

# **Responsible AI in Practice**

**A Practical Guide to Safe  
and Human AI**

**Toju Duke  
Paolo Giudici**

**Apress®**

## ***Responsible AI in Practice: A Practical Guide to Safe and Human AI***

Toju Duke  
Bedrock AI, London, UK

Paolo Giudici  
University of Pavia, Pavia,  
Italy, Via San Felice 7, 27100

ISBN-13 (pbk): 979-8-8688-1165-4  
<https://doi.org/10.1007/979-8-8688-1166-1>

ISBN-13 (electronic): 979-8-8688-1166-1

Copyright © 2025 by Toju Duke and Paolo Giudici

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director, Apress Media LLC: Welmoed Spahr  
Acquisitions Editor: Shaul Elson  
Development Editor: Laura Berendson  
Coordinating Editor: Gryffin Winkler

Cover image by Katrin Bolovtsova from Pexels (<https://www.pexels.com/>)

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail [orders-ny@springer-sbm.com](mailto:orders-ny@springer-sbm.com), or visit [www.springeronline.com](http://www.springeronline.com). Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail [booktranslations@springernature.com](mailto:booktranslations@springernature.com); for reprint, paperback, or audio rights, please e-mail [bookpermissions@springernature.com](mailto:bookpermissions@springernature.com).

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at <http://www.apress.com/bulk-sales>.

Any source code or other supplementary material referenced by the author in this book can be found here: <https://www.apress.com/gp/services/source-code>.

If disposing of this product, please recycle the paper

# Table of Contents

**About the Authors.....vii**

**Acknowledgments .....ix**

**Introduction .....xi**

**Part I: Introduction ..... 1**

**Chapter 1: Responsible AI and AI Governance .....3**

    Responsible AI..... 3

    AI Governance ..... 7

        AI Regulation and Policies ..... 7

    The SAFE-HAI Framework..... 18

**Part II: Technical Risks (Internal to an Organization).....25**

**Chapter 2: Accuracy .....27**

    Risk Management Framework..... 29

    How Will You Assess Model Accuracy? .....30

        Accuracy of Predictions..... 30

    Accuracy of Classifications ..... 34

    RGA: A Unified Measure of Accuracy..... 39

    Accuracy of Multidimensional Predictions ..... 47

    Accuracy of Textual Predictions ..... 49

    What Benchmarks Will You Use?..... 50

    How Will You Communicate This As Needed? ..... 53

TABLE OF CONTENTS

Scoring Rubric .....54

Mitigation .....54

**Chapter 3: Robustness.....57**

Measuring Robustness .....59

Model Robustness.....60

Model Comparison .....68

Adversarial Robustness Benchmark .....76

Scoring Rubric .....77

Mitigation .....77

**Chapter 4: Explainability .....81**

Measuring Explainability .....85

Model Explainability .....88

Scoring Rubric .....91

Mitigation .....92

**Part III: Ethical Risks (External) .....95**

**Chapter 5: Fairness and Human Rights .....97**

Measuring Fairness for Organizations .....99

Measuring Fairness for Individuals.....103

Model Fairness.....106

Scoring Rubric .....109

Mitigation .....110

**Chapter 6: Privacy .....111**

Scoring Rubric .....118

Mitigation .....119

<b>Chapter 7: Sustainability .....</b>	<b>121</b>
Environmental Sustainability .....	123
Social and Governance Sustainability .....	129
Model Sustainability.....	134
Scoring Rubric .....	135
Mitigation .....	135
Social and Governance.....	136
Economic and Environmental .....	136
<b>Chapter 8: Human-Centered AI .....</b>	<b>139</b>
Evaluating AI .....	140
Assessing AI .....	142
Improving AI .....	144
Scoring Rubric .....	149
Mitigation .....	149
<b>Part IV: Governance and Case Study.....</b>	<b>151</b>
<b>Chapter 9: Governance Processes .....</b>	<b>153</b>
Risks of AI Models and Applications .....	154
Governance Processes.....	157
<b>Chapter 10: Case Study .....</b>	<b>163</b>
Logistic Regression Models .....	167
Application of the Logistic Regression Model .....	168
Verification of the Significance of the Logistic Regression Model .....	169
Tree Models.....	171
Neural Networks .....	175
Model Comparison .....	176
SAFE-HAI Assessment.....	183

TABLE OF CONTENTS

Accuracy ..... 183

Explainability..... 185

Robustness ..... 187

Fairness ..... 188

**Appendix..... 191**

**Index..... 203**