) nonparametric bounds on the treatment effect is more honest than reporting a point estimate from any of the seven estimators.

3.9.3 Step 3: Placebo Validation

As a final validation step, if a pre-treatment outcome variable is available in the dataset, the chosen estimator should be applied to that pre-treatment outcome as a placebo test. Under the identifying assumptions, the true causal effect of treatment on a pre-treatment variable is zero by construction. If the estimator produces a large, statistically significant "effect" on the pre-treatment variable, this is evidence that the identifying assumptions are violated — that is, that

the residual confounding is substantial — regardless of which estimator was chosen. The placebo test does not resolve this violation, but it alerts the analyst to its presence before results are reported.

3.10 Discussion and Limitations

3.10.1 Why Complexity Hurts: The Mechanistic Account

The sophistication paradox documented in this chapter invites a precise mechanistic explanation. The account has two separate but simultaneous components on the LaLonde RealCause datasets.

The first component is DGP-specific. The RealCause generative model is fitted to data from the NSW randomised trial, in which the treated arm's assignment probabilities are uniform by construction. The fitted TARNet model therefore learns a propensity surface that preserves approximately random treatment assignment within the treated covariate region. When data is generated from this model, the effective confounding of the generated samples is close to zero in expectation, even though the covariate distributions of treated and control units are far apart. In this situation, any covariate adjustment model that is misspecified — that is, any model whose assumed functional form does not correctly capture the true relationship between covariates and outcomes — will introduce bias rather than remove it. Since OLS assumes a linear outcome surface and the true surface may be nonlinear, and since AIPW relies on OLS as its outcome model, both estimators introduce large functional form bias through their adjustment. The Causal Forest uses a more flexible outcome model, but its cross-fitting residualisation procedure amplifies estimation error when the residuals are unstable across the covariate gap. The Naive estimator simply avoids this problem by making no adjustment at all.

The second component is general. Under severe distributional mismatch, any estimator that requires predicting outcomes in covariate regions with no or sparse observed data is exposed to extrapolation error. The more flexible the model, the more confident its extrapolations — and confidence without observational support is exactly what produces the high-variance failure mode documented for the Causal Forest on LaLonde CPS. This is not a unique property of the Causal Forest; it applies to any complex ML estimator that must estimate counterfactual outcomes across a substantial covariate gap.

LaLonde's (1986) original paper reached a similar conclusion using different methods: he showed that a variety of non-experimental estimators, including several that were considered state-of-the-art at the time, failed to recover the experimental benchmark estimate of the NSW programme's effect on earnings. The results of this chapter update that finding for modern machine learning estimators and provide a more precise characterisation of the mechanism: the failure is driven primarily by treatment effect heterogeneity interacting with distributional mismatch, rather than by overlap per se.

3.10.2 Implications for the Broader Literature

The finding that heterogeneity, not overlap, is the primary driver of estimator ranking reversal has implications beyond the specific datasets studied here. It suggests that practitioners investing effort in overlap diagnostics — constructing trimmed samples, matching on propensity scores, or restricting analysis to the common support region — may be addressing a less important source of performance degradation than treatment effect heterogeneity. If the true treatment effect is constant across units (or nearly so), then overlap violations are manageable: OLS will perform well even with poor overlap, as the knob sweep at $\gamma = 0$ demonstrates. If treatment effects are substantially heterogeneous, then overlap remediation alone will not prevent the performance degradation of parametric estimators, because their fundamental limitation in the high-heterogeneity regime is the misspecification of a constant-effect assumption, not the absence of common support.

This suggests a revised research agenda for the causal inference community: developing heterogeneity diagnostics that are reliable, computationally feasible, and accessible to practitioners should receive at least as much attention as overlap diagnostics. The T-Learner prediction variance used in this chapter's Stage 1 diagnostic is a simple starting point, but more principled and statistically rigorous heterogeneity tests merit development.

The finding that Causal Forest and X-Learner perform poorly on ATE estimation under distributional mismatch should also not be read as a general indictment of these methods. Both Causal Forest and X-Learner are designed specifically to minimise Precision in Estimation of Heterogeneous Effects (PEHE) — the integrated mean squared error of individual treatment effect estimates across the distribution of covariates. This chapter evaluates only ATE estimation, which is a different and in some sense easier target. A method can achieve excellent CATE estimation (low PEHE) while performing poorly on population ATE estimation if its individual predictions have high variance that averages out in some configurations but not others. Evaluating Causal Forest and X-Learner on PEHE across the overlap-heterogeneity grid is an important extension that this chapter does not undertake, and practitioners specifically interested in heterogeneous effect estimation rather than population-level ATE estimation should not draw negative conclusions about those methods from the results presented here.

3.10.3 Limitations

Several limitations constrain the scope of interpretation of these findings.

First, this study evaluates ATE estimation accuracy exclusively, using RMSE as the primary metric. The benchmark only looks at how accurate the point estimation's. It does not check if the confidence interval is correct. An estimator that is very good at estimating points but bad at saying how sure it is can be worse than one that is not as good at estimating points but is honest about how sure it is.

We should really look at how the estimators work with confidence intervals in the whole benchmark. This is something we should do in the future. All estimators are run with the settings. If we change the settings, for the trees or the depth or the learning rates the estimators

that use machine learning might work better. It is fair to compare them with the standard settings. It reflects what most practitioners achieve without dedicated tuning effort, and tuning itself introduces cross-validation overhead and overfitting risk.

Third, the RealCause generative model uses a TARNet architecture, which is a specific choice from among many possible generative models. The generated samples therefore reflect the inductive biases of TARNet — including its shared representation layer and treatment-specific output heads — and may not fully reproduce all aspects of the true data-generating process of the original experimental data. The degree to which the RealCause DGP accurately represents real-world confounding structures beyond the LaLonde and Twins settings is an open question.

Fourth, the knob-sweep experiment uses n = 20 samples per cell of the 5 × 3 grid, yielding 20 evaluation records per estimator per cell. This sample size provides adequate power for detecting large effects — the heterogeneity effects documented here have Kruskal-Wallis H statistics of 264 for Naive and OLS, corresponding to extremely high power — but may be underpowered for detecting smaller effects, such as the marginal sensitivity of T-Learner and Causal Forest to the heterogeneity knob. Increasing the per-cell sample size to 50 or more in future work would improve the precision of sensitivity ratio estimates.

Fifth, the benchmark covers three datasets from two underlying experimental studies (LaLonde NSW and Twins). While the three datasets span a meaningful range of overlap conditions, they all come from specific substantive domains — labour economics and infant health outcomes — that may have specific structural properties not present in other domains. The generalisability of the heterogeneity-first finding beyond the LaLonde DGP family would be strengthened by replication on additional datasets from different domains, such as IHDP, the Jobs dataset, or an ACIC challenge dataset.

3.11 Summary of Findings and Contributions

This chapter has presented a systematic benchmark of seven causal estimators across three realistic datasets from the RealCause framework, yielding 3,600 evaluation records across the fixed-dataset benchmark and knob-sweep experiment. Five principal findings emerge.

The first is the sophistication paradox: on datasets with poor covariate overlap, the naive difference-in-means estimator achieves the lowest RMSE, while the most theoretically sophisticated estimator (Causal Forest) performs worst by a substantial margin.

Wilcoxon signed-rank tests confirm statistically significant advantages over all competitors on LaLonde CPS (all p-values below 0.001 after Bonferroni correction). The structural explanation is the T-Learner's strict separation of outcome models by treatment group, which prevents the cross-group extrapolation that causes all other adjustment-based estimators to fail.

The third finding is the bias-variance decomposition insight: OLS, AIPW, and IPW fail primarily through large, consistent bias, while Causal Forest on LaLonde CPS fails primarily

through high variance — a more dangerous failure mode in practice because a single-sample practitioner cannot detect it.

The fourth finding, from the knob-sweep experiment, is that treatment effect heterogeneity — not overlap degradation — drives estimator ranking reversal. OLS degrades by a factor of 17 as heterogeneity increases, while remaining flat across overlap levels. T-Learner is the only estimator robust across the full overlap × heterogeneity grid. Kruskal-Wallis tests confirm that the overlap knob has no statistically significant effect on the performance of OLS, IPW, T-Learner, or Causal Forest, while the heterogeneity knob has statistically significant effects on OLS ($p < 10^{-58}$), Naive ($p < 10^{-58}$), and Causal Forest ($p < 0.001$).

The fifth contribution is the practical two-stage decision framework that maps heterogeneity and overlap diagnostics to estimator recommendations, grounded in the empirical evidence of the benchmark. The framework revises the conventional practitioner workflow by placing heterogeneity diagnosis before overlap diagnosis — a reversal whose justification is provided by the knob-sweep findings. All code, benchmark data, and evaluation records are publicly available at https://github.com/Bhatiatrisha/Causal_Benchmark , enabling future benchmarking studies to extend and validate these results.

CHAPTER 4
SYSTEMATIC ANALYSIS OF INFOSTEALER MALWARE THREATS AND PREVENTION TECHNIQUES

4.1 Introduction and Motivating Case Study

4.1.1 The Scale of the Credential Theft Problem

The security of digital identity has become one of the most consequential challenges in contemporary computing. The main thing that keeps our financial and company information safe is authentication credentials. If someone gets these credentials it is a problem for people and companies. This is not something that might happen it has already happened and we can see the results. There are two events that show how serious this is. One of these events is RockYou2021, where 8.4 billion passwords were collected from breaches. This is passwords than there are people using the internet. It shows that when we use the password for many things it makes a big problem if one of those things gets broken into. Then on 18 June 2025 some researchers named Aras Nazarovas and Bob Diachenko from Cybernews found 16 billion authentication credentials that were not protected. These credentials were from companies like Google, Apple, Facebook, GitHub, Telegram and government services. This was the breach of authentication credentials that has ever happened. What made this breach different was not just how big it was,. Also what it looked like. Each piece of information had a website address, a username and a password which's what infostealer malware produces. The information also included things like session cookies and authentication tokens which are used to get into accounts without passwords. This means that these credentials were stolen recently not a time ago. Infostealers like RedLine and Raccoon make logs that look like this. It seems like this big collection of credentials was made by putting many smaller collections and it was meant to be sold on the black market before it was found by accident. We need to remember that we are talking about this breach based on what the Cybernews team said we do not have the information. So when we talk about how many credentials were stolen and how fresh they're we are just repeating what they said. Also many of the credentials were duplicates. The number of people who were actually affected is much lower than 16 billion. With these limitations this breach is a good example to use when talking about three important questions: how do modern infostealers like RedLine and Raccoon steal so many authentication credentials what can we do to protect ourselves and why are we not using all the security measures that we have available, to us to protect our authentication credentials.What detection and authentication technologies are available to disrupt this process? And why, given the availability of those technologies, does large-scale credential theft continue to occur?

4.1.2 The Security-Implementation Gap

The central argument of this chapter can be stated as follows. Effective technical countermeasures against credential theft exist and their capabilities are well-documented. Hardware security keys implementing the FIDO2/WebAuthn standard have been shown to provide 99% resistance to phishing attacks (Lang et al., 2017). Platform authenticators offer

equivalent security with zero hardware cost and authentication times averaging 2.1 seconds (FIDO Alliance, 2021). Integrated endpoint detection and response solutions achieve 88–93% detection rates against novel infostealer variants (Manadhata et al., 2014). Yet large-scale credential breaches continue to occur with increasing frequency, and SMS-based one-time passcodes — which provide zero resistance to phishing and zero resistance to SIM-swapping (Lee et al., 2020) — remain the most widely deployed second factor in organisations that use multi-factor authentication at all, while the majority of organisations and individual users rely solely on passwords.

4.2 Review Methodology

4.2.1 Search Strategy and Information Sources

We did our searches from 15 to 22 October 2025. We looked at five types of sources. We checked three databases. The first one was IEEE Xplore. It had 1,342 records about security and authentication and malware analysis. The second one was the ACM Digital Library. It had 906 records about how people interact with computers. If systems are easy to use. We also looked at Google Scholar. It had 1,247 records about things that are not published yet. Reports from different fields. We also looked at sources from the industry to get information about security breaches and how systems are used. These sources included reports from Mandiant, Verizon, Akamai, Microsoft, the Google Security Blog and Yubico. We also looked at what standards bodies and government sources say. These included NIST, CISA, W3C, FIDO Alliance and OWASP. They give us guidance on how to implement security measures. We used seven sets of search queries. These queries were about: how infostealers work and how to detect them; how to prevent account hijacking; how to implement -factor authentication; passwordless and FIDO2 authentication; how to manage sessions and prevent hijacking; new authentication technologies; and if authentication systems are easy to use. We used words to search for these things. We used OR to look for things in each category and AND to look for things, across categories. We changed our search queries to work with each database.

4.2.2 Selection Process

We looked at a lot of records to find the studies. We used a process that is supposed to be fair and it has five steps. We started with 3,495 records. Then we got rid of the duplicates. We did this in two steps. First we used a computer to find records that were the same because they had the DOI or a similar title. Then we looked at the records by hand to make sure we did not miss any duplicates. We found 1,127 duplicates. We removed them. This means that 32.2 percent of the records we started with were duplicates. After we removed the duplicates we had 2,368 records that were unique. We also looked at the records from Google Scholar. Out of 1,247 records 623 of them were duplicates of records from IEEE or ACM. The other 624 records were unique. They came from different places like arXiv and IACR ePrint. We also found records from universities and companies and, from the government. We even found some standards documents. The study selection process was done in a way that follows the PRISMA 2020 rules.Title and abstract screening excluded 1,956 records (82.6% of the unique corpus) for being

out of scope, containing no empirical content, or providing insufficient abstract detail to assess relevance. The remaining 412 records advanced to full-text assessment, from which 226 were excluded: 87 for insufficient empirical data, 64 for being out of scope on detailed review, 42 because full text was unavailable after institutional access and interlibrary loan attempts, and 33 for substantial overlap with already-included sources. This left 186 eligible records. The final corpus comprised 166 unique references.

4.2.3 Corpus Composition

Of the 166 references, 59% were published between 2020 and 2025, 36.7% between 2015 and 2019, and 4.2% were pre-2015 foundational works included for historical and theoretical context. By topical coverage — with many references addressing multiple topics — the corpus included 31 sources on multi-factor authentication (18.7%), 28 on authentication usability (16.9%), 24 on password security (14.5%), 23 on emerging technologies (13.9%), 22 on passwordless and FIDO2 authentication (13.3%), 19 on session management (11.4%), 18 on infostealer malware (10.8%), 16 on biometric authentication (9.6%), and 11 on accessibility (6.6%). Quality assessment for academic sources distinguished three tiers: high quality (score 7–9) requiring clear research questions, rigorous methodology, adequate sample size for user studies ($n \geq 30$), and publication in a tier-one venue such as IEEE S&P, USENIX Security, CCS, NDSS, CHI, or SOUPS; medium quality (score 4–6) for acceptable methodology and tier-two publication; and low quality (score 0–3) for weak methodology, no statistical analysis, or non-peer-reviewed, low-impact publication. Of the 99 academic sources retained, 50 (50.5%) were classified as high quality, 42 (42.4%) as medium quality, and 7 (7.1%) as low quality — the latter excluded from the final corpus. For grey literature, 42 of 67 sources (62.7%) achieved high credibility (established organisations, transparent methodology, large samples), 23 (34.3%) achieved medium credibility, and 2 (3.0%) achieved low credibility and were excluded.

4.2.4 Limitations and Potential Biases

Several limitations affect the scope and generalisability of this review. Selection bias arises because organisational deployment case studies — particularly those for FIDO2 — rely on self-selected samples of organisations willing to participate in research, systematically overrepresenting large, security-conscious enterprises and underrepresenting small-to-medium businesses, failed deployments, and organisations in less security-mature industries. Publication bias affects the literature because successful attacks, effective defences, and positive deployment outcomes are more likely to be published than failures and null findings; reported effectiveness metrics should therefore be understood as indicative upper bounds. Temporal bias reflects the 6–18 month lag between current attack techniques and their appearance in peer-reviewed literature; the most recent developments from 2024–2025 were addressed through preprints, industry reports, and technical blogs included through snowballing, but may lack the methodological scrutiny of peer-reviewed sources. Geographic and cultural bias reflects the concentration of the reviewed research in North American and Western European contexts; findings about usability, adoption, and threat landscape may not generalise to other regions. Participant demographics bias affects the usability literature, which

predominantly recruits young, technically sophisticated participants from universities or technology companies, potentially overestimating the usability of methods that are harder for older adults or less technically experienced users.

4.3 Infostealer Malware: Mechanisms and Evolution

4.3.1 Historical Development

The lineage of infostealer malware extends back to the late 1990s, when early credential-stealing tools relied primarily on keylogging — the recording of every keystroke entered by the user, from which passwords could be extracted by an analyst reviewing the logs. These early tools were simple in design, required significant manual post-processing of captured data, and made no attempt to selectively extract authentication credentials from the noise of general keyboard input. Their operational model required the attacker to maintain persistent remote access to retrieve the logged data, which was technically demanding and exposed the attacker to detection through network monitoring. FormGrab represented the first significant architectural advance in infostealer design (van Rijn, 2022). FormGrab was a change in the way things were done. It caught HTTP form data at the application layer. This means it got plaintext credentials before the browser encrypted them. It did not just log what people typed. Since HTTPS only encrypts data when it is moving not when it is in the browser this approach went around the security that's in place. It made records of credentials that were easy to use away. This is still the basis for how modern infostealer operations work. By the 2000s Zeus had made this idea even bigger. It combined form grabbing with web injection. This means it changed what people saw in their browsers in time. It added input fields and phishing content to real pages. It also redirected transactions without people knowing. SpyEye added a way to automatically try stolen credentials and a management console. This made it easier for people to run campaigns. Zeus and SpyEye made the infostealer model. It is a model that can do things is easy to use and does not require a lot of expertise. After 2010 the Malware-as-a-Service model made things more available. Platforms like Raccoon, RedLine and Vidar were available for a fee. People could get them through markets and Telegram channels. The people who made these platforms were separate from the people who used them. The skilled developers took care of the platform. The users just had to run the campaigns. They did not need to know a lot about technology. The platforms also came with customer support, tutorials and installation guides. These were sent through Telegram and Discord. All of this made it even easier for people to use these platforms. The result of all this was that a lot more people could run credential theft campaigns. This contributed to the number of breaches that happened like the one that was found in June 2025. The infostealer operations, like FormGrab and Zeus made it easy for people to steal credentials. This is still a problem today.

4.3.2 Core Credential Harvesting Mechanisms

Form grabbing intercepts form submission data at the application level, before encryption, using API hooking on browser processes. Modern implementations target all major browsers — Chromium-based browsers (Chrome, Edge, Brave, Opera), Firefox and its derivatives, and

others — and operate transparently to the user, who sees no indication that submitted form data is being copied (van Rijn, 2022).

Stored credential theft targets the local databases where browsers cache saved passwords for user convenience. Chrome stores credentials in an SQLite database (Login Data) protected by the DPAPI (Data Protection API) encryption mechanism, which ties decryption to the Windows user account. Modern infostealers defeat this protection by running under the credentials of the logged-in user — whose decryption context automatically unlocks DPAPI-protected data — and exporting the decrypted credential database. Firefox's logins.json file is similarly protected by a master password derivation scheme that infostealers circumvent when no master password has been set (the default configuration) (van Rijn, 2022).

Memory scraping extracts credentials directly from the process memory of browsers and other applications during active sessions, targeting the data structures where plaintext credentials are briefly held during processing. Memory scraping captures credentials that users have not saved in the browser, but requires process injection privileges that may trigger EDR alerts.

Session cookie theft is particularly damaging. Session cookies are authentication tokens issued after login; possessing one allows an attacker to access an authenticated session without knowing the underlying password and without triggering an MFA challenge, since MFA is only checked at initial login.

Target coverage has expanded steadily. Where early infostealers focused exclusively on browser credentials, current variants routinely collect FTP and email client credentials, VPN configurations, cryptocurrency wallet files, SSH private keys, API keys, and gaming platform credentials. AZORult, for instance, specialised in cryptocurrency wallet theft — reflecting the growing value of digital assets as a target category.

4.4 Defensive Detection Techniques

4.4.1 Endpoint Detection and Response Capabilities

The antivirus that uses signatures can find most of the stuff that is already known it can detect 85 to 95 percent of the known variants. However when it comes to ones it is not that good it can only detect 40 to 60 percent of them. This is because the people who make the stuff are always coming out with new versions they do this every 2 to 4 weeks, which makes it hard for the antivirus to keep up. The bad guys also use a trick called code this means they can make the bad stuff look different each time so the antivirus needs a new signature to detect it. The signature-based antivirus is good for a check but it is not enough on its own. There is another way to detect the stuff it is called Behavioural EDR this one looks at what the bad stuff is doing on the computer like trying to get into the password database or making suspicious calls to the system. It can also see if someone is trying to read a lot of files with passwords in them. This way of detecting the stuff can find 65 to 90 percent of it but sometimes it can make mistakes it can say something is bad when it is not this happens 5 to 12 percent of the time. This is a

problem because some good programs, like the ones that manage passwords or backup software can look like the stuff, to the Behavioural EDR.ML-based EDR achieves 85–92% detection against novel variants by recognising behavioural similarity to known malware families. The limitation is that attackers can probe deployed models and craft evasion strategies around their feature patterns, requiring continuous retraining to stay current.

Sandbox detonation achieves 90–95% detection against unsophisticated malware but is evaded by 60–70% of contemporary variants through VM detection, user interaction requirements, and time-delayed activation. Extended or bare-metal sandboxing helps but is computationally expensive. Integrated EDR platforms combining all four approaches achieve the best results — 88–93% detection against novel variants with 2–5% false positive rates. The residual 7–12% that evade even integrated detection represent the persistent vulnerability window within which large-scale credential theft, including the June 2025 breach, continues to occur.

4.4.2 Network-Based Detection

Domain reputation and blocklisting blocks network connections to known malicious C2 domains and IP addresses. Its fundamental limitation is reactivity: against DGA-based C2 architectures, blocklisting requires identifying, analysing, and registering or sinkholing algorithmically generated domains faster than the DGA generates them, which is operationally challenging. Antonakakis et al. (2010) found that domain reputation filtering achieved only 50–60% effectiveness against infostealers using DGAs, because the volume of newly generated domains consistently outpaced blocklist update cycles.

Traffic pattern analysis identifies C2 beaconing behaviour through detection of regular-interval network connections to external destinations, unusual data exfiltration volumes, and outbound connections to destinations with no previous connection history. Tegeler et al. (2012) demonstrated 65–75% effectiveness using traffic pattern analysis against botnet command-and-control traffic. The primary limitation is that modern infostealers route C2 traffic through Telegram, Discord, and cloud storage services, making their network traffic statistically indistinguishable from normal use of those services and defeating all pattern-based detection approaches simultaneously (Acarali et al., 2016).

TLS/SSL traffic inspection decrypts encrypted network communications to enable deep packet inspection of otherwise opaque C2 traffic. Durumeric et al. (2017) found that TLS inspection, when deployed, achieved 80–90% effectiveness in detecting malicious encrypted traffic. The significant limitation is a deployment rate below 30% of organisations, attributable to privacy and legal concerns regarding employee communication interception, the performance overhead of decryption at scale, and the technical complexity of certificate management for a TLS inspection infrastructure (O'Neill et al., 2016).

4.5 Account Hijacking and Session Hijacking: Definitions and Distinctions

4.5.1 Account Hijacking

Account hijacking is when someone uses stolen login details to pretend to be an user and get into protected services. This directly targets the way services check who you are. By giving them login details that they think are genuine. Guo et al. (2020) Said that account hijacking affects around 22% of internet users every year. This makes it one of the threats to online security. The ways attackers get login details for account hijacking vary a lot from methods to social tricks. Attackers get login details in bulk through malware like infostealers. These details are already. Ready to use. Some phishing campaigns target services to get login details for those platforms from users who have accounts there. When service providers databases are hacked password codes are exposed. These codes can be cracked offline. Some attackers use engineering to trick victims into giving them special codes. Credential stuffing is another issue. This happens when attackers use login details from one service on services. This works because many people use the password for different accounts (Wang et al., 2019). Thomas et al. (2019) Looked at over 4 billion attempts to stuff credentials. They found that 0.1-2% of these attempts worked. This might seem like a number but its actually very significant. For example if an attacker has 1 billion login details a 0.1% success rate means they can hijack 1 million accounts. Usually when an account is compromised it doesn't stay that way for long. Sivakorn et al. (2016) Found that 74% of compromised accounts led to takeovers within the same organisation.This happens when attackers use login details or single sign-on setups. Password spraying is also common. This is when attackers try a few passwords on many accounts to avoid getting locked out. Microsofts 2019 Security Intelligence Report found that this affected, over 60% of enterprise accounts. This shows that many people still use passwords even at work.

4.5.2 Session Hijacking

As Calzavara, Focardi, Squarcina, and Tempesta (2017) defined it: session hijacking attacks target the post-authentication state maintenance mechanisms rather than the authentication process itself, allowing attackers to assume control of already-authenticated sessions without knowledge of user credentials. Cross-Site Scripting (XSS) attacks inject malicious scripts into web pages that execute in the victim's browser context and extract session cookies through document.cookie or similar browser APIs. Stock et al. (2016) demonstrated widespread XSS vulnerabilities in major web platforms. Man-in-the-Middle (MITM) attacks intercept unencrypted network communications to capture session tokens in transit, most effective on unsecured wireless networks. Session fixation attacks force a victim to authenticate using a session identifier already known to the attacker, typically by embedding a pre-specified session token in a manipulated URL. Pellegrino et al., in work cited by Calzavara et al. (2017), identified session fixation vulnerabilities in 7% of the Alexa top 10,000 websites. Cross-Site Request Forgery (CSRF) exploits the browser's automatic inclusion of session cookies in all requests to their issuing domain, causing the victim's authenticated browser to execute attacker-specified actions. OWASP (2021) continues to list CSRF among the top ten web application security risks. Client-side malware — specifically infostealer session cookie theft, described in Section 4.3.2 — directly accesses browser cookie stores without requiring any of the above techniques, making it the most direct and difficult-to-defend against session hijacking vector. Calzavara et al. (2017) analysed the security of token-based authentication in 45 popular web applications and found that 31% implemented token management practices vulnerable to at

least one form of session hijacking, despite these applications representing some of the most technically capable development organisations on the web.

4.5.3 The Interrelationship Between Attack Vectors

Account hijacking and session hijacking are complementary rather than competing attack strategies, and modern infostealer families exploit this complementarity systematically. Whereas account hijacking provides durable, persistent access that survives session expiry (because static credentials remain valid until changed), session hijacking provides immediate access that can bypass MFA but expires when the session token becomes invalid. An attacker with both static credentials and active session tokens — precisely what the June 2025 breach dataset contained — can use the session tokens for immediate, frictionless access and the static credentials for durable long-term access after session expiry. Guo et al. (2020) analysed post-breach attack patterns and found that 47% of sustained unauthorised access cases involved initial account hijacking through credential stuffing followed by subsequent session hijacking techniques to maintain persistence and evade re-authentication events. This complementary exploitation pattern explains the design priority of contemporary infostealers: RedLine, Raccoon, and similar platforms target both static credentials and session cookies simultaneously as a matter of standard operational procedure (van Rijn, 2022).

4.6 Account Hijacking Prevention Techniques

46

4.6.1 Traditional Authentication and Its Documented Failures

Despite extensive documentation of their vulnerabilities, password-based authentication systems remain dominant in enterprise and consumer contexts. Wang et al. (2019) found that 86% of organisations still rely primarily on passwords as their primary authentication mechanism, supplemented by various enhancements. Password complexity requirements — mandating combinations of uppercase letters, lowercase letters, numerals, and special characters — were intended to increase the effective keyspace of user-chosen passwords. However, Ur et al. (2017) demonstrated that complexity requirements produce predictable patterns rather than genuinely random passwords: users subjected to complexity requirements consistently employ strategies such as capitalising the first letter, appending a number, and substituting common special characters for visually similar letters (@ for a, ! for i, etc.), which are well-known to password cracking tools and provide only marginal security improvement. Password reuse is pervasive and cognitively rational. Wang et al. (2019) found that 52% of users reuse passwords across multiple services, with a mean of 7.9 sites sharing the same password. Hanamsagar et al. (2018) characterised the specific transformation strategies users employ when they believe they are creating distinct passwords: 34% append service-specific suffixes to a base password (base_password → base_passwordgmail), enabling prediction of untested variants with approximately 28% accuracy when one variant is known; 19% use incremental numbering (MyPassword1, MyPassword2, MyPassword3); and 27% employ complexity adaptations that satisfy specific requirements while reusing a memorable base (password → p@ssword1). Even after experiencing a security incident, 70% of users continued

149

password reuse behaviour, suggesting that awareness alone does not change the behaviour (Wang et al., 2019). Hanamsagar et al. (2018) provide the appropriate framing for this behaviour: password reuse is cognitively rational rather than negligent. Users managing accounts across more than 100 services — a realistic estimate for contemporary internet users — face a genuine cognitive load problem that no mnemonic strategy can solve without some degree of reuse. The cognitive overload of managing unique credentials for all services is cited by 68% of users as a reason for reuse; concern about being locked out due to forgotten unique passwords is cited by 31%. Knowledge-based authentication questions, commonly used as account recovery mechanisms, provide substantially weaker security than passwords. Bonneau et al. (2015) found that 40% of security question answers could be guessed within five attempts using publicly available personal information (social media profiles, public records), and that knowledge-based authentication questions had significantly lower effective entropy than user-chosen passwords.

4.6.2 Multi-Factor Authentication: Comparative Analysis

Multi-factor authentication substantially reduces the effectiveness of credential theft attacks by requiring a second verification factor that the attacker must separately compromise. The major MFA implementations vary dramatically in their security effectiveness and usability, and the choice among them has consequences that dwarf the choice of how to implement single-factor authentication.

SMS-based OTP provides zero resistance to phishing — real-time proxies like Evilginx intercept and relay both the password and the code in the moment of submission. It offers zero resistance to SIM-swapping, with Lee et al. (2020) finding success rates of 39–77% across five major US carriers using basic social engineering. SS7 exploitation adds a third attack vector, with Positive Technologies (2019) intercepting codes for 33% of tested accounts. Authentication averages 8.2 seconds with a 12% error rate. NIST SP 800-63B has deprecated it for regulated applications. Despite all of this, SMS OTP accounts for 60–70% of MFA deployments in organisations using any second factor. That gap — between documented failure and continued dominance — is perhaps the clearest single illustration of the security-implementation gap this chapter addresses.

Email OTP delivers one-time codes via email rather than SMS, providing 100% resistance to SIM-swapping (since email accounts are not vulnerable to carrier-based social engineering) but 0% resistance to phishing (real-time relay attacks work equally well against email OTP as against SMS OTP) and approximately 50% resistance to MITM attacks. It requires no additional infrastructure beyond email service and has near-zero incremental cost, but has poor accessibility characteristics because it requires switching between applications and involves longer authentication times (12.5 seconds on average).

TOTP authenticator applications (Google Authenticator, Authy, Microsoft Authenticator) generate time-based codes using a shared secret seeded during enrolment, requiring no network connection to produce a code. TOTP provides 100% resistance to SIM-swapping and

approximately 95% resistance to MITM attacks (Dutson et al., 2019), but only approximately 50% resistance to phishing: a real-time phishing proxy can relay the entered TOTP code to the legitimate service within the 30-second validity window of the code. Authentication time averages 6.3 seconds with a 6% user error rate. TOTP requires no recurring cost and only moderate deployment complexity.

Push notification MFA sends an authentication request to a registered mobile application, which the user approves with a single tap. It provides approximately 60% phishing resistance (Altinkaynak and Nithyanand, 2022) — better than SMS or TOTP because modern push implementations include domain context that allows users to identify mismatched authentication requests — 100% SIM-swap resistance, and approximately 90% MITM resistance. Authentication time is 4.1 seconds with a 4% error rate. The primary limitation is susceptibility to push fatigue attacks, in which an attacker floods the victim with push requests until the victim approves one to stop the interruptions (Konoth et al., 2020).

FIDO2 hardware security keys implement the WebAuthn standard using public key cryptography bound to the origin domain of the requesting service. When the user authenticates to a service, the key generates a cryptographic signature using a private key that never leaves the hardware device, with the signature bound to the exact domain of the service. This origin binding makes FIDO2 100% resistant to phishing: a credential captured through a phishing site for evil-login.bank.com cannot be used against real login.bank.com because the origin in the signed challenge would not match. FIDO2 provides 100% SIM-swap resistance and 99.9% MITM resistance.

Platform authenticators implement the same FIDO2 protocol using biometric sensors or device PINs built into smartphones, laptops, and tablets (Windows Hello, Apple Face ID/Touch ID, Android biometrics), providing equivalent security to hardware keys — 99% phishing resistance, 99.9% MITM resistance — at zero marginal hardware cost. Authentication time of 2.1 seconds is the fastest of any MFA method, with a 1% user error rate. Platform authenticators are the highest-accessibility MFA option for sighted users. The FIDO Alliance (2021) reported that organisations implementing FIDO2 authentication (hardware keys or platform authenticators) reduced account takeovers by 99% compared to password-only systems.

The cost analysis over a three-year horizon for a 10,000-user organisation reveals a counterintuitive finding with respect to the security-cost tradeoff. Password-only systems have the lowest implementation cost ($25,000) but the highest ongoing operational cost due to password resets and account recovery, which account for 30–50% of all help desk tickets (Forrester, 2021). Three-year total cost of ownership for password-only systems is $580,000 ($58 per user). FIDO2/WebAuthn implementations have substantially higher initial costs ($175,000) but significantly lower ongoing operational costs, yielding a three-year TCO of $445,000 ($44.50 per user) — 22% lower than password-only systems. The Ponemon Institute (2021) found that credential-based breaches cost organisations an average of $146 per compromised record, and that MFA adoption reduced breach likelihood by 99.9%, making the return on investment for strong MFA adoption extremely favourable.

4.7 Session Hijacking Prevention Techniques

Network-level protections. HTTPS adoption grew from 40% to 80% of page loads between 2015 and 2017, substantially reducing passive interception. Adding HSTS — which forces browsers to reject unencrypted connections — cuts successful MITM attacks by around 95%. Certificate pinning provides additional MITM protection but is implemented by only 0.67% of Android applications despite being technically straightforward. Properly configured WAFs blocked 98% of automated session hijacking attempts in controlled conditions, though real-world effectiveness varies considerably with configuration quality and maintenance.

Application-level protections. Three cookie attributes address complementary attack vectors. HttpOnly blocks JavaScript from reading cookies, reducing XSS-based theft by 83%, though 23% of sites using it remain vulnerable to DOM-based sub-session attacks — meaning it must be paired with CSP headers and input sanitisation. The Secure flag restricts cookies to HTTPS connections, reducing interception-based hijacking by 91%. SameSite=Lax, Chrome's default since 2020, cut CSRF attacks by 94%; SameSite=Strict is recommended for high-security applications despite minor usability friction. Anti-CSRF tokens prevent 99.7% of CSRF attacks when correctly implemented. Session token rotation with 15-minute idle timeouts reduced banking session hijacking by 68%. Session binding to IP address and user-agent detected 92% of hijacking attempts, with some false positives from legitimate network switches.

User-level protections. Login notifications reduced account takeovers by 50%, and 76% of users who detected unauthorised access used session management interfaces to terminate suspicious sessions. Phishing awareness training helps but decays quickly — effectiveness drops by around 50% within four to six months, requiring ongoing reinforcement rather than one-time delivery. Alert fatigue is a parallel problem: users encounter an average of 45 security warnings per week and ignore 87% of them.

4.8 Comparative Analysis of Prevention Techniques

4.8.1 Security Effectiveness Against Modern Threats

The comparative effectiveness of authentication methods against the three primary credential-based attack vectors — phishing, infostealer malware, and credential stuffing — varies dramatically across methods and reflects the fundamental security architecture of each approach. Traditional passwords provide low protection against phishing (10–45% effectiveness, reflecting the fraction of phishing attempts that are caught by browser warning systems or user scepticism), low protection against infostealer malware (5–20%, reflecting only the subset of infections detected and blocked before credential harvest), and very low protection against credential stuffing (0–5%, reflecting only the subset of reused credentials that have not previously been exposed in any breach). SMS-based OTP provides moderate phishing protection (50–70%), low infostealer protection (15–30%), and high credential stuffing protection (85–95%) — the last because a time-limited code, even if the underlying password has been stolen, prevents immediate stuffing. Hardware security keys and FIDO2 passwordless

authentication provide very high to maximum protection across all three vectors (95–99%), with the origin-binding architecture specifically designed to make phishing and MITM attacks non-exploitable regardless of whether the underlying credential is ever exposed (Lang et al., 2017; FIDO Alliance, 2021; Konoth et al., 2020). No single session protection technique covers all attack vectors. HTTPS/TLS excels against MITM (90–99%) but offers nothing against client-side malware. HttpOnly handles XSS well (80–95%) but not network interception. SameSite addresses CSRF (85–95%) but little else. Device binding is the only technique with meaningful coverage against client-side malware — the primary vector for contemporary infostealers — achieving 70–90% effectiveness there, alongside moderate protection across other vectors. Effective session security requires combining all of these rather than relying on any one.

4.8.2 User Experience Metrics

The data challenges the assumption that security and usability trade off against each other. Across authentication time, error rate, and user satisfaction, the most secure methods also perform best. SMS OTP — the most widely deployed second factor — requires 25 seconds, has an 8–18% error rate, and 60% satisfaction, worse than passwords on every metric. FIDO2 passwordless authentication requires 4 seconds, has a 1–5% error rate, and 88% satisfaction — the fastest, most accurate, and most satisfying option in the comparison.

The reason is architectural. FIDO2 replaces the multi-step process of waiting for a code, reading it, and typing it with a single tap or biometric gesture. Security and speed improve together because the friction of OTP-based methods is eliminated entirely.

Organisational adoption barriers tell a different story, however. The top concerns cited by security professionals are user friction (74%), legacy system integration (68%), and help desk burden (63%). These are transitional barriers rather than steady-state usability problems, and they are substantially reduced by phased migration and platform authenticator deployment, which requires no hardware procurement at all.

4.8.3 Implementation Complexity and Total Cost of Ownership

The three-year TCO analysis presented by Forrester Research (2021) and described in Section 4.6.2 demonstrates that FIDO2's higher initial cost is more than offset by substantially lower operational costs, yielding a 22% TCO reduction compared to password-only systems for a 10,000-user organisation. The return on security investment calculation from the Ponemon Institute (2021) — where an expected annual loss reduction from credential-based breaches of approximately $420,000 to $420 (a factor of 1,000) easily justifies the implementation investment — provides the strongest economic argument for accelerating MFA adoption beyond regulatory minimum requirements.

4.9 Drawbacks and Vulnerabilities in Present Techniques

Technical limitations. Even strong authentication methods have exploitable weaknesses. Konoth et al. (2020) tested 14 MFA solutions and found 12 vulnerable to real-time relay attacks— proxies that intercept and forward both credentials and MFA codes simultaneously.

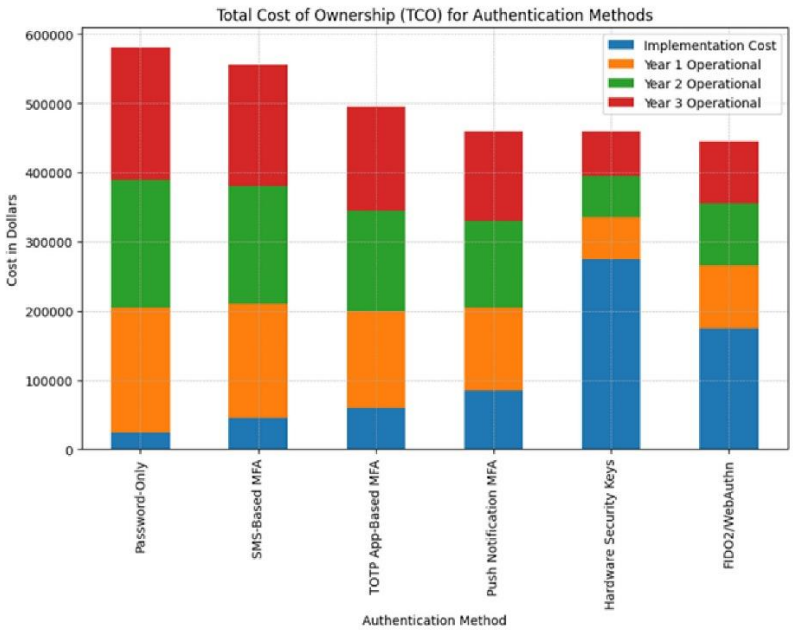

Figure 4.1: The stacked bar graph showcases the year-on-year Implementation Cost of various Authentication Methods

Only hardware-bound FIDO2 resisted, because its cryptographic signature is tied to the exact requesting domain and cannot be replayed elsewhere. SMS MFA faces an additional infrastructure-level weakness: SS7 exploitation enabled code interception for 33% of tested accounts. Session management is also poorly implemented in practice — 21% of major websites remained vulnerable to DOM-based XSS despite using HttpOnly cookies, and 67% of 405 analysed OAuth implementations contained at least one serious session hijacking vulnerability. Biometrics carry a unique risk: unlike passwords, compromised biometric data cannot be changed. The 2019 BioStar 2 breach exposed 28 million biometric records permanently.

Human factors. Technical protections are only as strong as the humans operating within them. In controlled experiments, 43% of employees handed over MFA codes to attackers impersonating IT support. SIM-swapping succeeded in 39–77% of attempts across five major US carriers using basic social engineering. Phishing training decays by 50% within four to six months and users ignore 87% of security warnings — not because they are careless but because the volume of alerts exceeds what human attention can sustain. As Herley (2014) noted, security measures that impose costs disproportionate to perceived benefit will be rejected. The most durable designs make the secure behaviour the easiest one.

Organisational barriers. Legacy system incompatibility affects 68% of organisations, and 72% have critical applications that cannot support modern authentication without expensive

rewrites. The average enterprise runs 4.3 separate authentication systems. Implementation cost is the primary barrier for 68% of organisations, with SMBs disproportionately affected. Skills shortages compound the problem — 70% of organisations report difficulty filling identity and access management roles. Perhaps most damaging is compliance-driven prioritisation: 58% of organisations prioritise meeting regulatory requirements over adopting more effective security measures when the two diverge, systematically underinvesting in areas that compliance frameworks have not yet caught up to.

4.10 Emerging Authentication Technologies

4.10.1 Evaluation Framework

This review evaluates three promising technologies — visual cryptography-based authentication, blockchain-based decentralised identity, and behavioural biometrics with continuous authentication — across seven dimensions: security effectiveness against documented attack vectors; usability (authentication time, error rates, user satisfaction, learning curve); deployability (integration complexity, infrastructure requirements, backward compatibility); scalability (throughput, latency, concurrent user support); cost (three-year TCO); maturity (standardisation status, vendor ecosystem, documented production deployments); and inclusivity (accessibility for diverse user populations including users with disabilities).

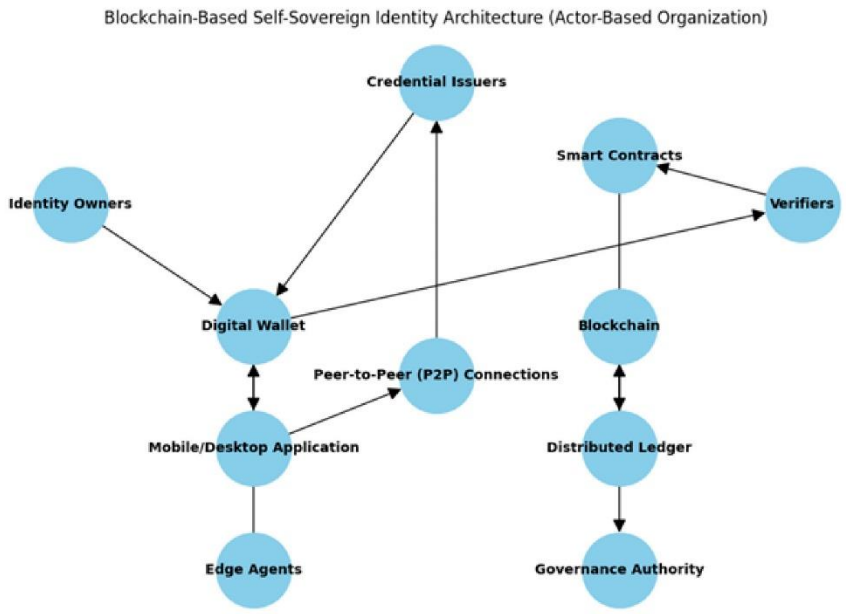

Figure 4.2: Flowchart showing the architecture of Block chain-Based Self-Sovereign Identity Architecture

4.10.2 Visual Cryptography-Based Authentication

Visual cryptography, introduced by Naor and Shamir (1994), applies secret sharing principles to authentication by splitting a visual verification image into multiple shares that individually reveal no information but jointly disclose the original image when physically or digitally overlaid. Recent authentication implementations adapt this principle to counter phishing by creating verification workflows that are fundamentally resistant to single-channel compromise. Wu et al. (2020) integrated visual cryptography with QR codes, creating an authentication system where the complete authentication QR code is only readable when a user-held physical transparency is overlaid on a server-displayed share. Visual cryptography shows strong security results in controlled studies — 99.7% replay attack resistance, 99.2% phishing resistance, and 98.7% verification accuracy after minimal training. However, its practical limitations are significant. No standardised protocol exists, custom development takes 6–12 months compared to 2–4 weeks for FIDO2, and only three commercial deployments have been identified in the literature. Most critically, the approach is fundamentally inaccessible to the approximately 2.2 billion people globally with visual impairments, and tactile alternatives remain experimental.

The overall assessment is that visual cryptography is viable only for niche high-security applications serving sighted users. Research investment priority is low until the accessibility barrier is resolved. Any deployment must provide alternative authentication methods for visually impaired users.

4.10.3 Blockchain-Based Decentralised Identity

Instead of authenticating through a centralised authority (Google, Facebook, a corporate LDAP server), users hold their own cryptographic identity keys and present verifiable credentials — digitally signed assertions issued by trusted entities — to authenticate without transmitting the underlying secret. Fedrecheski et al. (2020) documented that uPort, an Ethereum-based identity platform implementing the ERC-725/735 identity contract standards, reduced identity verification costs by 71% compared to traditional centralised systems while eliminating the risk of centralised database breaches. Dunphy and Petitcolas (2018) outlined the cryptographic foundations of blockchain-based identity, including the use of zero-knowledge proofs (ZKPs) to enable privacy-preserving authentication where users prove possession of credentials without revealing their contents. Lux et al. (2020) and Ferdous et al. (2019) document production-ready implementations using the Hyperledger Indy framework and the Sovrin network. The multi-dimensional assessment rates blockchain identity as excellent for security effectiveness and moderate for cost and maturity, but weak for usability, deployability, scalability, and inclusivity. The scalability limitation is severe: Sedlmeir et al. (2021) document that public blockchain networks process 7–15 transactions per second, orders of magnitude below the tens of thousands of transactions per second required for global authentication infrastructure. Permissioned blockchain implementations achieve 1,000–10,000 transactions per second (Hyperledger Indy) but at the cost of reduced decentralisation. The usability limitation is critical: Lockwood and Esteves-Verissimo (2020) found that 11–15% of blockchain wallet users permanently lost access due to private key mismanagement, and social recovery mechanisms introduced to address this (Markert et al., 2022) reduce the loss rate to approximately 5% but introduce new collusion vulnerabilities.

4.10.4 Behavioural Biometrics and Continuous Authentication

66

Keystroke dynamics captures the timing patterns of keyboard interaction: the dwell time (duration of each key press) and flight time (interval between consecutive key presses) for each key transition in a user's typing. Bailey, Okolica, and Peterson (2014) achieved 95% authentication accuracy with false acceptance rates below 0.1% using only 200 keystrokes — a small enough sample to capture from a single short document. Touch dynamics on mobile devices analyse the pressure, area, velocity, and trajectory of touch interactions. Li et al. (2019) demonstrated 97.8% accuracy by combining touch gesture analysis with accelerometer data, detecting unauthorised users within an average of 13 seconds of session assumption — before substantial damage could be done. Gait recognition, using accelerometer and gyroscope data from smartphones, achieved nearly 98% accuracy with near-zero false rejection rates within 30 seconds of walking in experiments by Fridman et al. (2017). Multi-modal behavioural systems combining keyboard, mouse, and application usage patterns achieved 98% accuracy with near-zero false rejections within 30 seconds (Fridman et al., 2017). The Facebook deployment documented by Freeman et al. (2016) represents the largest-scale production validation of behavioural biometrics: a system operating across more than two billion users achieved 99.5% account takeover detection rate — directly addressing the 47% of sustained breaches that involve post-authentication session-level exploitation — while keeping false positive interruptions to an operationally acceptable level. Zheng et al. (2021) demonstrated a system where behavioural biometrics verified high-value blockchain transactions, reducing unauthorised transaction rates by 99.7%. The multi-dimensional assessment rates behavioural biometrics as excellent for security effectiveness, usability, and scalability, and good for maturity and cost, with moderate ratings for deployability and inclusivity. Malekzadeh et al. (2021) demonstrated techniques for extracting authentication signals from behavioural data while provably preventing extraction of sensitive attributes including emotional state, stress levels, and medical conditions. Risk-based activation — triggering behavioural monitoring only during high-risk sessions rather than continuously — retains 99.2% of the security benefit while substantially reducing privacy exposure (Wiefling et al., 2021). Accuracy variation across age groups (Nicholson et al., 2019) and motor disability status (Dosono et al., 2019) requires adaptive modelling for equitable performance across diverse user populations. The overall assessment of behavioural biometrics is the most favourable of the three emerging technologies: production-ready at global scale, highest security effectiveness, best usability profile (fully transparent to the user), and scalable infrastructure. The research priorities rated highest are federated learning for privacy-preserving deployment, inclusivity studies across age and ability groups, and integration with FIDO2 initial authentication to create comprehensive defence-in-depth covering both initial authentication and ongoing session security.

4.11 Cross-Cutting Trends and Future Research Directions

4.11.1 Convergence Toward Passwordless Authentication

The literature reviewed in this chapter documents a clear and accelerating convergence toward passwordless authentication as the long-term trajectory of authentication technology. The

timeline of this convergence spans three periods: 2015–2018, characterised by FIDO U2F standardisation and early enterprise pilots with adoption below 5%; 2019–2021, characterised by the FIDO2/WebAuthn W3C recommendation and platform authenticator integration at scale achieving 15–20% enterprise adoption; and 2022–2025, characterised by the passkey ecosystem launch (with Apple, Google, and Microsoft committing to cross-vendor passkey

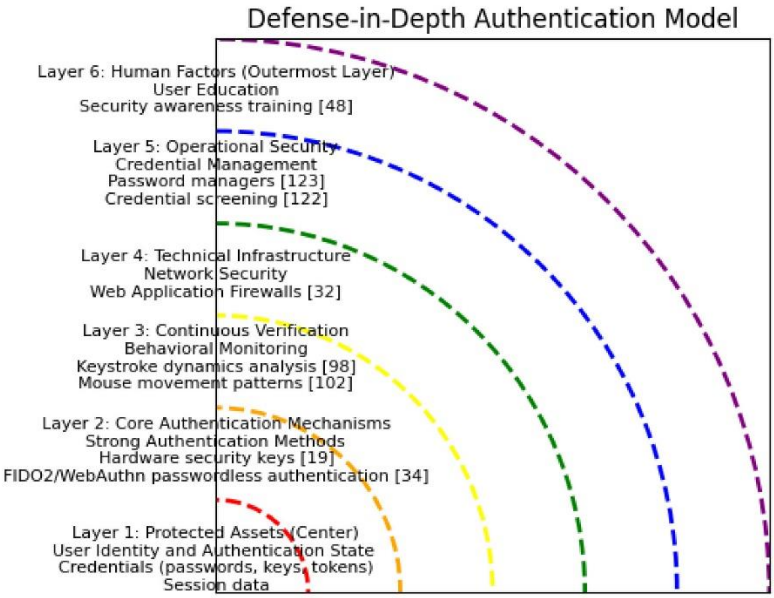

Figure 4.3: Multi-faceted and Multi-layered Approach to Authentication showcasing Defense-in-Depth.

synchronisation) and broad consumer deployments driving enterprise adoption to 35–40%. The primary barriers to reaching universal adoption are legacy system incompatibility (affecting approximately 40% of enterprise applications), deployment costs that average approximately 1.8 times initial estimates (Lassak et al., 2024), and user training requirements averaging 4.2 hours per employee.

Zero-trust shift. Perimeter-based security is giving way to zero-trust architectures built on continuous authentication, short-lived tokens (15–60 minutes), identity-aware proxies, and per-resource authorisation. The shift is driven by cloud migration (78% of workloads), widespread remote work, and high-profile breaches exposing the limits of perimeter models. Implementation challenges include a ~47ms per-check performance overhead, step-up authentication friction, and legacy integration complexity.

Layered defence. Security outcomes improve non-linearly with layering. MFA alone reduces account takeovers by ~96.3%; behavioural biometrics alone detect 91–95% of attacks; combining both reaches ~99.7%. The principle holds broadly — preventive, detective, and responsive controls each cover vectors the others miss, and their combination outperforms any single approach by a substantial margin.

Regulatory catalysis. Regulation accelerates adoption by roughly three to five years compared to voluntary timelines. PSD2's Strong Customer Authentication requirement is associated with 94% MFA adoption in regulated sectors versus 67% in unregulated ones. US Executive Order 14028 drove FIDO2 implementation across 78% of federal agencies by 2025, up from 12% in 2021.

Future research priorities. Key directions include: risk-based authentication models that reduce false positive friction; standardised MFA interaction patterns, which Colnago et al. (2018) estimate could cut errors by 54%; faster breach detection — currently averaging 197 days from compromise to discovery, reduced to 24 hours in ML-based evaluation by Pal et al. (2021); lightweight authentication protocols for IoT devices, where 73% of current protocols are incompatible with hardware constraints; and inclusive authentication design, given that 62% of visually impaired users currently rely on insecure workarounds due to inaccessible alternatives.

4.12 Conclusion

This chapter reviewed 166 sources on infostealer malware and credential theft prevention, selected through a PRISMA-compliant process from 3,495 initial records. The central finding is a persistent security-implementation gap. Modern infostealers have evolved from simple keyloggers into modular, subscription-based platforms that steal credentials, session tokens, and cryptocurrency wallets simultaneously, with 60–70% of current variants capable of bypassing sandbox detection. The 2–4 week window between new variant release and reliable detection coverage means credential theft continues largely unimpeded regardless of how sophisticated the detection infrastructure is. The June 2025 breach of 16 billion credentials reflects this reality directly.

The solution exists but is not being used. FIDO2 platform authenticators achieve 99–99.9% phishing resistance, faster authentication times, lower error rates, and 22% lower three-year total cost of ownership than password-only systems. SMS OTP — with zero phishing resistance — remains dominant because of legacy compatibility, compliance inertia, and organisational barriers that persist even when the economic case for upgrading is clear.

Among emerging technologies, behavioural biometrics are production-ready and warrant the highest research investment. Blockchain identity is viable for enterprise B2B contexts only. Visual cryptography is limited to niche applications and remains inaccessible to users with visual impairments.

The overarching lesson is consistent with the thesis's central argument: detection-centric strategies cannot close the persistent vulnerability window. The appropriate primary defence is credential-agnostic authentication — systems that remain secure even after credentials are stolen. Getting there requires organisations to diagnose their specific constraints before selecting solutions, the same context-first principle that drives the causal inference framework in Chapter 3 and the language strategy assessment in Chapter 5.

CHAPTER 5
ADVANCEMENTS AND CHALLENGES IN MULTILINGUAL AND MULTIMODAL SENTIMENT ANALYSIS

5.1 Introduction

Sentiment analysis has grown well beyond its origins as a single-language text classifier. Two expansions define its current trajectory: multilingual systems that handle diverse languages, dialects, and mixed-language input; and multimodal systems that integrate text, audio, and visual signals, recognising that emotion is expressed across multiple channels at once.

Both directions are driven by practical need. Social media generates opinion content in hundreds of languages. Mental health applications must interpret speech and gesture alongside text. Product review systems serve global markets where English is a minority language. A system that performs well only on English benchmark data is of limited real-world value.

This chapter presents a systematic review of the current state of multilingual and multimodal sentiment analysis, drawing on 18 primary studies selected from an initial pool of 29 through structured quality assessment. It synthesises model and dataset advances, documents the challenges constraining generalisation, and identifies the most promising research directions. Seven research questions, described in Section 5.2, structure the review.

5.2 Review Methodology

5.2.1 Research Questions

Seven research questions guide the review:

RQ1. What challenges does linguistic diversity pose for sentiment analysis, and how are current approaches addressing them?

RQ2. What are the most significant recent advances in ML and NLP for multilingual sentiment analysis?

RQ3. What real-world applications have emerged, and what are their implications for marketing, healthcare, and social media?

RQ4. What challenges affect data availability and quality, particularly for underrepresented languages?

RQ5. How is research addressing data scarcity in low-resource languages, and what dataset-building strategies are being used?

RQ6. How well do state-of-the-art models generalise across languages and modalities?

RQ7. What are the key limitations of current methods, and where is further research most needed?

5.2.2 Search Strategy and Selection

Searches across seven databases — ACM Digital Library, IGI Global, Elsevier, Springer, IEEE Xplore, Pertanika, and Wiley Online Library — using keyword combinations including "Multilingual Sentiment Analysis," "Multimodal Sentiment Analysis," "NLP," and "Deep Learning," supplemented by reference snowballing, yielded 29 initial studies.

Studies were included if they addressed multilingual or multimodal sentiment analysis, were published within the preceding five to seven years, were peer-reviewed, and used standard datasets and evaluation metrics. Single-language, single-modality, and unquantified studies were excluded. Quality assessment scored each study across nine criteria — methodological clarity, dataset diversity, domain generalisation, novelty, cultural sensitivity, and citation impact — on a 0–1 scale (maximum 9). Studies scoring 5 or above were retained, producing a final set of 18 primary studies, all published between 2020 and 2023.

5.3 Challenges of Linguistic Diversity (RQ1)

Language-specific sentiment expression. Sentiment is not language-neutral. Idiomatic expressions, cultural metaphors, and evaluative distinctions embedded in grammar vary across languages in ways that translation cannot resolve. Morphological structures — agglutination in Turkish and Finnish, tonal distinctions in Mandarin and Thai, script conventions in Arabic and Devanagari — break English-centric tokenisation pipelines. A model built for English is not merely less accurate in other languages; it is often systematically wrong in ways that go beyond simple performance degradation.

Code-switching. A substantial proportion of real-world multilingual text involves code-switching — the alternation between two or more languages within a single utterance. Hinglish (Hindi-English), Spanglish (Spanish-English), and analogous mixed-language varieties are ubiquitous on social media platforms in multilingual communities. Standard monolingual tokenisation pipelines fail on code-switched input because they cannot correctly segment tokens that span language boundaries, and embedding models trained on separate monolingual corpora cannot represent the cross-lingual interactions within a single sentence. Specialised tokenisation techniques and hybrid embedding models have been developed to address code-switching but remain limited in language pair coverage and domain generality.

Data imbalance. Existing sentiment datasets overwhelmingly favour high-resource languages, particularly English. Low-resource languages — including the majority of the approximately 7,000 languages spoken globally — lack labelled sentiment datasets of sufficient scale to train competitive models. This imbalance is self-reinforcing: models trained on English-dominant data perform poorly on low-resource languages, which reduces the incentive for investment in those languages' dataset construction, which perpetuates poor model performance. The result is

systematic underrepresentation of the linguistic majority of the world's population in sentiment analysis systems.

Partial solutions. The primary current responses to linguistic diversity challenges are: transfer learning with multilingual pre-trained models (mBERT, XLM-R) that learn shared cross-language representations; data augmentation through machine translation to generate parallel datasets; specialised tokenisation for code-switched input; and language-specific fine-tuning that incorporates morphological features and tonal distinctions. These approaches reduce the gap but do not eliminate it, particularly for typologically distant language pairs and genuinely low-resource languages where even the multilingual pre-training corpora are thin.

5.4 Advances in Machine Learning and NLP Techniques (RQ2)

5.4.1 Multilingual Pre-Trained Language Models

The most significant advance enabling multilingual sentiment analysis has been the development of large-scale pre-trained transformer models that learn cross-language representations through joint training on multilingual corpora. Three models dominate the reviewed literature.

Multilingual BERT (mBERT), pre-trained on Wikipedia text from 104 languages using masked language modelling, provides a shared transformer architecture that supports cross-lingual transfer. A model fine-tuned on English sentiment data can be applied to other supported languages through zero-shot transfer without language-specific training examples. The limitation is that mBERT's performance degrades substantially for low-resource languages, because their representation in the Wikipedia pre-training corpus is sparse. Languages with rich Wikipedia coverage receive strong representations; those without do not.

XLM and XLM-R improve on mBERT through the addition of Translation Language Modelling (TLM) as a pre-training objective, which uses parallel bilingual corpora to explicitly align representations across languages. XLM-R, trained on 2.5 terabytes of filtered Common Crawl text across 100 languages, substantially outperforms mBERT on low-resource language tasks by providing denser and more consistent representations across the language coverage space. XLM-T, a Twitter-domain fine-tuned variant of XLM-R, demonstrated F1 score improvements from 40.29% to 56.39% in cross-lingual sentiment classification on Twitter data, illustrating the additional gains available from domain-specific fine-tuning on top of multilingual pre-training.

mT5 frames all NLP tasks as text-to-text generation, handling classification, translation, and generation through a single interface. It performs well on sentence-level sentiment but less so on aspect-based tasks requiring precise opinion target localisation within complex syntax.

LASER and ALIGN produce sentence-level multilingual embeddings that support cross-lingual semantic alignment and sentiment classification without task-specific fine-tuning.

BYT5 and mBART extend pre-training to byte-level and sequence-to-sequence settings, improving coverage of morphologically complex languages and code-switched text. Adversarial training adds robustness against the noisy, abbreviated, and inconsistently punctuated input common in social media.

5.4.2 Multimodal Models and Performance

The tensor fusion approach models cross-modal interactions through outer products of unimodal representations, capturing pairwise and trimodal feature interactions that simple concatenation ignores. Attention-based fusion allows the model to dynamically weight modalities depending on their relative informativeness for a given input instance. Multilingual machine translation-augmented sentiment models averaged approximately 71% macro-F1 across multilingual tasks. In mean absolute error on regression-style sentiment intensity prediction, MTSA on CMU-MOSEI achieved the lowest MAE of 0.541, while other models on CMU-MOSI had a slightly higher MAE of 0.696. The key takeaway from the performance landscape is that the optimal choice of model depends on the available modalities and the specific task. Audio models outperform text-only models when audio data is available and of sufficient quality. Multimodal models outperform unimodal models when modalities are informative and aligned. Multilingual models with domain-specific fine-tuning outperform zero-shot transfer when small amounts of target-language labelled data are available.

Explainability tools including attention visualisation and SHAP values are increasingly applied to sentiment models to identify biases, improve reliability, and support interpretability for downstream applications in healthcare and policy contexts where black-box predictions are insufficient.

5.5 Real-World Applications (RQ3)

Multilingual and multimodal sentiment analysis has been deployed across a range of application domains, each with distinct technical requirements and implications.

Marketing and customer experience. Commercial tools including Brandwatch and Hootsuite monitor brand sentiment across social media platforms and languages, enabling rapid identification of shifts in customer perception and support for personalised advertising based on audience emotional profiles. Retailers use multilingual sentiment analysis on customer reviews for product recommendation optimisation and service delivery improvement. The primary benefit at the industry level is improved audience segmentation and faster response to reputational events.

Healthcare. Providers use multilingual sentiment analysis to analyse patient feedback and identify service gaps. Applications like Wysa and Woebot combine speech and facial expression analysis to detect early signs of psychological distress, supporting earlier intervention and more equitable mental health access across language groups.

Social media moderation. Platforms use multilingual sentiment analysis to detect hate speech, monitor public sentiment, and personalise content. The core challenge is cultural context — expressions that are offensive in one setting may be neutral in another, making cross-cultural generalisation of moderation models difficult.

Education. Platforms like Coursera apply sentiment analysis to multilingual student feedback to improve course design. Multimodal analysis of recorded sessions can track engagement and flag disengagement, enabling earlier intervention for at-risk students.

Entertainment and politics. Studios and streaming platforms use sentiment analysis to gauge audience responses and inform content decisions. Governments and campaigns use multilingual sentiment monitoring to track public opinion and adjust communication strategies in real time.

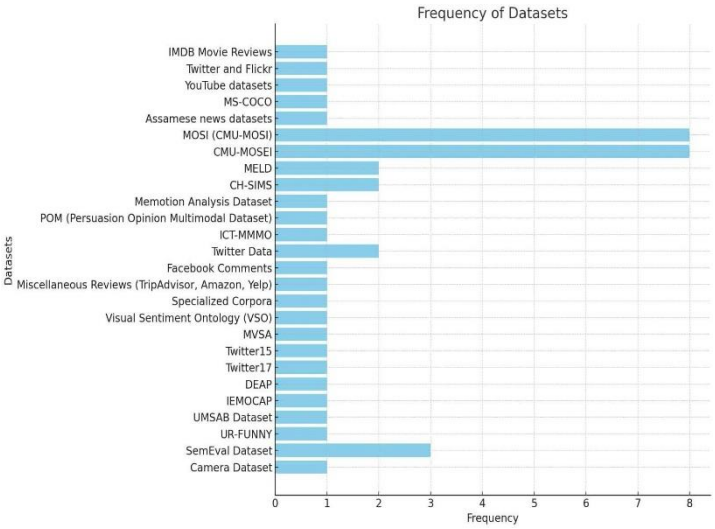

Figure 5.1 The figure summarises the datasets used across the 18 primary studies and how frequently each appears.

5.6 Data Availability and Quality Challenges (RQ4)

Annotation scarcity. English sentiment datasets contain hundreds of thousands of labelled examples; many low-resource languages spoken by tens of millions have fewer than a thousand, if any. This asymmetry constrains pre-training, fine-tuning, and evaluation across the board.

Dataset imbalance and cultural bias. Even within multilingual datasets, major languages dominate. Annotations reflect annotators' cultural perspectives, causing models to misinterpret sentiment expressions from other cultural contexts. A label assigned by one annotator may be systematically inconsistent with how speakers from a different background would characterise the same expression.

Noise and multimodal misalignment. Social media data contains misspellings, abbreviations, and ambiguous terms that introduce label noise. Multimodal datasets add further complications through timing mismatches between text, audio, and video, and inconsistent recording quality.

Temporal drift. Sentiment expressions evolve — slang changes, emoji semantics shift, cultural references move on. Datasets from even a few years ago may no longer represent current language use accurately, particularly for fast-moving social media registers. Regular refresh is necessary but costly.

Annotation cost. Multilingual annotation requires native-speaker expertise across languages; multimodal annotation adds alignment and non-verbal communication expertise on top. Costs scale multiplicatively with language and modality count, putting comprehensive dataset construction out of reach for most research teams without substantial institutional support.

5.7 Addressing Data Scarcity in Low-Resource Languages (RQ5)

Cross-lingual transfer. Models like mBERT and XLM-R enable zero-shot and few-shot transfer from high-resource to low-resource languages. A model fine-tuned on English can be applied directly to a supported low-resource language without any target-language labels, with performance degrading gradually as linguistic distance increases. Few-shot learning with as few as ten to a hundred target-language examples improves results substantially while keeping annotation costs low.

Synthetic data generation. Back-translation converts existing labelled data into low-resource languages via machine translation, generating synthetic training examples at minimal cost. Paraphrasing and synonym substitution add further variation. Translation noise is a limitation but synthetic examples still improve robustness and reduce overfitting on small datasets. Machine translation has been used to extend English sentiment datasets into languages like Swahili.

Crowdsourcing and community initiatives. Platforms like Amazon Mechanical Turk engage native speakers as annotators, providing cultural and linguistic authenticity at lower cost than expert annotation. Community-driven initiatives have produced some of the most impactful resources for low-resource language NLP: the Masakhane project focuses on African languages, AI4Bharat addresses Indian languages (including the IndicNLP dataset covering multiple Indian languages), and Mozilla Common Voice collects speech data across many languages. OPUS is a multilingual parallel corpus covering hundreds of language pairs. These initiatives demonstrate that community-driven approaches can achieve coverage that commercially motivated data collection does not reach.

Weak supervision and self-supervised learning. Emoji-based weak supervision treats emojis in social media posts as noisy proxy sentiment labels, enabling model training on large volumes of unlabelled data without manual annotation. Self-supervised tasks including masked language modelling allow models to learn language structure from raw text, reducing the dependence on labelled data for downstream sentiment classification.

Zero-shot and few-shot learning. Multilingual GPT and similar large-scale generative models demonstrate the ability to perform sentiment analysis in low-resource languages — including

Amharic — without any language-specific fine-tuning, by leveraging representations learned during multilingual pre-training. Active learning and adaptive sampling iteratively select the most informative examples for annotation, minimising labelling effort while maximising the information gained per annotation.

5.8 State-of-the-Art Model Performance and Generalisation (RQ6)

5.8.1 Multilingual Models

The cross-lingual generalisation capabilities of leading multilingual models follow a consistent pattern across the reviewed literature. Performance is strong for high-resource language pairs with structural similarity to the languages best represented in the pre-training corpus, moderate for structurally similar low-resource languages, and substantially weaker for typologically distant low-resource languages. mBERT's 104-language coverage provides reasonable cross-lingual transfer for most European languages but degrades noticeably for morphologically complex and typologically distant languages, reflecting the unequal representation of languages in Wikipedia-based pre-training. XLM-R's Common Crawl pre-training provides substantially better low-resource language coverage due to the greater volume and diversity of the crawled web text compared to Wikipedia.

Table 5.1. Performance of different types of models on popular datasets

Dataset	Model Type	Performance Metric	Value
General Textual Datasets	Textual Approaches (CNNs, LSTMs)	Accuracy	87.9%
Audio Sentiment Datasets	Audio Models (MFCC + CNNs)		92.7%
Microblog and Social Media Datasets	Joint Image-Text Models		87.6%
CMU-MOSI	Multimodal Approaches (Audio-Visual, Tri-Modal)		92.28%
CMU-MOSEI, CMU-MOSI	Tensor and Attention-based Fusion model		80%
CMU-MOSI	Self-MM	F1 Score	84.30%
Multilingual Tasks	XLM-T (Multilingual Tasks)		56.39%
Multilingual Setup	Machine Translation + Sentiment Models		71%
	MTSA (CMU-MOSEI)	MAE	0.541
	CMU-MOSI (other model)		0.696

5.8.2 Multimodal Models

VisualBERT and LXMERT align image region features with text tokens through shared transformer attention, performing well when visual and textual cues are complementary. Performance drops in low-resource language settings and when modalities carry conflicting sentiment signals. MMBT and UNITER use learned fusion layers to combine modalities and achieve competitive benchmark results. CLIP offers strong zero-shot text-image matching but does not extend to audio, limiting it to the text-image subset of multimodal tasks.

The fundamental limitation across all multimodal models is dependence on large aligned training datasets, which are almost exclusively available in English. Extending multimodal sentiment analysis to other languages requires both multilingual model capability and multimodal datasets in those languages simultaneously — a doubly demanding requirement that has kept multimodal approaches confined to the same high-resource languages that dominate unimodal research.

5.9 Key Limitations and Future Research Directions (RQ7)

5.9.1 Current Limitations

Computational cost. Large multilingual and multimodal transformers require resources that most research teams cannot access, particularly in lower-income countries and smaller institutions. This concentrates model development in a small number of well-resourced organisations, limiting both the diversity of perspectives in research and the geographic equity of deployment.

Evaluation fragmentation. No standardised evaluation framework exists for multilingual multimodal sentiment analysis. Studies use different datasets, metrics, and task formulations, making cross-paper comparison unreliable and genuine progress difficult to measure.

Cultural and linguistic nuance. Models consistently underperform on idiomatic, culturally specific, or context-dependent sentiment expressions. This is not purely a data problem — architectures that treat sentiment as a universal function of surface linguistic form will miss culturally embedded evaluative meaning regardless of how much data they are trained on.

Multimodal integration complexity. Timing mismatches between audio, video, and text, inconsistent recording quality, and missing modality channels all reduce real-world accuracy below what controlled benchmarks suggest.

5.9.2 Future Research Directions

Developing **smaller, distilled versions of multilingual and multimodal models** that run on resource-constrained hardware would make advanced sentiment analysis accessible beyond well-resourced institutions. Knowledge distillation, quantisation, and pruning techniques applied to models like XLM-R and multimodal transformers represent the most direct path to this goal.

Culturally and linguistically aware architectures. Models that explicitly incorporate sociolinguistic knowledge — regional dialect markers, cultural context embeddings, morphological features specific to target languages — have demonstrated improvements over generic multilingual models in targeted evaluations. Expanding this approach beyond individual well-studied cases toward a general framework for cultural contextualisation is a high-priority research direction.

Unified evaluation benchmarks. Developing standardised, cross-lingual, cross-modal evaluation frameworks — analogous to the role that CMU-MOSI and CMU-MOSEI play within multimodal English sentiment research — for multilingual and multimodal settings would enable reliable progress measurement and facilitate reproducible comparison across systems.

Dataset development for underrepresented communities. Open-source, balanced datasets covering a wider range of languages and modalities, developed in partnership with native-speaker communities, represent the most durable foundation for improving multilingual sentiment analysis. Community partnerships like Masakhane and AI4Bharat demonstrate that this is achievable with appropriate institutional support and incentive structures.

Domain adaptation. Models fine-tuned in one domain (news, social media) frequently generalise poorly to others (healthcare, legal, education) even within the same language. Systematic domain adaptation research — characterising the conditions under which cross-domain transfer succeeds or fails and developing adaptation methods appropriate for low-resource settings — would improve the practical deployability of sentiment systems.

5.10 Summary

This chapter has presented a systematic review of advancements and challenges in multilingual and multimodal sentiment analysis, drawing on 18 primary studies selected through a structured quality assessment from an initial pool of 29, all published between 2020 and 2023.

The central finding is a generalisation gap. Current models perform well on English-dominant benchmarks — audio models reach 92.7% accuracy, multimodal approaches 92.28% on CMU-MOSI, Self-MM 84.30% F1 on binary classification — but these results do not transfer reliably to the linguistic diversity, cultural specificity, and modality complexity of real deployment. The causes are structural: data scarcity for low-resource languages, annotation costs, culturally embedded evaluative language, and computational barriers outside well-resourced institutions. Mitigation strategies help but do not close the gap.

The practical lesson is the same one that runs through the entire thesis: understand the deployment context before selecting a model. A system that excels on English social media data is not a safe starting point for a healthcare application in a low-resource language. The deployment environment must drive the design choices, not the other way around.

CHAPTER 6
CROSS-DOMAIN SYNTHESIS AND DISCUSSION

6.1 Introduction

The three studies that constitute this thesis address problems in domains that share no surface-level resemblance. The first concerns the statistical estimation of causal effects from observational labour market and health data. The second concerns the security of digital authentication credentials against a specific class of malicious software. The third concerns the computational analysis of human sentiment across multiple languages and modalities. A reader encountering these three papers independently, without the framing provided by this thesis, would have no reason to connect them. They draw on different literatures, use different methods, evaluate different systems, and make different kinds of claims. Yet they are connected — not superficially, through the broad label of artificial intelligence or data science, but structurally, through a shared intellectual problem that each study instantiates in its own domain. That problem is the following: systems and methods that are designed and evaluated under idealised conditions exhibit systematic, consequential, and often underappreciated failures when those idealised conditions do not hold in practice.

6.2 The Shared Structural Problem: The Idealisation–Deployment Gap

Every method and system is built on assumptions about the environment it will operate in. In statistical settings these assumptions are stated explicitly; in engineering they are often implicit — embedded in training data choices, benchmark selection, or architecture decisions — but equally consequential when violated. The idealisation–deployment gap is the divergence between the conditions assumed during design and evaluation and those actually encountered in deployment. It is not a calibration problem that better engineering can eliminate. It is structural: the controlled conditions that make rigorous evaluation possible are precisely what makes the real world hard to represent in evaluation. The gap persists even when researchers act in good faith. What makes this gap consequential is that its effects are uneven. Some methods degrade gracefully when conditions depart from the ideal; others fail catastrophically. Which methods are robust and which are brittle is not reliably predictable from theoretical properties alone, and greater sophistication does not reliably confer greater robustness. Across all three domains in this thesis, the relationship between theoretical sophistication and practical performance is complicated, often inverted, and only explicable through careful analysis of how each method's design interacts with the specific violation it faces.

6.3 Three Manifestations of the Same Problem

6.3.1 The Sophistication Paradox in Causal Inference

In causal inference, the gap manifests as the sophistication paradox. Methods like Causal Forests, Double Machine Learning, and the X-Learner carry theoretical guarantees that hold

under adequate covariate overlap and well-behaved treatment effect surfaces — conditions that are often stated as asymptotic requirements and rarely met in practice. Applied causal inference routinely involves datasets where treated and control units come from substantially different populations and where treatment effects vary heterogeneously across units. The LaLonde datasets are the canonical example: treated units are NSW job training programme participants while control units are drawn from general labour force surveys, creating exactly the kind of distributional mismatch that breaks the assumptions underlying sophisticated estimators.

Under these conditions, the benchmark reveals a striking inversion: the naive difference-in-means estimator, which performs no covariate adjustment and has no theoretical justification in confounded observational settings, achieves RMSE of $1,132 on LaLonde PSID while the Causal Forest — the most theoretically sophisticated estimator — achieves RMSE of $15,850, a factor of fourteen worse. The pattern holds broadly: sophistication correlates negatively with performance under the overlap violation of the LaLonde datasets.

The Causal Forest's worst failures on LaLonde CPS are variance-dominated — it is approximately unbiased on average across the 100 samples but has a standard deviation of $20,299, meaning it is catastrophically wrong on individual samples in an unpredictable manner. Estimators that avoid cross-group prediction — specifically the T-Learner, which trains entirely separate models on each treatment arm — are the most robust, because they never require extrapolation across the covariate gap that causes the other methods to fail. The practitioner implication is a revised diagnostic workflow: diagnose treatment effect heterogeneity before overlap, because heterogeneity is the primary driver of performance degradation. This is counterintuitive — the standard applied causal inference workflow focuses on overlap diagnostics — but it is mandated by the empirical evidence.

6.3.2 The Security-Implementation Gap in Cybersecurity

In cybersecurity, the idealisation–deployment gap takes a different but structurally parallel form. The idealised conditions implicitly assumed by the security literature are that organisations deploy the most effective available defences and that users behave in ways consistent with security best practice. The deployment reality is that the most effective authentication technologies — FIDO2 hardware security keys and platform authenticators, which achieve 99% phishing resistance and represent the fastest and most user-friendly MFA methods in direct comparison — are deployed by a small fraction of organisations, while SMS-based OTP with zero phishing resistance remains dominant. The sophistication paradox in cybersecurity does not manifest at the level of individual technical methods — FIDO2 genuinely is more secure than SMS OTP under every measured dimension, unlike the causal inference case where the sophisticated method can be worse than the naive one. It manifests instead at the level of the gap between what the technical literature recommends and what is actually deployed. The detection literature provides a precise analogue, however: integrated EDR platforms are theoretically the most effective detection approach, yet 60–70% of contemporary infostealer variants evade sandbox detection, and the 2–4 week new variant release cycle continuously resets detection effectiveness to near zero for each release. The data

shows the opposite: FIDO2 platform authenticators have 2.1-second authentication times and 1% error rates, compared to 25 seconds and 8–18% error rates for SMS OTP. The security-implementation gap is partly sustained by a misunderstanding of the security-usability tradeoff that the empirical evidence directly refutes. The response to the security-implementation gap also requires a form of context-first thinking that parallels the causal inference diagnostic framework. An organisation's optimal authentication migration pathway depends on its specific legacy system constraints, user population characteristics, compliance requirements, and threat exposure profile. A standardised recommendation that ignores these contextual factors will fail to be adopted, because the barriers it does not address will prevent implementation regardless of how sound the technical recommendation is.

6.3.3 The Generalisation Gap in Multilingual Sentiment Analysis

In multilingual and multimodal sentiment analysis, the idealisation–deployment gap takes the form of what Chapter 5 terms the generalisation gap. The idealised conditions under which current state-of-the-art models are designed and evaluated are English-dominant, high-resource, single-domain benchmark datasets. CMU-MOSI, CMU-MOSEI, and the Twitter datasets used in the reviewed literature represent controlled evaluation environments with well-aligned multimodal data, clean annotations, and language distributions that over-represent English and a small number of other high-resource languages. The deployment reality is that sentiment analysis applications operate in environments characterised by linguistic diversity — hundreds of languages, many with minimal training data; code-switching between language pairs; culturally specific sentiment expressions that do not translate across cultural contexts; and multimodal data with timing mismatches, recording quality variation, and missing channels. The performance gap between benchmark conditions and these deployment conditions is large and persistent. Models that achieve 92.7% accuracy on audio sentiment benchmarks may fail substantially on code-switched social media text in a low-resource language, not because the model is poorly designed but because the benchmark conditions do not represent the deployment conditions. The mechanism in the NLP case is distributional shift at three levels simultaneously: linguistic shift (the target language differs from the training language), domain shift (the target domain differs from the training domain), and modality shift (the target application requires modalities not represented in training data). Unlike the causal inference case, where the failure is attributable primarily to a single mechanism (treatment effect heterogeneity driving cross-group extrapolation error), the NLP generalisation gap is multi-causal, and each contributing cause requires a distinct mitigation strategy. The response to the NLP generalisation gap also requires context-first thinking: understanding the linguistic profile, modality availability, cultural context, and computational constraints of the target deployment environment before selecting a model. A model selected purely based on benchmark performance without attention to the match between benchmark conditions and deployment conditions will generalise poorly, and the analyst who selects it will have misled themselves about the system's actual capability.

6.3.4 The Structural Identity of the Three Manifestations

Placing the three manifestations side by side reveals their structural identity. In each domain, there is a gap between the conditions under which methods or systems are designed and evaluated and the conditions under which they are deployed. In each domain, this gap is not primarily a problem of technical capability — the technical knowledge required to build better-performing systems is largely available — but a problem of the interaction between method design and deployment context. In each domain, the gap is sustained by a combination of factors that include both genuine technical barriers and factors that are better described as cognitive or organisational: misunderstandings about the implications of idealized evaluations, misplaced confidence in theoretical sophistication, legacy constraints, and the difficulty of diagnosing deployment context accurately before committing to a method. The following table summarises the parallel structure:

TABLE 6.1: Parallel structure of the idealisation–deployment gap across the three contributing studies of this thesis. Each column represents one domain; each row captures a dimension of the shared structural problem. The convergence of all three domains on an identical common prescription — diagnose context before selecting method — emerging independently from separate empirical evidence, constitutes the primary thesis-level contribution beyond the individual chapter findings.

Dimension	Causal Inference (Ch. 3)	Cybersecurity (Ch. 4)	Sentiment Analysis (Ch. 5)
Idealised condition assumed	Covariate overlap; homogeneous effects	Perfect detection; rational security adoption	High-resource language; aligned multimodal data
Deployment violation	Poor overlap; heterogeneous effects	Evasion-capable malware; weak authentication deployment	Linguistic diversity; data scarcity; cultural specificity
Manifestation	Sophistication paradox	Security-implementation gap	Generalisation gap
Most robust approach	T-Learner; naive estimator under near-zero confounding	FIDO2; credential-agnostic authentication	Cross-lingual transfer; community-driven data; few-shot learning
Key diagnostic	Heterogeneity index before overlap	Threat profile and legacy constraints before method selection	Linguistic and cultural profile before model selection
Common prescription	Diagnose context before selecting method	Diagnose context before selecting method	Diagnose context before selecting method

The convergence on the common prescription — diagnose context before selecting method — is not a post-hoc rationalisation imposed on disparate findings. It emerges independently from

each study's empirical results, and it represents the thesis's primary intellectual contribution beyond the individual domain-specific findings of its three papers.

6.4 Methodological Contributions of the Thesis

Beyond their individual empirical findings, the three studies collectively make methodological contributions that are worth examining as a group.

6.4.1 Benchmarking Under Realistic Conditions

All three studies share a methodological commitment to evaluation under realistic, rather than optimised, conditions. Chapter 3 uses the RealCause framework specifically because it provides data-generating processes grounded in real experimental data rather than fully synthetic distributions that may advantage methods with matching functional form assumptions. The 90–95% sandbox detection rate reported under non-evasive conditions is explicitly contrasted with the 60–70% evasion rate achieved by contemporary infostealers that specifically target sandbox environments — a realistic deployment condition that fundamentally changes the evaluation conclusion. Similarly, the comparative evaluation of MFA methods is conducted across security, usability, cost, and accessibility dimensions simultaneously, rather than on security effectiveness alone, because the deployment decision involves all four dimensions and a recommendation based on security alone will not survive contact with organisational reality. Chapter 5 selects primary studies specifically on the basis of whether they address generalisation across languages and modalities, and the quality assessment framework explicitly penalises studies that validate findings only on controlled datasets without demonstrating generalisation to real-world conditions or diverse domains. The emphasis throughout is on what models can do in deployment rather than what they can achieve under optimal benchmark conditions. This shared methodological orientation — realistic evaluation before recommendation — is itself a contribution to a literature that has been criticised across all three domains for over-reliance on controlled benchmarks that do not represent deployment conditions.

6.4.2 Mechanism Identification, Not Just Performance Comparison

All three studies go beyond establishing which method performs better to identify why. Chapter 3's knob-sweep experiment is designed explicitly to test the mechanism hypothesis: does overlap or heterogeneity drive the performance differences? The experiment's finding — that heterogeneity is the primary driver while overlap has no statistically significant independent effect — is not a performance observation but a mechanistic finding that has direct prescriptive implications for practitioner workflows. Without the knob sweep, the empirical performance differences would be attributable to multiple candidate mechanisms, and the diagnostic implications would be ambiguous. Chapter 4's detailed mapping of the detection-evasion arms race — specifically the temporal analysis showing that new infostealer variant release cycles (2–4 weeks) are shorter than the detection effectiveness recovery period — identifies a

structural asymmetry between attack and defence that explains why detection-centric strategies are insufficient as primary defences regardless of the sophistication of the detection technology.

6.4.3 Actionable Decision Frameworks

All three studies translate their findings into practical decision frameworks. Chapter 3 provides a two-stage diagnostic workflow — heterogeneity first, then overlap — with estimator recommendations grounded in benchmark evidence. Chapter 4 provides a multi-dimensional authentication evaluation framework and a phased migration pathway, each recommendation backed by quantitative evidence from the review. Chapter 5 provides model selection criteria based on linguistic profile, modality availability, and computational constraints, with strategy recommendations for different data scarcity and deployment conditions. The common thread is a view that a thesis-level contribution should go beyond documenting evidence — it should translate that evidence into guidance a practitioner can act on directly. This translation from empirical evidence to prescriptive guidance is where the practical value of empirical research is realised, and it is a contribution that the individual papers — constrained by journal and conference page limits — can only partially accomplish.

6.5 Implications for AI-Driven System Design

The synthesis across three domains supports a set of design principles for AI-driven systems intended for real-world deployment. These principles are not novel in isolation — each has been stated in various forms within individual domains — but their convergent derivation from independent evidence across three different fields strengthens their status from domain-specific heuristics to general principles for reliable system design.

6.5.1 Evaluate Against Deployment Conditions, Not Just Benchmark Conditions

The most direct implication of the idealisation–deployment gap across all three domains is that evaluation methodology shapes the conclusions drawn about a system's capability, and that evaluation on controlled benchmarks systematically overestimates performance in deployment. Causal Forest achieves RMSE ≈ 0.010 on the Twins positive control dataset but RMSE $15,850–$20,421 on the LaLonde datasets — a difference of three orders of magnitude that is entirely attributable to the difference in distributional conditions between the two evaluation environments. Sandbox detection achieves 90–95% effectiveness against non-evasive malware but is evaded by 60–70% of contemporary variants that specifically target sandbox environments — a difference attributable to the gap between controlled laboratory conditions and the adversarial deployment environment. Multilingual models achieve strong benchmark performance on high-resource language datasets but generalise substantially less well to low-resource languages, code-switched input, and non-English multimodal settings.

In each case, the benchmark result is not wrong — it accurately reflects performance under the benchmark conditions — but it is misleading when used to predict deployment performance, because the benchmark conditions do not represent the deployment conditions. System

designers who select methods based on benchmark performance without assessing the match between benchmark and deployment conditions will systematically make worse choices than those who insist on realistic evaluation. This is not a trivial insight: the academic incentive structure rewards benchmark performance improvement, creating systematic pressure toward evaluation under controlled conditions and away from realistic deployment assessment.

6.5.2 Prefer Robustness Over Optimality Under Idealized Conditions

Across all three domains, the methods that perform best in deployment are not those that achieve the highest performance under idealised conditions — they are those that degrade most gracefully when idealised conditions are violated. The T-Learner is not the best causal estimator under good overlap (Causal Forest and AIPW achieve equivalent performance on Twins); it is the most robust estimator when overlap is poor, because its design — training entirely separate models on each treatment arm — avoids the cross-group prediction that causes other methods to fail. FIDO2 platform authenticators are not the simplest authentication method; they are the most robust against the range of real-world attack vectors (phishing, SIM-swapping, MITM) that the deployment environment presents. Cross-lingual transfer learning with few-shot fine-tuning is not the highest-performing approach on high-resource language benchmarks; it is the most robust approach when the deployment context involves a low-resource language where large labelled datasets are unavailable. The preference for robustness over optimality under idealised conditions is a principle that runs counter to the standard academic evaluation framework, which measures performance under controlled conditions and rewards the method that achieves the highest controlled performance. A design framework oriented toward deployment reliability requires explicitly rewarding graceful degradation and penalising catastrophic failure, even when catastrophic failure occurs in conditions that are not represented in the standard benchmark.

6.5.3 Diagnose Context Before Selecting Methods

The prescriptive convergence across all three studies — diagnose context before selecting method — reflects a principle about the appropriate ordering of the design process. Method selection is not the first step in system design; it is a step that comes after the deployment context has been characterised with sufficient precision to make an informed selection. In causal inference, this means assessing treatment effect heterogeneity and overlap before selecting an estimator. In authentication security, it means assessing the organisation's legacy system constraints, user population characteristics, and threat exposure profile before selecting an authentication migration pathway. In sentiment analysis, it means characterising the target language profile, modality availability, cultural context, and computational constraints before selecting or building a model. This principle is violated routinely in practice, in all three domains. Practitioners use Causal Forests because they are state-of-the-art, without first checking whether their data has the overlap and heterogeneity properties for which Causal Forest is appropriate. Security teams deploy SMS OTP because it is the default recommendation in legacy compliance frameworks, without assessing whether their specific user population and threat profile make FIDO2 platform authenticators the better choice. NLP teams deploy

mBERT because it is widely available and well-documented, without assessing whether their target languages are well-represented in mBERT's training corpus or whether XLM-R's larger and more diverse pre-training would provide meaningfully better coverage. The commonality of this violation — and the consistent evidence across three domains that it leads to poor outcomes — suggests that there is a structural incentive problem in how technical methods are communicated and recommended. Methods are typically presented with evidence of their performance under favourable conditions and with theoretical arguments for their superiority, without adequate guidance on the conditions under which they fail and the diagnostic steps required to determine whether those conditions are present. Addressing this gap in how methods are communicated is as important as advancing the methods themselves.

6.5.4 Layer Defences and Approaches

A fourth design principle, which is most explicitly developed in Chapter 4 but present implicitly in all three studies, is that no single method provides comprehensive coverage of all relevant failure modes, and that layered approaches achieve non-linear gains relative to their individual components. In cybersecurity, MFA alone achieves approximately 96.3% account takeover reduction, behavioural biometrics alone achieve 91–95% attack detection, and their combination achieves approximately 99.7% — a non-linear improvement from layering. In causal inference, combining heterogeneity diagnostics, overlap diagnostics, and placebo validation provides substantially stronger assurance than any single check alone. In sentiment analysis, combining cross-lingual transfer with domain-specific fine-tuning and culturally aware augmentation produces more robust deployment performance than any single strategy. The same principle applies to evaluation. No single metric captures all relevant failure modes. RMSE reflects bias and variance but not confidence interval calibration. Accuracy reflects average performance but not failure distribution. Benchmark F1 reflects controlled-condition performance but not deployment generalisation. Meaningful evaluation requires multiple metrics chosen specifically to make different classes of failure visible.

6.6 Limitations of the Thesis

The three domains represented in this thesis were not selected through a systematic procedure designed to ensure that they constitute a representative or exhaustive sample of the domains in which the idealisation–deployment gap manifests. They were selected because they represent three independent research contributions made during the course of the doctoral programme. The convergent finding — that all three instantiate the same structural problem — may reflect something genuine about the ubiquity of the idealisation–deployment gap across AI and data-driven system domains, or it may partly reflect the fact that the three research projects were conducted by the same researcher and were influenced, even unconsciously, by shared theoretical commitments. Independent replication of the synthesis finding — demonstrating that the same structural problem appears in domains selected independently by different researchers — would strengthen the generality of the argument. The individual studies have limitations that Chapter 3, Chapter 4, and Chapter 5 each acknowledge in detail. Chapter 3's benchmark evaluates ATE estimation only, not CATE estimation or confidence interval

coverage, and is restricted to three datasets from two domains. Chapter 4's systematic review is subject to selection bias, publication bias, and geographic concentration in North American and Western European contexts, and its findings about emerging technologies are constrained by the early maturity of those technologies. Chapter 5's review is based on 18 primary studies from a three-year window, and the field is evolving rapidly enough that some findings may already be partially superseded by developments not captured in the review. The practical decision frameworks derived from all three studies require validation in deployment settings that differ from the evaluation conditions under which they were developed. The causal inference decision framework, for example, recommends using the T-Learner under moderate to high heterogeneity — but this recommendation is grounded in evidence from three datasets from two specific domains, and its generalisability to other domains and data-generating processes requires further empirical validation. Finally, the thesis does not address the ethical dimensions of the idealisation–deployment gap in depth, though they are significant. The generalisation gap in multilingual NLP has direct implications for linguistic equity: sentiment systems that perform well for English speakers and poorly for speakers of low-resource languages represent a form of technological discrimination that is invisible when evaluation is conducted exclusively on high-resource language benchmarks. The security-implementation gap has implications for the differential vulnerability of individuals and organisations with different levels of technical sophistication and resources. The causal inference sophistication paradox has implications for the reproducibility and reliability of policy evaluations that use sophisticated ML methods without first checking whether their data conditions are appropriate for those methods. These ethical dimensions deserve more systematic treatment than they receive in the individual chapters, and their integration into the design principles of Section 6.5 is incomplete.

6.7 Summary

This chapter has argued that the three studies constituting this thesis are unified by a single structural problem — the idealisation–deployment gap — that manifests in each domain as a specific, named phenomenon: the sophistication paradox in causal inference, the security-implementation gap in cybersecurity, and the generalisation gap in multilingual sentiment analysis. The parallel structure of these three phenomena is not superficial: in each case, the gap is sustained by the interaction between method design assumptions and deployment context violations; in each case, the most robust approaches are those designed to degrade gracefully rather than to achieve optimal performance under idealised conditions; and in each case, the appropriate response is context-first diagnostic thinking before method selection. The convergent prescription — diagnose context before selecting method — emerges independently from the empirical evidence of all three studies. It is this convergence, achieved across three genuinely different domains using genuinely different methods, that constitutes the thesis's primary intellectual contribution beyond the individual papers. It suggests that the idealisation–deployment gap is not a domain-specific technical problem requiring a domain-specific technical solution, but a general structural feature of the relationship between scientific method and real-world application that requires a general shift in how methods are designed, evaluated, and recommended

CHAPTER 7
CONCLUSIONS AND FUTURE WORK

7.1 Introduction

This thesis set out to examine a problem that is simultaneously methodological and practical: the conditions under which AI-driven systems and data-driven methods fail when the assumptions under which they were designed or evaluated are violated by real-world deployment conditions. Three independent research contributions — a causal inference benchmark, a systematic review of infostealer malware and authentication security, and a systematic review of multilingual and multimodal sentiment analysis — were brought together under a single analytical framework centred on the concept of reliability and resilience under realistic violations of idealised assumptions. The preceding chapters have presented each study in detail, reviewed the relevant literature across all three domains, and synthesised the findings into a set of design principles that span domains. This final chapter closes the thesis in three movements. It first summarises the contributions of each study and the thesis as a whole, reviewing what has been established and why it matters. It then revisits the four research objectives stated in Chapter 1 and provides explicit answers to each, anchored in the specific evidence produced by the studies. It concludes with a structured account of future research directions, organised both within each domain and at the intersection of the thesis's unifying theme.

7.2 Summary of Contributions

7.2.1 Contribution 1: The Causal Estimator Benchmark

Chapter 3 presented a systematic benchmark of seven causal estimators — Naive difference-in-means, Inverse Probability Weighting, OLS/ANCOVA, Augmented Inverse Probability Weighting, T-Learner, X-Learner, and Causal Forest with Double Machine Learning — evaluated across three datasets from the RealCause framework over 100 independently generated samples each, supplemented by a controlled knob-sweep experiment varying covariate overlap and treatment effect heterogeneity independently across a five-by-three parameter grid. The total benchmark corpus comprised 3,600 evaluation records. The primary contribution is the empirical establishment and mechanistic explanation of a sophistication paradox: on datasets with poor covariate overlap, the naive difference-in-means estimator achieves the lowest RMSE of all seven estimators on LaLonde PSID ($1,132) while the Causal Forest achieves the highest ($15,850) — a factor of fourteen difference in the direction opposite to what the theoretical sophistication ordering would predict. On LaLonde CPS, the T-Learner achieves the lowest RMSE of all seven estimators ($2,816), outperforming the Causal Forest ($20,421) by a factor of seven. On the Twins dataset, which has excellent overlap, the paradox completely disappears and all adjusted estimators converge to RMSE ≈ 0.010, confirming that the LaLonde differences are caused by distributional properties of the data rather than implementation artefacts. The controlled knob-sweep experiment provides the mechanistic explanation: it is treatment effect heterogeneity, not overlap degradation, that drives estimator

ranking reversal. OLS/ANCOVA degrades by a factor of 17 as heterogeneity increases from zero to its full value, while remaining essentially flat across all five overlap levels. Kruskal-Wallis tests confirm that the heterogeneity effect is highly significant for OLS ($H = 264.99$, $p = 2.88 \times 10^{-58}$) while the overlap effect is non-significant ($H = 4.65$, $p = 0.326$). The T-Learner is the only estimator for which neither the heterogeneity effect nor the overlap effect achieves statistical significance, making it the only genuinely robust estimator across the full parameter grid. The bias-variance decomposition reveals two qualitatively different failure modes: OLS, AIPW, and IPW fail primarily through large, consistent, systematic bias; Causal Forest on LaLonde CPS fails primarily through high variance (standard deviation $20,299), producing individual-sample estimates that are unpredictably unreliable despite approximate unbiasedness in expectation. The latter failure mode is more dangerous in practice because a single-sample practitioner — the typical applied setting — cannot detect it.

7.2.2 Contribution 2: The Infostealer Systematic Review

Chapter 4 presented a PRISMA-compliant systematic review of infostealer malware and credential theft prevention, drawing on 166 sources from 3,495 initial records across four research questions covering infostealer mechanisms, account hijacking prevention, session hijacking mitigation, and implementation barriers.

The central contribution is a structured mapping of the security-implementation gap. The problem is not a lack of effective technology — FIDO2 platform authenticators provide 99–99.9% phishing resistance, the fastest authentication times (2.1–3.2 seconds), the lowest error rates (1–2%), the highest user satisfaction (82–88%), and a three-year total cost of ownership 22% lower than password-only systems. The problem is that these technologies are not being used. SMS OTP — with zero phishing and zero SIM-swap resistance — remains dominant because of legacy compatibility and compliance inertia. Detection-centric strategies cannot close the persistent vulnerability window created by the 2–4 week new variant release cycle.

The emerging technology assessment concludes that behavioural biometrics are production-ready and warrant high investment priority; blockchain identity suits enterprise B2B contexts only; and visual cryptography is limited to niche applications, with a fundamental accessibility barrier affecting 2.2 billion people with visual impairments.

The broader significance is a reframing of the security problem. The limiting factor is not technical capability but the adoption gap — and closing it requires engaging with usability, organisational, and human factors that purely technical research cannot address.

7.2.3 Contribution 3: The Multilingual and Multimodal Sentiment Review

Chapter 5 presented a systematic review of multilingual and multimodal sentiment analysis, drawing on 18 primary studies selected from 29 initial studies published between 2020 and 2023, guided by seven research questions.

The central contribution is a structured characterisation of the generalisation gap — the divergence between strong benchmark performance on English-dominant, controlled datasets and substantially weaker performance under the linguistic diversity, data scarcity, and multimodal complexity of real deployment. Current performance figures give a sense of the ceiling: audio MFCC-CNN models reach 92.7% accuracy, multimodal approaches 92.28% on CMU-MOSI, Self-MM 84.30% F1 on binary classification, XLM-T improving from 40.29% to 56.39% F1 with fine-tuning, and MTSA achieving the lowest MAE of 0.541 on CMU-MOSEI. These are controlled-condition results, not typical deployment performance.

Five structural limitations are identified: low-resource data scarcity and imbalance; cultural and linguistic nuance failures in models that treat sentiment as language-neutral; multimodal integration challenges from timing mismatches and noise; computational barriers restricting access to large models; and evaluation fragmentation that makes cross-study comparison unreliable. For each, the review maps current mitigation strategies and their documented effectiveness.

The broader significance is synthesis across a fragmented field. NLP, speech processing, computer vision, and affective computing each engage with multilingual multimodal sentiment from different angles without fully engaging with each other. This review provides a unified view and names the generalisation gap as the structural limitation each community is encountering independently.

7.2.4 The Thesis-Level Contribution

Beyond the individual studies, the thesis makes a synthesis-level contribution: demonstrating that the sophistication paradox, the security-implementation gap, and the generalisation gap are structurally identical manifestations of the same underlying problem. In each domain, the gap lies between idealised evaluation conditions and realistic deployment conditions. In each, the most robust methods are those designed to avoid the specific assumption violation the deployment context induces. And in each, the appropriate response is to diagnose the deployment context before selecting a method rather than defaulting to the most sophisticated or best-benchmarked option.

This convergence across three independent empirical studies in three different fields elevates the prescription — diagnose context before selecting method — from a domain-specific heuristic to a general principle for reliable AI system design.

7.3 Answers to the Research Objectives

The first objective has been fully achieved. Across 3,600 evaluation records, the benchmark establishes that estimator performance rankings are strongly data-dependent, driven primarily by treatment effect heterogeneity rather than overlap. The T-Learner is the most robust ML estimator under distributional mismatch, with statistically confirmed advantages over all competitors on LaLonde CPS. The knob-sweep experiment isolates the mechanism: OLS

degrades by 17× as heterogeneity increases while remaining flat across overlap levels, and the T-Learner is the only estimator with non-significant Kruskal-Wallis results for both dimensions.

The decision framework translates these findings into a two-stage diagnostic procedure — heterogeneity before overlap — with specific estimator recommendations for each diagnostic outcome, operationalised through computable diagnostics that practitioners can apply before analysis. The original motivation — the absence of benchmarks simultaneously evaluating overlap and heterogeneity, and the gap between theoretical guarantees and real-world performance — is directly addressed.

The second objective was to conduct a systematic literature review of infostealer mechanisms and prevention techniques, characterise the security-implementation gap, identify the factors sustaining it, and evaluate emerging authentication technologies against a multi-dimensional framework. This objective has been fully achieved. The review maps the security-implementation gap with specificity across four dimensions: technical barriers (legacy system incompatibility affecting 72% of enterprises with critical applications unable to support modern authentication), usability misperceptions (74% of security professionals citing user friction as the primary adoption barrier, despite FIDO2 having faster authentication times and lower error rates than SMS OTP), organisational factors (58% of organisations prioritising compliance requirements over more effective security measures when they diverge), and human behavioural factors (52% of users reusing passwords, 43% of employees surrendering MFA codes to social engineers). The multi-dimensional evaluation framework is applied to visual cryptography, blockchain identity, and behavioural biometrics across seven dimensions, producing the tiered investment prioritisation described in Section 7.2.2.

The **third objective** has been fully achieved. The review synthesises the performance landscape across languages and modalities, characterises the generalisation gap between benchmark and deployment conditions, and analyses five structural limitation categories. Research directions — lightweight model development, culturally aware architectures, unified evaluation benchmarks, community-driven dataset construction, and domain adaptation — are grounded in the documented limitations and prioritised by their expected impact on the generalisation gap rather than incremental benchmark gains. The original motivation — fragmentation across research communities and the absence of a unified account of the generalisation gap — is addressed by the cross-community synthesis and the identification of the generalisation gap as the structural problem each community is independently encountering.

7.3.4 The Unifying Thesis Objective

A fourth objective runs implicitly through the thesis and is made explicit in Chapter 6: to show that the three studies share a common structural problem and to derive from that shared structure a general principle for reliable AI system design. This objective has been achieved through the synthesis in Chapter 6, which establishes the structural identity of the sophistication paradox, security-implementation gap, and generalisation gap, and derives from that identity the

principle of context-first diagnostic thinking. The principle is not only descriptively accurate — it correctly characterises what the evidence across all three studies shows — but it is prescriptively actionable: practitioners in each domain can apply it through the specific diagnostic procedures the respective chapters provide, and system designers working across domains can apply the general version without domain-specific translation.

7.4 Directions for Future Research

7.4.1 Future Work in Causal Inference

Confidence interval evaluation. The benchmark in Chapter 3 evaluates point estimation accuracy exclusively. Confidence interval coverage — the probability that the estimator's reported interval contains the true ATE — is arguably more important for scientific inference than point RMSE and can behave very differently. Causal Forest reports confidence intervals through its honest splitting procedure that are theoretically valid under regularity conditions; whether these intervals achieve nominal coverage under the overlap and heterogeneity conditions of the benchmark is unknown and represents an important extension. AIPW's doubly robust confidence intervals are also worth evaluating alongside its point estimates. A coverage-focused extension of the benchmark would require substantially larger per-cell sample sizes in the knob sweep to achieve reliable coverage estimates, but would provide a more complete picture of practical estimation reliability.

CATE evaluation. The benchmark evaluates ATE estimation only. The Causal Forest and X-Learner are designed specifically to minimise the Precision in Estimation of Heterogeneous Effects (PEHE) — the integrated mean squared error of conditional average treatment effect estimates — rather than ATE RMSE. These methods may perform substantially better on their intended target (CATE) than on ATE, even under the same distributional conditions. A parallel benchmark on PEHE across the same overlap–heterogeneity grid would clarify whether the methods' poor ATE performance under distributional mismatch reflects a fundamental robustness failure or a misalignment between the evaluation metric and the method's design target.

Extension to additional datasets and domains. The benchmark uses three datasets from two substantive domains: labour economics (LaLonde) and infant health (Twins). Whether the heterogeneity-first finding holds beyond labour economics and infant health — in healthcare, A/B testing, social programme evaluation, or educational research — requires replication on datasets with different confounding structures and outcome distributions. The RealCause framework supports this extension; the immediate next step is applying the benchmark to IHDP, the Jobs dataset, and the ACIC challenge datasets, each with a corresponding knob-sweep experiment.

Hyperparameter Analysis. All estimators in the benchmark are evaluated at default hyperparameter settings, which reflects realistic practitioner behaviour but limits the strength of conclusions about method performance under optimal configurations. A systematic

sensitivity analysis varying the key hyperparameters of each ML estimator — tree depth, number of estimators, learning rate, regularisation strength — would establish whether default-versus-tuned performance differences are large enough to change the benchmark rankings and therefore whether tuning effort is justified under different distributional conditions.

Heterogeneity diagnostic development. The practical decision framework uses the T-Learner prediction variance as a heterogeneity index for Stage 1 diagnosis. This is a reasonable starting point but is not the only possible approach, and its statistical properties — bias, variance, sensitivity to sample size — have not been formally characterised.

7.4.2 Future Work in Cybersecurity

Longitudinal deployment studies. The review relies mainly on single-organisation, large-enterprise case studies. Longitudinal research tracking how deployment rates, incident rates, and user acceptance evolve over time following different migration strategies would provide much stronger evidence about the dynamics of the security-implementation gap.

Behavioural biometrics privacy. The primary barrier to wider adoption is privacy — 73% of users report discomfort once the full scope of continuous monitoring is explained. Federated learning architectures that keep raw behavioural data on user devices and share only model gradients, combined with differential privacy mechanisms, represent the most direct path to resolving this barrier without sacrificing security effectiveness.

SMB-specific research. The reviewed literature skews heavily toward large enterprises. Small and medium businesses face different constraints — lower technical capacity, tighter budgets, different compliance requirements — that change the cost-benefit calculus for authentication upgrades. Guidance calibrated to SMB contexts would address a substantial and currently underserved segment.

Post-authentication attack surface. As initial authentication strengthens, post-authentication privilege escalation and lateral movement are becoming the primary attack surface. Research integrating behavioural biometrics with zero-trust per-resource authorisation to monitor authenticated sessions for anomalous access addresses the next phase of the attacker-defender dynamic.

IoT authentication. Seventy-three percent of current authentication protocols are incompatible with IoT device constraints. As IoT becomes embedded in critical infrastructure — healthcare monitoring, industrial control, smart buildings — authentication security for these devices becomes increasingly urgent. Lightweight protocols leveraging device-level behavioural signatures as continuous verification represent a promising direction.

7.4.3 Future Work in Multilingual Sentiment Analysis

Lightweight model development. Large multilingual transformers are too computationally expensive for most research teams and resource-constrained deployments. Knowledge

distillation applied to models like XLM-R — training smaller student models to replicate larger teacher models — could substantially reduce deployment costs while retaining most cross-lingual transfer capability. Quantisation and pruning offer complementary compression paths. Systematic evaluation of the performance-efficiency tradeoff across these approaches, with specific attention to low-resource language performance, would give practitioners without large-scale infrastructure a principled basis for model selection.

Cultural contextualisation architectures. Current models consistently fail on sentiment expressions that are idiomatic, context-dependent, or carry social and historical connotations not recoverable from surface linguistic form. Architectures that explicitly incorporate sociolinguistic knowledge — cultural context embeddings, native-speaker-derived sentiment lexicons, morphological features specific to target language families — have shown improvements in targeted evaluations and warrant systematic development as a general framework rather than language-specific engineering. Model architectures that explicitly incorporate sociolinguistic knowledge — regional and cultural context embeddings, sentiment lexicons developed with native-speaker input, morphological feature extraction calibrated to specific language families — have shown improvements in targeted evaluations. Developing a general framework for cultural contextualisation that can be applied across language families and cultural groups, rather than engineering bespoke solutions for individual well-studied cases, represents a high-impact research direction.

Unified multilingual multimodal benchmarks. The evaluation fragmentation documented in Chapter 5 — different datasets, metrics, task formulations, and language coverages across studies — makes reliable progress measurement difficult and wastes research effort on incomparable comparisons. Developing a standardised multilingual multimodal sentiment benchmark — analogous to the role GLUE and SuperGLUE play in English NLP evaluation — covering a representative sample of languages with balanced high-resource and low-resource representation, standardised audio-visual-text alignment, and unified evaluation metrics would provide the common ground necessary for systematic progress in the field. Community efforts along the lines of Masakhane and AI4Bharat provide a model for how such benchmarks can be developed through inclusive, community-driven partnerships.

Domain adaptation for specialised applications. The review documents that domain shift is a significant contributor to the generalisation gap independent of linguistic shift. Sentiment systems trained on social media data perform poorly in healthcare, legal, or educational contexts even within the same language. Systematic research on domain adaptation for multilingual sentiment — characterising which aspects of domain shift are most damaging, developing adaptation methods effective for small-data domain shift, and establishing evaluation protocols for cross-domain generalisation — would improve practical deployability across the application domains documented in Chapter 5.

Multimodal low-resource language development. The intersection of multilingual and multimodal challenges is the most demanding frontier in the field: building sentiment systems that can operate across multiple modalities in low-resource languages requires both the

multilingual model development and the multimodal dataset construction described separately, simultaneously. Community-driven multimodal data collection — recording native speakers of low-resource languages with annotated audio, video, and text — combined with weak supervision techniques that reduce the annotation burden, represents the most tractable path toward this goal.

7.4.4 Cross-Domain Future Work

Formalising the idealisation–deployment gap. Chapter 6 articulates the idealisation–deployment gap as a structural feature of AI-driven system design that manifests across multiple domains. This articulation is currently descriptive and domain-specific in its evidence base. A more formal treatment — developing a general theoretical framework that characterises the conditions under which the gap is large versus small, the properties of methods that make them robust versus brittle under deployment violations, and the diagnostic procedures appropriate for different classes of deployment violation — would provide a foundation for the general principle of context-first design that is independent of specific domain evidence. Such a framework could draw on connections to domain adaptation theory in machine learning, worst-case analysis in theoretical computer science, and robustness theory in statistics.

Realistic evaluation methodology. All three studies in this thesis share a commitment to evaluation under realistic conditions, but each implements this commitment differently and informally. A systematic methodology for realistic evaluation of AI systems — specifying how to characterise deployment distributions, how to design evaluation protocols that represent those distributions, and how to communicate the gap between benchmark performance and deployment performance — would benefit the broader AI research community and provide a common language for the kind of evaluation this thesis advocates.

7.5 Closing Remarks

This thesis began with an observation that is obvious in hindsight: every method performs better under the conditions it was designed for than those it will actually face. What is less obvious is precisely how this gap manifests across domains, what mechanisms drive the most consequential failures, and what diagnostic principles can address them.

The three studies arrive at these principles from genuinely different starting points — a statistical benchmark on labour market data, a 166-source cybersecurity review, and an 18-study NLP review. They differ as much as three contributions within a single thesis can. The fact that they converge on the same structural problem and the same prescription — diagnose context before selecting method — is not a coincidence of framing.

It reflects something real about how AI-driven systems fail in the world, and about what it takes to build systems that fail gracefully rather than catastrophically when they encounter conditions they were not designed for. The naive difference-in-means estimator outperforms the Causal Forest by a factor of fourteen on datasets with poor overlap, because it makes no

assumptions about the relationship between covariates and outcomes and therefore has no assumptions to violate. FIDO2 platform authenticators outperform SMS one-time passcodes on every dimension — security, speed, accuracy, cost, user satisfaction — because their cryptographic architecture renders stolen credentials non-exploitable rather than relying on detecting theft before exploitation. Multilingual models with community-driven low-resource training data outperform models with superior benchmark scores in deployment because the training data represents the actual deployment distribution rather than a convenient proxy for it. In each case, the robust solution is the one that was designed for the actual conditions of use rather than for the conditions of evaluation. The principle that emerges from this convergence is not a counsel of conservatism — it does not argue for simpler methods, weaker authentication, or smaller models.

The argument is not for simpler methods or lower ambition. It is for honesty about what evaluation conditions actually represent, rigour about what deployment conditions actually require, and the discipline to let those answers drive method selection rather than deferring to prestige, theoretical elegance, or benchmark rankings. That discipline needs to run through every stage of system design — from problem framing through evaluation protocol to how results and their limitations are communicated to practitioners.That is what this thesis has attempted to practise and to advocate. The specific findings — estimator rankings, detection rates, model performance figures — will be partially superseded as the fields develop. The structural principle will not: robustness requires context-first design, and the idealisation–deployment gap is the primary obstacle to reliable AI-driven systems.

REFERENCES

1. Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2), 1148-1178.
2. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68.
3. Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448), 1053-1062.
4. Gentzel, A., Pruthi, D., & Jensen, D. (2021). The case for evaluating causal models using interventional measures and empirical data. *Advances in Neural Information Processing Systems*, 34, 8522-8534.
5. Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.
6. Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10), 4156-4165.
7. LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4), 604-620.
8. Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems*, 30, 6446-6456.
9. Neal, B., Huang, C., & Raghupathi, S. (2021). RealCause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*.
10. Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.
11. Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846-866.
12. Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.
13. Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688-701.
14. Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469), 322-331.
15. Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.Akamai. (2021). *State of the Internet Security Report: Credential Stuffing in the Media Industry*. Akamai Technologies.
16. Binsalleeh, H., Ormerod, T., Boukhtouta, A., Sinha, P., Youssef, A., Debbabi, M., & Wang, L. (2010). On the analysis of the Zeus botnet crimeware toolkit. *2010 Eighth International Conference on Privacy, Security and Trust*, 31-38. IEEE.
17. Chandramohan, M., & Tan, H. B. K. (2014). Detection of mobile malware in the wild. *Computer*, 47(4), 60-63.
18. CISA. (2024). *Phishing-Resistant Multi-Factor Authentication Guidance*. Cybersecurity and Infrastructure Security Agency.

19. Continella, A., Guagnelli, A., Zingaro, G., De Pasquale, G., Barenghi, A., Zanero, S., & Maggi, F. (2016). ShieldFS: A self-healing, ransomware-aware filesystem. *Proceedings of the 32nd Annual Conference on Computer Security Applications*, 336-347.
20. Critical Appraisal Skills Programme (CASP). (2018). *CASP Checklists*. CASP UK.
21. Diachenko, B., & Nazarovas, A. (2025). 16 billion credentials exposed in massive data breach. *Cybernews Research Report*.
22. Dutson, J., Allen, D., Eggett, D., & Seamons, K. (2019). Don't punish all of us: Measuring user attitudes about two-factor authentication. *2019 IEEE European Symposium on Security and Privacy Workshops*, 119-128. IEEE.
23. FIDO Alliance. (2021). *FIDO2: Web Authentication (WebAuthn) Specification*. FIDO Alliance Technical Standards.
24. ISO/IEC 27001. (2022). *Information Security Management Systems – Requirements*. International Organization for Standardization.
25. Kharraz, A., & Kirda, E. (2017). Redemption: Real-time protection against ransomware at end-hosts. *International Symposium on Research in Attacks, Intrusions, and Defenses*, 98-119. Springer.
26. Kolodenker, E., Koch, W., Stringhini, G., & Egele, M. (2017). PayBreak: Defense against cryptographic ransomware. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 599-611.
27. Koth, A., Ott, J., & Holz, T. (2020). Evaluating the usability of hardware security keys. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-13.
28. Lang, J., Czeskis, A., Balfanz, D., Schilder, M., & Srinivas, S. (2017). Security keys: Practical cryptographic second factors for the modern web. *Financial Cryptography and Data Security*, 422-440. Springer.
29. Lassak, K., Hildebrandt, M., Golla, M., & Dürmuth, M. (2024). It's like flossing your teeth: On the importance and challenges of reproducible evaluation of authentication systems. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1-16.
30. Mandiant. (2024). *M-Trends 2024: A View from the Front Lines*. Mandiant Consulting.
31. NIST. (2020). *Digital Identity Guidelines: Authentication and Lifecycle Management (SP 800-63B)*. National Institute of Standards and Technology.
32. OWASP. (2023). *OWASP Top Ten 2023*. Open Web Application Security Project.
33. PCI Security Standards Council. (2022). *Payment Card Industry Data Security Standard (PCI DSS) v4.0*.
34. Polino, M., Scorti, A., Maggi, F., & Zanero, S. (2018). Jackdaw: Towards automatic reverse engineering of large datasets of binaries. *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, 121-143. Springer.
35. Scaife, N., Carter, H., Traynor, P., & Butler, K. R. (2016). CryptoLock (and drop it): Stopping ransomware attacks on user data. *2016 IEEE 36th International Conference on Distributed Computing Systems*, 303-312. IEEE.
36. van Rijn, D. (2022). The evolution of infostealer malware: From keyloggers to credential harvesting platforms. *Akamai Security Research*.
37. Verizon. (2024). *2024 Data Breach Investigations Report*. Verizon Enterprise Solutions.
38. W3C. (2021). *Web Authentication: An API for accessing Public Key Credentials Level 2*. World Wide Web Consortium.
39. Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7, 597-610.
40. Brandwatch. (2023). *Social Media Monitoring and Analytics Platform*. Brandwatch Inc.

41. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440-8451.
42. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186.
43. Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2), e19.
44. Hootsuite. (2023). *Social Media Management Platform*. Hootsuite Inc.
45. Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11), e12106.
46. Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4948-4961.
47. Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Ali, J. T., Abbott, J., Mabuya, V., Buzaaba, C., Kabenamualu, A., Kalipe, S., Kelechi, O., Chukwuneke, C., Onyefuluchi, N., Emezue, C., Dossou, B. F. P., Sibanda, M., Mukiibi, J., ... & Ezeani, I. (2020). Masakhane – Machine translation for Africa. *arXiv preprint arXiv:2003.11529*.
48. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748-8763. PMLR.
49. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI Technical Report*.
50. Sitaram, S., Chandu, K. R., Rallabandi, S. K., & Black, A. W. (2019). A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.
51. Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Gohneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., & Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 62-72.
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
53. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 483-498.
54. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103-1114.
55. Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L. P. (2018). Memory fusion network for multi-view sequential learning. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 5634-5641.

56. Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.
57. Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2236-2246.
58. Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 3-62.
59. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
60. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
61. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
62. Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583-621.
63. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
64. Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83.