University of Glasgow

Wu, Yichen (2026) *What are the facial signals of apology and how do they vary across moral domains?* PhD.

https://theses.gla.ac.uk/86390/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses
https://theses.gla.ac.uk/
research-enlighten@glasgow.ac.uk

What Are the Facial Signals of Apology and How Do They Vary Across Moral Domains?

Yichen Wu

B.Sc. Materials Science and Engineering
M.Sc. Applied Psychology
M.Sc. Psychology (Research)

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Psychology and Neuroscience
College of Medical, Veterinary & Life Sciences
University of Glasgow



August 2026

Abstract

Apologies are central to interpersonal repair, yet research has focused mainly on verbal content and has rarely established how receivers interpret dynamic visual behaviour after different kinds of wrongdoing. This thesis develops a receiver-centred account of apology as a multimodal social signal and examines how facial movements, gaze direction, and blushing shape judgements of apparent apology intensity, whether these signals vary across moral domains, and whether their organisation differs between Western and East Asian receiver groups. Across the empirical studies, a prototype-free generative method sampled visual signals in computer-generated faces presented after transgression scenarios, while participant-level models recovered the configurations associated with receivers' judgements.

An initial study with Western receivers showed that apparent apology depended on intensity-sensitive combinations of smile-related and strain- or negative-affect-compatible facial movements and gaze: lateral gaze was associated with weaker apology judgements, downward gaze with stronger judgements, and blushing showed no recurring population-level pattern. Subsequent studies examined visual apology following five moral domains, i.e., Fairness, Care/harm, Authority, Loyalty, and Sanctity/purity transgressions in Western receivers judging Western faces (WC) and East Asian receivers judging East Asian faces (EA). In both groups, the model-derived Not-sorry, Mild-sorry, and Strong-sorry response categories provided the broadest organisation, whereas moral context selectively differentiated particular configurations. Loyalty showed the clearest corrected differentiation in WC, while Fairness and Authority were most clearly separated in EA, supporting a shared, intensity-sensitive architecture with selective moral-domain tuning rather than five isolated domain-specific prototypes.

Direct comparison showed that complete Mild- and Strong-sorry configurations reliably differentiated WC and EA within every moral domain. Response category and receiver group organised the visual-signal space more consistently than moral domain, which contributed local information without producing five global clusters. Comparisons with emotion and social-signal reference configurations further indicated a graded mixture of affiliative, reward, dominance, guilt-, and appeasement-related correspondence rather than a one-to-one emotional code.

Together, these findings support nested or graded pluralism: apparent apology intensity supplies a common social-repair structure, receiver groups differ in how that structure is visually realised, and moral context selectively tunes distributed configurations. The results concern receiver interpretation rather than offenders' felt emotion, sincerity, or success in obtaining forgiveness. Visual apology is therefore best understood not as one fixed facial expression, but as a graded, multimodal, and contextually calibrated social signal.

Acknowledgements

I would first like to thank Rachael Jack, who supervised both my master's and my PhD, for her guidance and support throughout these projects. Her advice, responsiveness, and practical support with the experiments and thesis have been invaluable. I am especially grateful for everything she has taught me about conducting research and for helping me develop greater confidence and professionalism as a researcher.

I would also like to thank Chen Zhou for taking on my supervision during the final stages of my PhD, when things would have been somewhat hectic, and for helping make this transition so smooth. I am grateful to both Rachael and Chen for their continued guidance, their support in preparing for my viva, and the time they devoted to arranging meetings and discussing my work. Their care and encouragement have meant a great deal during this period.

I am grateful to my viva convener, Gregory Weir, my internal examiner, Martin Lages, and my external examiner, Hongbo Yu, for their thoughtful questions and engaging discussions. I particularly appreciate their willingness to accommodate the time difference, with Hongbo joining early in the morning in the United States and Greg and Martin agreeing to a 5 p.m. start at the end of their working day. Their flexibility and kindness helped make the viva a smooth and rewarding experience.

My thanks also go to my lab mates, Jonas, Valentina, Max, Monica, and Yuening, for their support, thoughtful feedback, and willingness to review and discuss my work. Their perspectives and encouragement have been a valuable part of my PhD experience.

I would like to thank my closest friends, Adrian, Zhuo, and Chris, for their friendship, encouragement, and emotional support. Thank you for cheering me up during difficult moments and for making life outside research so enjoyable and fulfilling. Having you alongside me has meant a great deal throughout this journey.

Finally, I would like to thank my cat, Mop, and my partner, Anran. Without them, I simply would not have made it here. This achievement is, above all, theirs.

Contents

Abstract i

Acknowledgements ii

List of Tables viii

List of Figures x

1 General Introduction 1

1.1 Definition of Apology . . . 1

1.2 What Constitutes an Apology? Strategies From the Linguistics Literature . . . 2

1.3 How Has Apology Been Studied Empirically? . . . 4

1.3.1 The Apologising Interaction . . . 4

1.3.2 The Offence and Its Context . . . 8

1.3.3 Outcomes of Apology . . . 9

1.3.4 Offenders, Receivers, and Relationship . . . 12

1.3.5 Apology by Artificial Agents . . . 14

1.4 Emotions and Nonverbal Behaviour in Apology . . . 14

1.4.1 The Role of Emotion . . . 14

1.4.2 The Role of Nonverbal Behaviour . . . 16

1.5 A Social-Signal Approach to Apology . . . 17

1.5.1 Social Signal Theory . . . 17

1.5.2 Multimodal Signals as Social Signals . . . 18

1.5.3 Apology as a Social Signal . . . 18

1.5.4 Thesis Overview . . . 19

2 Facial Expressions of Apology in a Western Sample 21

2.1 Chapter Introduction . . . 21

2.2 Methods . . . 22

2.2.1 Participants . . . 22

2.2.2 Procedure . . . 23

2.2.3 Stimuli . . . 24

2.2.4 Planned Data Analysis . . . 26

2.3 Results . . . 27

2.3.1 Action Units (AUs) . . . 29
2.3.2 Primary Analysis: Gaze PMI . . . 32
2.4 Discussion . . . 32
2.5 Limitations . . . 35

3 Visual Apology Signals Across Moral Domains: Western Receivers 37
3.1 Chapter Introduction . . . 37
3.1.1 Moral-Domain Theory . . . 37
3.1.2 Moral Domains and Apology . . . 39
3.2 The Current Study . . . 40
3.3 Methods . . . 41
3.3.1 Participants . . . 41
3.3.2 Procedure . . . 41
3.3.3 Stimuli . . . 41
3.4 Planned Data Analyses . . . 42
3.4.1 Overview and Analysis Hierarchy . . . 42
3.4.2 Stage 1: Model Construction and Population-Supported Inputs . . . 43
3.4.3 Stage 2: Population-Level Signal Description . . . 43
3.4.4 Stage 3: Primary Classification and Ablation . . . 43
3.4.5 Stage 4: Primary Signal-Specific Associations . . . 44
3.4.6 Stage 5: Cross-Method Integration . . . 44
3.5 Results . . . 45
3.5.1 Stages 1–2: Validated Inputs and Population-Level Signal Description . . . 45
3.5.2 Stage 3: Primary Classification and Ablation . . . 51
3.5.3 Stage 4: Primary Signal-Specific Associations . . . 56
3.5.4 Stage 5: Cross-Method Synthesis . . . 63
3.6 Discussion . . . 64
3.6.1 Summary of Findings . . . 64
3.6.2 Shared Response-Level Apology Architecture . . . 64
3.6.3 Selective Moral-Domain Differentiation . . . 65
3.6.4 Complementary Roles of Action Units and Gaze . . . 65
3.6.5 Theoretical and Methodological Implications . . . 66
3.6.6 Limitations Addressed by Chapters 4 and 5 . . . 66
3.6.7 Chapter Conclusion . . . 66

4 Visual Apology Signals Across Moral Domains: East Asian Receivers **68**
4.1 Chapter Introduction . . . 68
4.2 The Current Study . . . 69
4.3 Methods . . . 69
4.3.1 Participants . . . 69
4.3.2 Procedure . . . 70
4.3.3 Stimuli . . . 70
4.4 Planned Data Analyses . . . 70
4.5 Results . . . 71
4.5.1 Stages 1–2: Validated Inputs and Population-Level Signal Description . . . 71
4.5.2 Stage 3: Primary Classification and Ablation . . . 78
4.5.3 Stage 4: Primary Signal-Specific Associations . . . 80
4.5.4 Stage 5: Cross-Method Synthesis . . . 86
4.6 Discussion . . . 88
4.6.1 Summary . . . 88
4.6.2 Shared Response-Level Apology Pattern . . . 88
4.6.3 Selective Moral-Domain Differentiation . . . 89
4.6.4 Complementary Roles of Action Units and Gaze . . . 90
4.6.5 Theoretical and Methodological Implications . . . 90
4.6.6 Receiver-Group Comparison Deferred to Chapter 5 . . . 90
4.6.7 Chapter Conclusion . . . 91

5 Comparing Visual Apology Signals Across Western and East Asian Receiver Groups **92**
5.1 Chapter Introduction . . . 92
5.2 The Current Study . . . 92
5.3 Methods . . . 92
5.3.1 Participants . . . 92
5.3.2 Procedure . . . 92
5.3.3 Stimuli . . . 93
5.4 Planned Data Analyses . . . 93
5.4.1 Overview . . . 93
5.4.2 Stage 1: Chapter 3 and Chapter 4 Response-First Comparison . . . 93
5.4.3 Stage 2: Primary Receiver-Group Classification, Prevalence and Ablation . . . 94
5.4.4 Stage 3: Primary MI/PMI Associations and Cross-Domain Variability . . . 94
5.4.5 Stage 4: Exploratory Analyses: Emotion and Social-Signal Reference Correspondence . . . 95

5.4.6 Stage 5: Exploratory Analyses: Hamming Geometry and Clustering 95
5.5 Results 96
5.5.1 Stage 1: Chapter 3 and Chapter 4 Within-Group Response-First Comparison . . 96
5.5.2 Stage 2: Primary Direct WC–EA Receiver-Group Classification Results 97
5.5.3 Stage 3: Primary Ablation and MI/PMI Findings 103
5.5.4 Stage 4: Exploratory Emotion and Social-Signal Reference Correspondence . . . 113
5.5.5 Stage 5: Exploratory Hamming Geometry and Clustering 119
5.6 Discussion 124
5.6.1 Moral-Domain Organisation Across the WC and EA Receiver Groups 124
5.6.2 Apology-Response Organisation Across the WC and EA Receiver Groups 125
5.6.3 Chapter-Level Empirical Synthesis 127

6 General Discussion **128**
6.1 Overview 128
6.2 Apology Signals and the Organisation of Moral Domains 128
6.2.1 The Moral-Domain Debate: Shared and Domain-Specific Response Patterns . . 129
6.2.2 Cross-Cultural Moral Domains: Shared Organisation and Different Signal Configurations 131
6.2.3 Moral Domains and Moral Emotions: Contextual Correspondence Without One-to-One Labels 133
6.2.4 Moral-Domain Synthesis 134
6.3 Contributions to the Apology Literature 135
6.3.1 Smiling Versus Sad-Looking Apologies 135
6.3.2 Cross-Cultural Apologies 137
6.3.3 Apologies and Emotions 138
6.3.4 Apology-Literature Synthesis 140
6.4 Apology as a Social Signal: Social Communication, Function, and Evolutionary Origins 141
6.4.1 Receiver-Interpreted Social Communication 141
6.4.2 Multimodal and Multifunctional Signalling 143
6.4.3 Shared Social Functions and Cultural Calibration 144
6.4.4 Evolutionary and Functional Foundations 146
6.5 Limitations and Future Directions 148
6.5.1 Moral Domain and Scenario Are Not Separately Identified 148
6.5.2 Response Categories and Receiver-Side Construct Scope 149
6.5.3 Cultural, Identity, and Demographic Scope 150
6.5.4 Signal-Space and Interactional Coverage 150

6.5.5 An Integrated Programme for Future Research 151
6.6 Conclusion 151

References **153**

Appendix A: Supplementary Details for Chapter 3 **166**
A.1 Validated Participants and Models 166
A.2 Observed Screening Outcomes 166
A.3 Supplementary PMI Facemap Localisation 167

Appendix B: Supplementary Details for Chapter 4 **168**
B.1 Validated Participants and Models 168
B.2 Observed Screening Outcomes 168
B.3 Supplementary PMI Facemap Localisation 169

Appendix C: Supplementary Details for Chapter 5 **170**
C.1 Direct-Analysis Inputs 170
C.2 Supplementary One-Sided PMI Facemap Localisation 170
C.3 Design-Matched C2–D2–D5 Analysis 172
C.4 Literature Reference Mapping 173

Appendix D: Illustrative Videos of Facial Dynamics **175**

Glossary **176**

List of Tables

2.1 *Participant Characteristics for the Balanced Chapter 2 Sample.* 23
2.2 *Availability of Combined and Intensity-Specific Participant-Level Signal Models in the Unbinned Five-Response Analysis.* 28
2.3 *Most Prevalent AU Signals Across the Three Retained Sorry-Intensity Models.* 30
2.4 *Legacy Factorial ANOVA of Gaze PMI Values.* 32

3.1 *Mapping Between the Five Moral Domains and the Adapted Transgression Scenarios.* . . 42
3.2 *Top-Quartile AUs Among WC Receivers Across Moral Domains After Bayesian Thresholding.* 50
3.3 *Top-Quartile Gaze Indicators Among WC Receivers Across Moral Domains After Bayesian Thresholding.* 50
3.4 *Moral-Domain Classifier Summary for WC Receiver-Group Apology Models.* 53
3.5 *Diagonal Class-Specific Ablation Roles for WC Receiver-Group Moral-Domain Models.* 53
3.6 *Directed Off-Diagonal Ablation Roles for WC Receiver-Group Moral-Domain Models.* . 56
3.7 *Response-Level PMI Category Composition for WC Receiver-Group Signals.* 58
3.8 *PMI-Derived Signal-Type Counts by Response and Moral Domain for WC Receiver-Group Models.* 60
3.9 *Significant Positive and Negative PMI Associations by Response and Moral Domain for WC Receiver-Group Models.* 61
3.10 *Signal-Level Overlap Between WC Receiver-Group Ablation and Signed MI/PMI Results.* 63

4.1 *Top-Quartile East Asian AUs Pooled Across Moral Domains After Population-Prevalence Thresholding.* 76
4.2 *Top-Quartile East Asian Gaze Indicators Pooled Across Moral Domains After Population-Prevalence Thresholding.* 78
4.3 *Diagonal Class-Specific Ablation Roles for East Asian Moral-Domain Models.* 80
4.4 *Directed Off-Diagonal Ablation Roles for East Asian Moral-Domain Models.* 80
4.5 *Response-Level PMI Category Composition for East Asian Signals.* 82
4.6 *Positive and Negative PMI Association Counts by Response and Moral Domain for East Asian Models.* 84
4.7 *Significant Positive and Negative PMI Associations by Response and Moral Domain for East Asian Models.* 85
4.8 *Signal-Level Overlap Between East Asian Ablation and Signed MI/PMI Results.* 86

5.1 *Compact WC–EA Comparisons of Moral-Domain Classification and Signal Repertoires.* 96
5.2 *Direct WC–EA Receiver-Group Classification Within Each Moral Domain.* 97
5.3 *Largest Absolute Population-Thresholded WC–EA Prevalence Contrast Within Each Response and Moral Domain.* 99

5.4 *Primary Thresholded Panel-Level Permutation Tests of WC–EA Signal-Prevalence Differences.* . . . 102
5.5 *Ablation-Active Signals in the Direct Receiver-Group Classifiers, Distinguishing Class-Specific Diagonal Roles From WC–EA Pairwise Off-Diagonal Roles.* . . . 103
5.6 *Ablation-Active Signals and Their Significant Direct Receiver-Group Classification Cells.* 104
5.7 *Response-Level Direct MI/PMI Associations and Independently Derived Thresholded Presence.* . . . 106
5.8 *Signed MI/PMI Associations and Independently Thresholded Signal Presence in All 15 Domain-by-Response Panels.* . . . 108
5.9 *Western Receiver-Group (WC) Signed MI/PMI Associations by Response and Moral Domain.* . . . 109
5.10 *East Asian (EA) Signed MI/PMI Associations by Response and Moral Domain.* . . . 110
5.11 *Receiver-Group-Specific Cross-Domain Variability Averaged Across 47 Signals.* . . . 112
5.12 *Literature Source and Categorisation Map for Valence Categories (L1) and Emotion or Social-Signal Types (L2)* . . . 114
5.13 *Selected Thresholded-Primary Correspondence Estimates and Planned Contrasts for the Overall Response Profiles (M1)* . . . 115
5.14 *Hamming Distance Between Matched WC and EA Binary Profiles Within Each Response and Moral Domain.* . . . 121
5.15 *Participant-Level Threshold-Aware Permutation Tests of Hamming Alignment.* . . . 122
5.16 *Domain-Specific Receiver-Group Differences in Within-Participant Mild-to-Strong Hamming Change.* . . . 123

A1 Validated Western Receiver-Group Participant and Model Counts by Moral Domain. . . . 166
B1 Validated East Asian Participant and Model Counts by Moral Domain. . . . 168
B2 East Asian Model-Validation Performance by Moral Domain. . . . 168
C1 *Realised Input to the Direct WC–EA Receiver-Group Analyses.* . . . 170
C2 Design-Matched Classifier Performance and Corrected C2–D2 Inference for Population-Thresholded Vectors. . . . 172
C3 *Evidence Scope, Categorisation, and Signal Mapping for the Chapter 5 Literature References* . . . 173
D1 Video Index by Receiver Group, Moral-Domain Context and Response Category. . . . 175

List of Figures

1.1 *Explicit and Implicit Components Communicated Within an Apology Interaction.* 3
1.2 *Four Conceptual Models That Locate Apology Within Wider Social Processes.* 5
1.3 *Five-Element Framework for an Apology Interaction.* 6

2.1 *Trial Sequence and Session Structure.* 23
2.2 *Signal Channels Sampled in the Generated Stimuli.* 24
2.3 *Dynamic Stimulus Generation and an Example Experimental Trial.* 25
2.4 *Participant-Level MI and PMI Modelling Workflow.* 27
2.5 *Facemaps of Retained Facial-Movement Prevalence by Transgression and Sorry Response.* 28
2.6 *Prevalence of the 15 Most Frequent AU Signals Across the Three Retained Sorry Intensities.* 29
2.7 *Normalised AU Weights in the Two-Factor NNMF Solution.* 31
2.8 *Facemap Rendering of the Two Retained NNMF Patterns and the Descriptive Transgression Accents.* 33
2.9 *Mean Gaze PMI by Apparent Apology Intensity and Transgression.* 34

3.1 *WC Receiver-Group Population-Level Signal Prevalence by Response Category and Moral Domain.* 46
3.2 *WC Receiver-Group AU Prevalence After Bayesian Thresholding, Showing the Retained AU Effects.* 47
3.3 *WC Receiver-Group Gaze Prevalence After Bayesian Thresholding, Showing the Retained Gaze-Direction Effects.* 52
3.4 *Row-Normalised Classification of WC Receiver-Group Mild- and Strong-Sorry Models Across the Five Moral Domains.* 54
3.5 *Facemap Localisation of Ablation-Active Signals in the WC Receiver-Group Moral-Domain Classifier.* 55
3.6 *Diagonal Class-Specific Ablation Roles in the WC Receiver-Group Moral-Domain Classifier.* 57
3.7 *Directed Off-Diagonal Pair-Separation and Pair-Confusion Ablation Roles in the WC Receiver-Group Moral-Domain Classifier.* 57
3.8 *Facemap Localisation of Significant WC Receiver-Group ON-State PMI Associations by Response Category and Moral Domain.* 59
3.9 *Significant Positive and Negative PMI Associations Between Individual Signals and Moral Domains for WC Receiver-Group Models in Each Response Bin.* 62

4.1 *East Asian Population-Level Signal Prevalence by Response Category and Moral Domain.* 72
4.2 *East Asian AU Prevalence After Population-Prevalence Thresholding.* 73
4.3 *East Asian Gaze Prevalence After Population-Prevalence Thresholding.* 77
4.4 *East Asian Moral-Domain Confusion Matrix for the Mild- and Strong-Sorry Models.* . . 79

4.5 *Facemap Localisation of Ablation-Active Signals in the East Asian Moral-Domain Classifier.* . . . 81
4.6 *Diagonal Class-Specific Ablation Roles in the East Asian Moral-Domain Classifier.* . . . 82
4.7 *Directed Off-Diagonal Pair-Separation Roles in the East Asian Moral-Domain Classifier.* 82
4.8 *Facemap Localisation of Significant East Asian ON-State PMI Associations by Response Category and Moral Domain.* . . . 83
4.9 *Significant Positive and Negative PMI Associations Between Individual Signals and Moral Domains for East Asian Models in Each Response Bin.* . . . 87

5.1 *Row-Normalised Direct Receiver-Group Classification Matrices for the Five Moral Domains.* . . . 98
5.2 *Pooled Population-Level Signal Prevalence by Response Category and Receiver Group.* . 100
5.3 *WC Minus EA Population-Thresholded Prevalence for All 47 Signals in Each Response-by-Domain Panel.* . . . 101
5.4 *Direct Receiver-Group Ablation Roles Separated by Classifier-Cell Type.* . . . 103
5.5 *Facemap Localisation of Ablation-Active Signals in the Direct Receiver-Group Classifiers.* 105
5.6 *Facemap Localisation of Reciprocal WC–EA Receiver-Group MI/PMI Differences.* . . . 107
5.7 *Signed Receiver-Group-Associated Signal Presence and Absence Across All Moral-Domain and Response Panels.* . . . 111
5.8 *Thresholded-Primary Correspondence of the Overall Response Profiles (M1) With 17 Literature-Derived Emotion and Social-Signal Types.* . . . 117
5.9 *Thresholded-Primary Correspondence of Response Profiles by Receiver Group (M2-1) With Emotion or Social-Signal Types (L2), After Equal Averaging Over Moral Domains.* 117
5.10 *Thresholded-Primary Correspondence of Response Profiles by Moral Domain (M2-2) With Emotion or Social-Signal Types (L2), After Equal Averaging Over Receiver Groups.* 118
5.11 *Thresholded-Primary Correspondence of Full Condition Profiles (M3) With Emotion or Social-Signal Types (L2).* . . . 120
5.12 *Hamming-Distance Heatmaps and Average-Linkage Dendrograms for (a) the 30 Full-Response Receiver-Group-by-Response-by-Domain Profiles and (b) the 20 Mild- and Strong-Sorry Profiles.* . . . 121

A1 *Facemap Localisation of Significant WC OFF-State PMI Associations by Response Category and Moral Domain.* . . . 167
B1 *Facemap Localisation of Significant East Asian OFF-State PMI Associations by Response Category and Moral Domain.* . . . 169
C1 *Facemap Localisation of One-Sided WC–EA Receiver-Group MI/PMI Associations.* . . . 171

Chapter 1

General Introduction

1.1 Definition of Apology

Apologies are an important means of repairing social bonds after transgressions (Goffman, 1971). Academic definitions of apology emerged from sociology, especially Goffman's accounts of face and remedial interchange (Bippus & Young, 2020; Goffman, 1955, 1971). For Goffman (1971), apology is remedial work that makes an offence interactionally manageable. It has a twofold logic: the offender accepts blame for the act (the "bad self") while dissociating that act from an enduring identity (the "good self") by affirming sympathy with the victim and continued commitment to the violated norm. Apology thereby seeks to restore ritual equilibrium after an unsettling event.

Politeness theory offers a closely related account of apology as face repair (Brown & Levinson, 1987). Brown and Levinson distinguish positive face—the desire for approval—from negative face—the desire to act without imposition. Communicative acts that conflict with either desire are face-threatening acts (FTAs), and politeness strategies soften or repair those threats. Although both Goffman and Brown and Levinson treat apology as cooperative face work, Goffman situates it within the broader social organisation of offence and repair, whereas Brown and Levinson approach it primarily as a linguistic negative-politeness strategy through which a speaker redresses a threat to the hearer's negative face.

Later accounts refined this definition by emphasising a prior violation of social norms (e.g., Blum-Kulka & Olshtain, 1984; Scher & Darley, 1997) and the different functions apology serves for offenders and recipients. From the offender's perspective, apology can support impression management and image restoration; from the recipient's perspective, it addresses threatened face wants (Meier, 1998). Blum-Kulka and Olshtain (1984) similarly describe apology as involving face loss for the speaker and support for the hearer. This two-party account is consistent with Holmes (1990, p. 159), who defined apology as a speech act addressed to the recipient's face needs that remedies an offence for which the apologiser takes responsibility, thereby restoring equilibrium between the apologiser and the person offended.

More recent work takes a broader interactional view, treating apology not only as a speech act but as a form of social repair whose function depends on context. Some researchers explicitly describe apology as a social signal. Ohtsubo and colleagues analyse apology through costly-signalling theory, proposing that material or social cost can make conciliatory intent more credible (Ohtsubo & Watanabe, 2009; Ohtsubo et al., 2018, 2020). Syme and Hagen (2019) similarly considered whether some suicidal behaviours can function as costly signals after severe transgressions, while B. Ho (2012) modelled the value of apology in relation to its informational and relational consequences. Other accounts place apology within a broader repertoire of social-recovery signals (Ohtsubo & Watanabe, 2009; Poggi & D'Errico, 2011). For example, Poggi and D'Errico (2011) notes that a display of shame can signal a request for forgiveness even without

an explicit verbal statement such as "I'm sorry."[1, 2]

Everyday usage also complicates the assumption that apology always follows a clear violation or performs substantial repair (Rieger, 2017). Routine cases documented by Holmes (1990) included slight inconveniences, spatial intrusions treated as virtual offences, slips of the tongue, and social gaffes such as coughing while speaking, showing that apology need not presuppose a serious or unambiguous offence (Holmes, 1990, pp. 178–182). In a study of 3,070 dialogues from more than 1,700 speakers in the British National Corpus, Deutschmann (2003) found that approximately half of the apologies were formulaic, involving minimal offence severity and little display of regret. Saying sorry after accidentally obstructing another person's path is one such routine use.

No single comprehensive definition of apology is shared across research fields, but the accounts converge on apology as a remedial act through which an offender addresses actual or possible disruption following a transgression or breach of social norms. Some theories restrict apology to a speech act, whereas others treat it as a broader social-repair signal that can take several forms. Wrongdoing often involves moral responsibility, but routine apologies also occur when responsibility and offence severity are minimal.

1.2 What Constitutes an Apology? Strategies From the Linguistics Literature

Goffman (1971) described the elements of a full apology as expressions of embarrassment or chagrin, clarification of the conduct that should have occurred, willingness to accept punishment, acknowledgement that the behaviour was wrong, commitment to future conformity, and willingness to make amends. Building on this interactional account, Brown and Levinson (1987) described several negative-politeness strategies through which a speaker can communicate regret or reluctance: admitting the imposition, indicating reluctance, providing compelling reasons, and asking for forgiveness.

Blum-Kulka and Olshtain (1984) provided cross-linguistic evidence about the composition of apology. Participants completed dialogues in which an apology was required, allowing Blum-Kulka and Olshtain to identify recurrent strategies. The most direct was an illocutionary force indicating device (IFID): a conventional expression such as "sorry," "apologise," or "excuse me." Four additional strategies were an explanation of the offence, an acknowledgement of responsibility, an offer of repair, and a promise of forbearance. Later work extended this repertoire to include directly seeking forgiveness (Schlenker & Darby, 1981), expressing concern for the offended person (Blum-Kulka et al., 1989, as cited in Rieger, 2017), redirecting attention from the offence (Blum-Kulka et al., 1989, as cited in Chang, 2008), displaying emotion, and checking whether harm occurred (Meier, 1992, as cited in Meier, 1998).

As shown in Figure 1.1, according to Harland et al. (2025), apology can contain nine explicit and three implicit components. The explicit components are a cue or statement of apology (e.g., "I'm sorry"; i.e., an IFID), responsibility, affirmation of the issue and its impact, explanation, moral admission, regret, reform, repair, and a petition for forgiveness. The implicit components are recipient-centred dialogue (e.g., prompts that facilitate engagement and attention to the recipient's responses), affective engagement (e.g., empathy, respect, and concern for the recipient's wellbeing), and behavioural evidence that commitments to reform

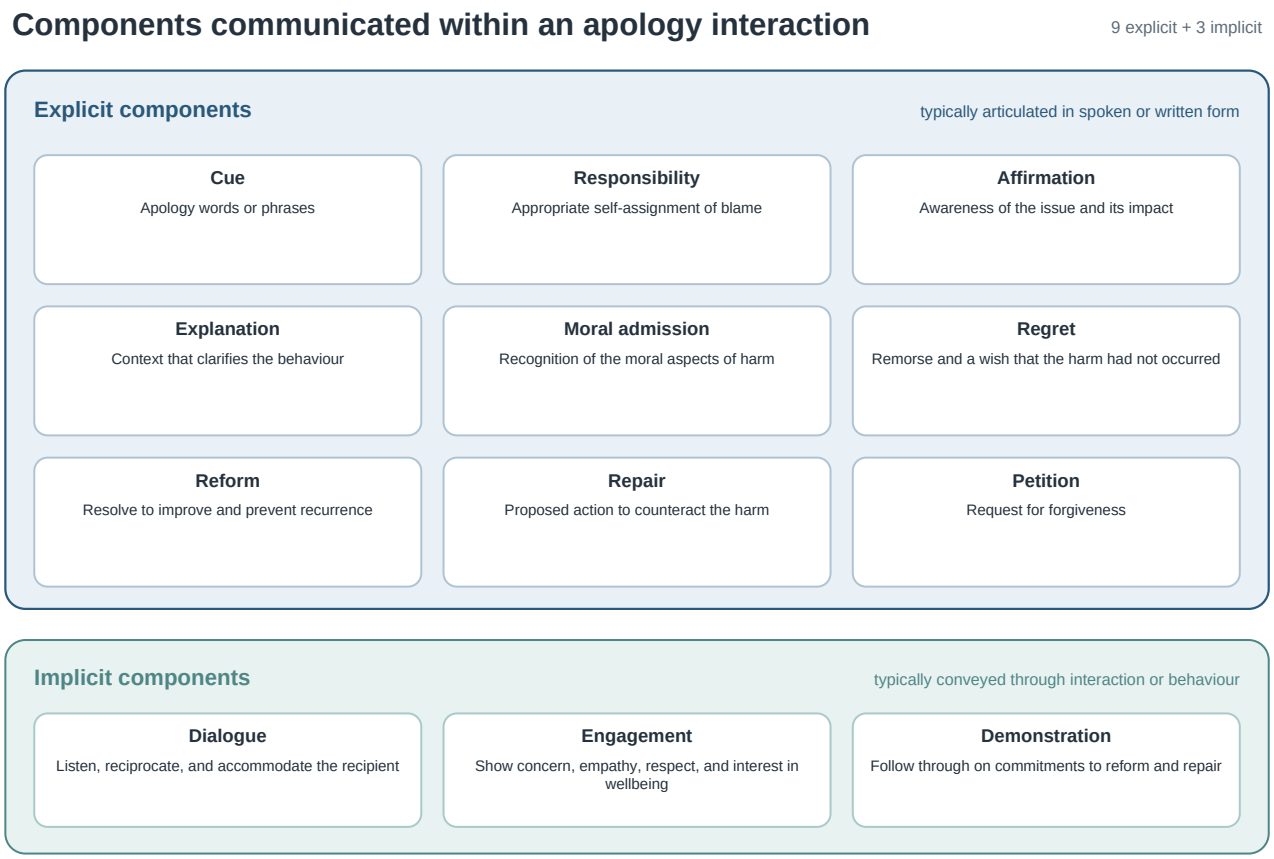
[1] The broad social-recovery framing used here is not identical to the face-redressive mechanisms described in the preceding sociological and politeness accounts. The former covers repair across a range of social interactions; the latter specify how an apology addresses actual or potential face loss and helps restore interactional equilibrium (Brown & Levinson, 1987; Goffman, 1971).

[2] *apology* and *sorry* are used interchangeably in this thesis, while we acknowledge that some researchers do not conceptually treat them as identical. Chapter 2 explains the choice of task wording, and Section 6.5.2 returns to our interpretation limits.

and repair have been enacted. Harland et al. treat explicit and implicit as the components' most likely forms rather than rigid categories. Not all strategies are necessary in every context (Robbennolt, 2003; Schlenker & Darby, 1981); indeed, a "full" apology can be less effective than a partial apology when disproportionate to a minor transgression (Newman & Kraynak, 2013), or be interpreted as overpolite or self-mocking (Aijmer, 2019). These components are therefore neither exclusive nor exhaustive, but an open-ended repertoire that can be elaborated, supplemented, or reconfigured according to the interactional context and the functions an apology must serve.

Figure 1.1

Explicit and Implicit Components Communicated Within an Apology Interaction.



Note. Original visual synthesis of the twelve components in Table 1 of Harland et al. (2025). The definitions are paraphrased, and explicit versus implicit denotes the components' typical mode of communication rather than a rigid classification.

Across different decompositions of apologies, emotional expression is a recurrent apology strategy whose role extends beyond the linguistic formula itself. Harland et al. (2025) includes regret for the harm among the explicit components of apology, while Hareli and Eisikovits (2006) emphasises the importance of communicating the transgressor's feelings about the wrongdoing. Expressions of remorse likewise contribute to the interpretation of the wrongful act (Ogiermann, 2009; O'Hara & Yarn, 2002; Slocum et al., 2011). The broader literature identifies several apology-relevant emotions, including embarrassment (Chang, 2008; Dijk et al., 2009; Schlenker & Darby, 1981; Thorstenson et al., 2020), regret (García, 1989; Harland et al., 2025; Holmes, 1990; Slocum et al., 2011), remorse (Blatz & Philpot, 2010; Kirchhoff et al.,

2012; Nunney & Manstead, 2021; Schlenker & Darby, 1981; Slocum et al., 2011), sorrow (Schlenker & Darby, 1981), guilt (Hareli & Eisikovits, 2006; Jeter & Brannon, 2018), shame (Dijk et al., 2009; Hareli & Eisikovits, 2006; Kirchhoff et al., 2012), and pity (Hareli & Eisikovits, 2006).

Contemporary research increasingly treats apology as a speech event or form of social action embedded in an interaction, rather than as an isolated utterance, i.e., a speech act (Chang & Haugh, 2011; Leech, 2014). Several frameworks capture different aspects of this complexity. Figure 1.2 compares the four earlier frameworks, showing how they move from the internal structure of a prototypical apology to responsibility exchange, mediated outcomes, reconciliation, and trust rebuilding. Deutschmann (2003) described the prototypical apology through four components: offender, offended person, offence, and remedy. Blatz and Philpot (2010) modelled intergroup apology as influencing forgiveness, attitudes, and trust through mediators including remorse, sincerity, empathy, and responsibility attribution, with offence, apology, intergroup, and intragroup characteristics acting as moderators. Chaudhry and Loewenstein (2019) located apology within responsibility exchange: after a negative outcome, apologising transfers blame from the receiver to the originator, whereas blaming transfers it in the opposite direction. Tomlinson et al. (2004) linked apology, timeliness, sincerity, and relationship characteristics to willingness to reconcile and the rebuilding of trust. Later on, as shown in Figure 1.3, Harland et al. (2025) offers a broader five-element framework comprising the offender, recipient, offence and its context, interaction, and outcomes. This framework treats apology as an interaction between parties concerning an offence and directed towards an outcome, rather than as a single self-contained act.

No single set of components exhaustively defines apology; what an apology should contain depends on the interactional context. Across disciplines, however, researchers repeatedly identify several broad types of content that apologies may include: acknowledging the offence and its consequences, accepting responsibility, expressing emotion, offering an explanation, repairing the harm, committing to change, and responding to the recipient's needs or perspective. Contemporary research has consequently moved beyond cataloguing isolated components towards understanding apology as a social-repair signal whose interpretation and effectiveness depend on context, relational dynamics, and the conduct of the parties involved.

1.3 How Has Apology Been Studied Empirically?

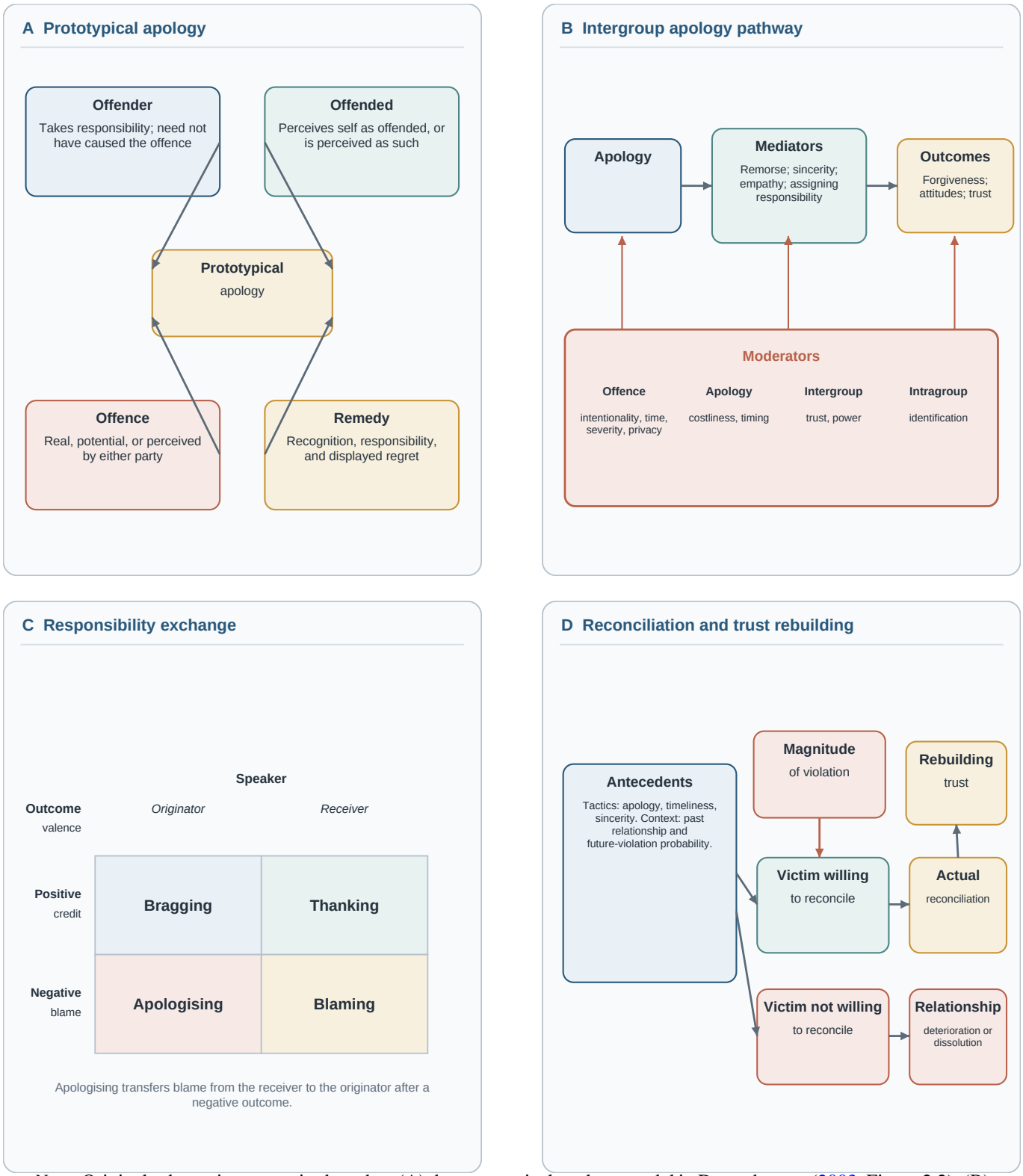
Apology research originated in sociological and politeness accounts (Brown & Levinson, 1987; Goffman, 1971), but increasingly treats apology as a communicative act of social repair. The review therefore follows the five-part interactional framework of Harland et al. (2025): the offence and its context, the apologising interaction, the offender, the recipient, and the outcome. This structure accommodates verbal and embodied signals while preserving the distinction between signaler and receiver, including receiver perspectives beyond the direct victim (Russell et al., 2025). That distinction is empirically necessary: perpetrators and victims differ in when they expect apology and in the emotions underlying those expectations (Leunissen et al., 2013), while an apology can separately convey the offender's emotion and address the victim's emotion (Kirchhoff et al., 2012). Figure 1.3 shows the structure used to organise the remainder of the empirical review of this section.

1.3.1 The Apologising Interaction

A substantial body of work has tested the effects of particular apology components and expanded strategic taxonomies derived from different theories.

Figure 1.2

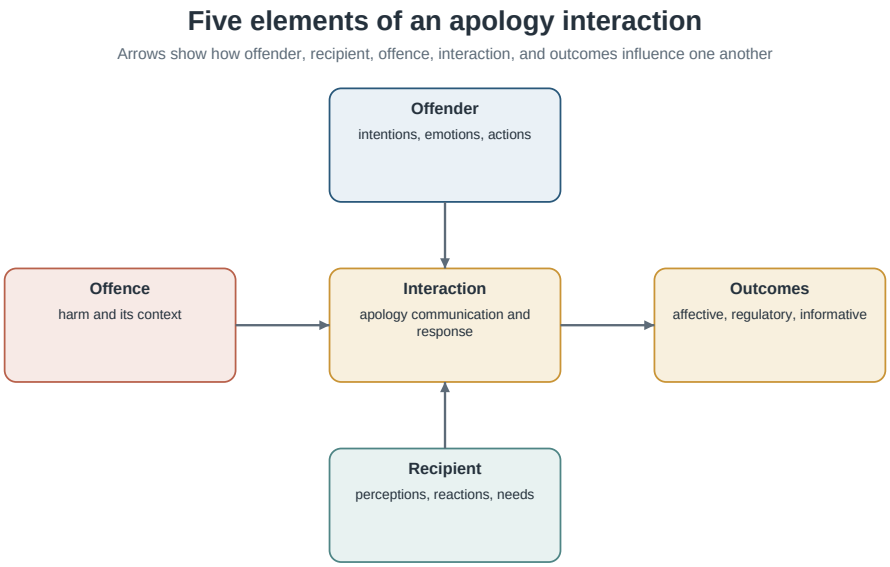
Four Conceptual Models That Locate Apology Within Wider Social Processes.



Note. Original schematic summaries based on (A) the prototypical apology model in Deutschmann (2003, Figure 2.2); (B) the intergroup apology framework in Blatz & Philpot (2010, Figure 1); (C) responsibility exchange theory in Chaudhry & Loewenstein (2019, Figure 1); and (D) the reconciliation and trust-rebuilding model in Tomlinson et al. (2004, Figure 1). The diagrams were redrawn with consistent terminology for comparison.

Figure 1.3

Five-Element Framework for an Apology Interaction.



Note. Adapted from Figure 1 in "AI apology: a critical review of apology in AI systems," by H. Harland et al. (2025), *Artificial Intelligence Review, 58*, Article 369, https://doi.org/10.1007/s10462-025-11305-8. CC BY 4.0. The diagram was redrawn and the annotations were shortened.

Early qualitative work established the strategic repertoire of apology. In the Cross-Cultural Study of Speech Act Realisation Patterns, participants from eight language groups completed discourse sequences eliciting requests or apologies, producing a cross-linguistic taxonomy of five apology strategies (Blum-Kulka & Olshtain, 1984). Conversational role-play similarly identified combinations of regret, accounts, self-effacement, and redress derived from interactional and politeness frameworks (Brown & Levinson, 1987; García, 1989; Goffman, 1971). A corpus of 183 New Zealand apologies then showed that explicit apology was nearly universal, whereas combining several strategies was not more common than using one (Holmes, 1990).

Quantitative studies then tested how established apology components relate to outcomes (Harland et al., 2025). More severe transgressions were generally judged to require more elaborate, multi-strategy apologies (e.g., Bippus & Young, 2020; Kirchhoff et al., 2012; Lewicki et al., 2016; Ohbuchi et al., 1989; Schlenker & Darby, 1981), while a simple apology statement remained the most common component (Park & Guan, 2009). The relative value of particular strategies nevertheless varied by study and context. After severe transgressions, four components—expressing emotion, admitting fault, using an IFID, and explaining—were especially effective (Kirchhoff et al., 2012). Among four restoration strategies assessed through questionnaires and structured interviews, perceived true sorriness predicted third-party forgiveness more strongly than excuse, an apology expression, or admission of guilt (Allan et al., 2006). Across minor and major transgressions, accepting responsibility, articulating an apology, and identifying the specific wrongdoing were rated more important than the other five of the eight strategies tested (Bippus & Young, 2020). Compensation was also important, whether as the most effective of seven tested strategies (Jeter & Brannon, 2018) or as a strategy valued more by residentially mobile than residentially stable partici-

pants (Komiya et al., 2020). Collectively, these findings show that apology effectiveness depends on the configuration and contextual fit of its components, not simply their number.

Research on emotional expression has expanded beyond regret to remorse, guilt, shame, embarrassment, sadness, and empathy. Negative emotion after a transgression can indicate that the outcome was distressing and unexpected for the offender, while embarrassment can reduce dispositional blame after a clumsy act (Scher & Darley, 1997). Apologies motivated by shame or guilt, but not pity, were more likely to elicit forgiveness (Hareli & Eisikovits, 2006); guilt displays supported judgements of future morality through perceived moral values and commitment to change (Bosche et al., 2026); and shame reduced the insult perceived after an out-group offence (Giner-Sorolla et al., 2008). Remorse also predicted forgiveness beyond group status, harm, and personal relevance (Gold & Weiner, 2000), while offender empathy was associated with greater recipient satisfaction (Roschk & Kaiser, 2013). Collectively, the effect of emotion depends on whether the display provides contextually appropriate evidence of concern and repair.

Emotional information in apology is also carried through nonverbal behaviour, especially the face (see Ekman, 1993; Nunney & Manstead, 2021; O'Hara & Yarn, 2002; ten Brinke & Adams, 2015; Yamamoto et al., 2021). Negative rather than positive emojis increased customer satisfaction in one service context (Ma & Wang, 2021); in public apologies, sadness-related behaviour was treated as norm-congruent, whereas smiling reduced perceived sincerity and investor confidence (ten Brinke & Adams, 2015). A data-driven study likewise found that mental representations of an apologetic face corresponded most strongly with sadness (George et al., 2023). Blushing extends this evidence beyond facial movement: it increased perceived embarrassment or shame and was associated with more favourable post-transgression evaluations, greater perceived apology sincerity, and a higher likelihood of forgiveness (Dijk et al., 2009; Thorstenson et al., 2020).

The meta-analysis by Russell et al. (2025) indicates that expressing emotion in apology can facilitate forgiveness when the display fits normative expectations for a moral offender; it does not imply that any negative emotion will be beneficial in every context.

Research beyond fixed component inventories shows that delivery and interactional placement shape apology meaning. In medical disclosure, physicians used smiling and pleasant vocal delivery, and recipients evaluated physicians and errors more favourably when nonverbal behaviour accompanied the disclosure (Hannawa, 2011; Hannawa et al., 2016); an affiliative smile may therefore function as appeasement rather than inappropriate positive affect (Hess et al., 2002). Timing is similarly conditional. Later apologies can be more effective when victims have time to speak and the conflict is resolved (Ebesu Hubbard et al., 2013; Frantz & Bennigson, 2005), whereas earlier apology produced greater satisfaction in a service-recovery vignette without a discussion period (Roschk & Kaiser, 2013). Sequential position and response pairing also matter (Robinson, 2004), as does assigning agency ("I'm sorry I offended you" rather than "I'm sorry you were offended") (Wang & McGlone, 2020). Even ironic or sarcastic apologies can maintain affiliation among teenagers when their interactional function is mutually understood (Aijmer, 2019). Collectively, these findings show that apology signals acquire meaning from their placement within an unfolding relationship.

Costly signalling theory offers one framework for explaining apology credibility (see Zahavi, 1975). It predicts that cost can strengthen an apology when it provides harder-to-fake evidence of conciliatory intent, not that greater cost is invariably more effective. Ohtsubo and Watanabe (2009) manipulated whether an

apology carried a material cost and found that costly apologisers were judged more sincere and elicited fewer complaints. Ohtsubo et al. (2018) found that costly apologies elicited stronger activation in theory-of-mind regions, consistent with greater inference about the apologiser's communicative intent. Syme and Hagen (2019) found that suicidal behaviours were associated with transgression, punishment, and shame, but did not show that they facilitated victims' forgiveness. Expectancy Violation Theory offers a different contextual account: a positive violation of expectations can improve evaluation (Burgoon, 2015). Consistent with this possibility, Polin et al. (2024) found that counter-stereotypical apologies—more communal for men and more agentic for women—were perceived as more effective.

Together, this literature moves from cataloguing apology strategies to explaining how their linguistic, emotional, nonverbal, temporal, and costly components function within interaction. The effects of these components are often context-dependent because receivers interpret the act as configured social evidence about the offender and the relationship; the next section therefore examines how the offence and its setting constrain that interpretation.

1.3.2 The Offence and Its Context

The form and effectiveness of apology vary with the transgression. Bippus and Young (2020) found that articulating an apology was especially important after inappropriate or inconsiderate conduct, whereas requesting forgiveness mattered more after a broken promise; these differences were more pronounced when consequences were severe. Together with other offence-comparison studies (e.g., A. D. Cohen et al., 1986; Syme & Hagen, 2019), this evidence shows why the features of the transgression must be considered alongside the apology itself.

The most extensively studied offence feature is severity. More severe transgressions generally require more apology components (e.g., Bippus & Young, 2020; Kirchhoff et al., 2012; Lewicki et al., 2016; Ohbuchi et al., 1989; Schlenker & Darby, 1981). A perfunctory apology such as "pardon me" may suffice for a minor consequence, whereas a major consequence may require a more elaborate apology that expresses remorse, offers help or repair, accepts responsibility, explains the violation, and requests forgiveness (Kirchhoff et al., 2012; Schlenker & Darby, 1981). Contextual fit nevertheless matters: an elaborate apology after a minor transgression may appear overpolite (Chang & Haugh, 2011) or ironic (Aijmer, 2019). Bennett and Earwaker (1994) found that more severe transgressions elicited greater experienced and expressed anger and a stronger demand for justification (see also Ohbuchi et al., 1989). Tomlinson et al. (2004) found that the perceived probability of future violation mediated the relationship between apology and willingness to reconcile, although this pathway weakened as severity increased. Severity effects do not generalise uniformly across offences. A. D. Cohen et al. (1986), for example, found that participants used some strategies, including offers of repair, less often as severity increased in one scenario than in another, while Wang and McGlone (2020) found no evidence that severity changed the positive evaluation of apologising offenders. Severity must therefore be interpreted together with offence type and the form of repair.

Intentionality and responsibility further constrain repair (Fehr et al., 2010). Intentional offences are less readily forgiven (Struthers et al., 2008); apology reduced punishment when intent was ambiguous but not when it was clearly intentional (Fischbacher & Utikal, 2013); and, in a social-rejection paradigm in which rejection could be perceived as intentional, apology increased expressed hurt and aggression (Freedman

et al., 2017). Elaborating an apology helped children chiefly after unintentional harm (Darby & Schlenker, 1982). In recalled transgressions, victims expected apology more after intentional harm, whereas offenders were more willing to apologise after unintentional harm, with anger and guilt mediating their respective perspectives (Leunissen et al., 2013). In imagined scenarios, greater offender responsibility predicted victims' anger, demands for justification, and rejection (Bennett & Earwaker, 1994), while taking the offender's perspective in high-responsibility, high-consequence offences prompted more self-castigation and explicit requests for forgiveness (Schlenker & Darby, 1981).

Ambiguity and violation type introduce further constraints. When it was unclear whether wrongdoing had occurred, an apology that explicitly admitted fault could create a less favourable impression than an apology without explanation or admission (Newman & Kraynak, 2013). Lewicki et al. (2016) also found that apologies were more effective after competence-based than integrity-based violations. At the other end of the continuum, apology can smooth interaction even when no clear transgression has occurred, including after slips of the tongue (Deutschmann, 2003), social gaffes (Holmes, 1990), or a service failure for which the speaker was not personally responsible (Yamamoto et al., 2021).

Although this thesis focuses on interpersonal apology, the wider literature spans intergroup, organisational, commercial, public, and legal settings. Marketing and business research examines outcomes such as customer satisfaction and market response (e.g., Kraig et al., 2022; Ma & Wang, 2021; Shao et al., 2022); public-apology research includes government misconduct, corporate hazards, and social-media communication (e.g., Hornsey et al., 2020; Ohtsubo et al., 2020; Sandlin & Gracyalny, 2018; ten Brinke & Adams, 2015; Wenzel et al., 2017); and legal research examines dispute resolution, sentencing, criminal justice, and litigation (J. R. Cohen, 1999; Levi, 1997; Petrucci, 2002; Wagatsuma & Rosett, 1986). Accepting responsibility can improve victims' evaluations and settlement prospects (Robbennolt, 2003), while apology has also been associated with victim empowerment and reduced recidivism (Petrucci, 2002). These settings share some repair processes with interpersonal apology but differ in offences, audiences, outcomes, and the structure of interaction.

The effectiveness and appropriate form of apology therefore depend jointly on offence type, severity, intentionality, responsibility, ambiguity, and institutional setting. Findings from interpersonal, intergroup, commercial, public, and legal contexts cannot be treated as interchangeable. Because this thesis focuses on interpersonal transgressions, the next section distinguishes the receiver-side outcomes through which repair is assessed.

1.3.3 Outcomes of Apology

Apology outcomes can be organised into three broad categories: actions, emotions, and evaluations.

Actions. Forgiveness is the most frequently examined action because it is the victim-side response most directly paired with the offender's attempt at repair (Beyens et al., 2015; Fehr et al., 2010). It is also unusually portable across children's judgements, workplace offences, medical errors, intergroup apologies, and AI-authored crisis communication (Darby & Schlenker, 1982; Grunenberg et al., 2026; Hannawa et al., 2016; Lim et al., 2025; Nunney & Manstead, 2021), while narrower endpoints, including patient satisfaction, willingness to continue a relationship, and confidence in a communicator, remain context-specific (Hannawa et al., 2016; Lim et al., 2025; Wenzel et al., 2017). I therefore treat forgiveness as a primary, but not automatic, indicator of effectiveness within the sequence offence → apology → mediators and

moderators → forgiveness.

In general, apology and interpersonal forgiveness have a moderate association between ($r = 0.42$), indicated by a meta-analysis (Fehr et al., 2010). At the same time, intent, empathy, and anger showed similar or stronger associations.

Apology can increase forgiveness directly. More elaborate apologies increased children's forgiveness (Darby & Schlenker, 1982), while a compensation-oriented apology increased both self-reported and behavioural forgiveness in a resource-allocation task (Jeter & Brannon, 2018). The direct effect is nevertheless limited: in a social-rejection experiment, apology increased pressure to express forgiveness without increasing felt forgiveness (Freedman et al., 2017).

Apology also influences forgiveness through several mediating pathways (Fehr et al., 2010). Perceived relationship value mediated apology effects in a double-randomisation experiment (Forster et al., 2021); anger reduction mediated the effect of apology completeness (Kirchhoff et al., 2012); greater empathy and less rumination mediated associations between perceived sincerity and forgiveness after medical error (Hannawa et al., 2016); and perceived possibility of reconciliation and hope mediated collective-apology effects (Wenzel et al., 2017). Struthers et al. (2014) further found that victim embarrassment mediated reported but not experienced forgiveness, showing that a mediator can explain a socially communicated response without accounting for the underlying motivational change.

Several factors moderate apology outcomes. Intentionality and severity change how apology relates to forgiveness and how much completeness matters (e.g., Kirchhoff et al., 2012; Struthers et al., 2008); defensive self-esteem alters the apology–forgiveness relationship (Eaton et al., 2007); high prior relationship value reduces the additional information supplied by an apology (Forster et al., 2021); perceived insincerity can make apology no more effective than no apology (Grunenberg et al., 2026); and residential mobility changes the relative effectiveness of verbal apology and compensation (Komiya et al., 2020). Moderation also occurs within indirect pathways: reconciliation desirability changes the relationship between perceived possibility and hope of reconciliation (Wenzel et al., 2017), while relational tone changes indirect authorship effects on forgiveness through emotion and sincerity (Lim et al., 2025). Higher offender responsibility and offence severity can also increase the likelihood that an apology is rejected (Bennett & Earwaker, 1994).

Another extensively studied action outcome is how victims manage their relationship with the offender. In business cooperation, apology sincerity and the prior relationship were strongly associated with willingness to reconcile (Tomlinson et al., 2004); after workplace offences, apology increased willingness to restore the relationship and reduced avoidance (Grunenberg et al., 2026); and in romantic relationships, readiness to repair correlated with perceived apology and communal qualities (Schumann et al., 2023). The preferred form of repair also depended on social ecology: residentially mobile participants valued compensation more than residentially stable participants (Komiya et al., 2020). Even among strangers assigned to a shared task, apology increased perceived relationship value (Forster et al., 2021). Collectively, apology can support reconciliation across relationships, but the effective form of repair depends on relational and social context.

Negative actions such as revenge, punishment, and aggression provide a third class of behavioural outcomes. Apology reduced willingness to seek revenge, although not to the level observed when no offence occurred (Grunenberg et al., 2026). It also reduced imagined victims' severe physical and verbal aggression through more favourable impressions, less negative emotion, and reduced desire for an apology (Ohbuchi et

al., 1989). More elaborate apologies elicited lower third-party punishment recommendations from children (Darby & Schlenker, 1982), while Wooten (2009) found that punishment effects were mediated by the perceived appropriateness of accepting the apology and expectations about future intent. These effects remain conditional: apology reduced punishment when intentionality was ambiguous but not when the offence was clearly intentional (Fischbacher & Utikal, 2013), and it could improve impressions without reducing punishment (Wang & McGlone, 2020).

Interactive games allow apology effects to be measured through consequential behaviour. In Beyens et al. (2015), participants could punish two opponents after previously receiving punishment from them; one opponent apologised before the roles reversed and the other did not. Participants behaved less aggressively towards the apologising opponent, and the apology also altered neural responses under low-punishment conditions. B. Ho (2012) used a repeated trust game in which agents could send a costly "I'm sorry" message after learning whether a joint project had succeeded. Apology frequency declined as its price increased and as the relationship approached its final round, but apologies were more likely after failure and among agents who returned a larger share. Receiving an apology also predicted greater investment in the next round after controlling for project outcome, prior investment, and average return. These studies show how apology can change subsequent decisions rather than only retrospective ratings.

Game-based apology experiments remain less common than vignette and recall studies, but they offer two important design advantages. First, they measure consequential behaviour—such as punishment, investment, or cooperation—rather than relying only on recalled experiences, hypothetical judgements, or self-reported intentions. Second, controlled manipulations of apology, counterpart behaviour, and interaction history permit stronger causal tests and, in repeated games, show whether repair changes subsequent decisions as an interaction unfolds (Beyens et al., 2015; Fischbacher & Utikal, 2013; B. Ho, 2012).

Emotions. Emotion enters apology in two places: the offender's felt or displayed emotion as interpreted by the victim, and the victim's own emotional response. Across studies, remorse and related offender signals can reduce victim anger or negative affect, which can in turn support forgiveness and other repair outcomes.

Much of this research focuses on emotions expressed by the offender, consistent with definitions that make remorse central to apology (Meier, 1992, as cited in Meier, 1998); see also Harland et al. (2025). The most relevant moral emotions are remorse, guilt, shame, and embarrassment. Remorse has been manipulated as an apology component (Wang & McGlone, 2020), and expressing shame or remorse was among the most important elements after severe transgressions (Kirchhoff et al., 2012). This sensitivity develops early: six-year-old children inferred more remorse from elaborate apologies (Darby & Schlenker, 1982), and children preferred remorseful offenders and assigned them less punishment (Darby & Schlenker, 1989). Remorse in confession also improved imagined victims' forgiveness and expectations of the offender (Gold & Weiner, 2000). Other evidence links facial redness to embarrassment and apology sincerity (Thorstenson et al., 2020), guilt to willingness to apologise (Leunissen et al., 2013), and apology to children aged 4–9 attributing sadness and guilt to a wrongdoer (Smith et al., 2010).

Receiver-side moral emotions have received less attention than offenders' emotions. After an offence, recipients may experience embarrassment or awkwardness because apology creates social pressure to express forgiveness (Freedman et al., 2017; Risen & Gilovich, 2007). Anger is nevertheless the most prominent receiver-side response. Apology can reduce anger, which in turn supports forgiveness (Kirchhoff et

al., 2012), and can reduce reactive aggression (Beyens et al., 2015). These effects are conditional: apology reduced anger more effectively after minor than severe offences (Bennett & Earwaker, 1994), while Kubo et al. (2012) found that apology suppressed revenge motivation without eliminating physiological or subjectively experienced anger.

Evaluations. The final broad component of apology outcomes is evaluation, directed either towards the apology or the offender. Researchers do not always separate these targets, but the distinction clarifies which part of the interaction is being judged. Some studies first ask how sorry or apologetic the offender appears (Darby & Schlenker, 1982; George et al., 2023). The most frequently examined evaluation is apology sincerity. Costly apologies were perceived as more sincere than no-cost apologies (Ohtsubo & Watanabe, 2009; Ohtsubo et al., 2018); sincerity also appears in evaluations of apology components and remedial accounts (Kirchhoff et al., 2012; Schmitt et al., 2004). In interactional settings, sincere apologies predicted greater communication satisfaction, empathy, and forgiveness and fewer hard negative emotions (Ebesu Hubbard et al., 2013; Hannawa et al., 2016). Emotional and embodied cues matter too: guilt and shame increased perceived sincerity (Hareli & Eisikovits, 2006), while facial redness increased perceived apology sincerity (Thorstenson et al., 2020). Related outcomes include satisfaction (Frantz & Bennigson, 2005; Giner-Sorolla et al., 2008; Schumann et al., 2023), feeling cared for or respected (Giner-Sorolla et al., 2008; Schumann et al., 2023), apology efficacy (Lewicki et al., 2016; Polin et al., 2024), politeness (Haugh & Chang, 2019), and evaluations of the offender's traits, morality, and overall impression (Darby & Schlenker, 1982; Dijk et al., 2009; Gold & Weiner, 2000; Nunney & Manstead, 2021; Ohbuchi et al., 1989; Risen & Gilovich, 2007; Schmitt et al., 2004; Struthers et al., 2008).

In conclusion, apology outcomes can be organised as actions, emotions, and evaluations. Forgiveness is the most common action outcome; remorse and anger anchor the offender- and victim-side emotional pathways; and sincerity and satisfaction are central evaluations of the apology and offender. Across all three, the decisive distinction is between the signaler's conduct and the perceiver's interpretation, which need not coincide. The next section accordingly examines how offender and victim characteristics shape those two sides of the process.

1.3.4 Offenders, Receivers, and Relationship

The Offender. Apologies are not produced automatically after harm, but depend on the offender's disposition, motive, emotional state, perceived cost, social position, and expectation that the relationship can still be repaired. Ohtsubo et al. (2015) developed and validated a Japanese Proclivity to Apologise Measure and found that this disposition predicted whether participants had apologised after a recent transgression. The results suggest that apologising has a trait-like component rather than being only a situational speech act. This trait account is supported by Howell et al. (2011), who developed the Proclivity to Apologise Measure and found it associated with agreeableness, self-esteem, compassion, and lower narcissism. Mu and Bobocel (2019) further specified the motivational structure behind apology: self-blame, relational value, personal expedience, and fear of sanctions, and importantly, apology motives also shaped transgressors' own perceptions of victim forgiveness and reconciliation. Offence perception and self-regulation also shape apology production: Schumann and Ross (2010) found that men apologised less often partly because they had a higher threshold for what counted as offensive behaviour; Guilfoyle et al. (2019) found that high self-control increased apologetic and reduced nonapologetic responses; Rizvi et al. (2022) found that trait

and experimentally induced mindfulness promoted apologetic behaviour; and Guilfoyle et al. (2022) found that high-power transgressors were less willing to apologise and more willing to offer nonapologies unless they adopted an other-focus. Receivers' judgements of offenders' social traits also matter: Schumann et al. (2023) found that high-baseline apologisers were judged higher in communion, which indirectly improved reactions to a specific apology, although low-quality frequent apologies could reduce agency judgements; Polin et al. (2024) found that counter-stereotypical apologies were evaluated as more effective, with communal apologies by men and agentic apologies by women receiving relative benefits; and De Cremer and Schouten (2008) showed that apologies for unfair treatment increased fairness perceptions only when the authority was respectful, making respectful conduct another factor in interpreting an apology as meaningful and sincere. B. Ho (2012) gives a signalling interpretation of why the traits of offenders matter: an apology updates the receiver's beliefs about the offender's disposition and is more meaningful when future interaction remains valuable.

The Receiver. Apology effectiveness depends on who receives the signal and what the receiver already believes about the self, the offender, and the relationship. Eaton et al. (2007) found that participants with defensive self-esteem were least forgiving and most avoidant or vengeful after receiving an apology, which means an apology can fail when it confirms ego threat more strongly than remorse. Ohtsubo et al. (2015) likewise treat forgiveness as partly dispositional, showing that trait forgivingness predicted forgiveness in a recalled workplace offence, while proclivity to apologise predicted apology-making after one's own transgression. Komiya et al. (2020) show that receivers' social environments also matter: apologies were expected to be especially effective in residentially stable contexts, while compensation was more effective for residentially mobile participants, suggesting that receivers evaluate repair signals partly through the social ecology of future interaction. Forster et al. (2021) identify a more direct mechanism: apologies promoted forgiveness by increasing perceptions of the offender's relationship value, but apology effects were weaker for already high-value transgressors because the apology added less new information.

Relationship Conditions. Apology works because both parties are embedded in a continuing social relation, and the value, stability, status, cultural norm, and group boundary of that relation alter what the apology means. Forster et al. (2021) provide the clearest mechanism, showing experimentally that apologies promote forgiveness by increasing perceived relationship value. B. Ho (2012) similarly predicts that apologies are more frequent in longer relationships, early in relationships, and between better-matched partners, because apology has signalling value when the receiver's updated beliefs affect future cooperation. Tomlinson et al. (2004) support this relationship-centred view in professional trust repair: victim willingness to reconcile after a broken promise depended not only on offender tactics such as apology, timeliness, and sincerity, but also on relationship characteristics such as past relationship and future violation probability. Guan et al. (2009) add that apology obligations and intentions vary by national culture and relationship type, with participants showing stronger obligation and intention to apologise to a stranger than to a friend. Gonzales et al. (1990) show the same general point in an account-making setting: sex, relative status, and consequence severity shaped the length and concessionary content of remedial accounts after a gaffe.

Taken together, offender and victim characteristics and the relationship between them shape how apology functions as social repair; its meaning cannot be understood outside this broader interactional context. The final subsection extends this framework to the emerging case of apology by artificial agents.

1.3.5 Apology by Artificial Agents

As artificial intelligence becomes embedded in workplaces, services, corporate communication, and human–robot interaction, apology is increasingly studied as a repair signal from an artificial social actor (Harland et al., 2025). Text-based research shows that both authorship and relational delivery matter: complete reliance on AI-mediated workplace apology reduced perceived authenticity and indirectly reduced willingness to forgive (Glikson & Asscher, 2023), while a warm relational tone reduced negative emotion and supported trust and forgiveness after AI-authored crisis apologies (Lim et al., 2025). Embodied studies add the same configurational lesson. Two-robot apologies were evaluated more positively than one-robot apologies, particularly when paired with reparative action (Okada et al., 2023); humorous apologies worked better for warm robots, whereas polite apologies worked better for competent robots, especially when failure severity was low (Cho et al., 2026). Taxonomies of virtual-assistant repair likewise place apology within a progression from no reparation to apology and compensation (Galdon & Wang, 2020).

Collectively, this literature extends the human-apology story: receivers evaluate authorship, tone, embodiment, and action together rather than treating the apology statement as sufficient. AI systems add distinct design questions—when a need to apologise should be detected, who bears responsibility, whether an algorithmic decision should be explained, and whether the system should adapt its later behaviour (Harland et al., 2023). By identifying facial configurations that receivers associate with apparent apology, this thesis could inform research on and the design of facial apology displays for embodied AI agents, including virtual characters and robots with expressive faces. How people interpret and respond to these displays could then be tested in human–agent interactions.

1.4 Emotions and Nonverbal Behaviour in Apology

The preceding sections reviewed verbal apology strategies and the conditions shaping their effects. This section broadens the discussion to multimodal communication, examining how words, emotion and nonverbal behaviour contribute to receivers' interpretations of an apology.

1.4.1 The Role of Emotion

Much of the apology research focuses on emotions expressed by the offender. Apology is not only a linguistic formula but a remedial act that asks the victim to reinterpret the offender after a transgression. A minimally successful apology, therefore needs to communicate two related things: that the offender accepts responsibility and that the offender's internal state is appropriate to the harm. Wang and McGlone (2020) make this link explicit by treating responsibility and remorse as core features of apology; responsibility acknowledges the broken norm, while remorse conveys the offender's emotional state. Similarly, Kirchhoff et al. (2012) treat the communication of emotion, such as shame and regret, as an apology component and show that apology completeness predicts forgiveness partly through anger reduction, especially for severe offences. Taken together, Wang and McGlone (2020) and Kirchhoff et al. (2012) show that emotion is not decorative in apologies; it is part of what makes the apology interpretable as sincere.

The emotions most relevant to apology are moral emotions: remorse, guilt, shame, and embarrassment. Moral emotions matter because they connect an offender's self-evaluation to social norms and the welfare of others. Haidt (2003) defines moral emotions as emotions linked to the interests or welfare of society and other people, while Tangney et al. (2007) argue that moral emotions connect moral standards to moral behaviour. This makes them especially suitable for apology research: a transgression creates doubt about

the offender's moral orientation, and an apology attempts to repair that doubt by showing that the offender has recognised the moral meaning of the act.

The developmental evidence shows that this interpretation begins early. Darby and Schlenker (1982) found that children inferred greater remorse from more elaborate apologies and usually recommended less punishment. Darby and Schlenker (1989) further showed that children responded more positively to offenders who appeared remorseful rather than happy and unremorseful after a transgression. Smith et al. (2010) similarly found that children used apology to revise their expectations about the wrongdoer's emotion, attributing negative feelings such as sadness, feeling sorry, and guilt to the offender. Taken together, Darby and Schlenker (1982, 1989) and Smith et al. (2010) support the idea that apology is understood not merely as a conventional phrase but also as evidence about the offender.

Remorse is the most direct moral emotion in this literature because it is usually tied to recognising wrongdoing, accepting responsibility, and wanting repair. Gold and Weiner (2000) found that manipulated remorse shaped perceptions of causal stability, moral character, punishment, and forgiveness in confession scenarios. Tangney et al. (2011) also describe remorse and regret as central to guilt, and link guilt to making amends. In apology terms, this matters because victims are not only asking whether the offender feels bad now. They are also asking whether the offender is the kind of person who will do it again and change for the better. The central place of offender-centred emotions in social repairing is the reason why they are important for the present thesis. The thesis is interested in how victims perceive the offender, so the relevant properties are the emotions the offenders' displays and the emotions victims infer from the offenders.

A short terminology clarification is useful here because remorse, guilt, shame, and regret are sometimes used unevenly across the literature. Tangney et al. (2011) explicitly use guilt as an umbrella term that subsumes remorse until practically important distinctions are identified. Bhushan et al. (2020) show why this needs care: they note that remorse has been described as an offence-specific version of guilt, but their empirical mapping found shame and remorse overlapping and distinct from guilt. Zeelenberg and Breugelmans (2008) further distinguish regret from guilt by showing that guilt is mainly tied to interpersonal harm, whereas regret can refer to both interpersonal and intrapersonal negative outcomes. In this section, remorse is therefore used as an apology-relevant signal of repentance and responsibility; guilt as a behaviour-focused emotion that can motivate reparation; shame as a self-, status-, and face-focused emotion; and regret as a broader negative evaluation that may or may not imply interpersonal responsibility.

Guilt and shame therefore extend this point in different ways. Leunissen et al. (2013) showed an apology mismatch: victims' need for apology was mediated by anger, while perpetrators' willingness to apologise was mediated by guilt. Howell et al. (2012) similarly found that guilt-proneness was consistently related to apology willingness and apology elements, whereas shame measures were unrelated or inversely related depending on the measure. Guilt is therefore especially relevant on the offender side because it can motivate apologising, whereas victims' anger can motivate wanting an apology. Shame should be handled more cautiously: Tangney et al. (2011) distinguish shame from guilt through the self-versus-behaviour focus, with shame linked to hiding or escaping and guilt linked to making amends. D. Y.-F. Ho et al. (2004) add that shame, embarrassment, and losing face are tied to public judgement, visual exposure, and social standing in Chinese and Japanese cultural evidence. At the same time, shame displays are not simply harmful. Giner-Sorolla et al. (2008) found that, in outgroup reparations, expressions of shame reduced recipients' insult more than guilt or no emotion. The cautious thesis claim is therefore that guilt and shame

are not interchangeable "negative emotions": guilt more directly supports reparative motivation, while shame cues may help when victims read them as self-abasement, responsibility, or appeasement rather than as self-focused distress.

Embarrassment is particularly important for a multimodal apology section because it is often visible. Keltner (1995) describes embarrassment as having a distinct nonverbal display and an appeasement function. Thorstenson et al. (2020) provide direct evidence that facial redness increases perceived embarrassment and related social evaluations such as apology sincerity and forgiveness likelihood. This makes embarrassment a key bridge between moral emotion and nonverbal behaviour: it is not only an emotion that can be verbally claimed, but a state victims may infer from visible facial cues.

Methodologically, apology research has often represented emotion as a sentence-level component by manipulating whether the offender said sorry, accepted responsibility, offered repair, promised forbearance, or expressed regret (Kirchhoff et al., 2012; Wang & McGlone, 2020). Eye-tracking, nonverbal apology behaviour, embodied remorse, and facial-representation studies instead examine emotion as social evidence distributed across face, body, gaze, timing, and voice (Bonensteffen et al., 2020; George et al., 2023; Hornsey et al., 2020; Yamamoto et al., 2021); George et al. (2023), for example, identified sadness as a key component of the mental representation of an apologetic face. Collectively, this work shifts the scientific question from apology as text to apology as multimodal interaction.

1.4.2 The Role of Nonverbal Behaviour

Nonverbal behaviour matters because receivers use it to infer an offender's otherwise inaccessible internal state. Facial expression is a central emotion channel and can be coded systematically through the Facial Action Coding System, but display rules, individual differences, and context prevent it from serving as a transparent readout (Ekman, 1993). Facial cues may instead support, weaken, or contradict what an offender says (Bonensteffen et al., 2020; George et al., 2023). Incongruent expression can damage an apology (George et al., 2023); medical-disclosure meaning is distributed across verbal and nonverbal channels (Hannawa, 2011); and eye-tracking shows that victims attend to offenders' nonverbal behaviour when judging sincerity (Bonensteffen et al., 2020).

Facial expressions contribute to how receivers interpret apologies, but their meaning depends on context. George et al. (2023) found that apologetic faces, without any context provided, were strongly associated with sadness, regret, remorse, and submissiveness. Yamamoto et al. (2021) made the same link more directly in apology behaviour by operationalizing apology facial expression as sadness displays, measuring AU1, AU2, AU4, and AU15 together with gaze direction during role-played apologies. In that role-play study, Yamamoto et al. (2021) found longer upper-face display duration in instrumental than sincere apologies and interpreted direct gaze plus facial displays as a way of conveying "the feeling of apology". ten Brinke and Adams (2015) similarly found, across archival and experimental public-apology evidence, that sadness or normative affect supported sincerity and confidence, whereas deviant smiling was read as insincerity. Hornsey et al. (2020) add bodily evidence: embodied remorse increased perceived remorse, positive appraisals, response satisfaction, empathy, and perceived likelihood of reoffending, although direct forgiveness effects were small or negligible. Embarrassment and blushing evidence extends this negative-emotion pattern beyond a sad brow or downturned mouth. Dijk et al. (2009) manipulated blushing on neutral, shameful, and embarrassed expressions in transgression and mishap vignettes, finding that blushing actors were

evaluated more favourably. Thorstenson et al. (2020) then manipulated facial redness experimentally and found that redder faces increased perceived embarrassment, apology sincerity, and forgiveness likelihood, making facial redness relevant to apology-related social evaluation. Taken together, the empirical work shows that apology-relevant negative emotion is not only a verbal claim. It is also tested through visible sadness, remorse, gaze direction, blushing, and bodily submission.

The smile literature creates an apparent contradiction, because smiles can either violate or support apology norms. On the risk side, ten Brinke and Adams (2015) show that smiling as deviant happiness in public apologies can signal insincerity, and digital evidence shows similar boundary problems: Xie et al. (2025) found that negative emojis in chatbot service recovery can increase customer forgiveness whereas positive emojis can be ambiguous; Yang and Qian (2024) found that the smiling-face emoji can reduce perceived politeness in Chinese WeChat contexts. On the support side, smile research shows that a smile is not a single signal. J. Martin et al. (2017) distinguish reward, affiliation, and dominance smiles; Ambadar et al. (2009) show that observers distinguish amused, polite, and embarrassed/nervous smiles; and Ekman (1993) notes that several positive states can share a smiling signal and need contextual knowledge to distinguish. Thus, a smile in apology is not automatically inappropriate; its meaning depends on morphology, timing, channel, relationship, and the social function it performs. This helps explain why smile-like cues appear positively in apology research. Park and Guan (2009) included smiling among nonverbal apologising strategies and found in their Study that Chinese participants reported using smiling as an apology strategy more often than U.S. American participants. Mapson (2014) gives a sign-language version of the same issue: in British Sign Language apologies, non-manual features were central, and a polite grimace mouth gesture was used especially in personally oriented apologies such as losing a pen or damaging a car. Dijk et al. (2009) also matter here because their embarrassed expression was a non-Duchenne smile, while added blushing made the display more remedial in transgression and mishap contexts. Jucker (2023) similarly treats the apologetic smile as a recurrent apologetic gesture and reports that most apology-related gesture descriptions use smile-related wording. The contradiction is therefore not "sadness versus smile" as simple opposites. It is a distinction between a smile read as pleasure, amusement, or dominance, which can undermine remorse, and a smile-like cue read as polite, affiliative, embarrassed, or appeasing, which can help manage the interaction.

Taken together, this evidence supports treating apology as a multimodal, receiver-interpreted social signal (e.g., Bonensteffen et al., 2020; Ekman, 1993; George et al., 2023; Jucker, 2023; Mapson, 2014; Park & Guan, 2009; ten Brinke & Adams, 2015; Wang & McGlone, 2020; Yamamoto et al., 2021). Receivers do not directly decode an offender's inner state; they interpret how verbal content and embodied behaviour are configured as social evidence. The present thesis therefore asks how visual signals—facial movements, gaze direction, and blushing—shape judgements of apparent apology intensity, rather than whether an apology contains the correct words.

1.5 A Social-Signal Approach to Apology

1.5.1 Social Signal Theory

This thesis treats apology as a receiver-interpreted social signal: a perceivable cue through which a victim infers how an offender understands the transgression, the relationship, and possible future conduct. Poggi and D'Errico (2011) define social signals as physical stimuli, produced by a sender or perceived by a receiver, that provide information about social interactions, relationships, attitudes, and emotions. This

definition is useful for apology research because it does not restrict meaning to explicit words. A signal may be intentionally communicative, but it may also be informative when the receiver draws social meaning from different sources of behaviours or contexts.

This distinction matters after a transgression. An offender may say "I am sorry", but the victim must still decide what that utterance means. The same words can signal remorse, politeness, social obligation, self-protection, or strategic damage control. A social-signal account therefore shifts attention from apology production alone to apology interpretation: what information does the victim perceive, and why does that information seem credible or not?

Credibility is central because many apology signals are easy to produce. Costly-signalling theory gives one explanation for why some signals are trusted more than others. Zahavi's (1975) handicap principle argues that costly displays can become credible because lower-quality or insincere signalers cannot afford to produce them in the same way. McAndrew (2019) summarises the same logic for human behaviour: signals are persuasive when they are observable, costly, and difficult to fake. In apology contexts, the question is not only whether the offender apologises, but whether the apology, considered alongside other information, provides credible evidence of conciliatory intent.

1.5.2 Multimodal Signals as Social Signals

Social signals are often multimodal. Social Signal Processing research treats nonverbal behaviour as observable evidence through which people understand social interaction, including facial expression, gesture, posture, and relational attitude (Vinciarelli et al., 2009). Facial-expression research makes the same point at the level of the face: the face is a visible source of information about social intentions, motivations, identity, emotion, status, pain, and deception (Jack & Schyns, 2015; Schmidt & Cohn, 2001). Social-functional emotion theory also supports this view as emotions help coordinate interactions and relationships rather than merely expressing private internal states (Keltner & Haidt, 1999).

Multimodal cues are especially important when sincerity is uncertain. Poggi and D'Errico (2010) note that distinguishing controllable from uncontrollable social signals can help separate deceptive from sincere signals. Blushing offers a clear example: Castelfranchi and Poggi (1990) describe it as a non-intentional and counter-voluntary signal that can communicate sensitivity to others' judgements and shared values. Embodied remorse offers another example. Hornsey et al. (2020) found that physical displays such as kneeling or crying could increase perceived remorse in public apology contexts, even though these displays did not necessarily increase forgiveness. Smile perception gives an indirect and more cautionary example: Ambadar et al. (2009) show that the perceived meaning and genuineness of smiles depend on morphology and timing, while also showing that perceivers are imperfect decoders of spontaneous expression.

These examples justify a multimodal approach without assuming that nonverbal cues are transparent proof of inner remorse. Instead, nonverbal behaviour supplies additional social evidence. A victim may treat tears, gaze, posture, facial tension, blushing, or smile timing as cues that strengthen, weaken, or contradict the verbal apology. The theoretical point is therefore not that the body is always honest, but that receivers often use embodied behaviour when judging whether an apology is sincere.

1.5.3 Apology as a Social Signal

Apology is a social signal because it appears after a norm violation and asks the victim to update their understanding of the offender. B. Ho (2012) modelled apology as a signal in a trust-game setting: after a

negative outcome, an apology could affect whether the receiver treated the offender as trustworthy enough for future interaction. This fits the broader social-signal framework because apology communicates about the state of the relationship, the offender's attitude toward the transgression, and the likely direction of future interaction.

The main difficulty is that an apology can be cheap. Chaudhry and Loewenstein (2019) describe apologies as apparently materially costless communications, creating a "cheap talk" puzzle, while arguing that such acts can still carry social image costs. Ohtsubo and Watanabe (2009) make the credibility issue more direct: when exploitation was possible, merely saying "I am sorry" was not enough unless the apology was backed by a signal that made sincerity harder to fake. Their experiments showed that costly apologies were judged more sincere than no-cost apologies.

Later costly-apology work extends this mechanism. Ohtsubo et al. (2018) found that costly apologies, compared with non-costly apologies and no apology, more strongly activated theory-of-mind regions, suggesting that receivers process cost as evidence of conciliatory intention. Ohtsubo et al. (2020) found a similar logic for group apologies: costly group apologies increased perceived sincerity and intention-processing activity. Together, these studies support the conclusion that the victim is not simply receiving an apology but evaluating whether the signal is credible.

Cost is therefore one route to credibility, not the only one. Across this literature, receivers treat apology content, relational context, cultural expectations, and nonverbal presentation as converging or conflicting evidence about remorse, responsibility, relationship value, and future risk. This receiver-side integration explains why conventional words can fail and why multimodal cues matter when repair appears strategic or insincere.

Yet existing research does not establish how receivers combine dynamic visual signals to infer apology following explicit transgressions, or whether the resulting receiver-interpreted social signal is organised mainly by apparent apology intensity or is also sensitive to moral and receiver-group context. The empirical task isolates a narrower part of that problem: without a spoken or written apology, how sorry does a generated face appear after a transgression, and which visual signals support that receiver judgement? This thesis addresses that gap with a prototype-free, participant-level generative approach that tests configurations of visual signals across apparent apology intensity and moral domain in two receiver groups whose stimuli were ethnically matched: Western receivers judging Western faces (WC) and East Asian receivers judging East Asian faces (EA).

1.5.4 Thesis Overview

Chapter 2 establishes the receiver-side repertoire of visual signals and foundational method; Chapters 3 and 4 test how that repertoire is shared and selectively differentiated across moral domains within the WC and EA receiver groups; and Chapter 5 compares the two groups directly. Together, these studies develop a receiver-centred account of visual apology as graded, multimodal, and context-sensitive rather than one fixed set of facial expressions or five isolated moral-domain prototypes.

Together, this chapter establishes the thesis's scientific premise: apology is a receiver-interpreted, multimodal social signal whose meaning depends on how cues are configured, because nonverbal social information can be interpreted without speech (Vinciarelli et al., 2009), emotional responses can be rapid and relatively automatic (Keltner & Haidt, 1999), and blushing can provide counter-voluntary social evidence

(Castelfranchi & Poggi, 1990). The unresolved question is how receivers combine facial movements, gaze direction, and blushing when judging apparent apology intensity after a transgression. Chapter 2 therefore provides the necessary first empirical test by identifying the dynamic visual-signal configurations used by Western receivers before later chapters examine moral-domain sensitivity and compare the WC and EA receiver groups.

Chapter 2

Facial Expressions of Apology in a Western Sample

2.1 Chapter Introduction

Apologies are usually studied as verbal acts, yet receivers encounter them as multimodal social signals. The offender's words occur together with facial movement, gaze, timing, vocal delivery, and bodily conduct. These channels matter because an apology is not evaluated only for whether the conventional words were spoken. Receivers also judge whether the offender appears to recognise the transgression, accept its interpersonal implications, and display an appropriate orientation towards repair. In this chapter, I examine which dynamic facial signals make a person appear more or less sorry.

Existing work provides competing expectations about the appearance of apology. One possibility is that an apologetic face should resemble negative-affect displays associated with sadness, shame, guilt, remorse, or distress. Another is that smile-related movement can contribute to affiliation, reassurance, or politeness and can therefore also support apology judgements. These possibilities need not be mutually exclusive. Facial cues do not form a context-free dictionary: similar movements can acquire different meanings depending on their configuration, temporal dynamics, and the social event in which they occur (Ekman, 1993). The theoretically useful unit is therefore not a single facial movement treated as a complete apology signal, but a configuration of movements interpreted in relation to an offence.

George et al. (2023) used pixel-based reverse correlation to estimate mental representations of apologetic faces and found that sadness was prominent in the resulting templates. This work provided direct evidence that receivers have visual expectations about apologetic appearance. Its static classification images could not, however, separate facial movement from identity morphology, represent the temporal organisation of facial action, or test additional channels such as gaze and blushing. The task was also comparatively context-neutral. Without an explicit transgression, receivers may rely more heavily on familiar emotion categories when deciding what an apologetic face should look like.

In the present experiment, the investigation of multimodal apology signals follows the same analytical logic as research on apology more broadly. Earlier work first decomposed apologies into their constituent components and then examined how individual components, and combinations of components, shaped the interpretation and outcomes of the apology as a whole. Applying this component-based approach to multimodal apology, the first step is to identify the verbal and nonverbal signals from which it is constructed. The next step could be to investigate how these signals, both individually and in combination, influence the interpretation, evaluation, and broader outcomes of the apology.

As an initial step, the present experiment addressed the limitations in apology literature using dynamic, computer-generated faces embedded in transgression scenarios. A generative platform sampled facial Action Units (AUs), gaze directions, and blushing independently, allowing their associations with sorry judgements to be estimated without beginning from a fixed emotion prototype. Models were built separately for each receiver. This individual-level approach preserved between-person variation and made it possible to ask which effects recurred across receivers rather than treating a group-average face as the only representation of the phenomenon (Ince et al., 2021).

This chapter provides the foundational empirical test of Chapter 1's receiver-interpreted social-signal account by testing how facial movements, gaze directions, and blushing states are associated with apparent apology without imposing an emotion prototype. Establishing this baseline repertoire of visual signals is necessary before later chapters can test whether its organisation varies with moral domain or between the Western and East Asian receiver groups.

Participants judged how "sorry" each face appeared rather than how "apologetic" it appeared. Uses of *sorry* are frequent in ordinary British English and extend across routine conversational settings (Aijmer, 2019; Deutschmann, 2003). The wording therefore directed participants towards an everyday interpersonal judgement while avoiding the implication that they should evaluate a formally complete verbal apology. The central questions were: which facial movements and gaze directions were associated with sorry judgements; whether these signals changed across apparent apology intensity; and whether any descriptive variation was evident across three transgression contexts.

2.2 Methods

The method was designed to identify visual cues associated with apology without first specifying what an apologetic face should look like. Dynamic faces varied facial movements, gaze and blushing within a transgression context, allowing their associations with sorry judgements to be estimated against randomness. Models, facial signals that were relevant to apology, were first estimated for each participant to preserve differences between receivers/participants, then compared across participants, apparent apology intensities, and transgressions. The Planned Data Analysis section below first explains how signal–response associations were identified and combined and then explains how the achieved AU and gaze patterns were examined.

2.2.1 Participants

The analysis set comprised 60 White participants aged 18–35 years ($M = 20.95$, $SD = 3.51$). The sample contained 30 women and 30 men and was balanced across transgressions, with 20 participants assigned to each scenario (Table 2.1). Each transgression group contained 20 participants (10 women and 10 men), reducing unequal cell size and gender composition as additional sources of descriptive condition differences without implying that transgression caused any observed contrast or that equal gender composition is universally required. The analysis used the first balanced set of 60 participants. Two later records, participant 93 and participant 101, were not included in this balanced analysis set.

Participants were recruited through the University of Glasgow Subject Pool or social media. Eligibility criteria excluded self-reported learning difficulties and psychiatric, psychological, neurological, visual-processing, or hearing conditions that could interfere with the task. Recruitment also sought participants with limited extended exposure to other cultural environments so that the chapter could first characterise apology judgements within a comparatively homogeneous Western sample.

Testing took place in a purpose-built booth in the School of Psychology and Neuroscience at the University of Glasgow. Participants provided written informed consent and were paid £9 per hour. The study was approved by the University of Glasgow College of Medical, Veterinary and Life Sciences Ethics Committee (application 200220328).

Table 2.1

Participant Characteristics for the Balanced Chapter 2 Sample.

Measure	Value
Participants	60
Gender	30 female; 30 male
Age	M = 20.95, SD = 3.51, range 18–35
Transgression allocation	20 foot; 20 marathon; 20 pen

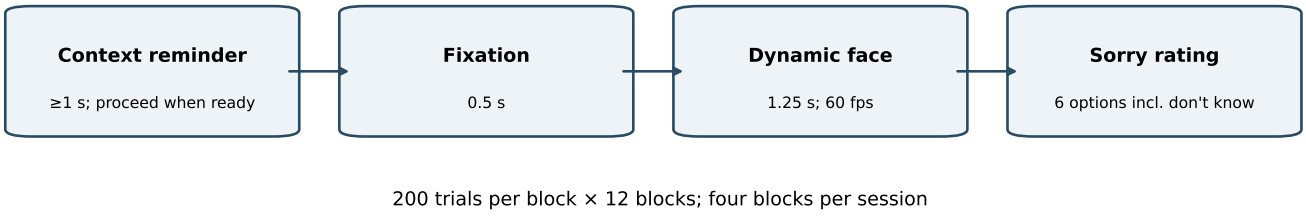
2.2.2 Procedure

Participants received written and verbal instructions, could ask questions, and completed six practice trials with the experimenter present. They were seated in a darkened testing room and used a chin rest, mouse, keyboard, and graphical response interface. The experiment used a 51-cm-wide, 1920 × 1080-pixel monitor at a viewing distance of 71 cm, with stimuli subtending around 14.37° vertically and 8.79° horizontally, approximating the apparent size of a human face during typical social interaction (Jack et al., 2012).

Figure 2.1 summarises a trial. At the start of each block, participants read a transgression scenario presented for at least 1 s and clicked a button when ready. A central fixation cross then appeared for 0.5 s, followed by a dynamic face. Participants rated how sorry the person appeared using one of five substantive responses: *not sorry at all*, *a little sorry*, *somewhat sorry*, *very sorry*, or *extremely sorry*. A sixth *don't know* response was available and was excluded from modelling. After selecting a response, participants pressed the space bar to begin the next trial.

Figure 2.1

Trial Sequence and Session Structure.



Note. The transgression remained constant for each participant, whereas face sex alternated by block and was counterbalanced across participants.

Each block contained 200 trials, with the transgression reminder repeated after every 50 trials. Participants completed 12 blocks, giving 2,400 trials in total. Four blocks formed a session, so the task comprised

three sessions. A mandatory two-minute break separated blocks, and sessions were separated by at least one hour. Participants completed the three sessions over two or three consecutive days. The transgression was constant for a participant and therefore constituted a between-participant factor. Face sex alternated across blocks and was counterbalanced.

2.2.3 Stimuli

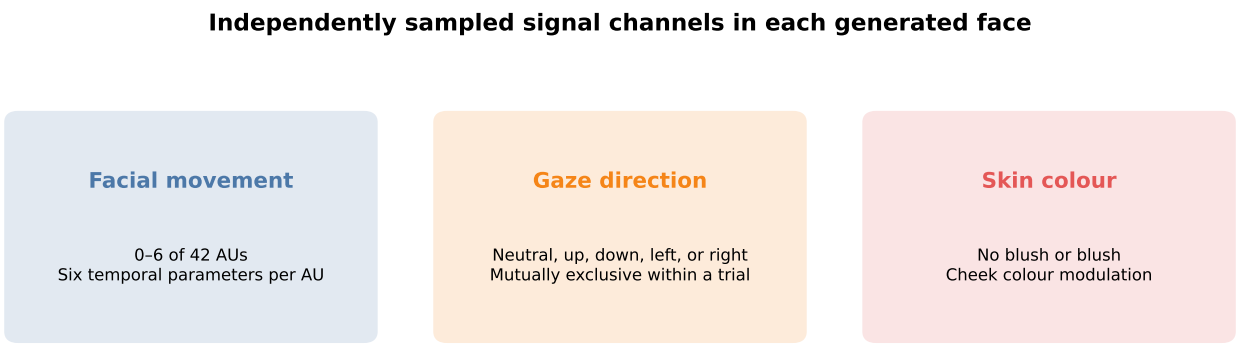
The 2,400 dynamic stimuli were generated with the Generative Face Grammar (GFG), a perception-driven platform for synthesising photorealistic facial movements (Yu et al., 2012). The base set contained 1,200 female and 1,200 male faces. The same underlying stimulus space was used across transgression groups, with presentation order randomised across participants. Videos were generated at 60 frames per second with a duration of 1.25 s, with a resolution of 600 × 800 pixels.

Short Introduction to GFG. GFG represents facial expression using Action Units (AUs), anatomically defined facial movements described by the Facial Action Coding System (Ekman & Friesen, 1978). It can generate biologically plausible combinations of AUs and vary how each movement unfolds over time. Each AU selected for the current animation trial was specified by six temporal parameters controlling onset, acceleration, peak timing and amplitude, deceleration, and offset. This parameterisation allowed the experiment to sample dynamic facial behaviour rather than presenting a small set of fixed photographs or emotion prototypes.

The experiment manipulated three signal channels: facial movement, gaze direction, and blushing (Figure 2.2). Figure 2.3 shows, on the left side, the facial signal synthesis procedure; on the right side, an example facial stimulus with the response options and transgression type being presented. Head movement was not sampled because rotation could obscure the visibility of facial and gaze signals. The channels were analysed separately before their implications were integrated.

Figure 2.2

Signal Channels Sampled in the Generated Stimuli.

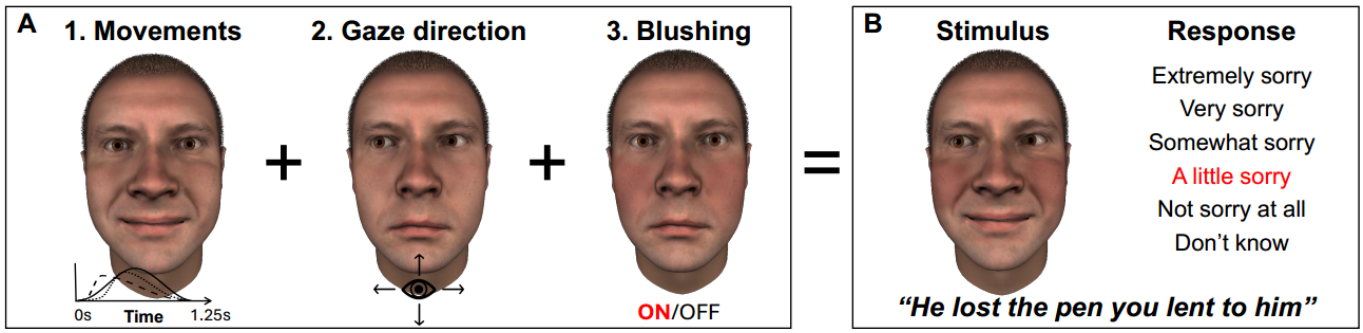


Note. Facial movements could co-occur, whereas gaze directions were mutually exclusive within a trial.

Action Units (AUs). For each stimulus, the number of AU draws (zero to six) followed a binomial sampling rule with $n = 6$ and $p = .5$. Labels were then drawn with equal probability from the set of 42

Figure 2.3

Dynamic Stimulus Generation and an Example Experimental Trial.



Note. Facial movements, gaze direction, and blushing were sampled as separate channels and combined in the rendered stimulus.

AUs, with replacement; repeated labels activated the same AU only once. Each selected AU received independently sampled values for its six temporal parameters. Consequently, a trial could contain several simultaneous facial movements, and both the composition and dynamics of the movement pattern varied across trials.

Gaze Direction. Each trial contained one of five gaze states: neutral, up, down, left, or right. Gaze was mutually exclusive by design, so a face could not look in two directions within the same stimulus. The gaze onset was fixed in the stimulus configuration to reflect natural gaze aversion rather than sampled with the six AU parameters. This difference in sampling space motivated a separate gaze analysis.

Blushing. The skin-colour channel contained two states: no blushing or blushing. Blushing was applied to the cheek region using a fixed temporal profile. It was sampled independently of AUs and gaze, allowing individual-level associations with sorry responses to be assessed even when blushing did not recur at the population level.

Facial Identities. Each trial presented a different facial identity. The identity generator varied three-dimensional face shape and two-dimensional texture while controlling categorical attributes such as age, sex, and ethnicity (Zhan et al., 2019). The stimuli used White identities aged approximately 18–35 years to match the participant population. Random identity components, hair properties, eye colour, and facial texture produced 2,400 stimulus identities, divided equally between female and male faces. Varying identity across trials reduced the risk that an association between a facial movement and a response was driven by one repeatedly presented person.

Transgression Selection. Participants judged one of three scenarios: an offender accidentally stepped on the participant's foot while queueing; an offender took a shortcut in a marathon in which the participant also competed; or an offender lost a pen borrowed from the participant. The first scenario foregrounded harm or discomfort, the second foregrounded unfairness, and the third provided an everyday property-loss comparison. Harm and fairness are central dimensions in pluralist accounts of the moral domain (Graham et al., 2011); the inclusion of more than one context also reflects evidence that apology expectations vary with offence and relationship characteristics (Guan et al., 2009; Mapson, 2014). These scenarios were

included to broaden contextual sampling, not to provide a complete or confirmatory test of moral-domain differences.

2.2.4 Planned Data Analysis

The analysis separated model construction, descriptive summaries, primary analyses and exploratory analyses. Participant-level MI/PMI identified the associations between signals and responses used to construct each receiver's models; this was distinct from the response-model validation introduced in Chapter 3. Prevalence summaries described the retained signals. Gaze-PMI comparisons addressed the primary question of apparent apology intensity, while NNMF and its condition-specific AU accents were exploratory.

MI/PMI modelling used MATLAB scripts; NNMF and gaze ANOVA used the Statistics and Machine Learning Toolbox. Python scripts using NumPy and SciPy provided NNMF diagnostics and participant-clustered gaze sensitivity analyses.

Modelling. Mutual information (MI) asked whether knowing a signal's state helped predict a receiver's response. For a signal S and response R, it quantified how much the signal reduced uncertainty about the response (Ince et al., 2017):

$$I(S;R) = \sum_{s,r} p(s,r) \log_2 \left(\frac{p(s,r)}{p(s)p(r)} \right). \tag{2.1}$$

MI was calculated separately for every AU, gaze direction and blushing state within each participant. Shuffling response labels 1,000 times provided a chance reference for each signal; the model retained signals whose observed MI met or exceeded that signal's 95th-percentile threshold. This was a separate threshold for each signal, without correction across all tested signals.

The first model collapsed responses into *not sorry* and *sorry*, with all four sorry intensities combined. For a significant signal, the direction of its association was determined from how signal presence changed the response distribution. This produced one generic not-sorry and one generic sorry signal vector for every participant.

After identifying a signal related to the response, pointwise mutual information (PMI) asked which sorry category it supported. PMI described the association between a particular signal state s and response r:

$$\mathrm{PMI}(s,r) = \log_2 \left(\frac{p(s,r)}{p(s)p(r)} \right), \tag{2.2}$$

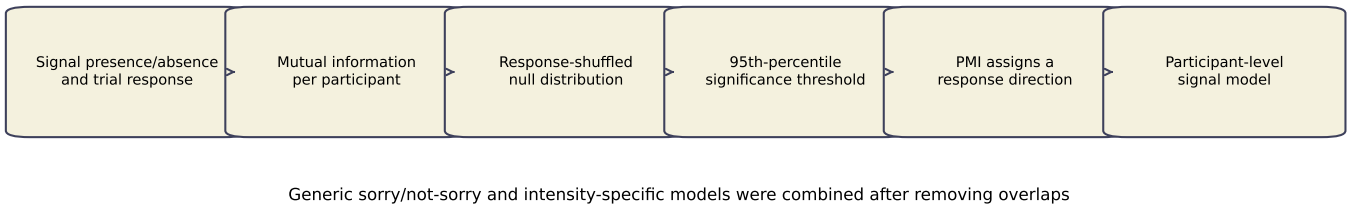
so that positive values indicated that a signal and response co-occurred more often than expected under independence. For each significant signal, the response with the largest relevant PMI value defined its intensity-specific assignment.

Generic sorry signals were combined with intensity-specific signals, with overlapping signals retained only in their intensity-specific category. Each participant's final model recorded which of the 42 AUs and five gaze directions were associated with each response. These 47-signal binary vectors were compared across participants and conditions. Blushing was separated post hoc into a parallel channel-specific array. Figure 2.4 summarises the modelling sequence.

In conclusions, our modelling method allowed as to retrieve unbiased signals (from any signal prototypes), for each participant, that were used to significantly indicate the given level of apparent sorriness.

Figure 2.4

Participant-Level MI and PMI Modelling Workflow.



Note. Generic and intensity-specific signal sets were combined after overlapping signals were assigned to the intensity-specific set.

Analysis. After model construction, prevalence summaries described which signals recurred across receivers and how their frequency differed across apparent apology intensities and transgressions. For AUs, non-negative matrix factorisation (NNMF) provided a description of movements that occurred together across the participant-level models (Lee & Seung, 1999). Its factors described recurring configurations in certain transgressions or apology intensities.

Gaze was analysed separately because only one gaze state was shown per stimulus, whereas AUs could be combined. Gaze PMI quantified whether a direction was associated with a particular sorry response. A factorial ANOVA examined variation across gaze direction, sorry intensity and transgression, and a participant-clustered sensitivity analysis checked whether conclusions depended on treating repeated values from the same participant as independent.

2.3 Results

Table 2.2 separates the availability of combined response models from the intensity-specific evidence contributing to them. Models were initially constructed for 60 participants across five response categories (300 models in total). However, extremely-sorry responses were rare: only 2,120 of 144,000 trials (1.47%) received this response, and 13 participants never used it: only 18 participants had any signal specific to that intensity. On the other hand, intensity-specific signals were available for 58, 56, and 54 participants at the three retained sorry levels, respectively. The extreme-sorry category was therefore excluded from any afterwards analyses. The retained analysis contained 180 sorry models: three intensities × three transgressions × 20 participants. Figure 2.5 shows the retained condition-by-response facemaps for these models.

Blushing did not produce a population-level pattern. No participant had a generic not-sorry blushing signal. Four participants had a somewhat-sorry-specific blushing signal and two had a very-sorry-specific signal, but no blushing effect recurred sufficiently often to support a group-level interpretation. Thus, blushing remained an individual-level association rather than a recurring population-level pattern; the population-level results therefore focus on AUs and gaze direction.

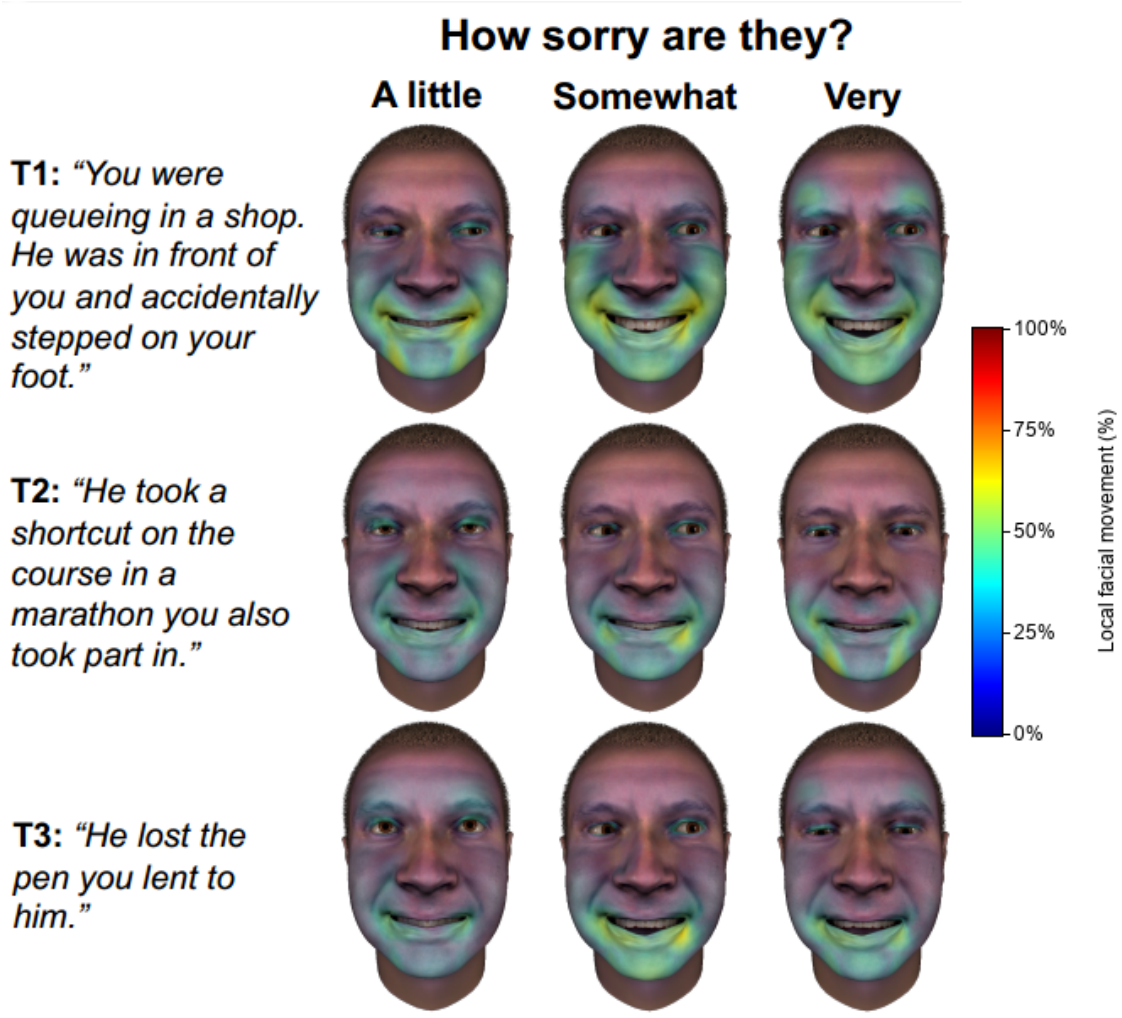
Table 2.2

Availability of Combined and Intensity-Specific Participant-Level Signal Models in the Unbinned Five-Response Analysis.

Response	Combined model	Empty	AU	Gaze	Blushing	Intensity-specific
Not sorry	60	0	60	57	0	n/a
A little	60	0	60	56	0	58
Somewhat	60	0	60	51	4	56
Very	60	0	60	56	2	54
Extremely	60	0	59	40	0	18

Figure 2.5

Facemaps of Retained Facial-Movement Prevalence by Transgression and Sorry Response.



Note. Colour highlights facial changes from AUs weighted by the proportion of participants whose models contained each AU. Each face is a group summary and displays only the most prevalent gaze direction. The bar shows the normalised local facial movement scale (0–100%); colours fade into the facial texture at lower values.

2.3.1 Action Units (AUs)

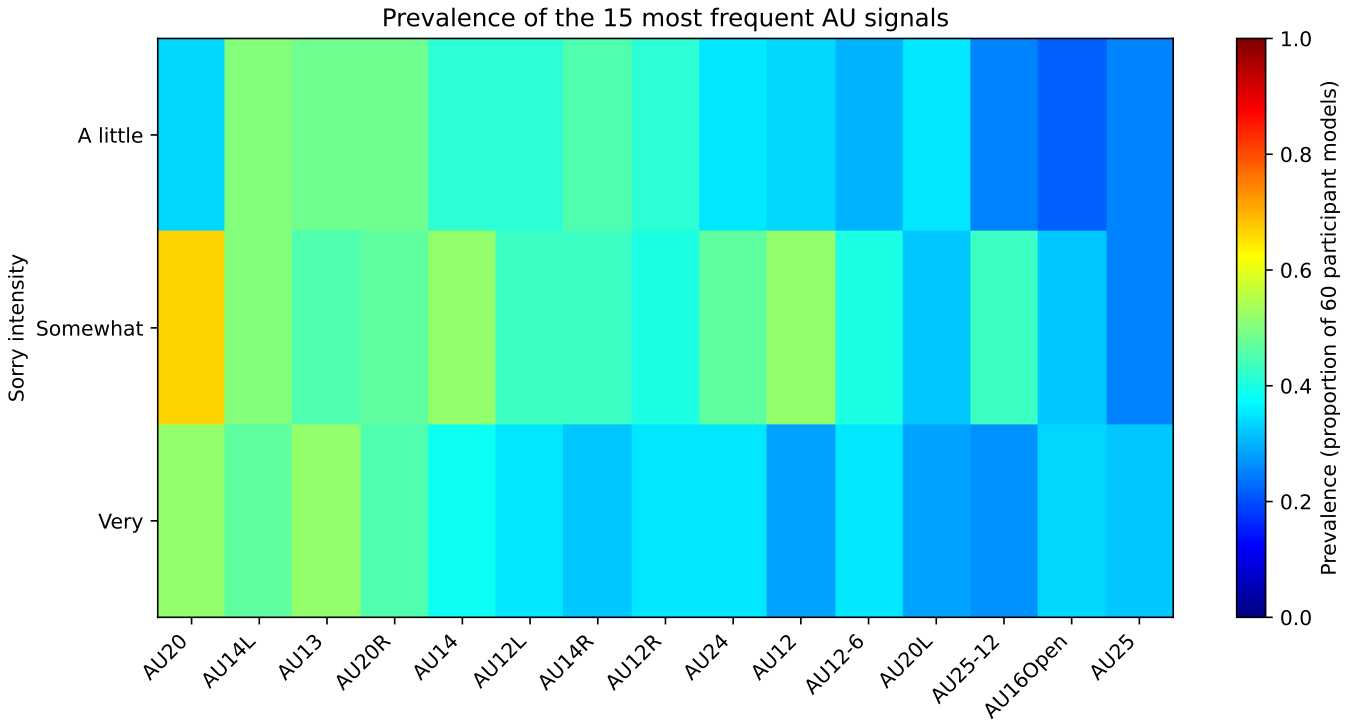
Across not-sorry and the three retained sorry levels, all 42 sampled AUs appeared in at least one participant model. Not-sorry models contained 29 unique AUs, whereas all 42 appeared somewhere across the sorry models. This difference indicates greater signal diversity across the three sorry intensities, but it does not imply that every AU was prevalent or psychologically diagnostic at the population level.

Not-sorry judgements were dominated by AU43 in 48 of 60 participant models, followed by AU9 in 38 and AU10Open in 31. The sorry models had a different lower-face emphasis. Across the 180 retained sorry models, the most frequent movements were AU20 (91 models), AU14L (88), AU13 (87), AU20R (84), AU14 (79), and AU14R (72). AU12, the bilateral Lip Corner Puller, appeared in 68 models, alongside unilateral AU12L and AU12R variants in 72 and 70 models. Thus, smile-related movements were recurrent, but the most prevalent configuration also included lip stretching, dimpling, and other mouth-shape changes rather than a single canonical smile.

As Figure 2.6 and Table 2.3 show, the intensity profiles were not monotonic for every movement. AU20 increased from 20 of 60 a-little-sorry models to 40 somewhat-sorry models and remained present in 31 very-sorry models. AU12 appeared in 18, 31, and 19 models across those intensities. The central AU result was therefore configuration-level: related lower-face movements recurred across the retained sorry levels, but their prevalence shifted with apparent apology intensity rather than defining one fixed set of facial expressions.

Figure 2.6

Prevalence of the 15 Most Frequent AU Signals Across the Three Retained Sorry Intensities.



Note. Each cell is the proportion of the 60 participant models containing the AU at that response level.

Table 2.3

Most Prevalent AU Signals Across the Three Retained Sorry-Intensity Models.

AU	Not-sorry models	Sorry models	Proportion of 180
AU20	0	91	50.6%
AU14L	2	88	48.9%
AU13	0	87	48.3%
AU20R	0	84	46.7%
AU14	1	79	43.9%
AU14R	3	72	40.0%
AU12L	0	72	40.0%
AU12R	2	70	38.9%
AU24	1	70	38.9%
AU12	3	68	37.8%

Exploratory Analysis: NNMF. Non-negative Matrix Factorisation (NNMF) was used to identify recurring AU configurations (Lee & Seung, 1999). Let V denote the non-negative 42-AU by 180-model matrix. NNMF estimates non-negative factor weights W and model loadings H such that

$$V \approx WH, \qquad W \geq 0, \quad H \geq 0. \tag{2.3}$$

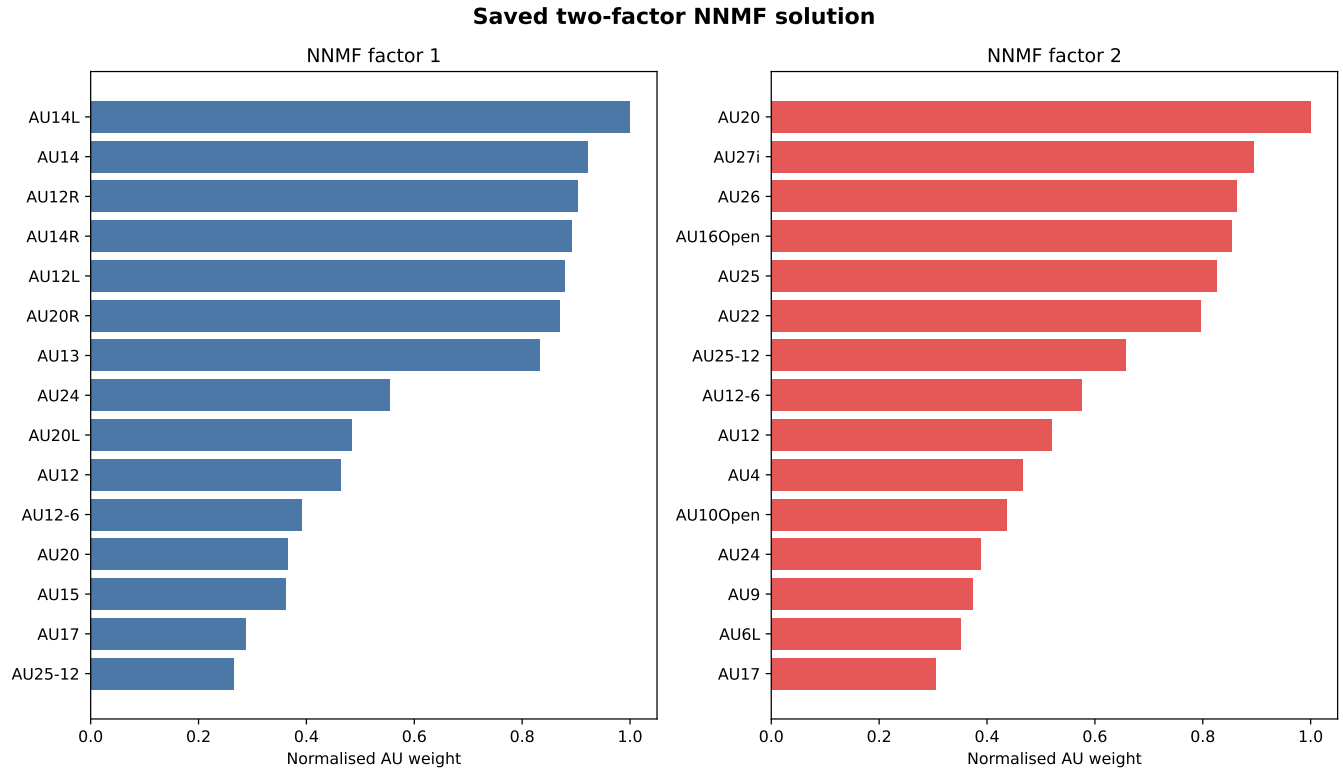
The original analysis fixed the rank at two after exploratory inspection of reconstruction error and returned two distinct AU configurations. A subsequent corrected rank search showed decreasing reconstruction error at every tested rank from one through six, so it did not uniquely select rank two; alternative-initialisation stability for the retained two-factor solution was nevertheless near-perfect. Its statistical adequacy and generalisability are considered in the Limitations.

Factor 1 weighted AU14L, AU14, unilateral AU12 variants, AU20R, AU13, AU24, and AU12 most strongly. This factor combined dimpling and lip-corner-pulling movements with other lower-face actions and is consistent with a smile-related configuration. Factor 2 placed its largest weights on AU20, AU27i, AU26, AU16Open, AU25, and AU22, together with AU4 and some AU12-containing movements. This second configuration emphasised lip stretching, mouth opening, and jaw movement and was more compatible with a strained or negative-affect display. Figure 2.7 shows the 15 largest weights for each factor.

Median factor loadings varied across the nine intensity-by-transgression conditions. Factor 1 declined from .728 to .491 across intensities in the foot scenario, fell and then rose in the marathon scenario (.319, .299, .425), and rose and then fell in the pen scenario (.352, .504, .270). Factor 2 increased across intensities in the foot (.065, .345, .490) and pen (.106, .319, .361) scenarios, but was non-monotonic in the marathon scenario (.214, .226, .158). Neither factor therefore showed a uniform intensity trend across all three contexts. The residual "accent" procedure likewise identified small and irregular sets of condition-specific AUs. These accents were defined by a 75th-percentile factor-weight filter and a prevalence threshold, not by an independent inferential test. They are therefore retained as exploratory visual descriptions rather than evidence that each transgression possesses a distinct apology grammar. The two retained factors therefore captured recurring but context-variable facial-movement configurations, not two discrete apology emotions or transgression-specific grammars. Figure 2.8 shows how the two patterns and the residual accents con-

Figure 2.7

Normalised AU Weights in the Two-Factor NNMF Solution.



Note. Factor 1 emphasised smile-related lower-face actions, whereas Factor 2 emphasised lip stretching, mouth opening, and jaw movement.

tributed across the nine condition profiles.

2.3.2 Primary Analysis: Gaze PMI

The gaze models could retain associations with more than one of the five gaze states. In not-sorry models, upward gaze was most frequent, occurring for 47 of 60 participants; neutral gaze occurred for 23, and each remaining direction occurred for five or fewer participants. The a-little-sorry models instead contained right gaze for 39 participants, upward gaze for 36, and left gaze for 35. At somewhat sorry, right gaze remained most frequent (33 participants), followed by left (23) and down (13). At very sorry, down gaze became the most frequent direction (34 participants), followed by right (26) and left (17).

The PMI profiles clarified this shift. Averaged across transgressions, lateral gaze had positive PMI at the a-little-sorry level (left = .016; right = .017). Downward gaze increased from approximately -.003 at a little sorry to .020 at somewhat sorry and .049 at very sorry. Upward gaze became increasingly negative across the same levels (-.005, -.023, and -.028). Thus, downward gaze was increasingly associated with stronger sorry judgements, whereas upward gaze was associated with not-sorry or weaker apology interpretations.

As Table 2.4 and Figure 2.9 show, the factorial ANOVA reproduced the original analysis values exactly. There was a main effect of gaze direction, $F(4,871) = 60.59$, $p < .001$, $f = .527$, and a gaze-by-intensity interaction, $F(8,871) = 17.31$, $p < .001$, $f = .399$. The ANOVA also returned a gaze-by-transgression interaction, $F(8,871) = 5.77$, $p < .001$, $f = .230$. A post-hoc participant-clustered sensitivity check assessed whether these conclusions depended on treating the 900 PMI observations as independent (60 participants $\times$ 3 apology intensities $\times$ 5 gaze directions $=$ 900). The check kept the same estimated gaze, intensity and transgression effects, but recalculated their uncertainty while allowing the 15 values from each participant to be related. The 60 participants, rather than the 900 PMI values, were treated as independent units. It retained the gaze and gaze-by-intensity effects but gave $p = .065$ for gaze by transgression. The shared conclusion is that gaze direction differentiated apparent apology intensity, especially through the association between downward gaze and stronger sorry judgements. Because the participant-aware analysis did not retain gaze by transgression, the chapter does not treat gaze as robustly context-specific.

Table 2.4

Legacy Factorial ANOVA of Gaze PMI Values.

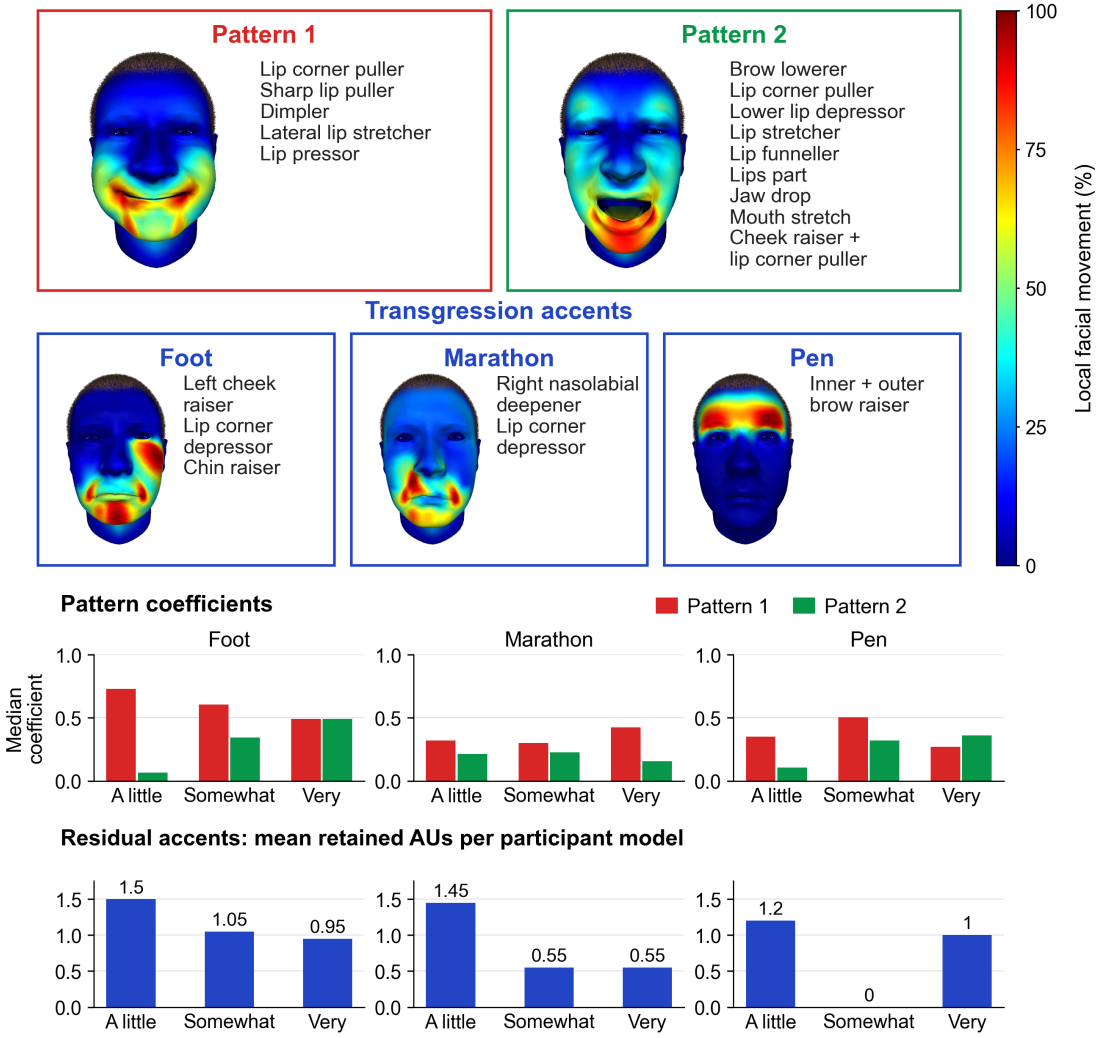
Effect	*df*	*F*	*p*	Partial eta squared	Cohen *f*
Gaze direction	4	60.59	< .001	0.218	0.527
Intensity	2	1.08	= .340	0.002	0.050
Transgression	2	1.67	= .188	0.004	0.062
Gaze × Intensity	8	17.31	< .001	0.137	0.399
Gaze × Transgression	8	5.77	< .001	0.050	0.230
Intensity × Transgression	4	0.23	= .920	0.001	0.033

2.4 Discussion

This chapter asked which dynamic visual signals lead Western receivers to judge an offender as more or less sorry. Considered together, the channel analyses show that apparent apology was associated with

Figure 2.8

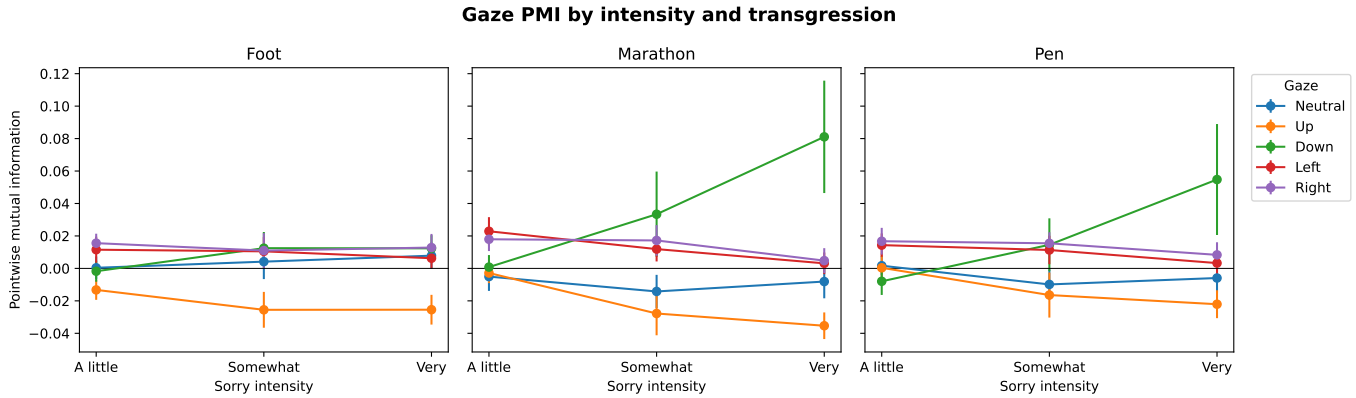
Facemap Rendering of the Two Retained NNMF Patterns and the Descriptive Transgression Accents.



Note. Factor faces use the saved maximum-normalised AU weights; accent faces locate retained AUs at maximum rendering amplitude. Colour shows normalised local facial movement (0–100%). Red and green bars give median factor coefficients; blue bars give the mean number of retained residual AUs per participant model after filtering, on a separate axis. These are descriptive summaries.

Figure 2.9

Mean Gaze PMI by Apparent Apology Intensity and Transgression.



Note. Error bars are 95% normal-approximation intervals based on the 20 participants within each transgression. Positive values indicate more frequent signal–response co-occurrence than expected under independence.

intensity-sensitive facial-movement configurations and gaze-direction patterns rather than one fixed set of facial expressions. Smile-related and strain- or negative-affect-compatible facial movements contributed to the configurations, lateral gaze direction was associated with lower-intensity judgements, downward gaze direction became increasingly associated with greater apparent apology intensity, and blushing did not recur at population level.

The AU results help reconcile competing expectations in the apology literature. The repeated appearance of AU12 and its unilateral variants shows that lip-corner pulling can contribute to how sorry an offender appears. This need not mean that receivers interpreted the display as happiness. In combination with AU14 dimpling, AU20 lip stretching, mouth opening, and contextual information about the offence, smile-like movements may communicate affiliation, appeasement, nervousness, or an attempt to reduce interpersonal threat. The second NNMF configuration further shows that mouth stretch and jaw-opening movements co-existed with the smile-related pattern. Apology appearance was therefore organised around a mixture of affiliative and negatively valenced cues rather than a pure sadness or pure smile display.

This configuration account extends the static reverse-correlation result reported by George et al. (2023). Their sadness-like template identified an important negative-affect component of apology expectations. The present data show that receivers also use dynamic lower-face and gaze information when making the judgement in an explicit transgression context. Because the stimuli varied identity on every trial, these associations are more readily attributed to controlled movements than to stable facial morphology.

Gaze carried the clearest graded pattern. Upward gaze was common in not-sorry models and showed negative PMI for the retained sorry levels. Lateral gaze was most strongly associated with a-little-sorry judgements, whereas downward gaze increased with apparent apology intensity. Downward gaze may be compatible with shame, submission, reduced social dominance, or withdrawal, but the present task did not ask participants to label those states. The defensible conclusion is therefore perceptual: receivers used gaze direction to differentiate apparent apology intensity, especially by associating downward gaze with stronger sorry judgements.

The three scenarios broadened the contexts in which signals were sampled, but Chapter 2 does not establish distinct signals for harm, fairness, and everyday loss. The NNMF loadings and accents varied descriptively, yet their changes were not uniform. The gaze-by-transgression effect was also not retained by the participant-clustered sensitivity analysis. These results motivated the more systematic moral-domain comparisons developed in the following chapter rather than providing a definitive domain account here.

More generally, the models describe receiver expectations, not the movements that apologising offenders necessarily produce. A signal can make a generated face look sorry without being a deliberate or universal apology behaviour. The findings therefore concern the perceptual code used by this Western sample under these task conditions. They identify candidate facial and gaze cues through which receivers infer apology, while leaving production, sincerity, cultural generality, and downstream forgiveness to later work.

2.5 Limitations

The response distribution was strongly uneven. Not-sorry, a-little-sorry, and somewhat-sorry responses dominated the data, whereas extremely-sorry responses constituted 1.47% of trials and were absent for 13 participants. Combining generic sorry signals with intensity-specific signals meant that a participant could have a non-empty extreme-sorry model even without reliable evidence specific to that response. Excluding the extreme category improved comparability across retained models but did not remove the underlying imbalance. Future work should design or re-bin the scale so that intensity comparisons have adequate within-participant support.

The MI analysis used a separate signal-wise 95th-percentile permutation threshold for each predictor, but did not provide a single family-wise correction across all AUs and channels. The broad signal coverage—including at least one occurrence of every AU in a retained sorry model—should therefore not be interpreted as evidence that all 42 AUs are population-level apology cues. Prevalence, effect magnitude, and replication across receivers remain more informative than isolated significance.

The NNMF analysis for AU yielded two formal factors, but rank two was selected by inspection rather than a preregistered criterion, and sensitivity analysis across ranks one to six found no clear elbow supporting this choice. The rank-two solution showed modest fit (RMSE = 0.353; relative Frobenius error = 0.745; 28.4% of variance explained). Nevertheless, 30 random initialisations produced virtually identical factors (mean cosine similarity = 0.999664), indicating negligible random-start instability. The principal limitations are therefore rank selection, low explained variance, and uncertain generalisability beyond the stimuli and scenarios examined. Likewise, the 75th-percentile accent rule was descriptive, and its apparent condition differences require validation across a broader range of transgressions and cultures.

The ANOVA analysis for gaze direction was limited by treating multiple PMI values from the same participant as independent observations, despite their within-participant dependence. When uncertainty was recalculated to allow PMI values from the same participant to be related, the gaze-by-transgression interaction was no longer statistically significant ($p = .065$). The check did not estimate participant or face-identity random effects or compare alternative models. Future work could use hierarchical models to examine participant differences directly.

Finally, the sample represented one Western population, the generated identities were restricted to the same broad ethnicity, and the scenarios covered only three contexts rather than the full moral-domain space. Differences among the scenarios could reflect severity, familiarity, intentionality, interpersonal

consequences, or moral content. Within these perceptual and design boundaries, the chapter answers its research question: Western receivers associated distinct facial-movement configurations and gaze-direction patterns with apparent apology intensity, while blushing showed no recurring population-level pattern. Because the exploratory scenarios cannot isolate moral-domain organisation, Chapter 3 next tests whether the baseline visual-signal architecture is systematically tuned across five moral domains.

Chapter 3

Visual Apology Signals Across Moral Domains: Western Receivers

3.1 Chapter Introduction

3.1.1 Moral-Domain Theory

Apology becomes theoretically important when an action is understood as a transgression. The offender is not only describing what happened but also inviting the victim to reinterpret the offender, the offence, and the future relationship. B. Ho (2012) models apology as a signal that can restore damaged relationships, while Ohtsubo and Watanabe (2009) shows empirically that receivers are sensitive to whether an apology is sufficiently costly to appear sincere. Before asking whether apology signals differ across offences, it is therefore necessary to clarify how moral psychology has conceptualised the kinds of wrongs that require repair.

Contemporary discussions of moral domains often begin with Moral Foundations Theory (MFT), developed by Haidt, Graham, and their colleagues. Haidt and Joseph (2004) proposed that morality rests on intuitive, partly prepared systems that cultures elaborate into local virtues, and Graham et al. (2013) later formalised MFT as a pluralist account of moral judgement. This account emerged from a broader theoretical history. Kohlberg's (1971) cognitive-developmental work placed justice, rights, reciprocity, and principled reasoning at the centre of moral development, while Kohlberg (1969) described social conflict and experience as drivers of change between structures of moral judgement. Gilligan (1982) challenged the assumption that justice exhausts moral maturity, arguing that care, responsibility, communication, and relationship maintenance can organise moral reasoning in ways that a justice-only model misses. Social-domain work associated with Turiel (1983) also made the moral–conventional distinction central to developmental psychology, particularly in relation to welfare, justice, rights, and harm (Buchtel et al., 2015; Graham et al., 2011; Gray et al., 2012).

A second lineage is functional and evolutionary. Hamilton's (1964) account of inclusive fitness explained how behaviour benefiting relatives can be selected despite costs to the actor, while Trivers (1971) extended evolutionary analysis to reciprocal altruism among non-kin, including cheating detection, moralistic aggression, gratitude, and guilt-linked reparation. These accounts frame morality not only as abstract reasoning but also as a set of evolved and socially regulated responses to recurrent problems of care, cooperation, coalition, hierarchy, and contamination. Shweder et al. (1997) added a cultural-pluralist account organised around the ethics of autonomy, community, and divinity, while Chiu et al. (1997) showed that implicit theories of social-moral reality shaped whether people emphasised duties within an existing order or moral principles and individual rights. Together, these strands provided the background for MFT's claim that harm and fairness are important but incomplete.

The Development of Moral-Domain Theory. MFT's distinctive contribution was to turn this broader lineage into a vocabulary of moral domains. Haidt and Joseph (2004) argued that the mind contains an "intuitive ethics": prepared sensitivities that cultures can elaborate into different virtues, stories, and institutions. Haidt and Joseph drew on developmental psychology, anthropology, primatology, evolutionary theory, and cultural psychology to identify recurrent moral concerns that appeared both psychologically pre-

pared and culturally elaborated. Graham et al. (2011, 2013) later formalised five initial candidate foundations: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation. These foundations were not intended as an arbitrary list of topics or merely as factors discovered in questionnaires. They are candidate systems that respond to recurrent adaptive challenges and acquire culturally specific forms (Graham et al., 2013; Haidt & Joseph, 2004).

The rationale for each foundation follows from this adaptive logic. In Graham et al.'s (2013) theoretical account, the care/harm foundation concerns suffering, need, vulnerability, and protection, and is linked to the problem of caring for children and vulnerable others. The fairness/cheating foundation concerns cooperation, reciprocity, cheating, proportionality, and trust, and draws strongly on reciprocal-altruism logic. Loyalty/betrayal concerns coalitions and group commitment, authority/subversion concerns hierarchy and ordered relationships, and sanctity/degradation concerns contamination, disgust, sacredness, and the protection of spiritual or symbolic purity. In this sense, Graham et al. (2013) present the five domains as a pluralist but constrained account: multiple foundations are expected because human social life contains multiple recurring problems, but the list remains open to revision when new candidate foundations meet the criteria for "foundationhood".

Moral-domain theory also required measurement. Graham et al. (2009) tested how liberals and conservatives differed in their endorsement of moral and immoral statements, finding that liberals emphasised harm and fairness more strongly whereas conservatives drew on the five foundations more evenly. Graham et al. (2011) subsequently operationalised MFT through the Moral Foundations Questionnaire (MFQ). Later validation supports the usefulness of moral-domain measures while showing that their structure should not be treated as fixed. Dogruyol et al. (2019) found that a five-factor model was broadly stable across WEIRD and non-WEIRD samples although item loadings differed; Atari et al. (2023) developed the MFQ-2 across 25 populations and found cultural variation in the foundations' nomological networks; and Costa and Moreira (2024), using the Moral Foundations Sacredness Scale, found stronger support for a two-factor individualising/binding model. These findings justify using moral domains as an organising framework while treating their exact measurement structure as an empirical question.

Once moral domains could be measured, researchers began using them to examine other aspects of social life. For example, Buchtel et al. (2015) showed that the meaning of "immoral" could depend on whether a culture treats harm or incivility as the more prototypical violation. Hui and Bond (2009) showed that face loss after relational transgression affected forgiveness and repair-related motivation, especially among Hong Kong Chinese participants. Laham et al. (2010) showed that reactions to autonomy and community transgressions differed across India and Britain, with autonomy violations more strongly linked to anger and community violations more strongly linked to contempt. These studies do not all use MFT in the same way, but they converge on the same implication: the judged meaning of a violation shapes emotional, relational, and behavioural responses. This links moral-domain research with moral transgressions. Moral domains are not only abstract concerns; they are also reflected in daily interactions where violations take place. Hence, investigating moral domains also requires understanding how people interpret concrete acts as particular kinds of wrongdoing.

From Moral Domains to Moral Transgressions. Moral-transgression research moves from abstract domain labels to concrete violations. Clifford et al. (2015) developed the Moral Foundations Vignettes, a standardised scenario set designed to represent different foundation violations while reducing confounds

such as scenario length, complexity, and required background knowledge. These vignettes make moral domains more empirically tractable by allowing responses to recognisable transgressions to be compared.

Moral emotions provide a further bridge between transgression and repair. Tangney et al. (2007) reviewed guilt, shame, embarrassment, pride, gratitude, empathy, and related emotions as forces that can motivate both avoidance of wrongdoing and constructive repair. Landmann and Hess (2018) found partial support for emotion-specific foundations: care was especially associated with compassion and purity with disgust, whereas fairness, authority, and loyalty showed less specific patterns. Sunar et al. (2021) similarly found that anger, disgust, contempt, shame, and guilt varied with the relational model being violated. The moral meaning of a transgression therefore depends not only on a domain label but also on emotional and relational interpretation; these levels are related but not interchangeable.

Contemporary Monist and Pluralist Accounts. The contemporary debate can be framed as a disagreement between Gray's harm-based account and Graham's pluralist account. Gray et al. (2012) argue that morality is rooted in mind perception: people understand moral wrongs through a dyad comprising an intentional agent and a suffering patient. In this account, even apparently harmless violations can involve perceived, potential, symbolic, or spiritual harm. Gray and Keeney (2015) sharpened the challenge by arguing that widely used purity scenarios may be confounded with weirdness and severity rather than isolating a distinct foundation. Schein and Gray (2018) later formalised this position as the Theory of Dyadic Morality, in which moral condemnation is driven by norm violation, negative affect, and perceived dyadic harm.

The pluralist response accepts that severity, typicality, and domain-general processes matter but argues that they do not explain away all differences between harm and impurity judgements (Graham, 2015). Confounds in some stimuli do not establish that loyalty, authority, sanctity, or fairness are merely disguised forms of harm; instead, they demonstrate the need for better measures and more careful tests. MFT is therefore best understood as pragmatic pluralism: shared processes can operate across morality while domain-specific concerns retain explanatory value (Graham, 2015; Graham et al., 2013).

Other theories complicate the debate rather than choosing one side. Beal (2020) argues that monism and pluralism are often framed too narrowly as a question about the number of irreducible moral values, when moral cognition may depend on more basic interpretive frames. Zhang et al. (2023) likewise describe the debate as unresolved, with MFT and dyadic morality representing major positions. Moral judgement can also be organised through social-domain, face, relational, dyadic, and cultural accounts (Buchtel et al., 2015; Gray et al., 2012; Hui & Bond, 2009; Schein & Gray, 2018; Sunar et al., 2021). For this thesis, moral domains are therefore theoretically informed categories of violation, not proof that morality contains exactly five natural kinds. This position remains compatible with MFT while allowing the empirical studies to inform the monist–pluralist debate.

3.1.2 Moral Domains and Apology

Direct work linking moral domains to apology is still limited. As a measure-development and correlational study, Howell et al. (2011) developed the Proclivity to Apologise Measure and found that willingness to apologise correlated positively with care as a moral foundation, compassion, and other positive emotional dispositions. In a validation and behavioural-prediction study, Ohtsubo et al. (2015) validated a Japanese Proclivity to Apologise Measure and found that it predicted whether participants had apologised after a recent transgression. These studies show that apology can be studied alongside moral dispositions, but they

do not yet explain whether different moral domains require different apology signals.

Nonetheless, previous work shows that moral domains and apology are closely linked. Apologies matter because they are social signals after wrongdoing: B. Ho (2012) modelled apology as a signal that can restore relational beliefs. Drawing on Graham et al.'s (2013) theoretical account, Tangney et al.'s (2007) moral-emotion review, and empirical emotion findings from Landmann and Hess (2018) and Sunar et al. (2021), a harm violation may require concern for the victim's suffering; a fairness violation may require responsibility and compensation; a loyalty violation may require recommitment to the group; an authority violation may require deference or recognition of status; and a purity, sanctity, or civility violation may require shame, restraint, or reaffirmation of shared values. As the five moral domains are open to revision for different foundations, the requirements to resolve each moral domain are also not settled conclusions, but theory-led hypotheses derived from the reviewed domain, emotion, and apology literatures.

One or Multiple Moral Domains? Visual Apology Signals as a Test. The thesis uses visual apology signals to test the multi-domain versus uni-domain contrast because these signals are less tied to explicit verbal labels. A pluralist prediction, consistent with the theoretical claims of Graham et al. (2013) and Graham (2015), is that different moral violations should be repaired through somewhat different visual-signal configurations because they implicate different damaged values and expected emotions. Empirical findings that emotional responses vary by moral and relational context support this possibility (Landmann & Hess, 2018; Sunar et al., 2021). Visible concern and remorse may be especially important after harm, whereas deference, embarrassment, or restrained shame may matter more after authority, community, purity, or face-related violations (Hui & Bond, 2009; Sunar et al., 2021; Tangney et al., 2007). If receivers judge different configurations as more appropriate after different moral violations, that pattern would support psychologically differentiated repair requirements without presuming that the domains are wholly distinct.

An account based on harm instead proposes that moral judgements across domains share a common basis. If moral violations are understood through agency, harm, and patient suffering (Gray et al., 2012; Schein & Gray, 2018), the same core visual apology pattern may operate across domains—for example, visible remorse and acknowledgement that the victim or moral patient was harmed. Apparent domain differences could then reflect severity, weirdness, cultural familiarity, relational context, or perceived harm rather than distinct moral foundations (Gray & Keeney, 2015; Schein & Gray, 2018). Similarities and differences across moral domains can therefore indicate whether visual apology repair is organised primarily by a common social-repair architecture or by domain-sensitive expectations, a contrast motivated by Gray et al. (2012) and Graham (2015).

3.2 The Current Study

This study asks whether Western receivers organise visual apology signals primarily through a shared response-level architecture or also tune those configurations to the moral domain of the transgression. Social-signal theory treats observable behaviour as evidence about relationships, attitudes, emotions, and interaction (Poggi & D'Errico, 2011; Vinciarelli et al., 2009); in apology contexts, less controllable or embodied cues—including blushing, facial redness, facial movements, and bodily remorse—can provide additional sincerity-relevant evidence under uncertainty (Castelfranchi & Poggi, 1990; Hornsey et al., 2020; Ohtsubo & Watanabe, 2009; Ohtsubo et al., 2018; Thorstenson et al., 2020). Using prototype-free participant-level models, the chapter first identifies configurations of facial movements and gaze direction

associated with the Not-sorry, Mild-sorry, and Strong-sorry response categories, then tests whether complete configurations and particular signal contributions carry moral-domain information. The scientific contrast is therefore between a common visual repair pattern and selective domain-sensitive tuning, not between five presumed moral-domain prototypes.

3.3 Methods

The method extended Chapter 2 from identifying apology-related signals to asking whether their organisation varied across moral domains. The same participant-level approach achieved apparent apology signals per participant and hence preserved differences between receivers, while five transgression scenarios provided a structured comparison of context. The analyses first described recurring signal patterns and then tested whether the facial signal configurations distinguished the domains. To assess each signal's contribution, we removed it and refitted the classifier with the other signals available. In parallel, MI analysis was conducted to locate facial signals that were significantly associated with a given domain, provided a complementary way to locate those facial signal differences.

3.3.1 Participants

The recruitment and screening procedure followed Chapter 2, except that each participant was assigned to one of five moral-domain transgressions. Seventy-seven participants were recruited for the Western receiver group (WC). Seventy-six were included in the modelled dataset ($M_{\text{age}} = 20.70$, $SD = 3.17$; 64.47% women), and all 76 met the model-validation criterion, contributing 228 response-specific models. The domain-level sample, model, and threshold counts are reported in Appendix A. The matched East Asian sample is reported in Chapter 4.

3.3.2 Procedure

The procedure of Chapter 3 was the same as Chapter 2. Participants viewed dynamically generated face videos after reading a transgression scenario and judged how sorry the face appeared. As in Chapter 2, each participant completed 2,400 trials. These ratings were used to estimate a separate signal model for each participant and response category.

3.3.3 Stimuli

The facial stimulus-generation procedure followed Chapter 2. The analysis included 47 binary predictors: 42 facial action units (AUs) and five gaze indicators representing no gaze movement, looking up, looking down, looking left, and looking right. The generated videos contained no head movement, spoken apology, or audio. AU temporal parameters and blushing information were available from stimulus generation but were not included in the moral-domain analyses. Facial identity was a rendering characteristic rather than a signal predictor, and the face-identity set was matched to the receiver sample.

Chapter 2 used three transgressions, whereas Chapter 3 used five scenarios corresponding to Fairness, Care/harm, Authority, Loyalty, and Sanctity/purity. The Fairness scenario retained the earlier marathon-shortcut context and reused the earlier WC response observations. The other scenarios were adapted from the Moral Foundations Vignettes developed by Clifford et al. (2015). Candidate scenarios were screened for cross-cultural comprehensibility, a direct offender–victim relation, approximately equal and otherwise unspecified social status, and the absence of strongly gendered offender roles. Equal status was held constant because apology practices can vary with partner status (Barnlund & Yoshioka, 1990), while social power

was not a target factor in this experiment. The final wording placed the participant in the affected person's position rather than that of a bystander. Table 3.1 reports the resulting scenario set.

Table 3.1

Mapping Between the Five Moral Domains and the Adapted Transgression Scenarios.

Moral domain	Original vignette	Adapted vignette
Fairness	You see a runner taking a shortcut on the course during the marathon in order to win.	He/she took a shortcut on the course in a marathon you also took part in.
Care/harm	You see a boy placing a thumbtack sticking up on the chair of another student.	He/she placed a thumbtack sticking up on a chair you were about to sit on.
Authority	You see a staff member talking loudly and interrupting the mayor's speech to the public.	He/she talked loudly and interrupted you while you were speaking at the Mayor's event.
Loyalty	You see an employee joking with competitors about how bad his company did last year.	He/she is an employee from the company where you work. He/she joked with competitors about how bad your company did last year.
Sanctity/purity	You see a man searching through the trash to find women's discarded underwear.	He/she was searching through your rubbish to find your discarded underwear.

3.4 Planned Data Analyses

3.4.1 Overview and Analysis Hierarchy

The analysis hierarchy separates model construction and validation (Stage 1), descriptive prevalence summaries (Stage 2), primary classification and ablation (Stage 3), primary signal-wise MI/PMI (Stage 4), and cross-method integration (Stage 5). Primary analyses address the central research questions:

1. construct and validate a response-specific signal model for each participant;
2. describe which signals recurred across receivers, response categories and moral domains after population-prevalence thresholding;
3. test whether complete apology configurations distinguished the moral domains, then remove signals to locate their contributions to correct identification or confusion (ablation analysis);
4. examine whether each signal, considered separately, was associated with a moral domain within each response category (MI analysis); and
5. integrate these results while preserving the distinction between configuration-level, conditional and marginal evidence.

Moral-domain classification and ablation therefore used only validated, population-thresholded mild- and strong-sorry models. MI/PMI used the same validated and thresholded representation but was estimated separately for not sorry, mild sorry, and strong sorry.

MATLAB scripts implemented MI/PMI, resampling and Bayesian prevalence thresholding. Response-model validation, classification and ablation used classifiers from the Statistics and Machine Learning Toolbox. Descriptive summaries and figures were prepared in MATLAB and Python, with GFG (an in-house tool) used for the facemaps.

3.4.2 Stage 1: Model Construction and Population-Supported Inputs

Participant-specific ratings were rebinned into not-sorry, mild-sorry and strong-sorry categories, each represented by a 47-signal model. Response-model validation used 20 repeated participant-grouped 80/20 holdouts to check whether the three categories could be distinguished in models from receivers held out from classifier fitting. All response models from one receiver stayed together. Mean error below 30% was an operational screening rule of thumb. When screening required an exclusion, candidate removals were selected adaptively using the validation scores. No additional outer test independently evaluated this adaptive screening. Bayesian population-prevalence thresholds ($\alpha = .05$, $\beta = 1$, with correction across the three response categories) then retained signals recurring across receivers within each moral domain and response category (Ince et al., 2021, i.e., the signals being kept should be representative). The 228 validated, thresholded models supplied the descriptive prevalence and MI/PMI analyses; the 152 mild- and strong-sorry models supplied classification and ablation.

3.4.3 Stage 2: Population-Level Signal Description

Descriptive Prevalence Summaries. Prevalence summaries asked which AU and gaze signals recurred across receivers and whether their frequency changed with response category or moral domain. All AUs and gaze directions tied for the highest prevalence within a panel were retained. Upper-quartile rankings summarised the more frequent signals, while mild–strong contrasts were described by raw prevalence differences and Cohen's *h*. Cohen's *h* summarised the size of the difference between two prevalence proportions; zero indicated equal prevalence and larger absolute values indicated larger differences. These descriptive summaries characterised the composition of the retained signal sets. A descriptive domain-sensitivity flag required an upper-quartile frequency in one domain, absence or a lower-quartile frequency in another, and a range of at least 30 percentage points.

3.4.4 Stage 3: Primary Classification and Ablation

Moral-Domain Classification. The moral-domain classifier asked whether complete apology configurations contained information about the transgression context. It distinguished Fairness, Care/harm, Authority, Loyalty and Sanctity/purity using the mild- and strong-sorry models. Not-sorry models were excluded because this comparison concerned the organisation of judged apology configurations. Predictors that were zero across all retained apology rows were removed before fitting.

Classification used repeated-holdout cross-validation with 50 participant-level stratified 80/20 splits, keeping each participant's mild- and strong-sorry models together. Participants could recur across repetitions. Population thresholds had been fitted to the complete retained sample before these splits, so the evaluation described discrimination among the constructed representations rather than independent predic-

tion by the full modelling procedure. Balanced accuracy summarised correct identification across the five domains; the confusion matrix showed which domains were identified or confused. Rows of the confusion matrix were normalised to 100%. With the true-class counts fixed, row and whole-matrix normalisation gave the same cell-wise significance results.

A bagged-tree classifier used uniform class priors, 50 learning cycles and at most 20 splits per tree. An omnibus test compared balanced accuracy with 1,000 participant-label permutations, keeping each receiver's two apology rows together and using plus-one Monte Carlo values.

The 25 cell-wise follow-ups used Bonferroni correction to limit the chance of one or more false positives across the full confusion matrix. They located reliable correct identifications and directed confusions. A significant diagonal cell concerned correct identification of one domain; a significant off-diagonal cell concerned one observed-to-predicted domain pair.

Ablation Analysis. Ablation assessed each signal's conditional contribution to held-out classification. Each signal was removed and the classifier was refitted on the same participant splits. Changes in the confusion matrix identified signals supporting correct identification or contributing to confusion. The removal-and-refitting comparison followed the leave-one-covariate-out rationale (Lei et al., 2018).

A signal could help identify one domain, distinguish one directed pair, or make the classifier confuse that pair. Full-minus-ablated differences were tested two-sided with 1,000 paired sign flips on the same participant splits. Bonferroni correction covered the 25 cells for each predictor separately; there was no additional correction across predictors.

3.4.5 Stage 4: Primary Signal-Specific Associations

MI/PMI Signal-Specificity Analysis. MI/PMI asked a different question: whether the presence or absence of one signal was associated with moral domain when that signal was considered separately. The analysis was run within each response category, using one row per participant. MI identified a signal–domain dependency; PMI located the domain and signal state contributing to it.

Positive associations indicated overrepresentation of signal presence, whereas negative associations indicated overrepresentation of signal absence. Tests used 1,000 participant-label permutations and each signal's own 95th-percentile threshold; PMI followed the MI gate. No correction across all 47 signals was applied, so these primary signal-wise analyses identify candidate associations rather than a jointly corrected set of moral-domain markers.

3.4.6 Stage 5: Cross-Method Integration

Agreement across methods was evaluated only where their inputs and roles were comparable. Classification addressed complete configurations, ablation addressed a signal's contribution with other predictors available, and MI/PMI addressed the signal's separate association with a domain. Convergence could strengthen the description of a signal's relevance.

For illustration only, suppose the correct-identification rates for the five domains were .40, .50, .60, .70 and .80. Balanced accuracy would be $(.40+.50+.60+.70+.80)/5 = .60$, giving each domain equal weight. If one diagonal cell fell from .60 to .55 after removing a signal on the same splits, the full-minus-ablated difference would be .05, a positive contribution to identifying that domain. If signal presence and a domain co-occurred with probability .20 when independence predicted .10, PMI would be $\log_2(.20/.10) = 1$ bit, indicating overrepresentation of signal presence.

3.5 Results

3.5.1 Stages 1–2: Validated Inputs and Population-Level Signal Description

All 76 WC participants entering modelling passed the validation criterion and were retained, producing 228 response-specific models. Mild sorry was the least readily distinguished response category, whereas not sorry was the most readily distinguished. Appendix A reports the response-validation outcomes. The main results therefore begin with the retained AU and gaze direction patterns before testing moral-domain information in the mild- and strong-sorry models.

Population-Level Signal Prevalence. Population-prevalence thresholding removed low-prevalence effects in every moral domain while retaining all five gaze indicators within each domain-level union. The analyses below focus on the retained response- and domain-level signal patterns.

Figure 3.1 summarises the retained AUs and displays one gaze state per panel. The most prevalent gaze state is shown; ties are drawn with neutral-looking eyes.

Action Units. The population-supported AU panels showed a clear response structure (Figure 3.2). AU43 led the Fairness, Care/harm, and Loyalty not-sorry panels, whereas AU9 led Authority and Sanctity/purity. Mild-sorry peaks varied across domains, while strong-sorry panels were more concentrated: Fairness peaked at AU13 and the other four domains peaked at AU20.

Table 3.2 pools retained AU frequencies across the five moral domains within each response category and lists the upper quartile of nonzero pooled AUs. The upper quartile contained three of 11 eligible not-sorry AUs, eight of 30 mild-sorry AUs, and six of 23 strong-sorry AUs. The selected not-sorry AUs had a mean pooled prevalence of 74.6%, compared with 58.8% for strong-sorry AUs and 43.1% for mild-sorry AUs. Not-sorry models were led by AU43, AU9, and AU10Open; mild-sorry models had a broader and less concentrated set; and strong-sorry models were centred on AU20, AU13, AU20R, AU24, AU14, and AU14R.

Across individual moral-domain panels, not-sorry peaks were strongest for AU43 and AU9, strong sorry peaks were strongest for AU20 and AU13, and mild-sorry peaks were less uniform.

The pooled and panel-level summaries also show that some AUs were domain-sensitive rather than uniformly frequent. For this descriptive check, an AU was treated as domain-sensitive when it appeared in the within-panel upper quartile for at least one moral domain, was absent or in the within-panel lower quartile for another, and showed a cross-domain prevalence range of at least 30 percentage points. Under this rule, mild-sorry AU20L and AU14 varied strongly across domains, with ranges above 60 percentage points, whereas strong-sorry AU24 showed a smaller but still substantial domain range above 40 percentage points. Thus, while Table 3.2 identifies response-level regularities, Figure 3.2 shows that some AUs were more tied to particular moral domains.

Figure 3.2 shows that mild- and strong-sorry models were not scaled versions of the same AU pattern. Across domains, mild sorry was especially separated by AU9, AU43, and AU10-related signals, which were also prominent in the not-sorry profile. Strong sorry was more clearly separated by AU4, AU20-related signals, AU24, AU16Open, and AU27i.

Gaze Indicators. The population-supported gaze panels showed a simpler response structure than the AU results (Figure 3.3). Every not-sorry panel peaked at looking up, mild-sorry peaks were distributed across upward, leftward, and rightward gaze, and strong-sorry panels mainly peaked at looking down.

Figure 3.1

WC Receiver-Group Population-Level Signal Prevalence by Response Category and Moral Domain.



Note. Rows show not sorry, mild sorry, and strong sorry; columns show Fairness, Care/harm, Authority, Loyalty, and Sanctity/purity. Colour highlights facial changes from AUs weighted by the proportion of participants whose models contained each AU. The bar shows the normalised local facial movement scale (0–100%); colours fade into the facial texture at lower values.

Figure 3.2

WC Receiver-Group AU Prevalence After Bayesian Thresholding, Showing the Retained AU Effects.

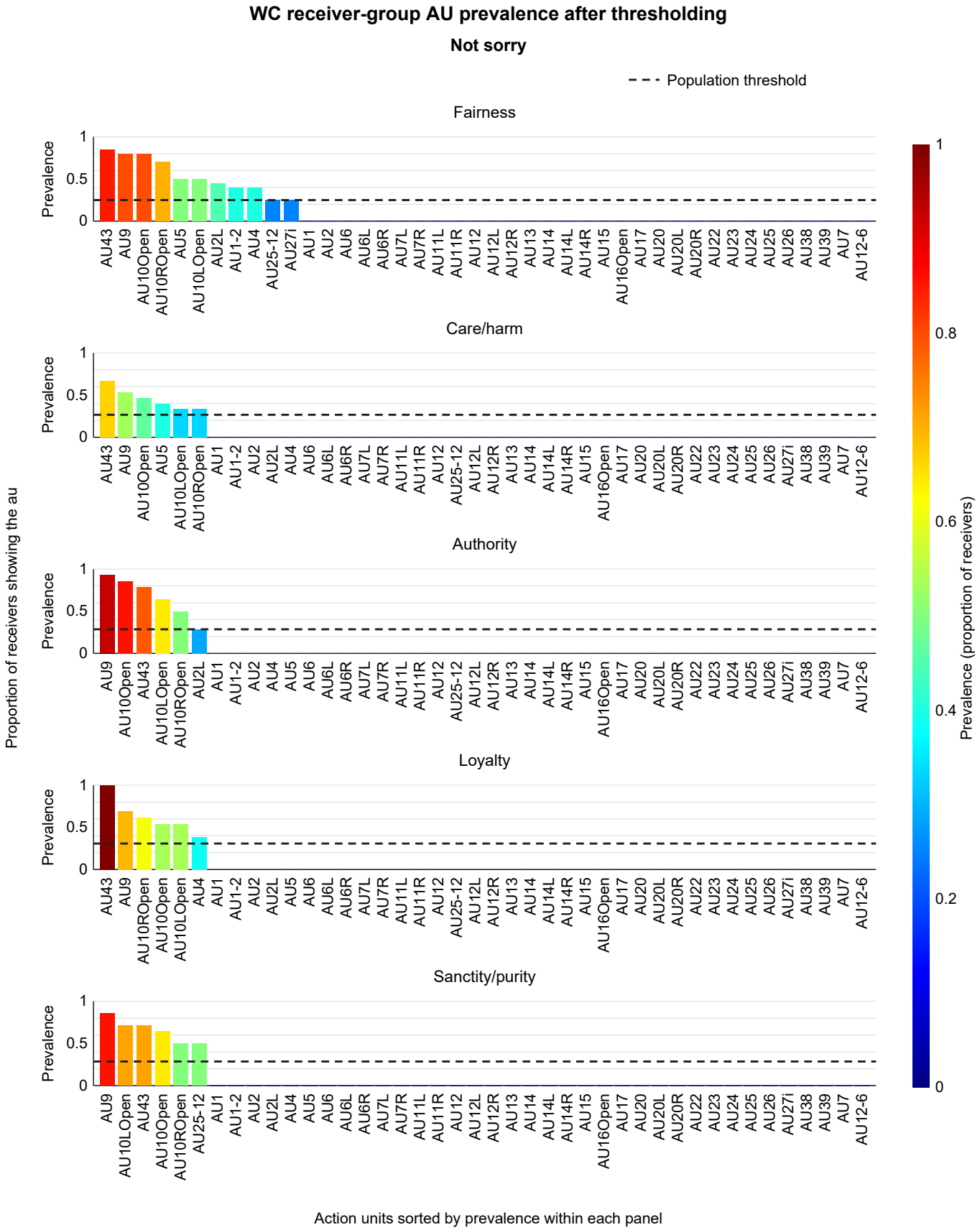


Figure 3.2 (continued)

WC Receiver-Group AU Prevalence After Bayesian Thresholding, Showing the Retained AU Effects.

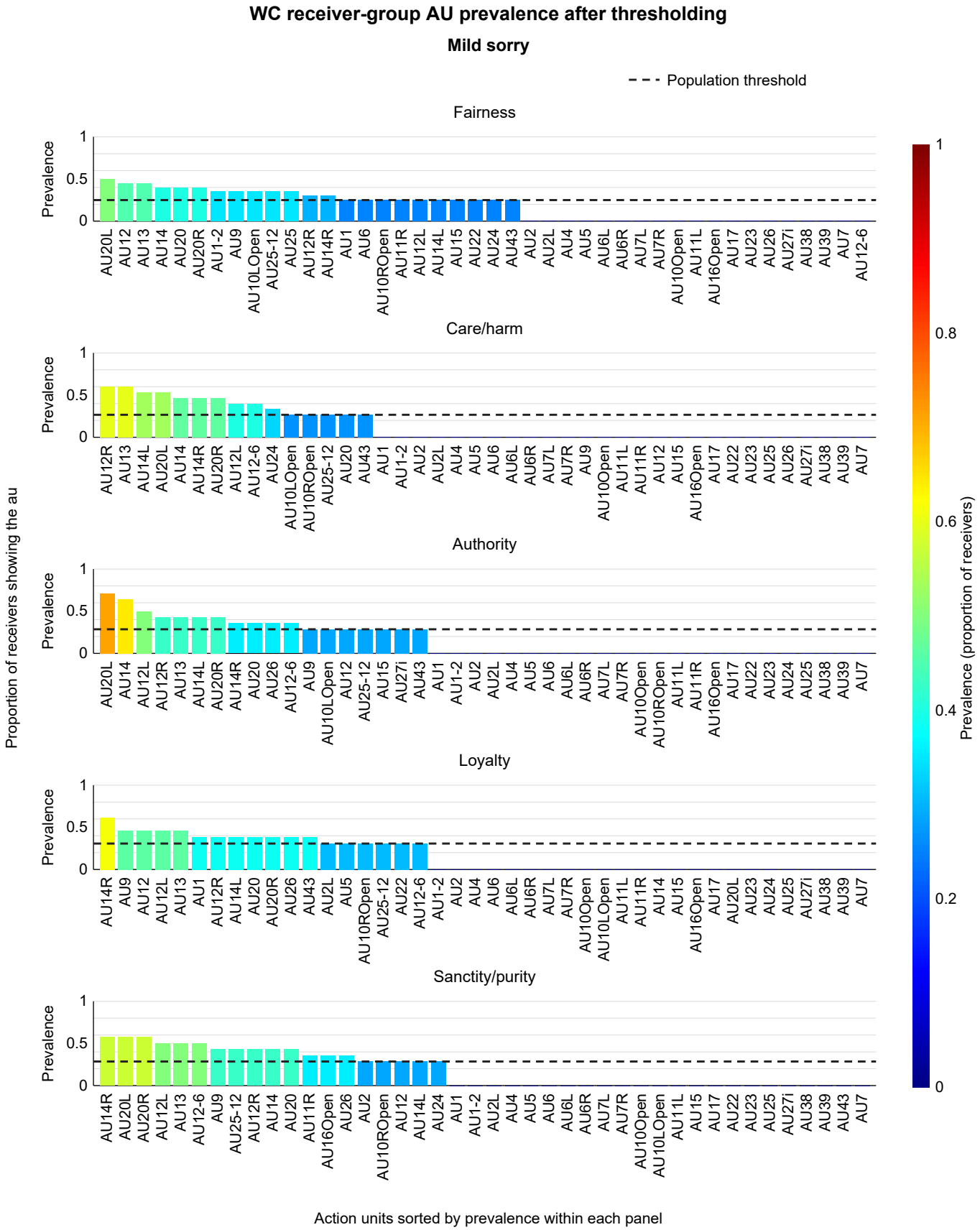


Figure 3.2 (continued)

WC Receiver-Group AU Prevalence After Bayesian Thresholding, Showing the Retained AU Effects.

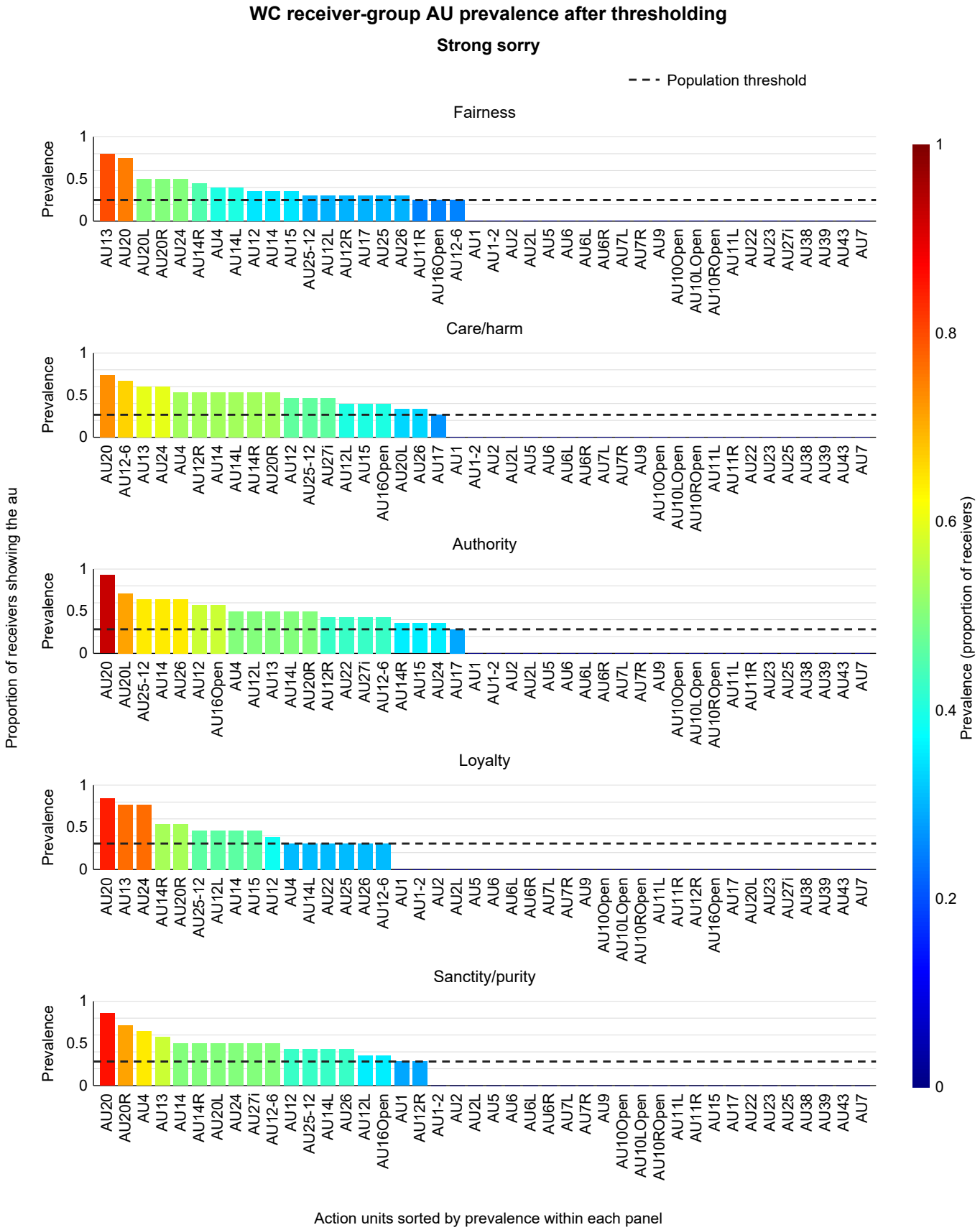


Table 3.2

Top-Quartile AUs Among WC Receivers Across Moral Domains After Bayesian Thresholding.

Response category	AU	Frequency
Not sorry	AU43	61/76 (80.3%)
Not sorry	AU9	58/76 (76.3%)
Not sorry	AU10Open	51/76 (67.1%)
Mild sorry	AU13	37/76 (48.7%)
Mild sorry	AU20L	36/76 (47.4%)
Mild sorry	AU20R	34/76 (44.7%)
Mild sorry	AU14R	34/76 (44.7%)
Mild sorry	AU12R	32/76 (42.1%)
Mild sorry	AU12L	31/76 (40.8%)
Mild sorry	AU14	30/76 (39.5%)
Mild sorry	AU20	28/76 (36.8%)
Strong sorry	AU20	62/76 (81.6%)
Strong sorry	AU13	50/76 (65.8%)
Strong sorry	AU20R	42/76 (55.3%)
Strong sorry	AU24	41/76 (53.9%)
Strong sorry	AU14	37/76 (48.7%)
Strong sorry	AU14R	36/76 (47.4%)

Note. Frequencies are pooled across the five moral domains within each response category.

Figure 3.1 shows only the selected gaze state, so other prevalent directions may not be visible.

Table 3.3 applies the pooled upper-quartile rule to the gaze results. Two nonzero pooled directions were eligible for not sorry and four for each sorry response, so one dominant direction was selected per response category. Looking up led both not-sorry and mild-sorry models, although mild-sorry panel peaks varied more across domains, while looking down led strong-sorry models.

Table 3.3

Top-Quartile Gaze Indicators Among WC Receivers Across Moral Domains After Bayesian Thresholding.

Response category	Gaze direction	Frequency
Not sorry	Looking up	62/76 (81.6%)
Mild sorry	Looking up	52/76 (68.4%)
Strong sorry	Looking down	46/76 (60.5%)

Note. Frequencies are pooled across the five moral domains within each response category.

Across individual moral-domain panels, looking up dominated not sorry, mild-sorry peaks were divided among leftward, rightward, and upward gaze, and strong-sorry peaks were mainly downward.

Compared with AUs, the gaze profiles were less domain-sensitive under the same upper-quartile/lower-quartile and 30-percentage-point range rule. The clearest domain-sensitive gaze was mild-sorry looking left,

with a cross-domain range of 33.9 percentage points. Mild-sorry looking right was also frequent, appearing in 51/76 receiver-domain observations (67.1%) and peaking in Authority and Loyalty. Overall, gaze showed a clearer response-level than domain-level structure: upward gaze marked not sorry and remained important for mild sorry, whereas downward gaze marked strong sorry.

Figure 3.3 shows that mild- and strong-sorry gaze profiles were not scaled versions of the same pattern. Looking up was the strongest mild-sorry-separated direction in every domain, whereas looking down was the strongest strong-sorry-separated direction in every domain.

Overall, the AU channel contained a larger and more differentiated repertoire of signals, whereas gaze showed greater consistency across moral domains and a clearer relationship with response category. The upper-quartile AUs were most prevalent for not sorry, with a mean pooled prevalence of 74.6%, followed by strong sorry at 58.8% and mild sorry at 43.1%. Not-sorry responses were characterized primarily by AU43, AU9, and AU10Open, together with a strong upward-gaze tendency, whereas strong-sorry responses showed a contrasting profile centred on AU20, AU13, and related AU14 and AU20 variants, alongside downward gaze. Mild sorry was more heterogeneous across moral domains and occupied an intermediate position between these profiles: its most common AUs included AU13, AU20 and AU14 variants, as well as the smile-related AU12R and AU12L signals. More broadly, AU12-related signals ranged from 28.9% to 42.1% in mild sorry and from 31.6% to 44.7% in strong sorry, suggesting that smile-related activity was a meaningful component of both apology conditions and was somewhat more pronounced in strong apologies. By comparison, only AU25-12 remained nonzero in the thresholded not-sorry models, at 15.8%, indicating that this smile-related component was more closely associated with apology than with its absence. The overlap between mild sorry and not sorry was stronger in the gaze channel: looking up remained common in mild sorry at 68.4%, compared with 81.6% in not sorry. However, mild sorry was not defined by upward gaze alone, because looking right and looking left occurred at similarly high rates (67.1% and 65.8%, respectively). Thus, mild sorry combined a broadly apology-related AU pattern, including a substantial smile-related component, with a partially not-sorry-like upward-gaze pattern, whereas strong sorry produced the clearest coordinated shift toward a distinct AU configuration and downward gaze.

3.5.2 Stage 3: Primary Classification and Ablation

Moral-Domain Classification. The classifier asked whether complete population-thresholded mild- and strong-sorry configurations carried moral-domain information beyond the shuffled-label distribution.

As Table 3.4 shows, the classifier used 152 WC model rows, comprising 76 mild-sorry and 76 strong-sorry models. Removing global-zero predictors left 37 of the original 47 predictors: 32 AUs and five gaze indicators. The observed balanced accuracy was .4492. None of the 1,000 balanced accuracies reconstructed from the participant-label permutations was as large, giving a plus-one Monte Carlo result of $p = .001$. The complete thresholded apology vectors therefore distinguished moral domains above the permutation null distribution.

Figure 3.4 shows the complete row-normalised matrix. The diagonal values were .398 for Fairness, .283 for Care/harm, .460 for Authority, .815 for Loyalty, and .290 for Sanctity/purity. Loyalty-to-Loyalty classification was the only cell exceeding its Bonferroni-corrected permutation threshold. The directed errors show relative overlap rather than symmetric distance: Care/harm was classified as Authority at .287, approximately its correct-classification rate; Sanctity/purity was classified as Care/harm and Authority at

Figure 3.3

WC Receiver-Group Gaze Prevalence After Bayesian Thresholding, Showing the Retained Gaze-Direction Effects.

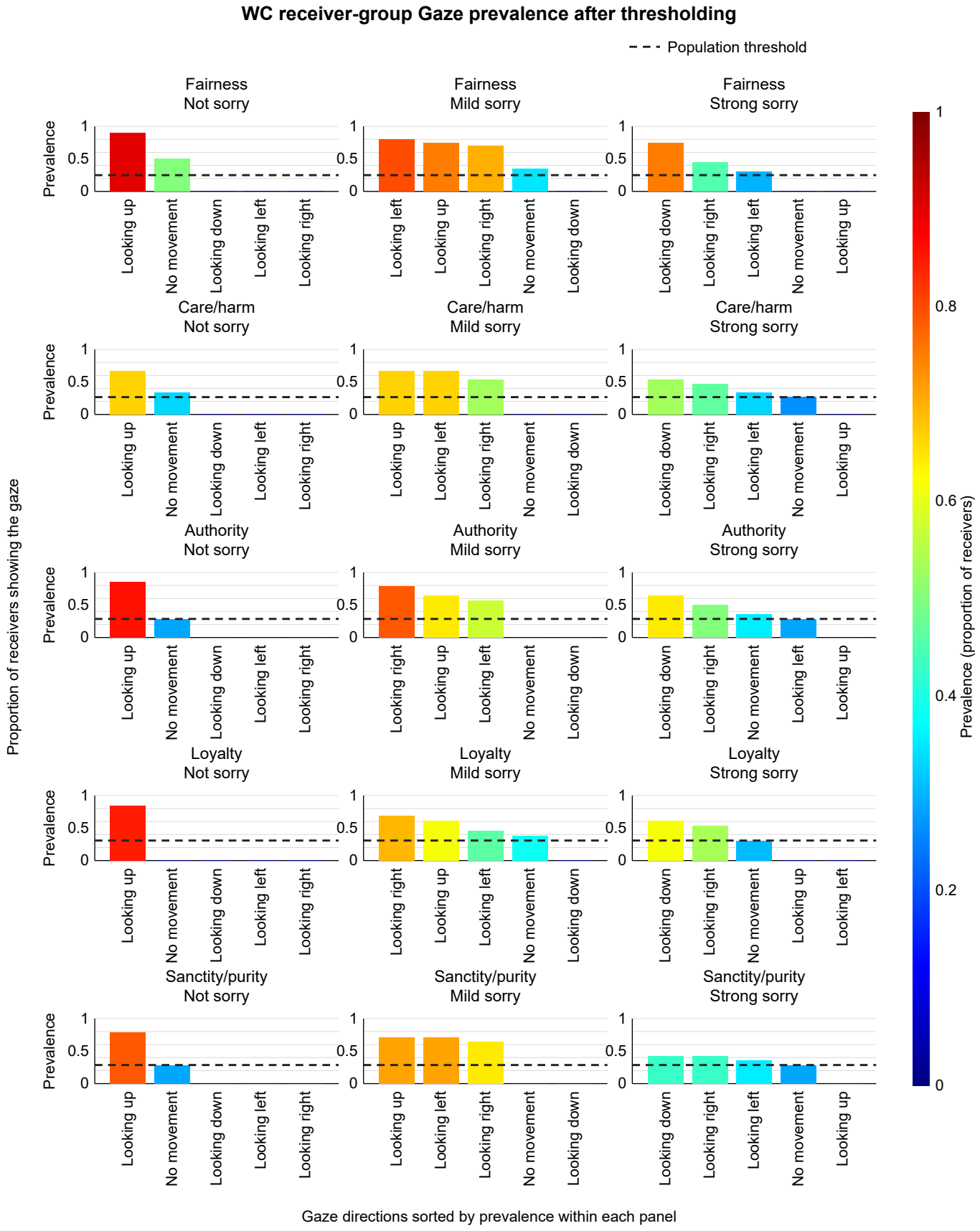


Table 3.4

Moral-Domain Classifier Summary for WC Receiver-Group Apology Models.

Receiver group	Model rows	Balanced accuracy	Predictors retained	Predictors removed	Significant diagonal domain
WC	152	.4492*	37	10	Loyalty (.815)

Note. *Omnibus plus-one permutation $p = .001$ from 1,000 participant-label permutations. The cell-wise follow-up retained Bonferroni correction across 25 confusion-matrix cells.

.215 in each of these two outgoing cells; and Fairness was most frequently assigned to Loyalty (.240). The significant omnibus result therefore reflected uneven domain information rather than five uniformly separable classes.

Ablation Analysis. Ablation asked which retained signals contributed to the classifier's decisions while the other signals remained available. A signal was ablation-active if removing it and refitting reliably changed at least one classifier cell. A class-specific identifier supported correct domain classification: removing it reduced correct classifications. A class-specific confuser showed the opposite pattern: removing it improved correct classifications. Pair-separation and pair-confusion roles instead concerned one directed misclassification between domains. Removing a pair-separation signal increased that error; removing a pair-confusion signal reduced it. These are conditional classifier contributions, not causal or universal domain markers.

Seventeen of the 37 classifier predictors had at least one significant raw ablation cell; the remaining 20 were globally non-used. The active set comprised 16 AUs and one gaze indicator. Reading each significant cell according to its sign and matrix location produced four diagonal identifiers, three diagonal confusers, five directed pair-separation effects, and 15 directed pair-confusion effects.

The four class-specific identifiers all occurred on the Loyalty diagonal: AU12R, AU16Open, AU20L, and AU27i. The class-specific confusers were AU6 for Fairness and AU14R and AU20L for Care/harm. Table 3.5 reports these diagonal roles and their same-domain signed-PMI context.

Figure 3.5 localises the ablation-active signals on the face while preserving the distinction between class-specific identifiers, class-specific confusers, directed pair separation, and directed pair confusion.

Table 3.5

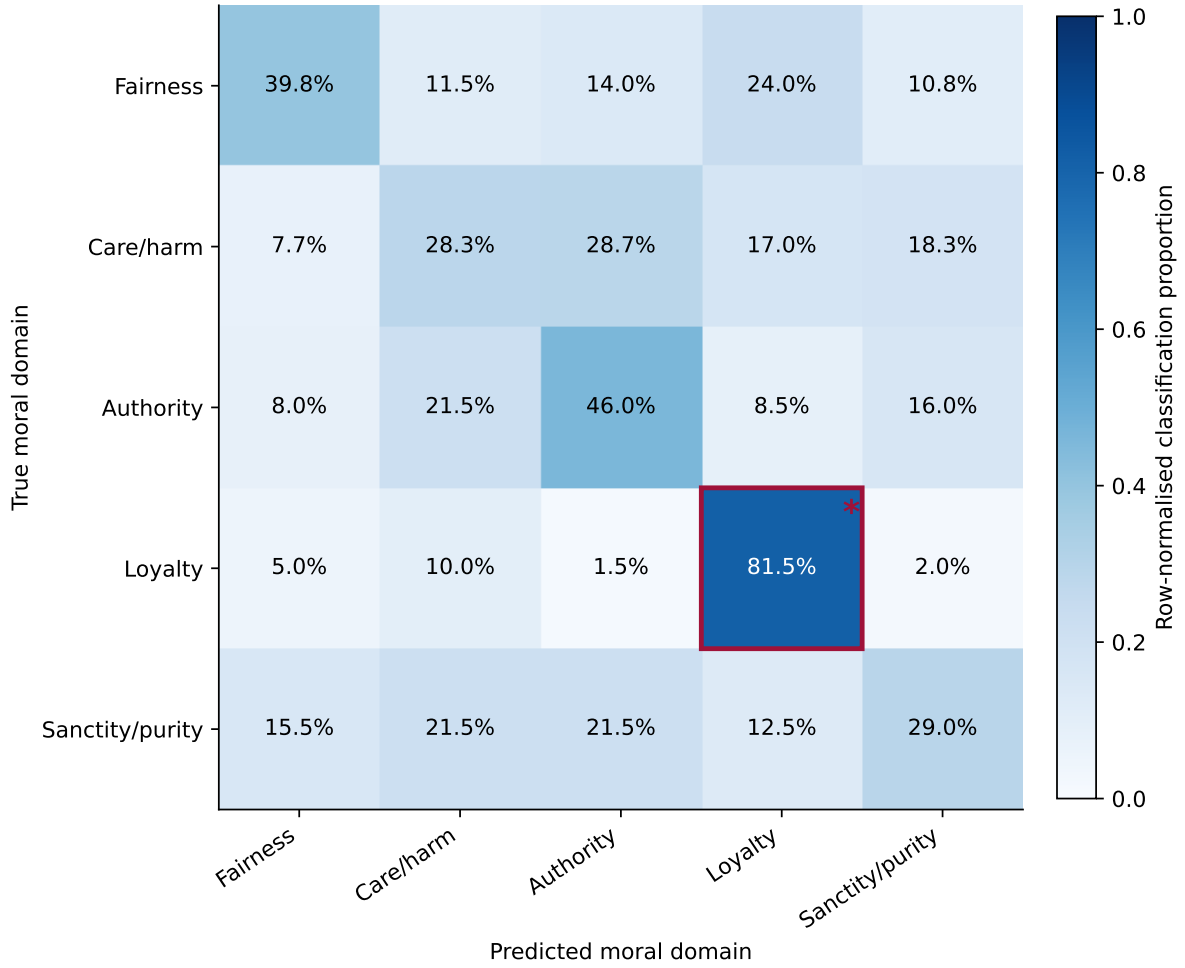
Diagonal Class-Specific Ablation Roles for WC Receiver-Group Moral-Domain Models.

Domain	Signal	Role	Positive PMI	Negative PMI
Fairness	AU6	Confuser	Mild sorry	None
Care/harm	AU14R	Confuser	None	None
Care/harm	AU20L	Confuser	None	None
Loyalty	AU12R	Identifier	None	Strong sorry
Loyalty	AU16Open	Identifier	None	Strong sorry
Loyalty	AU20L	Identifier	None	Mild and strong sorry
Loyalty	AU27i	Identifier	None	Strong sorry

Note. Diagonal roles concern correct classification of one domain. Removing an identifier reduced correct classification; removing a confuser improved it. Positive and negative PMI columns report aligned signed associations for the same signal and domain in the apology-response bins.

Figure 3.4

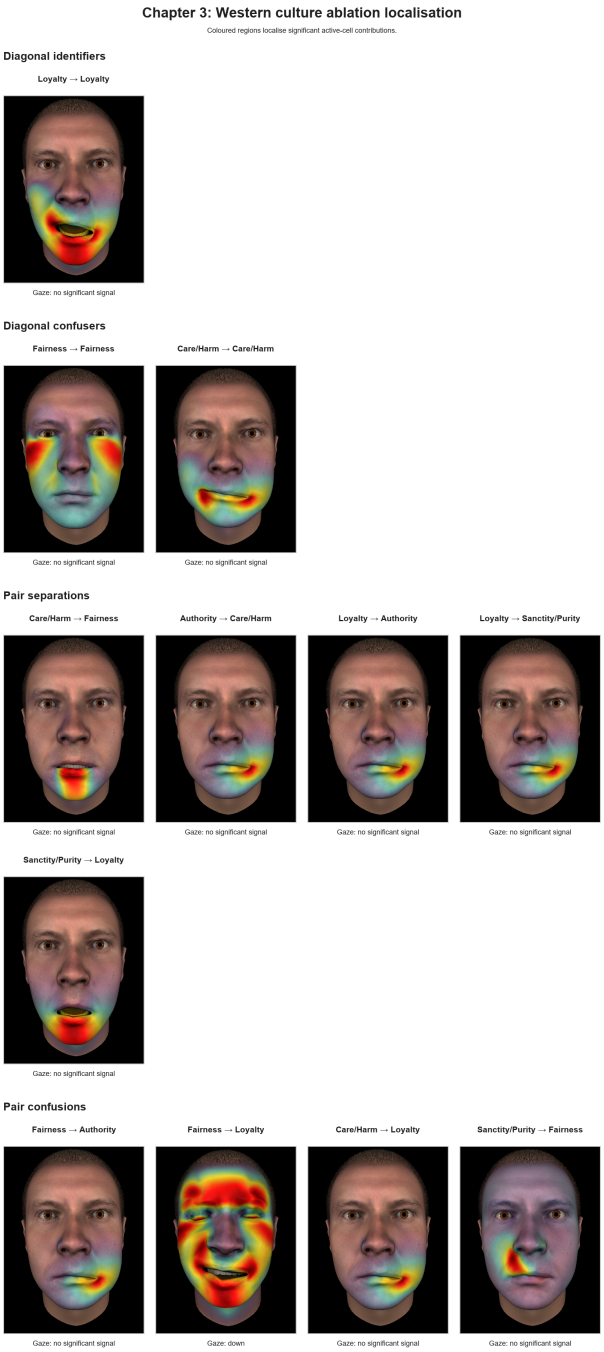
Row-Normalised Classification of WC Receiver-Group Mild- and Strong-Sorry Models Across the Five Moral Domains.



Note. The outlined Loyalty diagonal was the only cell exceeding its Bonferroni-corrected permutation threshold. Rows show the predicted classifications within each true moral domain (100% per row).

Figure 3.5

Facemap Localisation of Ablation-Active Signals in the WC Receiver-Group Moral-Domain Classifier.



Note. Active signals reliably changed classification after removal and refitting. Colours locate the selected AUs, each displayed at maximum rendering amplitude. Panels distinguish class-specific and directed pair roles.

The 20 off-diagonal effects were retained as directed pairwise contrasts. Twelve of the 15 pair-confusion effects occurred in the Fairness-to-Loyalty cell, showing that a broad shared signal set contributed specifically to that directed boundary. The five pair-separation effects were AU20L for Authority to Care/harm, Loyalty to Authority, and Loyalty to Sanctity/purity; AU25 for Care/harm to Fairness; and AU27i for the directed contrast from Sanctity/purity to Loyalty. Thus, Fairness to Loyalty contained the broadest set of conditional pair-confusion effects, whereas the smaller set of conditional pair-separation effects localised a limited number of more differentiated directed pair boundaries. Table 3.6 and Figures 3.6–3.7 keep the class-specific and pairwise results separate.

Table 3.6

Directed Off-Diagonal Ablation Roles for WC Receiver-Group Moral-Domain Models.

Signal	Observed → predicted cell	Pairwise role
AU1	Fairness → Loyalty	Confusion
AU2	Fairness → Loyalty	Confusion
AU4	Fairness → Loyalty	Confusion
AU6	Fairness → Loyalty	Confusion
AU10ROpen	Fairness → Loyalty	Confusion
AU11R	Sanctity/purity → Fairness	Confusion
AU12	Fairness → Loyalty	Confusion
AU12R	Fairness → Loyalty	Confusion
AU15	Fairness → Loyalty	Confusion
AU17	Fairness → Loyalty	Confusion
AU20L	Fairness → Authority	Confusion
AU20L	Care/harm → Loyalty	Confusion
AU20L	Authority → Care/harm	Separation
AU20L	Loyalty → Authority	Separation
AU20L	Loyalty → Sanctity/purity	Separation
AU25	Care/harm → Fairness	Separation
AU27i	Fairness → Loyalty	Confusion
AU27i	Sanctity/purity → Loyalty	Separation
AU43	Fairness → Loyalty	Confusion
Looking down	Fairness → Loyalty	Confusion

Note. Separation denotes a negative off-diagonal full-minus-ablated effect; confusion denotes a positive off-diagonal effect. Pairwise roles retain their observed-to-predicted direction and are not duplicated into endpoint-domain assignments.

3.5.3 Stage 4: Primary Signal-Specific Associations

MI/PMI Signal Specificity. MI/PMI then asked a different question: within each response category, in which domains was a signal's ON state overrepresented or underrepresented, with the latter corresponding to overrepresentation of its OFF state?

Across the 141 signal–response combinations formed by 47 predictors and three response bins, 67 (47.5%) were non-used, 35 (24.8%) were present but not significant, and 39 (27.7%) had at least one significant signed association with a moral domain. Among the 39 signed combinations, 31 had at least one significantly positive association, 24 had at least one significantly negative association, and 16 had both positive and negative associations across different domains. These combinations generated 74 significant signal–domain associations, divided evenly between 37 positive and 37 negative associations. Moral-

Figure 3.6

Diagonal Class-Specific Ablation Roles in the WC Receiver-Group Moral-Domain Classifier.

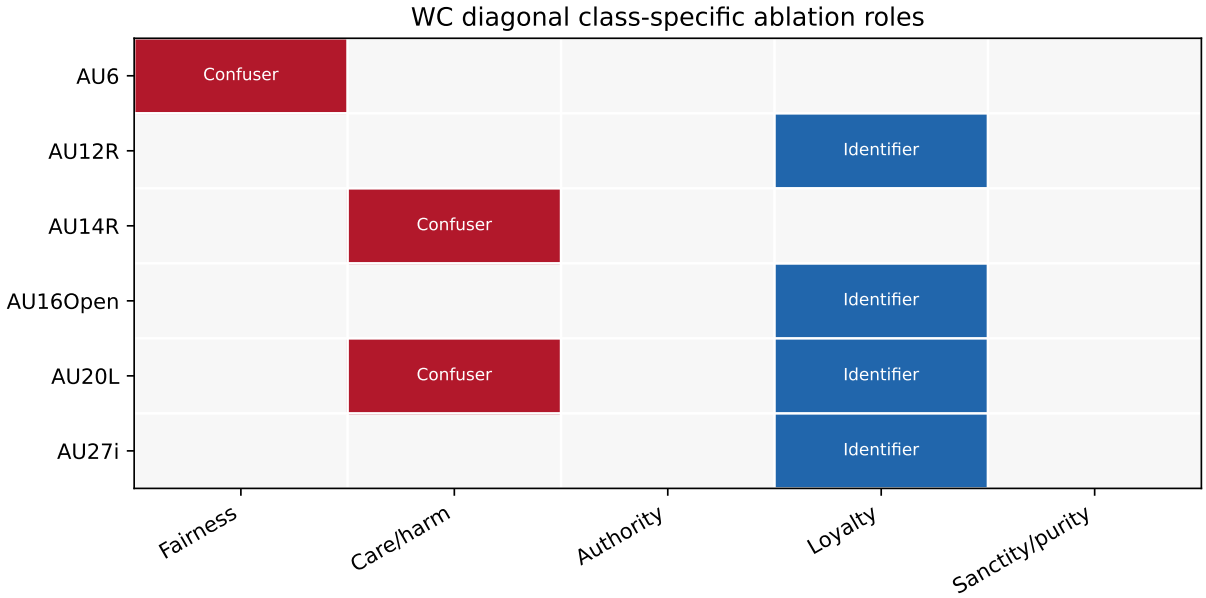
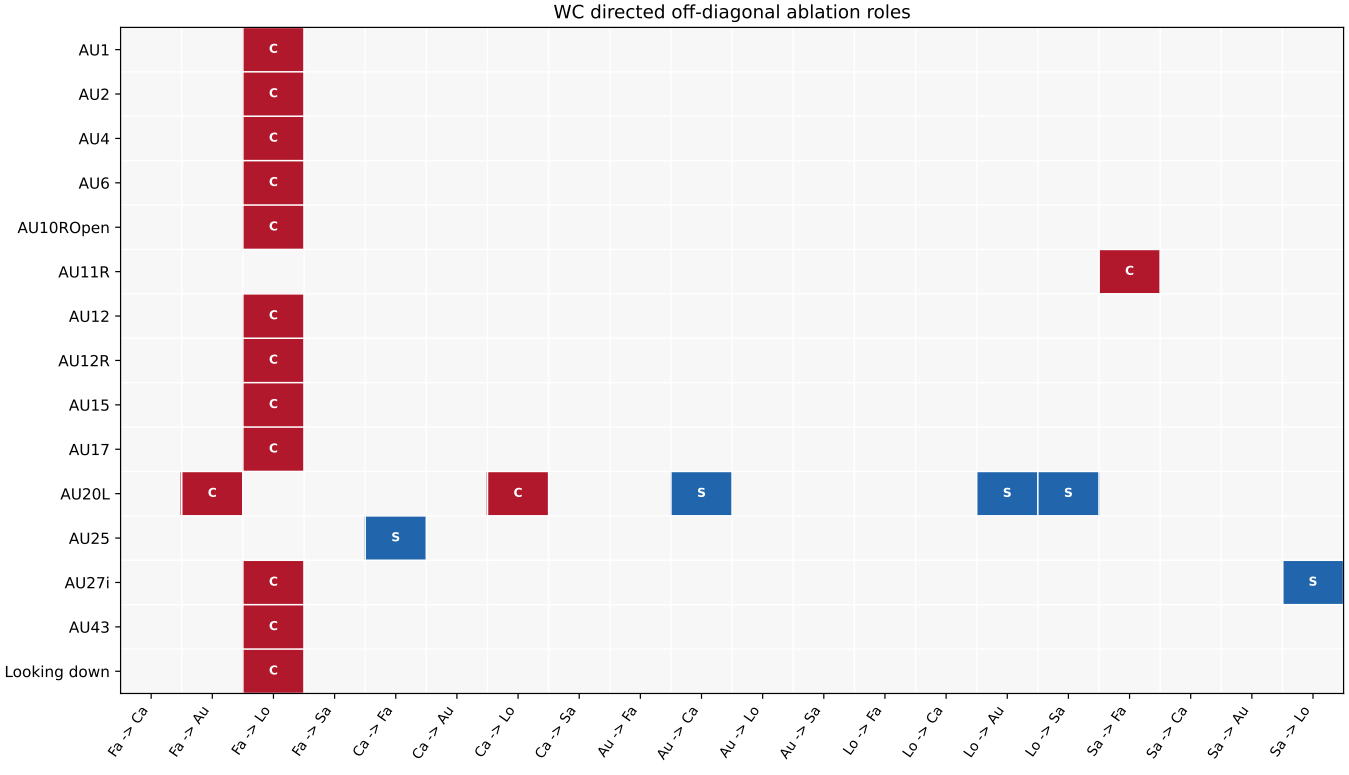


Figure 3.7

Directed Off-Diagonal Pair-Separation and Pair-Confusion Ablation Roles in the WC Receiver-Group Moral-Domain Classifier.



S = supports separation; C = contributes to confusion. Cells retain observed-to-predicted direction.

Fa = Fairness; Ca = Care/harm; Au = Authority; Lo = Loyalty; Sa = Sanctity/purity.

domain specificity therefore involved both overrepresentation and underrepresentation of signal presence. Table 3.7 reports the response-level composition.

Figure 3.8 localises the ON-state associations retained for the main chapter. The corresponding OFF state associations concern selective signal absence and are shown separately in Appendix A.

Table 3.7

Response-Level PMI Category Composition for WC Receiver-Group Signals.

Response	Non-used	Present/not significant	Any significant	Positive	Negative	Both directions	Signal–domain associations (+/−)
Not sorry	34	6	7	6	5	4	8/13
Mild sorry	13	13	21	17	10	6	18/13
Strong sorry	20	16	11	8	9	6	11/11
Total	67	35	39	31	24	16	37/37

Note. Within each response row, Non-used, Present/not significant, and Any significant are mutually exclusive and sum to 47 predictors. Positive and Negative count unique signals with at least one association in the corresponding direction and can overlap; Both directions is their intersection. Percentages reported in the text use 47 predictors within response rows and 141 signal–response combinations overall. Signal–domain associations count significant signal–domain cells, with positive counts before the slash and negative counts after it. A signal contributes one association for every moral domain in which it is significant, so this count can exceed the number of unique positive or negative signals. A negative association denotes significant OFF-state PMI and therefore underrepresentation of signal presence, not a numerically negative PMI value.

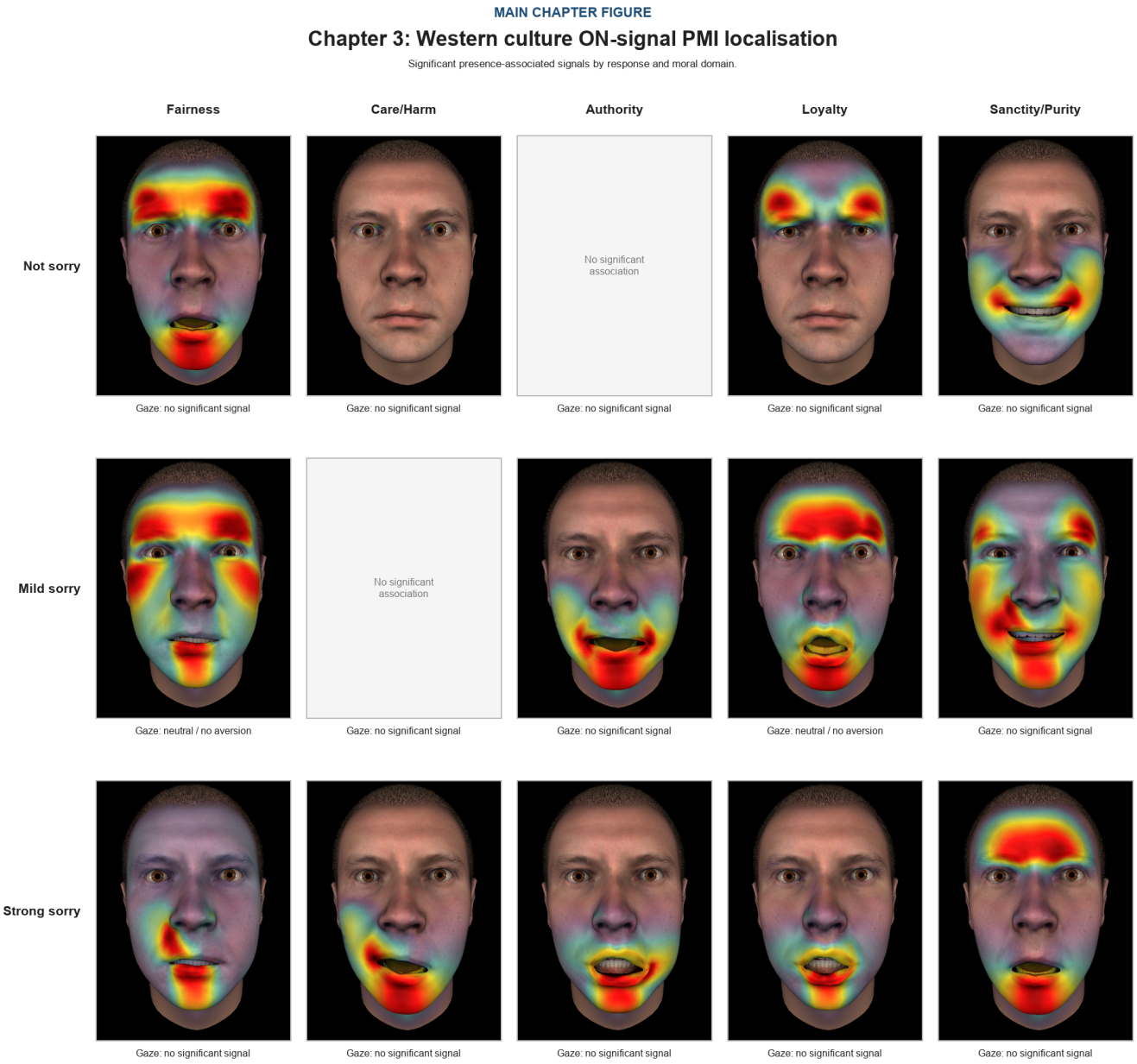
The signed repertoire was broadest for mild sorry. Not sorry contained seven signals with at least one signed association (14.9%), comprising six positive signals and five negative signals, four of which carried both directions across domains. These signals generated eight positive and 13 negative domain associations. Mild sorry contained 21 signed signals (44.7%): 17 positive, ten negative, and six carrying both directions; these generated 18 positive and 13 negative associations. Strong sorry contained 11 signed signals (23.4%): eight positive, nine negative, and six carrying both directions; these generated 11 associations in each direction. Mild sorry therefore showed the broadest moral-domain differentiation at the individual-signal level, whereas the not-sorry repertoire remained the most restricted despite containing more negative than positive domain associations. Table 3.8 resolves these categories by response and domain, and Table 3.9 lists every significant AU and gaze association.

Action Units. In not-sorry models, the eight positive and 13 negative associations formed a complementary pattern. Fairness contained five positive AU associations and no negative associations. By contrast, Care/harm, Authority, Loyalty, and Sanctity/purity each contained three or four negative associations. AU4 was positively associated with Fairness and Loyalty but negatively associated with Care/harm, Authority, and Sanctity/purity. AU5 was positively associated with Fairness and Care/harm but negatively associated with Authority, Loyalty, and Sanctity/purity. AU2L and AU25-12 showed related reversals across domains. The not-sorry differences therefore depended partly on the selective absence of otherwise shared signals.

Mild sorry contained 18 positive and 13 negative associations. Loyalty had the largest signed profile, with six positive and four negative associations, followed by Fairness with four associations in each direction. Authority and Sanctity/purity each had four positive and one negative association, whereas Care/harm had no positive associations and three negative associations. AU20L and AU14 were positively associated with Authority but negatively associated with Loyalty; AU26 was positively associated with Loyalty but negatively associated with Fairness and Care/harm; and AU2, AU27i, and AU12-6 also reversed direction across domains. Mild-sorry differentiation therefore involved both domain-specific enrichment and selec-

Figure 3.8

Facemap Localisation of Significant WC Receiver-Group ON-State PMI Associations by Response Category and Moral Domain.



Note. Coloured regions locate the selected AUs whose presence was overrepresented in the panel. Each AU is displayed at maximum rendering amplitude; blank panels contain no retained ON-state association.

Table 3.8

PMI-Derived Signal-Type Counts by Response and Moral Domain for WC Receiver-Group Models.

Receiver group	Response	Domain	Not-sig	Sig-pos	Sig-neg	Non-used
WC	Not sorry	Fairness	6	5	0	34
WC	Not sorry	Care/harm	6	1	3	34
WC	Not sorry	Authority	6	0	3	34
WC	Not sorry	Loyalty	6	1	4	34
WC	Not sorry	Sanctity/purity	6	1	3	34
WC	Mild sorry	Fairness	13	4	4	13
WC	Mild sorry	Care/harm	13	0	3	13
WC	Mild sorry	Authority	12	4	1	13
WC	Mild sorry	Loyalty	13	6	4	13
WC	Mild sorry	Sanctity/purity	12	4	1	13
WC	Strong sorry	Fairness	16	2	4	20
WC	Strong sorry	Care/harm	16	2	0	20
WC	Strong sorry	Authority	16	3	0	20
WC	Strong sorry	Loyalty	15	2	5	20
WC	Strong sorry	Sanctity/purity	16	2	2	20

Note. Not-sig denotes a signal that was present but had no retained signed association. Sig-pos and Sig-neg denote signal–domain cells that passed the signal-level MI gate and the ON-state or OFF-state PMI threshold, respectively. Sig-neg therefore indicates underrepresentation of signal presence. Non-used is defined at response level and is repeated across the five domain rows. Because a signal can be positive for one domain and negative for another, signed domain counts do not form a mutually exclusive partition of the 47 predictors.

tive underrepresentation within the broadest response-level repertoire.

Strong sorry contained 11 positive and 11 negative associations. Positive associations occurred in all five domains, whereas negative associations were concentrated in Fairness, Loyalty, and Sanctity/purity. Loyalty had two positive and five negative associations, the largest signed profile in this response bin, while Authority had three positive and no negative associations. AU16Open and AU20L were positively associated with Authority but negatively associated with Loyalty. AU27i was positively associated with Care/harm and Sanctity/purity but negatively associated with Fairness and Loyalty, while AU22 was positively associated with Authority and Loyalty but negatively associated with Fairness. These opposing associations sharpened particular domain contrasts without producing five mutually exclusive signal sets.

Gaze Indicators. No movement was the only gaze state with a significant signed PMI association. It was positively associated with mild-sorry Fairness and Loyalty, negatively associated with not-sorry Loyalty, and negatively associated with strong-sorry Fairness. Its direction therefore depended on both response category and moral domain. Looking up remained present but non-significant across all five not-sorry domains. Looking up, Looking left, and Looking right were present but non-significant across the mild-sorry domains, while downward and lateral gaze were broadly present but non-significant in strong sorry. The signed PMI results therefore confirm that gaze carried less moral-domain specificity than the AU repertoire, although No movement contributed to selected contrasts. Figure 3.9 shows the complete signed PMI membership by response and domain.

Table 3.9

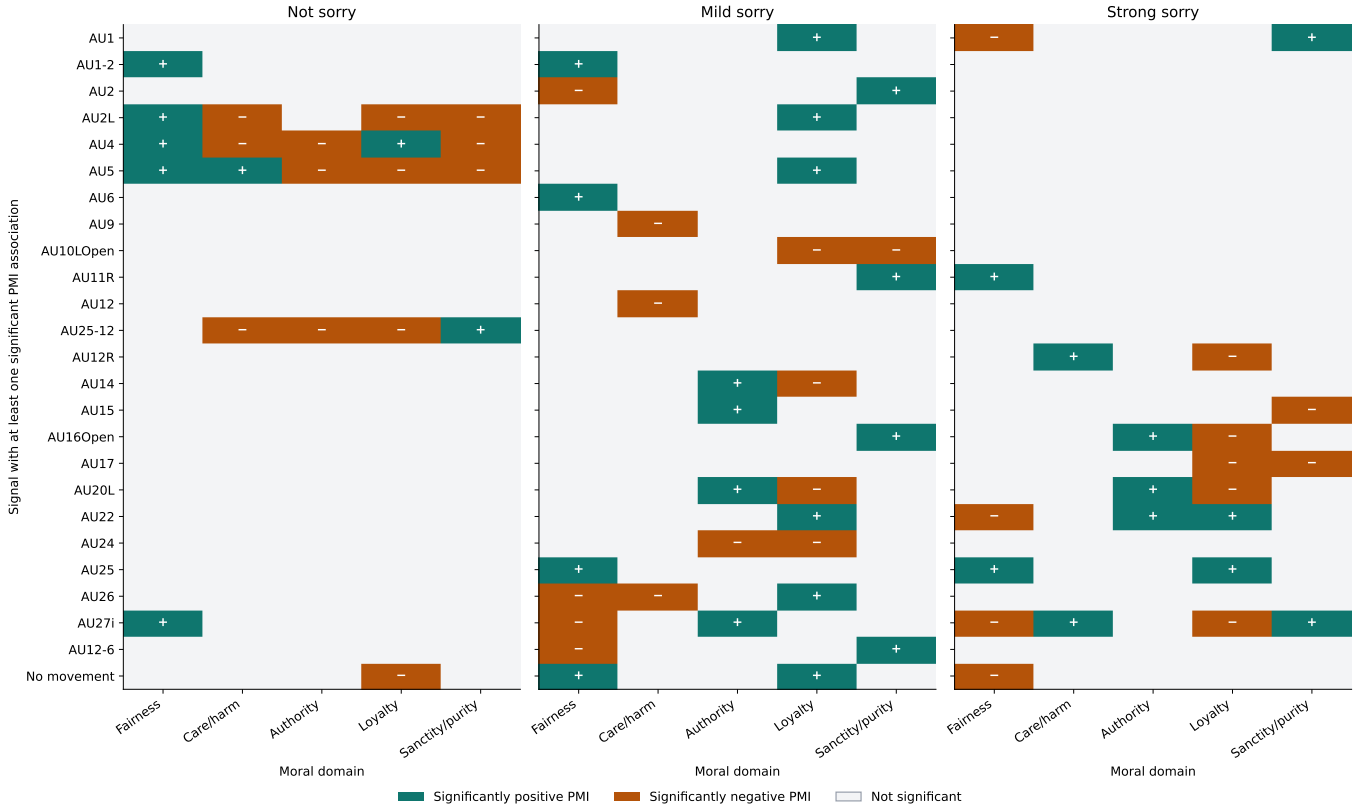
Significant Positive and Negative PMI Associations by Response and Moral Domain for WC Receiver-Group Models.

Response	Domain	Positive AUs	Positive gaze	Negative AUs	Negative gaze
Not sorry	Fairness	AU1-2, AU2L, AU4, AU5, AU27i	–	–	–
Not sorry	Care/harm	AU5	–	AU2L, AU4, AU25-12	–
Not sorry	Authority	–	–	AU4, AU5, AU25-12	–
Not sorry	Loyalty	AU4	–	AU2L, AU5, AU25-12	No movement
Not sorry	Sanctity/purity	AU25-12	–	AU2L, AU4, AU5	–
Mild sorry	Fairness	AU1-2, AU6, AU25	No movement	AU2, AU26, AU27i, AU12-6	–
Mild sorry	Care/harm	–	–	AU9, AU12, AU26	–
Mild sorry	Authority	AU14, AU15, AU20L, AU27i	–	AU24	–
Mild sorry	Loyalty	AU1, AU2L, AU5, AU22, AU26	No movement	AU10LOpen, AU14, AU20L, AU24	–
Mild sorry	Sanctity/purity	AU2, AU11R, AU16Open, AU12-6	–	AU10LOpen	–
Strong sorry	Fairness	AU11R, AU25	–	AU1, AU22, AU27i	No movement
Strong sorry	Care/harm	AU12R, AU27i	–	–	–
Strong sorry	Authority	AU16Open, AU20L, AU22	–	–	–
Strong sorry	Loyalty	AU22, AU25	–	AU12R, AU16Open, AU17, AU20L, AU27i	–
Strong sorry	Sanctity/purity	AU1, AU27i	–	AU15, AU17	–

Note. Positive associations indicate that signal presence (ON state) was overrepresented in the domain. Negative associations indicate that signal absence (OFF state) was overrepresented and hence signal presence was underrepresented; they are not numerically negative PMI values. All entries passed both the signal-level MI gate and the relevant state-specific PMI threshold. A dash denotes no significant association in that direction.

Figure 3.9

Significant Positive and Negative PMI Associations Between Individual Signals and Moral Domains for WC Receiver-Group Models in Each Response Bin.



Note. A + cell denotes overrepresentation of signal presence; a − cell denotes overrepresentation of signal absence and therefore underrepresentation of signal presence. Both directions passed the signal-level MI gate and the relevant state-specific PMI threshold; blank cells show no retained signed association.

3.5.4 Stage 5: Cross-Method Synthesis

Integration Across Prevalence, Classification, Ablation, and PMI. Taken together, the analyses place moral-domain information within a shared WC apology architecture rather than in five isolated moral-domain prototypes. Population thresholding established the supported signal representation, and response category supplied its clearest organisation. Within that architecture, complete-vector classification detected omnibus moral-domain information, with corrected cell-wise localisation only for Loyalty; ablation identified class-specific and directed conditional contributions; and MI/PMI identified marginal signal–domain associations within response categories. These levels are complementary but not interchangeable.

As Table 3.10 shows, over the aligned mild- and strong-sorry inputs, ablation identified 17 active signals and signed PMI identified 23. Jaccard overlap expresses the shared signals as a percentage of all distinct signals identified by either method, counting each signal once (Jaccard, 1912). Twelve AUs occurred in both sets, so the Jaccard overlap was $12/28 = 42.9\%$. The overlap covered 70.6% of ablation-active signals and 52.2% of signed-PMI signals. Looking down was the only active gaze indicator in ablation, while No movement was the only signed-PMI gaze indicator; the gaze sets therefore did not overlap.

Table 3.10

Signal-Level Overlap Between WC Receiver-Group Ablation and Signed MI/PMI Results.

Channel	Ablation active	Signed PMI	Shared	Union	Shared/ablation	Shared/PMI	Jaccard
Action Units	16	22	12	26	75.0%	54.5%	46.2%
Gaze	1	1	0	2	0.0%	0.0%	0.0%
All signals	17	23	12	28	70.6%	52.2%	42.9%

Note. Signed PMI includes positive or negative associations in mild or strong sorry. Shared counts signals identified by both methods; Union counts signals identified by either method, counting each signal once. Jaccard is Shared/Union, expressed as a percentage (Jaccard, 1912). These sets pool mild and strong sorry across moral domains; the overlap therefore concerns signal membership, not matching domain-specific roles. The shared signals were AU1, AU2, AU6, AU11R, AU12, AU12R, AU15, AU16Open, AU17, AU20L, AU25, and AU27i.

Same-domain integration was restricted to diagonal ablation roles. All four Loyalty identifiers had negative Loyalty PMI: AU12R, AU16Open, and AU27i in strong sorry, and AU20L in both apology bins. AU6, the Fairness diagonal confuser, had positive mild-sorry Fairness PMI. The Care/harm diagonal confusers AU14R and AU20L had no signed Care/harm PMI. These patterns describe whether a class-specific conditional role also had a marginal association; they do not establish signal exclusivity.

Off-diagonal roles were integrated at the level of their two-endpoint contrast. For example, AU20L supported Authority-to-Care/harm, Loyalty-to-Authority, and Loyalty-to-Sanctity/purity separation, while its signed PMI was positive for Authority and negative for Loyalty. AU27i supported Sanctity/purity-to-Loyalty separation and was positively associated with strong-sorry Sanctity/purity but negatively associated with strong-sorry Loyalty. By contrast, AU25 supported Care/harm-to-Fairness separation without signed PMI at either endpoint. The pairwise evidence therefore remained conditional even when marginal endpoint associations helped interpret a contrast.

No same-domain Jaccard was calculated from the pairwise roles. Such a statistic would convert directed contrasts into duplicated domain assignments and would not compare like with like.

Moral-Domain Similarities and Differences. Loyalty was the clearest complete WC configuration: it was the only significant diagonal cell and all four class-specific identifiers occurred on its diagonal. The negative Loyalty PMI for those identifiers indicates that their absence contributed marginal information, while the classifier result remained multivariate.

The pairwise results also showed substantial overlap. Twelve signals contributed to Fairness-to-Loyalty confusion, including Looking down, while Fairness had an AU6 diagonal confuser. Care/harm had two diagonal confusers and was linked to Fairness by an AU25 separation effect. Authority and Sanctity/purity had no diagonal class-specific role, but AU20L and AU27i supported selected directed separations involving those domains.

The formal domain analyses therefore refined rather than replaced the response-level prevalence result. Response category remained the clearest organisation of the WC signal space, while moral context selectively changed complete configurations and particular directed boundaries within a shared repertoire.

3.6 Discussion

3.6.1 Summary of Findings

This study asked whether visual apology signals form a common repair pattern across moral violations or vary by moral domain among WC receivers. All five domain-specific model sets met the validation criterion, retaining all 76 WC participants and 228 response-category models. Overall, the results show that response category across the Not-sorry, Mild-sorry, and Strong-sorry models supplied the shared, intensity-sensitive architecture, while moral context selectively differentiated complete configurations and directed pair boundaries. The omnibus classifier, conditional ablation effects, and marginal MI/PMI associations therefore converge on neither five isolated moral-domain prototypes nor a wholly domain-general signal; they show a common visual-signal repertoire with selective moral-domain tuning.

3.6.2 Shared Response-Level Apology Architecture

The shared not-sorry–mild-sorry–strong-sorry progression extends the central finding of Chapter 2 that apology is represented through dynamic, multimodal signal configurations rather than one fixed set of facial expressions. Chapter 3 tested whether this response-level organisation remained visible when receivers judged apologies following five morally different violations. Its persistence across the five domains shows that response category provided a broader organising structure than the more selective moral-domain information. The analyses do not establish a temporal order in receiver processing.

Mild sorry was the most heterogeneous and least sharply bounded response category. Its mean validation accuracy was lower than that of not sorry and strong sorry, and its upper-quartile AU set was broader and less concentrated. Mild-sorry models also retained some features prominent in the not-sorry profile, including AU9, AU43, AU10-related activity, and upward gaze. This overlap does not invalidate the modelling category, but neither does it establish one uniformly intermediate state between not sorry and strong sorry.

At the population level, strong sorry showed a more concentrated AU prevalence profile centred on AU20-related activity, alongside AU13, AU24, and AU14-related movements; downward gaze was also the dominant gaze indicator. This profile is compatible with receivers interpreting a stronger reparative stance. The result concerns a population-level pattern and should not be reduced to a one-to-one mapping between

an individual AU or gaze indicator and an emotion.

The response-level prevalence pattern also qualifies the simple contrast between *sad-sorry* and *smile-sorry* displays introduced in Chapter 2. Strong-sorry panels showed a population-level tendency toward AU20-related movements and downward gaze, closer to the sad-sorry profile, while AU12-family movements remained prevalent in mild- and strong-sorry panels and provided a smile-related tendency. These labels are shorthand for separate population-level profiles, not evidence that any one AU reveals sadness or happiness. The pooled summaries neither establish two discrete subtypes nor show that the movements co-occurred within individual displays; they show that receiver preferences encompassed both tendencies across the generated displays.

The not-sorry models supplied an important contrast for understanding this response-level progression. Their highly consistent AU and upward-gaze profile helped distinguish the absence of apology from the two apology responses. However, not-sorry models were excluded from the formal moral-domain classifier. They therefore support the shared response pattern but do not provide formal evidence that the moral domains themselves were separable.

3.6.3 Selective Moral-Domain Differentiation

The moral-domain classifier produced a balanced accuracy of 44.92%, and none of the 1,000 balanced accuracies from the participant-label permutations was as large ($p = .001$). The complete mild- and strong-sorry vectors therefore contained omnibus moral-domain information. The corrected cell-wise tests showed that this information was unevenly distributed: Loyalty-to-Loyalty classification at .815 was the only cell exceeding its Bonferroni-corrected threshold.

Loyalty was consequently the clearest candidate for a selectively differentiated WC repair configuration. Its four diagonal identifiers—AU12R, AU16Open, AU20L, and AU27i—all had negative Loyalty PMI in at least one apology bin. Their absence supplied consistent marginal information. However, the correct-classification rate of 81.5% for Loyalty was obtained using all facial signals included in the classifier; it does not describe the performance of any single signal.

The other domains are better described through directed classification overlap and conditional pair effects than as failed classes or symmetric distances. Fairness-to-Loyalty confusion was supported by 12 conditional pair-confusion effects, and Care/harm-to-Authority classification was approximately as frequent as Care/harm's correct classification. Care/harm, Authority, and Sanctity/purity were also linked by smaller sets of directed pair-separation and pair-confusion effects. Only the Loyalty diagonal survived the corrected cell-wise test. These results support pragmatic pluralism: a shared, intensity-sensitive response-level pattern coexisted with selective information about the type of moral violations.

3.6.4 Complementary Roles of Action Units and Gaze

Four AU predictors supported correct identification of Loyalty. AU6 hindered correct identification of Fairness, while AU14R and AU20L hindered correct identification of Care/harm. Facial AUs also accounted for 19 of the 20 effects/instances that increased or reduced misclassification between domains. Looking down was the only gaze indicator with an ablation effect, contributing to Fairness models being classified as Loyalty.

Ablation and MI/PMI answered different questions. Their valid signal-level sets overlapped by 42.9%, but the only active ablation gaze indicator did not overlap with the signed-PMI gaze set. Diagonal same-

domain convergence was strongest for the four Loyalty identifiers, whereas pairwise roles required two-endpoint interpretation. Agreement strengthens selected conditional contributions, while non-overlap indicates conditional versus marginal information rather than failed replication.

The distinction also prevents a movement from becoming a fixed domain label. AU20L was a Loyalty identifier and Care/harm confuser on the diagonal, supported three directed separations, and contributed to two directed confusions. Its signed associations also changed by domain. Signal meaning therefore depended on its position within a response- and domain-conditioned configuration.

3.6.5 Theoretical and Methodological Implications

The results support a nested account of visual apology perception. The main pattern across the models reflected whether faces were judged not sorry, mildly sorry or strongly sorry, while information about the damaged value or relationship was represented more selectively in complete configurations of present and absent AU and gaze-direction components. This representational hierarchy explains how stable response-level organisation can coexist with uneven moral-domain separability; it does not establish the order in which receivers processed the information.

Methodologically, the participant-level, prototype-free generative approach provides a way to estimate context-conditioned receiver representations without imposing one canonical apology expression for each domain. The analysis retains variation between receivers, establishes which effects reach a population-prevalence criterion, and then examines both configuration-level classification and signal-level associations. This approach is particularly useful when the theoretical question concerns the balance between a common social signal and contextual tuning.

3.6.6 Limitations Addressed by Chapters 4 and 5

Chapter 3 establishes the response-level and moral-domain pattern within the WC receiver group only. It cannot determine whether the shared intensity-sensitive pattern, moral-domain differences, and different contributions of AUs and gaze direction also appear among EA receivers. Chapter 4 addresses this question through the corresponding within-EA analysis of the same response-category and moral-domain structure.

A separate within-group analysis also cannot establish that the WC and EA receiver groups differ. In particular, an effect reaching a significance threshold in one receiver group but not the other is not itself evidence of a group difference. Chapter 5 therefore compares the groups directly to test whether apology configurations, moral-domain separability, and AU/gaze-direction contributions differ.

Chapter 3 should consequently be treated as the WC baseline for the thesis's cumulative argument. Its common-core-plus-selective-tuning conclusion remains group-bounded until Chapter 4 reports the corresponding EA analysis and Chapter 5 directly compares the WC and EA receiver groups.

3.6.7 Chapter Conclusion

Among WC receivers, visual apology signals were organised primarily by response category: a shared, intensity-sensitive architecture structured apparent apology across moral violations, while moral context selectively tuned complete configurations and directed pair boundaries. Loyalty supplied the clearest corrected evidence of differentiation, whereas the broad Fairness-to-Loyalty confusion pattern showed that Fairness and Loyalty overlapped conditionally rather than forming five isolated moral-domain prototypes. The chapter therefore answers its research question with a common-core-plus-selective-tuning account;

Chapter 4 tests whether the same hierarchy appears among EA receivers, and Chapter 5 determines which WC–EA differences survive direct comparison.

Chapter 4

Visual Apology Signals Across Moral Domains: East Asian Receivers

4.1 Chapter Introduction

The East Asian receiver (EA) study was designed as a matched extension of the Western-receiver study because moral evaluation, apology practice, and attention to nonverbal signals can vary across cultural settings. The chapter therefore tests whether the response-level and moral-domain organisation established in Chapter 3 is reproduced when receivers and rendered face identities are matched to an East Asian sample. Chapter 3 examined whether Western observers use a common visual apology signal across moral domains or instead expect different signal configurations after different kinds of wrongdoing. The present chapter extends the same question to East Asian observers. This extension is theoretically informative because the meaning of a transgression, the relational damage it creates, and the conduct expected from an offender are partly organised by culture. Moral domains may provide a broadly useful framework across populations, but this does not require each domain to carry identical social meaning or weight in every cultural setting. For example, Chinese participants have been found to moralise uncivilised conduct more strongly than Western participants, while face loss and relational concerns can play a particularly important role in responses to transgression in Chinese contexts (Buchtel et al., 2015; Hui & Bond, 2009). East Asian data therefore provide a meaningful test of whether the moral-domain structure identified in Chapter 3 is also reflected in the apology signals perceived by a different population of receivers.

Culture may also shape what an apology is understood to communicate. Apology can signal remorse, responsibility, relationship value, or an intention to repair, but the relative importance of these meanings varies across social systems. Japanese and American participants, for instance, can interpret and use apologies differently, and apology may function differently from compensation as a means of maintaining relationships across cultural settings (Komiya et al., 2020; Maddux et al., 2011). Research on Chinese, Japanese, and American apology practices similarly suggests cultural variation in the verbal and nonverbal strategies used to display sincerity, deference, and relational concern (Barnlund & Yoshioka, 1990; Park & Guan, 2009). These findings motivate examining receivers' perceptual expectations rather than assuming that a facial configuration judged apologetic in one population will necessarily carry the same repair value in another.

Nonverbal signals are especially suitable for this purpose. Facial expressions draw on shared perceptual and communicative resources, yet their production and interpretation are shaped by cultural display rules and culturally learned patterns of attention. Cross-cultural work has shown differences in facial representations of emotion and in how observers attend to the eyes, mouth, and gaze of another person (Jack et al., 2012; D. Matsumoto, Yoo, et al., 2008; Senju et al., 2013). Gaze direction may therefore contribute differently to apology judgements across populations, while facial action units may combine into culturally familiar patterns of remorse, embarrassment, concern, or restraint.

The category "East Asian" is used here to describe the recruited participant group, not to imply that East Asian cultures are homogeneous. The present study is a targeted extension of the Chapter 3 design rather than a comprehensive test of all cultural explanations. Its immediate purpose is to establish the signal structure perceived by East Asian receivers within each moral domain. Direct statistical comparisons

between cultural groups and the broader interpretation of cultural similarities and differences are reported in Chapter 5.

4.2 The Current Study

The current study asked whether EA receivers' visual apology signals were organised primarily by a shared response-level pattern or selectively differentiated across five moral domains. Participants evaluated dynamically generated facial movements and gaze directions after imagining themselves as the victim of a Fairness, Care/harm, Authority, Loyalty, or Sanctity/purity transgression. As in Chapter 3, data-driven reverse correlation estimated, for each receiver and response category, the AU and gaze-direction configuration associated with each level of apparent apology intensity. Because the procedure did not impose one fixed set of facial expressions, shared and domain-sensitive configurations could emerge from receivers' judgements.

The analyses were interpreted as one cumulative test rather than as separate endpoints. Population prevalence established the shared response-level architecture; multivariate classification tested whether complete visual-signal configurations also carried moral-domain information; ablation isolated conditional contributions to that classification; and MI/PMI localised marginal signal–domain associations within response categories. Together, these analyses test whether EA signals are organised mainly by apparent apology intensity while preserving distinct configuration-level, conditional, and marginal evidence for selective moral-domain differentiation.

4.3 Methods

The East Asian study used the same receiver-centred procedure as Chapter 3 to ask whether apparent apology was organised by apology intensity across moral domains in an EA receiver group. Varying facial movements and gaze across many trials allowed each receiver's signal configuration to be estimated without imposing a fixed apology expression. The analyses first established which configurations characterised not sorry, mild sorry, and strong sorry, then asked whether the apology configurations also distinguished transgressions, i.e., moral domains. Whole-configuration classification, removal of individual signals, and signal-specific associations answered different parts of that question. This chapter focused on EA participants, while the direct WC–EA comparisons were reserved for Chapter 5.

4.3.1 Participants

The recruitment and screening procedure followed the same design as Chapter 3. The initial East Asian dataset contained 83 participants: 52 women and 31 men. Participants were assigned to one of five moral-domain conditions, with 17 participants in fairness, 18 in care/harm, and 16 each in authority, loyalty, and sanctity/purity. Each participant completed 2,400 trials.

The analysed sample comprised 80 participants: 52 women and 28 men, aged 18–35 years ($M = 23.28$, $SD = 2.78$). Retained domain samples ranged from 14 participants in fairness to 18 in care/harm, and each participant contributed one model for each of the three response categories, giving 240 participant-by-response models. Appendix B reports the domain-specific sample, model, and prevalence-threshold counts.

4.3.2 Procedure

The procedure was the same as in Chapter 3. Participants first read a transgression scenario and were instructed to imagine themselves as the person affected by the offence. They then viewed dynamically generated face videos and judged how sorry each face appeared. The experiment used repeated high-trial sampling so that a response model could be estimated separately for each participant. The moral-domain condition remained constant for a participant, whereas the facial movements and gaze indicator varied across trials according to the stimulus-generation procedure.

4.3.3 Stimuli

The five transgression scenarios and the facial stimulus-generation procedure were the same as in Chapter 3. The scenarios represented fairness, care/harm, authority, loyalty, and sanctity/purity. They had previously been screened for comprehensibility across Western and East Asian contexts, direct relevance to an identifiable victim, approximately equal and unspecified social relationships, and the absence of strongly gendered offender roles.

The analysis included 47 binary predictors: 42 facial action units (AUs) and five gaze indicators representing no gaze movement, looking up, looking down, looking left, and looking right. The generated videos contained no head movement, spoken apology, or audio. AU temporal parameters and blushing information were available from stimulus generation but were not included in the moral-domain analyses. Facial identity was a sample-matched rendering characteristic rather than a signal predictor: WC receivers viewed the matched Western face-identity set and EA receivers viewed the matched East Asian set.

4.4 Planned Data Analyses

The EA study followed the same five-stage procedure as Chapter 3, Section 3.4, with 47 signals and three response categories:

The analyses used the software described in Chapter 3.

1. **Construct the model inputs.** Not-sorry, mild-sorry and strong-sorry models underwent the same participant-grouped validation, adaptive screening and population-prevalence thresholding. Table B1 reports the retained counts.

2. **Describe the retained signals.** Prevalence summaries described recurring AUs and gaze directions, domain variation and mild–strong differences using the Chapter 3 criteria.

3. **Test moral-domain classification and conditional contributions.** Primary classification distinguished five domains from mild- and strong-sorry configurations; ablation removed and refitted each predictor on the same splits. Classification used an overall balanced-accuracy permutation test and Bonferroni correction across 25 cells; ablation corrected 25 cells per predictor, without across-predictor correction.

4. **Test marginal signal associations.** Primary MI/PMI used each signal's own 95th-percentile threshold within all three responses, with PMI following the MI gate. Without across-signal correction, these remained signal-wise candidates.

5. **Integrate the evidence.** Signal-set overlap used aligned mild- and strong-sorry inputs. Same-domain comparisons used diagonal ablation roles; directed off-diagonal roles retained their two-endpoint interpretation.

Holdouts, permutations, settings and interpretation followed Chapter 3, Sections 3.4.2–3.4.6. Classifier evaluation remained conditional on thresholds fitted to the retained sample before splitting; screening and broader validity limits are described in Section 3.4.2 and Chapter 6, Section 6.5.

4.5 Results

4.5.1 Stages 1–2: Validated Inputs and Population-Level Signal Description

Validated Models and Response Distribution. The validation procedure retained 80 East Asian participants and 240 response-bin models, distributed across Fairness, Care/harm, Authority, Loyalty, and Sanctity/purity as 14, 18, 16, 16, and 16 participants, respectively. Mild sorry was the least readily distinguished response category, matching the ordinal validation pattern in the WC analysis despite differing absolute accuracies. Appendix B reports the retained samples and domain-specific validation outcomes.

Population-Level Signal Prevalence. The population-prevalence threshold was four receivers in each moral domain. Thresholding primarily reduced the larger AU predictor space; all five gaze indicators remained represented in every domain-level union.

Figure 4.1 summarises the retained AUs and displays one gaze state per panel. The most prevalent gaze state is shown; ties are drawn with neutral-looking eyes.

Action Units. Figure 4.2 presents the retained AU profiles. The post-threshold domain unions contained 20, 26, 24, 25, and 26 AUs in Fairness, Care/harm, Authority, Loyalty, and Sanctity/purity, respectively.

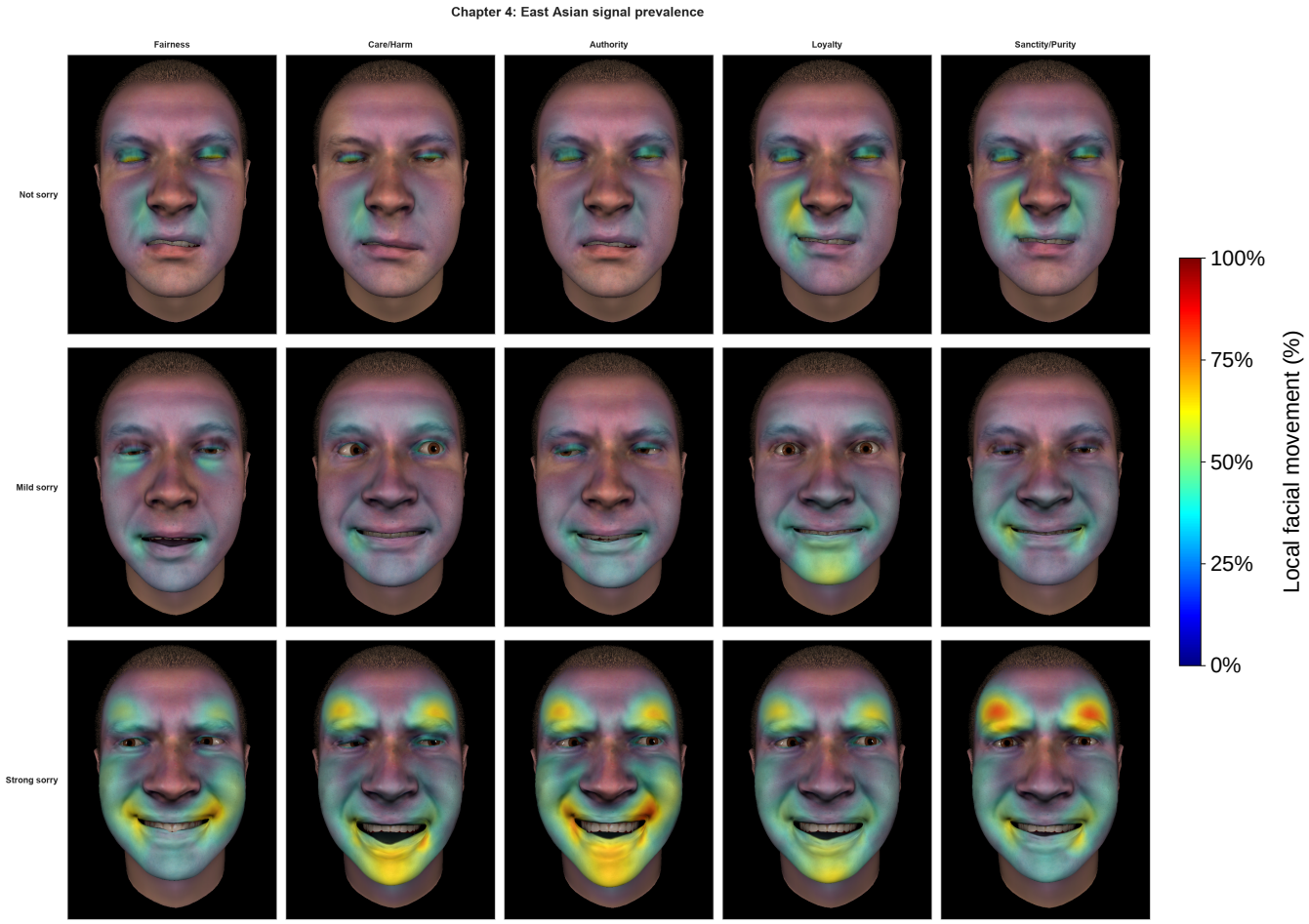
Table 4.1 pools retained AU frequencies across moral domains within each response category and reports the upper quartile of nonzero pooled signals. The upper quartile contained three of ten eligible not-sorry AUs, eight of 29 mild-sorry AUs, and seven of 26 strong-sorry AUs. Their mean pooled prevalence was 48.3%, 27.5%, and 57.9%, respectively. Not sorry was led by AU43, AU10ROpen, and AU9. Mild sorry contained a broad, relatively low-prevalence set centred on AU14, whereas strong sorry was more concentrated around AU4, AU14, AU12, AU25-12, AU20, AU12-6, and AU27i.

The retained profiles showed both shared response structure and selective domain variation. AU43 was the not-sorry peak in every domain. Mild-sorry peaks differed across domains, whereas strong sorry produced larger strict-majority sets centred on AU12/AU25-12 in Fairness, AU20 in Care/harm, AU25-12 in Authority, AU4/AU24 in Loyalty, and AU4 in Sanctity/purity.

The descriptive domain-sensitivity rule identified AU10Open in not sorry, with a 37.5-percentage-point range. Domain variation was broader in mild sorry. AU24 had the largest mild-sorry range (56.3 points), reaching 9/16 in Loyalty, 4/18 in Care/harm, and 4/16 in Sanctity/purity but not surviving thresholding in Fairness or Authority. AU14R varied by 44.4 points, and AU12, AU14L, AU17, AU20, and AU12-6 each varied by 43.8 points. AU43 and AU7 both peaked descriptively in Fairness and varied by 42.9 points. In strong sorry, AU26 varied by 66.7 points, AU24 by 62.5 points, AU16Open by 50.0 points, AU27i by 43.8 points, and AU20 by 43.7 points. These ranges indicate selective domain tuning within a shared response level repertoire.

Figure 4.1

East Asian Population-Level Signal Prevalence by Response Category and Moral Domain.



Note. Rows show not sorry, mild sorry, and strong sorry; columns show Fairness, Care/harm, Authority, Loyalty, and Sanctity/purity. Colour highlights facial changes from AUs weighted by prevalence. The bar shows the normalised local facial movement scale (0–100%); colours fade into the facial texture at lower values.

Figure 4.2

East Asian AU Prevalence After Population-Prevalence Thresholding.

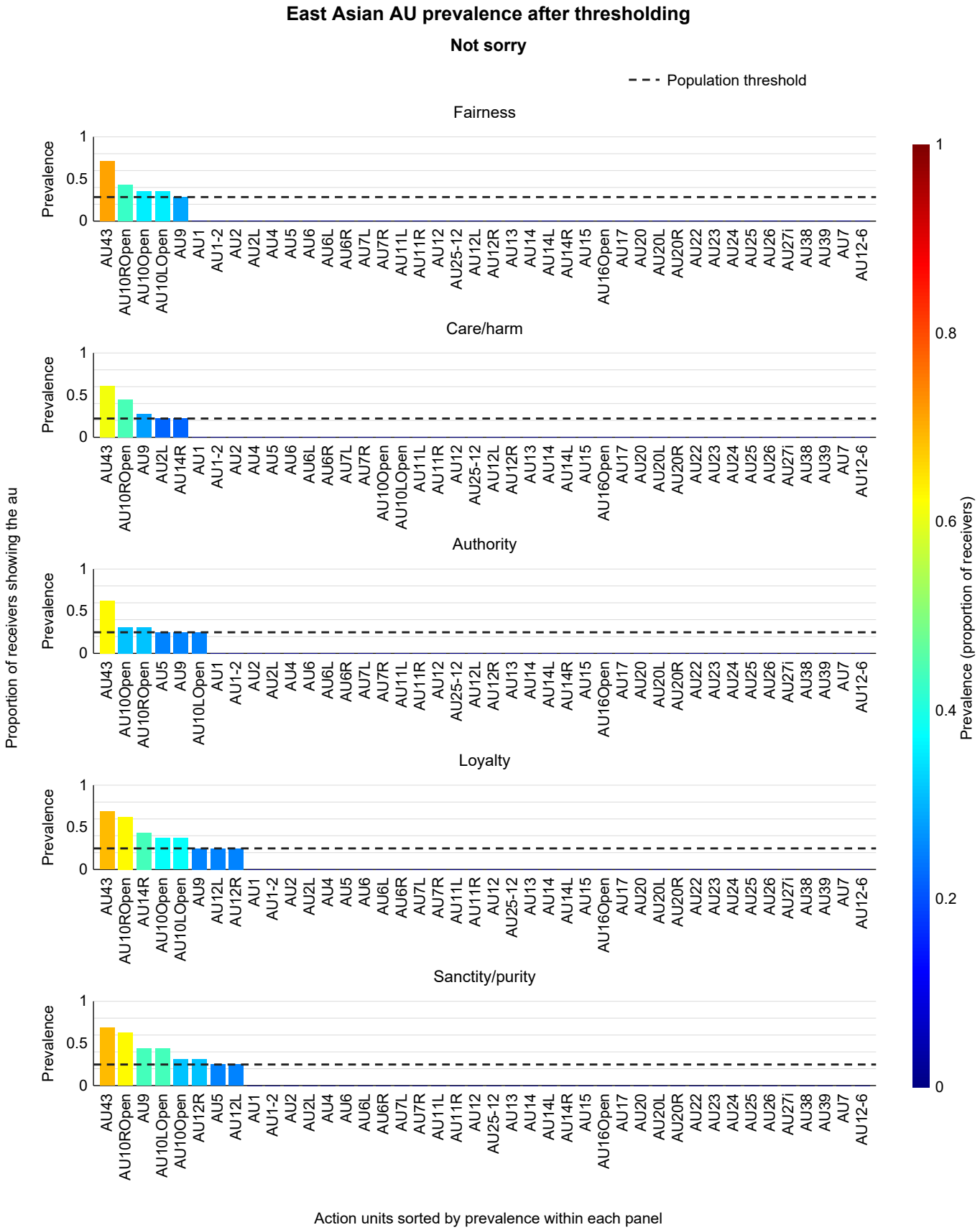


Figure 4.2 (continued)

East Asian AU Prevalence After Population-Prevalence Thresholding.

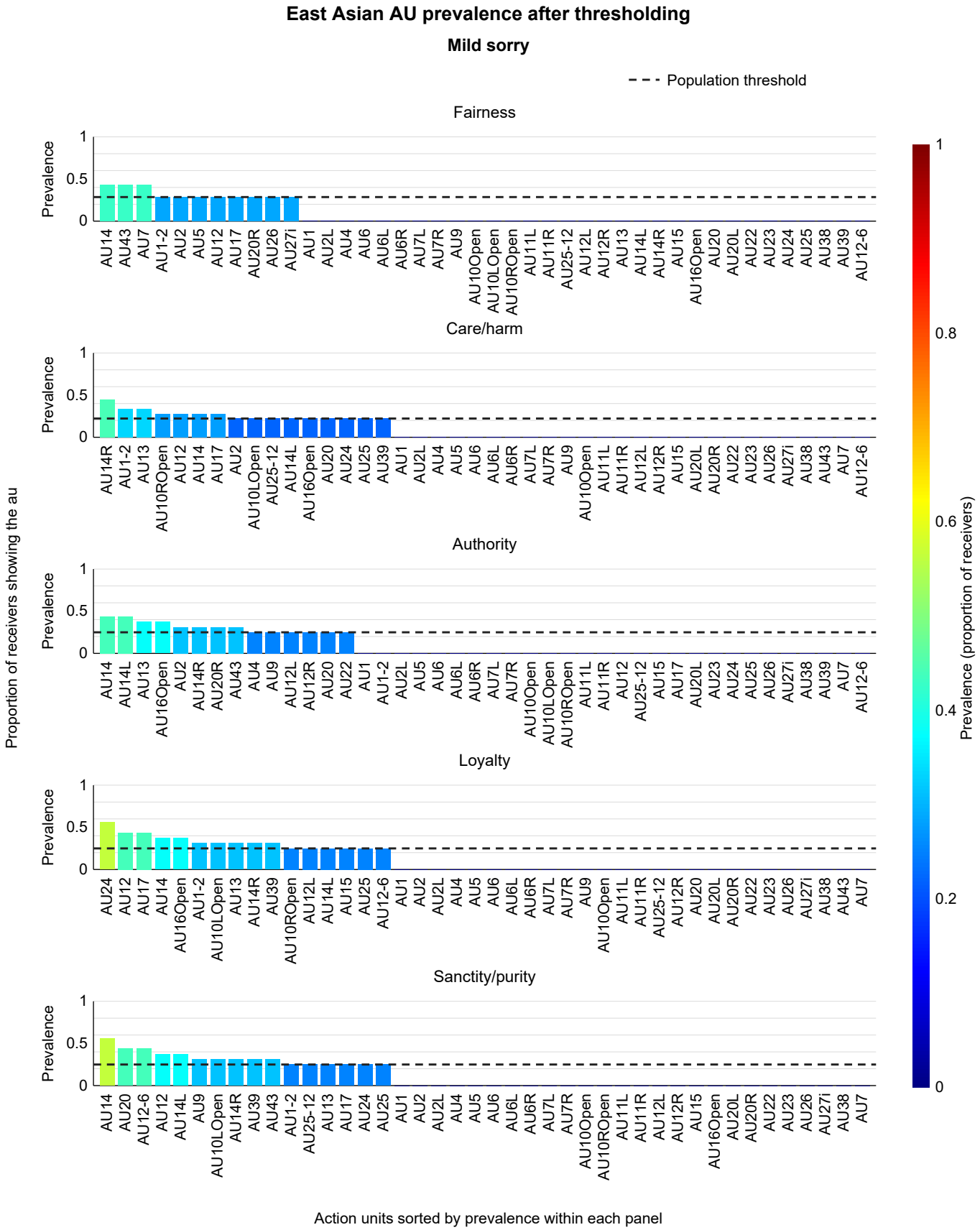


Figure 4.2 (continued)

East Asian AU Prevalence After Population-Prevalence Thresholding.

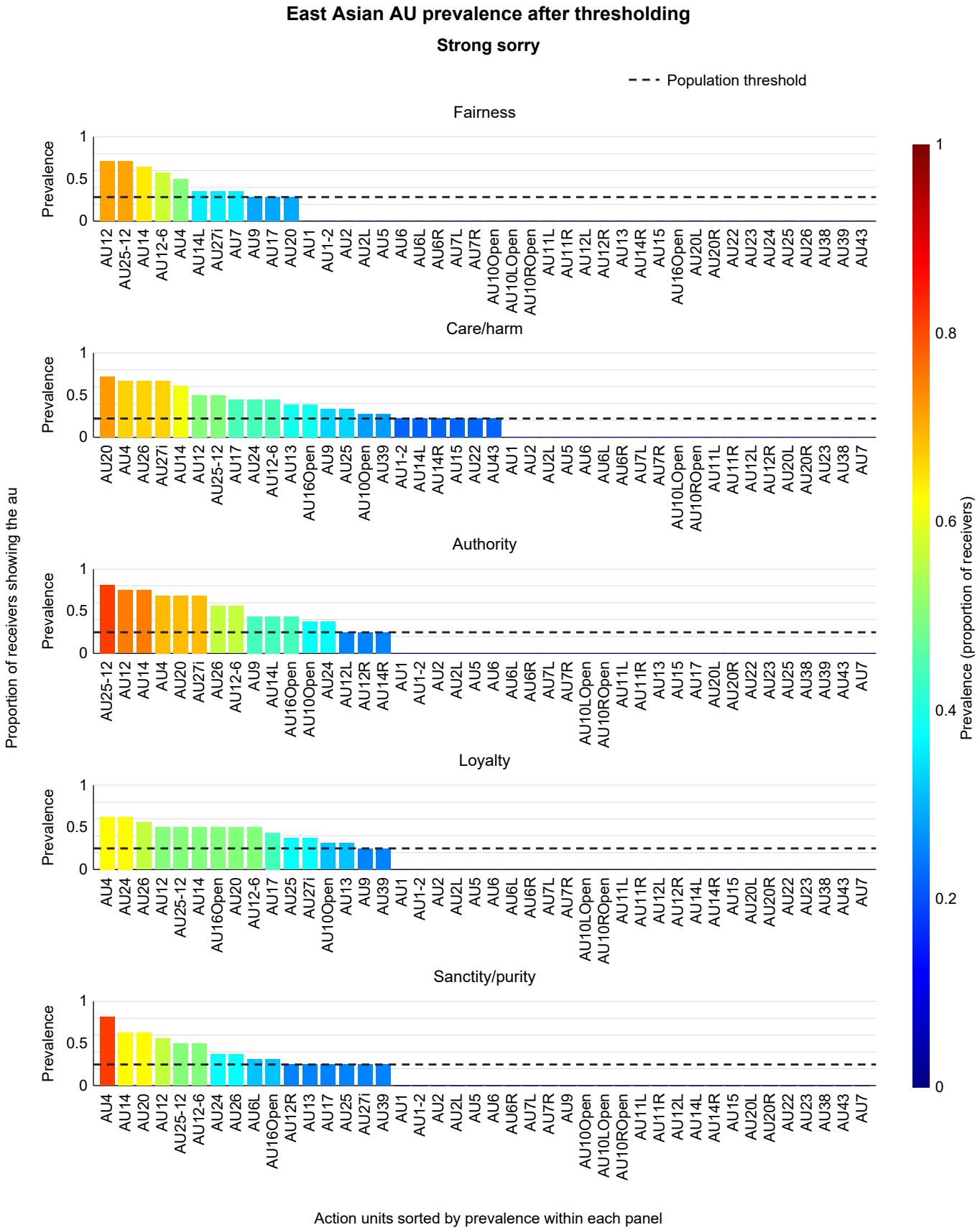


Table 4.1

Top-Quartile East Asian AUs Pooled Across Moral Domains After Population-Prevalence Thresholding.

Response category	AU	Frequency
Not sorry	AU43	53/80 (66.3%)
Not sorry	AU10ROpen	39/80 (48.8%)
Not sorry	AU9	24/80 (30.0%)
Mild sorry	AU14	33/80 (41.3%)
Mild sorry	AU14R	23/80 (28.8%)
Mild sorry	AU12	22/80 (27.5%)
Mild sorry	AU13	21/80 (26.3%)
Mild sorry	AU14L	21/80 (26.3%)
Mild sorry	AU17	20/80 (25.0%)
Mild sorry	AU1-2	19/80 (23.8%)
Mild sorry	AU24	17/80 (21.3%)
Strong sorry	AU4	53/80 (66.3%)
Strong sorry	AU14	50/80 (62.5%)
Strong sorry	AU12	48/80 (60.0%)
Strong sorry	AU25-12	48/80 (60.0%)
Strong sorry	AU20	46/80 (57.5%)
Strong sorry	AU12-6	41/80 (51.3%)
Strong sorry	AU27i	38/80 (47.5%)

Strong-sorry AU separation was generally larger than mild-sorry separation: AU25-12 and AU12 distinguished Fairness and Authority, while AU4 separated Care/harm, Loyalty, and Sanctity/purity.

The EA not-sorry profile descriptively corresponded with WC in the prominence of AU43. The leading strong-sorry components differed: the EA upper-quartile profile centred more strongly on AU4, AU14, AU12/AU25-12, and AU20, whereas the WC profile was led by AU20 and AU13. These are descriptive correspondences between separate within-group analyses rather than tested receiver-group effects.

Gaze Indicators. Figure 4.3 presents the retained gaze profiles. The smaller five-direction space produced a clearer response-level pattern than the AU channel. Looking up led four of the five not-sorry panels, mild-sorry panels were led by upward or lateral gaze, and all five strong-sorry panels were led by a lateral direction.

The pooled upper-quartile results in Table 4.2 identified looking up as the dominant not-sorry direction, looking right and looking left as the leading mild-sorry directions, and looking right as the dominant strong-sorry direction. Upward gaze therefore distinguished the not-sorry baseline, while lateral gaze became more prominent as apologies strengthened.

The panel maxima retained a compact response pattern. Not sorry peaked at looking up in four domains but at No movement in Care/harm; mild-sorry peaks varied among upward, lateral, and no-movement signals; and strong sorry peaked laterally in every domain.

Gaze showed fewer domain-sensitive patterns than AUs. No movement varied across not-sorry domains by 32.5 percentage points. In mild sorry, No movement and looking left each varied by 37.5 points and looking up varied by 31.3 points. No strong-sorry gaze indicator met the domain-sensitivity rule. The leading lateral direction varied across strong-sorry domains, but rightward gaze remained sufficiently common

Figure 4.3

East Asian Gaze Prevalence After Population-Prevalence Thresholding.

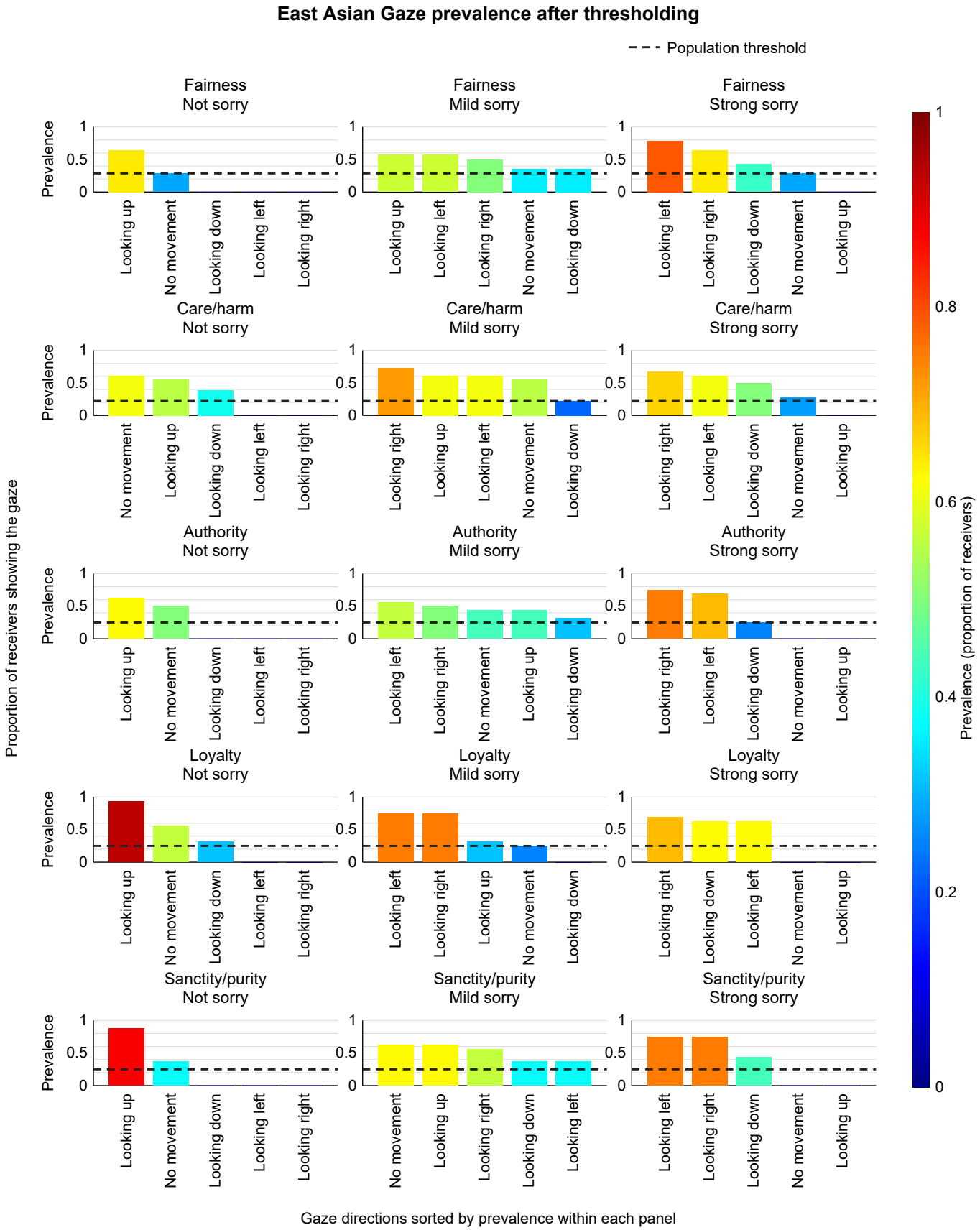


Table 4.2

Top-Quartile East Asian Gaze Indicators Pooled Across Moral Domains After Population-Prevalence Thresholding.

Response category	Gaze direction	Frequency
Not sorry	Looking up	58/80 (72.5%)
Mild sorry	Looking right	49/80 (61.3%)
Mild sorry	Looking left	46/80 (57.5%)
Strong sorry	Looking right	56/80 (70.0%)

that the cross-domain range did not reach the prespecified descriptive criterion.

Looking up provided the clearest mild-sorry separation in Fairness, Care/harm, and Loyalty, and tied with No movement in Authority and Sanctity/purity. The leading strong-sorry direction varied by domain rather than forming one uniform downward-gaze pattern.

Upward gaze led not-sorry models in both EA and WC. The strong-sorry profile was instead led by lateral gaze in EA and downward gaze in WC. Mild sorry formed the broadest and most heterogeneous signal-level repertoire in both within-group analyses. Across the EA results, AUs supplied the larger differentiated repertoire, whereas gaze showed a more compact progression from an upward not-sorry tendency toward increasingly lateral apology displays.

4.5.2 Stage 3: Primary Classification and Ablation

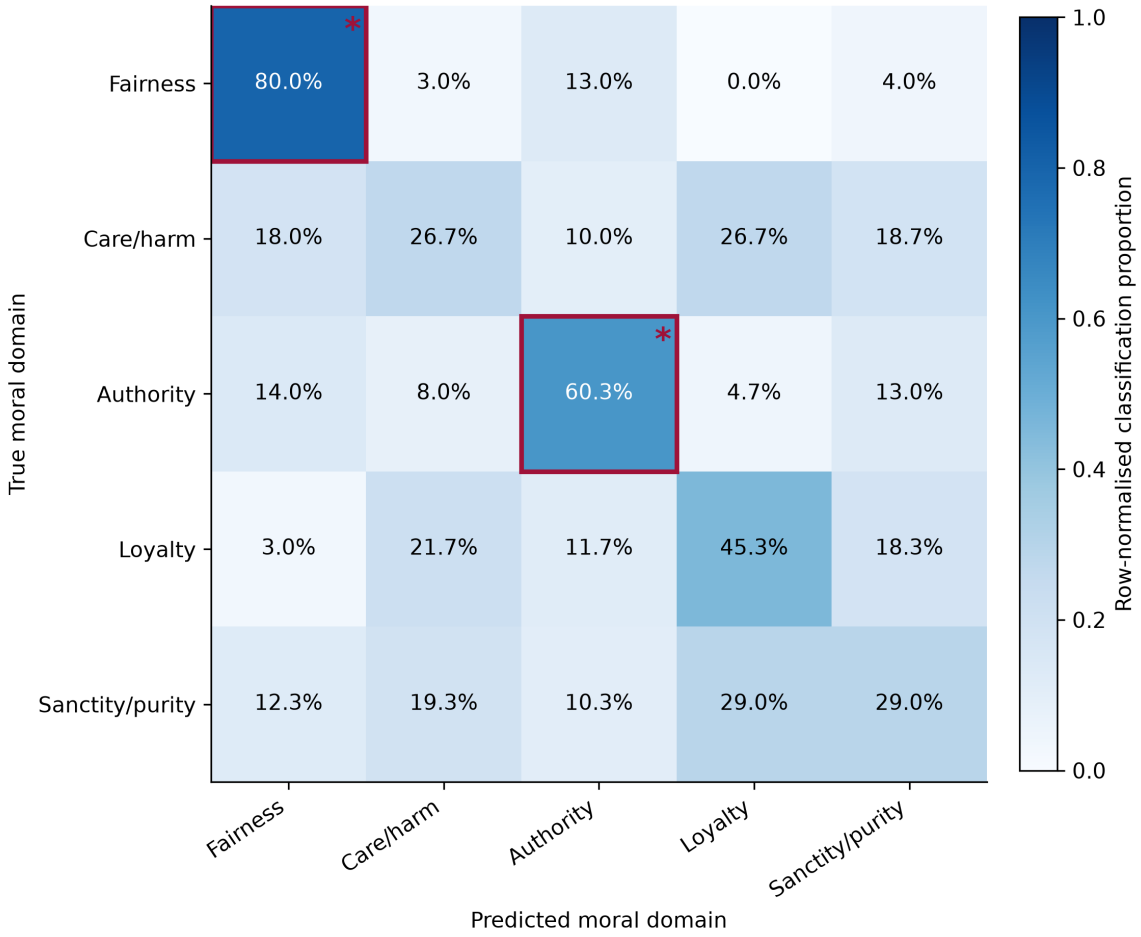
Moral-Domain Classification. The classifier used 160 EA mild- and strong-sorry models and 36 predictors retained in at least one of those models. Its balanced accuracy was .4827. None of the 1,000 balanced accuracies reconstructed from the participant-label permutations was as large, giving a plus-one Monte Carlo result of $p = .001$. The complete thresholded apology vectors therefore distinguished moral domains above the permutation null distribution.

Figure 4.4 presents the complete matrix. The diagonal values were .800 for Fairness, .267 for Care/harm, .603 for Authority, .453 for Loyalty, and .290 for Sanctity/purity. Fairness and Authority were the only diagonal cells exceeding their Bonferroni-corrected permutation thresholds. Care/harm was assigned equally often to Care/harm and Loyalty (.267 each), while Sanctity/purity was assigned equally often to Sanctity/purity and Loyalty (.290 each). The significant omnibus result therefore reflected selective rather than uniform domain separability.

Ablation Analysis. Ablation-active signals were retained predictors whose removal and refitting reliably changed at least one classifier cell, with the other signals still available. An identifier helped correct domain classification, so removing it reduced correct classifications; a confuser showed the opposite pattern. A pair-separation signal helped distinguish one directed pair, so removing it increased that particular misclassification. As Table 4.3 and Figure 4.6 show, only AU7, AU12R, and AU24 had a significant ablation cell; the other 33 classifier predictors were globally non-used. The raw cell locations separated one diagonal class role from four directed pairwise roles. AU24 was a positive Loyalty diagonal and therefore the only class-specific identifier. No diagonal class-confuser, off-diagonal pair-confusion, or gaze effect was detected.

Figure 4.4

East Asian Moral-Domain Confusion Matrix for the Mild- and Strong-Sorry Models.



Note. Rows denote observed domains and columns predicted domains. Each row sums to 100%. Highlighted diagonal cells passed the Bonferroni-corrected cell-wise permutation threshold.

As Table 4.4 and Figure 4.7 show, the remaining effects supported directed pair separation. Removing AU7 increased Care/harm-to-Fairness, Authority-to-Fairness, and Sanctity/purity-to-Fairness misclassification. Removing AU12R increased Sanctity/purity-to-Authority misclassification. These effects show that AU7 and AU12R helped distinguish specific directed pairs; they do not make either signal an independent class-specific marker of both endpoint domains.

Figure 4.5 localises the sparse ablation repertoire on the face while retaining the separation between the Loyalty identifier and the four directed pair-separation effects.

Table 4.3

Diagonal Class-Specific Ablation Roles for East Asian Moral-Domain Models.

Domain	Signal	Role	Positive PMI	Negative PMI
Loyalty	AU24	Identifier	Mild and strong sorry	None

Note. AU24 was the only diagonal class-specific role: removing it reduced correct Loyalty classification, so it was an identifier. Its positive PMI associations involved the same signal and domain.

Table 4.4

Directed Off-Diagonal Ablation Roles for East Asian Moral-Domain Models.

Signal	Observed → predicted cell	Pairwise role
AU7	Care/harm → Fairness	Separation
AU7	Authority → Fairness	Separation
AU7	Sanctity/purity → Fairness	Separation
AU12R	Sanctity/purity → Authority	Separation

Note. Separation denotes a negative off-diagonal full-minus-ablated effect: removing the signal increased that directed misclassification. Pairwise roles are not duplicated into endpoint-domain assignments.

4.5.3 Stage 4: Primary Signal-Specific Associations

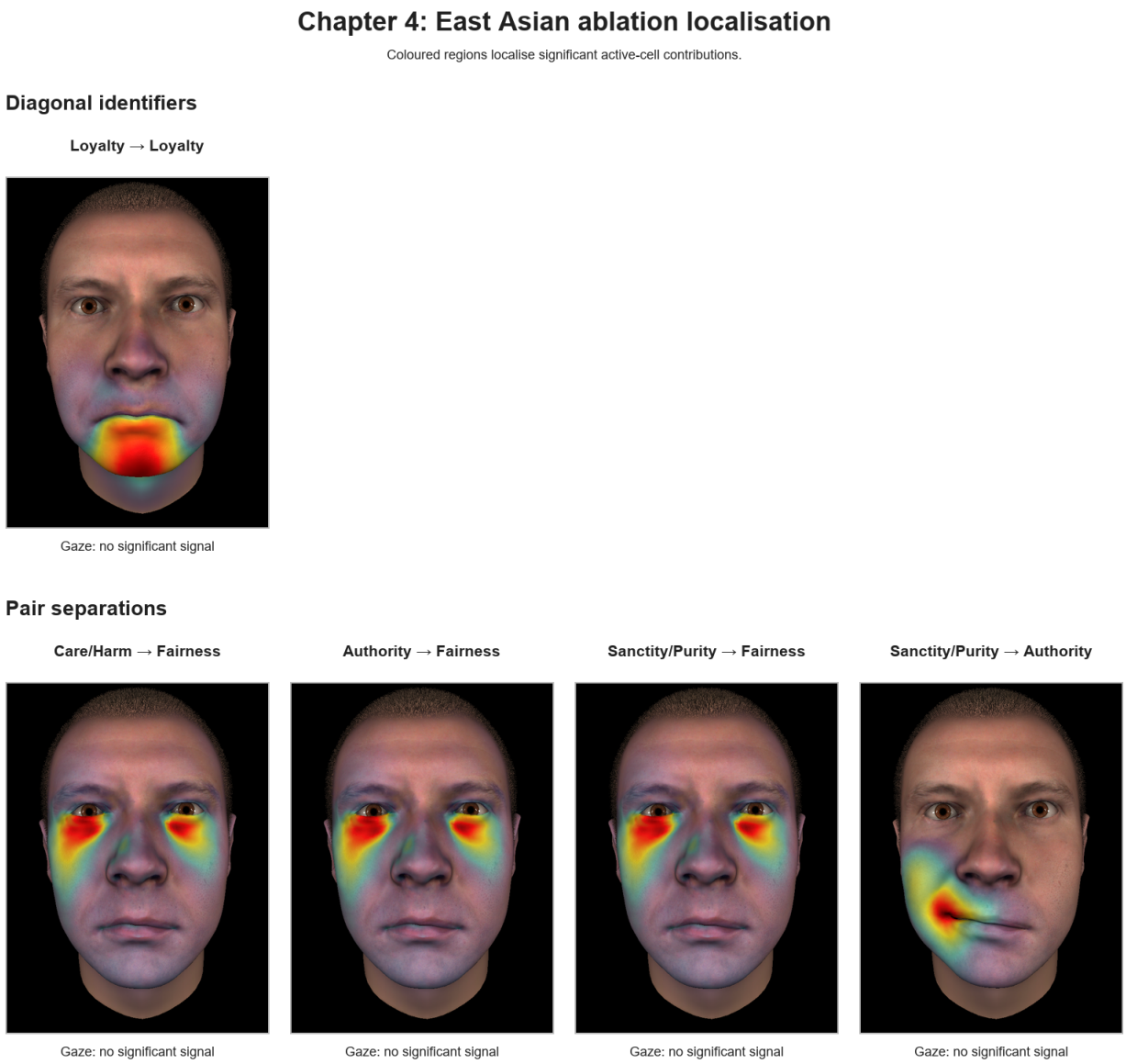
MI/PMI Signal Specificity. Across the 141 signal–response combinations formed by 47 predictors and three response bins, 64 (45.4%) were non-used, 21 (14.9%) were present but not significant, and 56 (39.7%) had at least one significant signed association with a moral domain. Among the 56 signed combinations, 41 had at least one positive association, 27 had at least one negative association, and 12 carried both directions across different domains. These combinations generated 93 significant signal–domain associations: 52 positive and 41 negative. Moral-domain specificity therefore involved both overrepresentation and underrepresentation of signal presence. Table 4.5 reports the response-level composition.

Figure 4.8 localises the ON-state associations retained for the main chapter. The corresponding OFF-state associations concern selective signal absence and are shown separately in Appendix B.

The signed repertoire was smallest for not sorry and broadest for mild sorry. Not sorry contained eight signed signals (17.0% of 47), comprising six positive signals, three negative signals, and one carrying both directions; these produced ten positive and five negative domain associations. Mild sorry contained 27 signed signals (57.4%), comprising 19 positive signals, 13 negative signals, and five carrying both direc-

Figure 4.5

Facemap Localisation of Ablation-Active Signals in the East Asian Moral-Domain Classifier.



Note. Removing the Loyalty identifier reduced correct Loyalty classification; removing a pair-separation signal increased the named directed error. No confuser, pair-confusion, or gaze role was retained. Colours locate the selected AUs, each displayed at maximum rendering amplitude.

Figure 4.6

Diagonal Class-Specific Ablation Roles in the East Asian Moral-Domain Classifier.

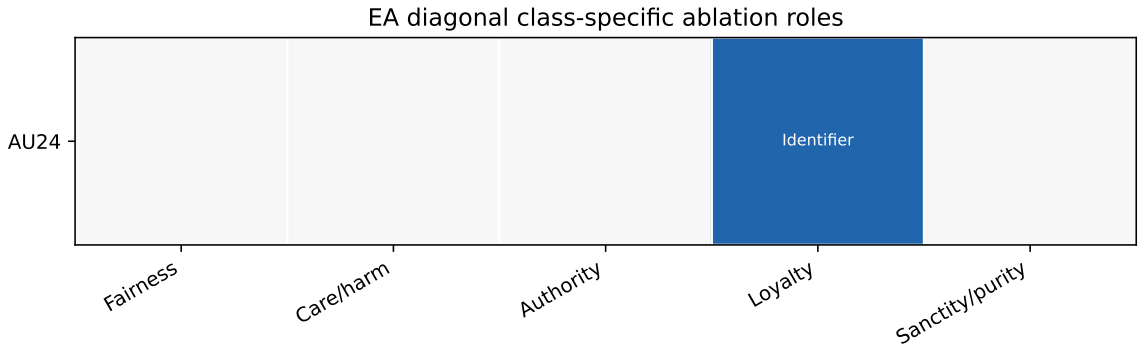


Figure 4.7

Directed Off-Diagonal Pair-Separation Roles in the East Asian Moral-Domain Classifier.

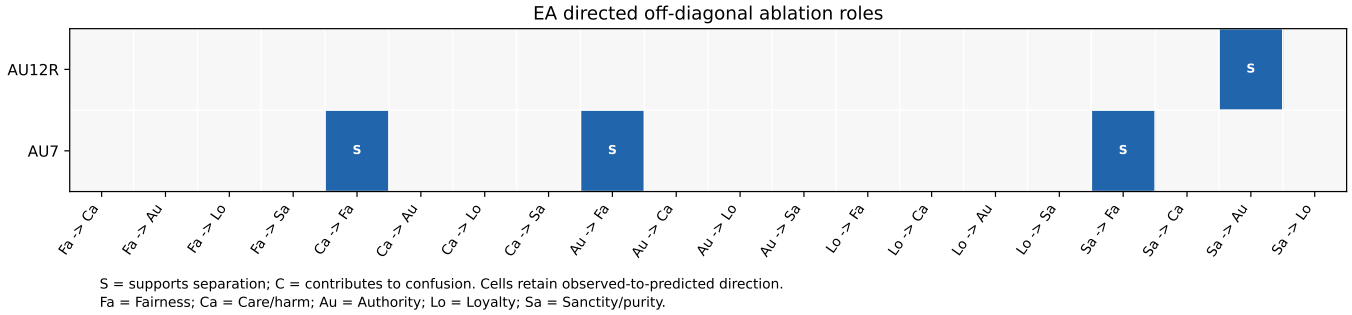


Table 4.5

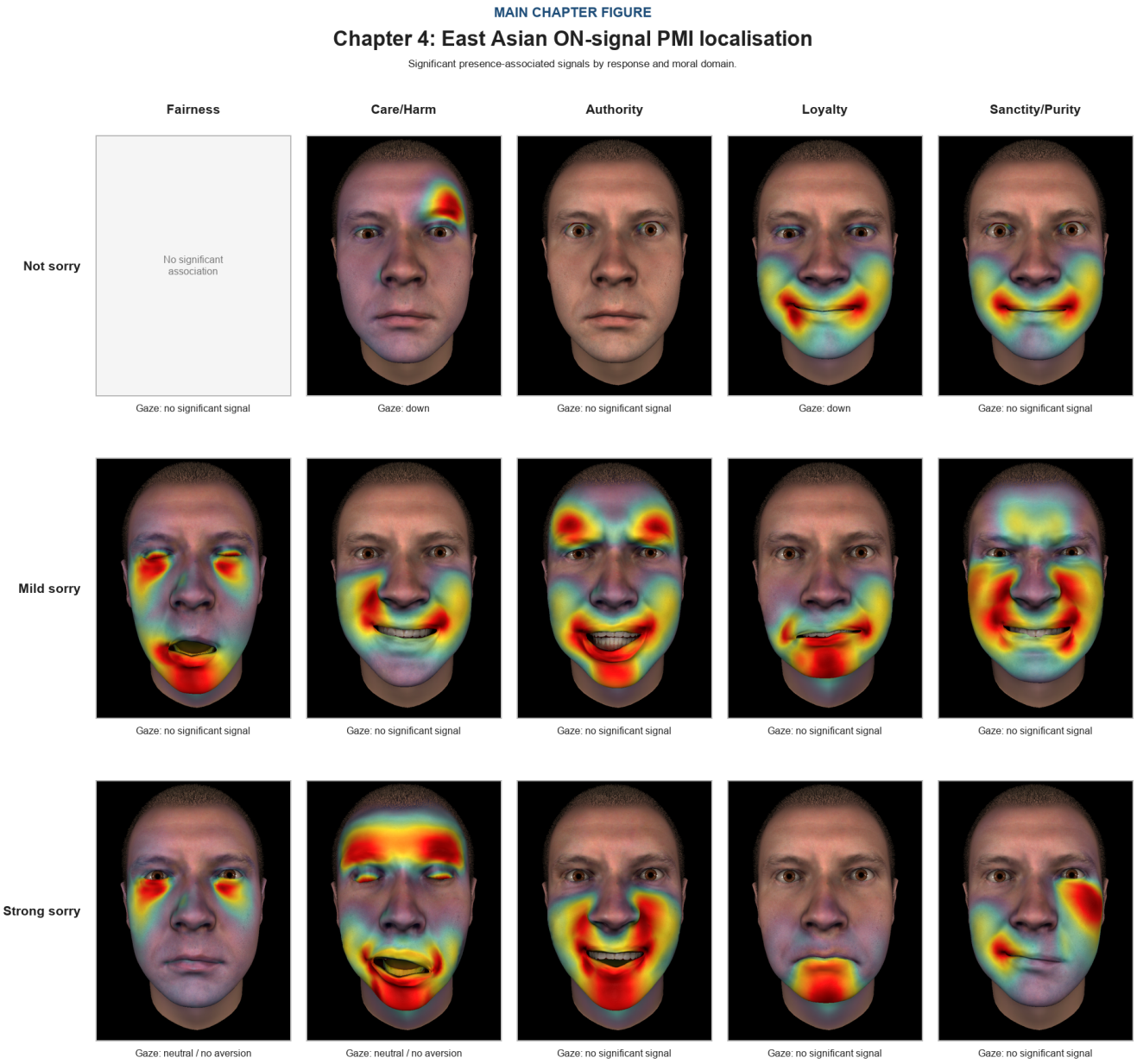
Response-Level PMI Category Composition for East Asian Signals.

Response	Non-used	Present/not significant	Any significant	Positive	Negative	Both directions	Associations (+/−)
Not sorry	34	5	8	6	3	1	10/5
Mild sorry	13	7	27	19	13	5	23/20
Strong sorry	17	9	21	16	11	6	19/16
Total	64	21	56	41	27	12	52/41

Note. Within each response row, Non-used, Present/not significant, and Any significant are mutually exclusive and sum to 47 predictors. Positive and Negative count unique signals with at least one association in the corresponding direction and can overlap; Both directions counts a signal that is positive for at least one domain and negative for at least one other domain. Associations count significant signal–domain cells, with positive counts before the slash and negative counts after it. A signal contributes one association for every moral domain in which it is significant, so the association count can exceed the number of unique positive or negative signals. A negative association denotes significant OFF-state PMI and therefore underrepresentation of signal presence, not a numerically negative PMI value.

Figure 4.8

Facemap Localisation of Significant East Asian ON-State PMI Associations by Response Category and Moral Domain.



Note. Coloured regions locate the selected AUs whose presence was overrepresented in the panel. Each AU is displayed at maximum rendering amplitude; blank panels contain no retained ON-state association.

tions; these produced 23 positive and 20 negative associations. Strong sorry contained 21 signed signals (44.7%), comprising 16 positive signals, 11 negative signals, and six carrying both directions; these produced 19 positive and 16 negative associations. Mild sorry therefore showed the broadest marginal moral-domain differentiation, consistent with its heterogeneous population-level repertoire.

Table 4.6 resolves the association totals by response and moral domain. Fairness had the largest signed profile in mild sorry, with six positive and eight negative associations. In strong sorry, Authority combined six positive and four negative associations, while Care/harm had eight positive and no negative associations. The complete signal membership is reported in Table 4.7.

Table 4.6

Positive and Negative PMI Association Counts by Response and Moral Domain for East Asian Models.

Response	Moral domain	Positive associations	Negative associations
Not sorry	Fairness	0	1
Not sorry	Care/harm	2	2
Not sorry	Authority	1	1
Not sorry	Loyalty	4	0
Not sorry	Sanctity/purity	3	1
Mild sorry	Fairness	6	8
Mild sorry	Care/harm	2	1
Mild sorry	Authority	6	5
Mild sorry	Loyalty	5	4
Mild sorry	Sanctity/purity	4	2
Strong sorry	Fairness	2	7
Strong sorry	Care/harm	8	0
Strong sorry	Authority	6	4
Strong sorry	Loyalty	1	1
Strong sorry	Sanctity/purity	2	4

Note. Counts are significant signal–domain cells. Positive associations indicate overrepresentation of signal presence; negative associations indicate overrepresentation of the OFF state and hence underrepresentation of signal presence.

Action Units. In not-sorry models, AU2L was positively associated with Care/harm, AU5 with Authority and Sanctity/purity, and AU12L/AU12R with Loyalty and Sanctity/purity. AU14R was additionally positive for Loyalty. The only negative AU associations occurred in Care/harm, where AU10Open and AU10LOpen were underrepresented. Not-sorry domain tuning was therefore comparatively sparse and depended partly on gaze, despite the strong shared prevalence of AU43 and upward gaze.

Mild sorry contained the largest AU-specific profile. Fairness combined six positive and eight negative AU associations; Authority combined six positive and five negative associations; Loyalty combined five positive AU associations with three negative AU associations and one negative gaze association. Several signals reversed direction across domains. AU20R was positive for Fairness and Authority, whereas AU24 was positive for Loyalty but negative for Fairness and Authority. AU16Open was positive for Authority and Loyalty but negative for Fairness and Sanctity/purity. AU43 was positive for Fairness but negative for Care/harm and Loyalty. AU20 was positive for Sanctity/purity but negative for Fairness and Loyalty. Thus, the broad mild-sorry repertoire was differentiated by opposing patterns of enrichment and underrepresentation rather than isolated one-domain signals.

Strong sorry also contained extensive AU differentiation. Care/harm had seven positive AU associations

Table 4.7

Significant Positive and Negative PMI Associations by Response and Moral Domain for East Asian Models.

Response	Domain	Positive AUs	Positive gaze	Negative AUs	Negative gaze
Not sorry	Fairness	–	–	–	Looking down
Not sorry	Care/harm	AU2L	Looking down	AU10Open, AU10LOpen	–
Not sorry	Authority	AU5	–	–	Looking down
Not sorry	Loyalty	AU12L, AU12R, AU14R	Looking down	–	–
Not sorry	Sanctity/ purity	AU5, AU12L, AU12R	–	–	Looking down
Mild sorry	Fairness	AU5, AU20R, AU26, AU27i, AU43, AU7	–	AU10LOpen, AU13, AU14L, AU14R, AU16Open, AU20, AU24, AU39	–
Mild sorry	Care/harm	AU10ROpen, AU25-12	–	AU43	–
Mild sorry	Authority	AU4, AU12L, AU12R, AU16Open, AU20R, AU22	–	AU10LOpen, AU12, AU17, AU24, AU39	–
Mild sorry	Loyalty	AU12L, AU15, AU16Open, AU17, AU24	–	AU2, AU20, AU43	Looking down
Mild sorry	Sanctity/ purity	AU9, AU25-12, AU20, AU12-6	–	AU2, AU16Open	–
Strong sorry	Fairness	AU7	No movement	AU10Open, AU13, AU16Open, AU24, AU25, AU26, AU39	–
Strong sorry	Care/harm	AU1-2, AU13, AU14R, AU15, AU22, AU26, AU43	No movement	–	–
Strong sorry	Authority	AU10Open, AU12L, AU12R, AU14L, AU14R, AU27i	–	AU13, AU17, AU25, AU39	–
Strong sorry	Loyalty	AU24	–	AU14L	–
Strong sorry	Sanctity/ purity	AU6L, AU12R	–	AU9, AU10Open, AU14L, AU27i	–

Note. Positive associations indicate that signal presence was overrepresented in the domain. Negative associations indicate that signal absence was overrepresented and hence signal presence was underrepresented. All entries passed the signal-level MI gate and the relevant state-specific PMI threshold. A dash denotes no significant association in that direction.

and no negative associations; Authority had six positive and four negative AU associations; Fairness had one positive and seven negative AU associations. AU13 and AU26 were positive for Care/harm but negative for Fairness, while AU13 was also negative for Authority. AU14R was positive for Care/harm and Authority, whereas AU14L was positive for Authority but negative for Loyalty and Sanctity/purity. AU24 was positive for Loyalty but negative for Fairness, and AU12R was positive for Authority and Sanctity/purity. These reversals distinguish moral-domain relations within the shared strong-sorry configuration.

Gaze Indicators. Gaze contributed a smaller signed set. Looking down was involved in every not-sorry domain: it was positively associated with Care/harm and Loyalty but negatively associated with Fairness, Authority, and Sanctity/purity. In mild sorry, looking down was negatively associated only with Loyalty. In strong sorry, No movement was positively associated with Fairness and Care/harm. No other gaze indicator produced a signed association, despite the broad population prevalence of upward and lateral gaze. Figure 4.9 presents the complete signed PMI pattern.

Signed PMI identified its broadest signal set in mild sorry in both EA and WC: 27 signals in EA and 21 in WC. AUs supplied most of the domain-specific information in each analysis. The correspondence concerns the breadth and channel distribution of the signed repertoires; the individual associations require direct receiver-group comparison before they can be interpreted as receiver-group differences.

4.5.4 Stage 5: Cross-Method Synthesis

Integration Across Prevalence, Classification, Ablation, and PMI. Across methods, the EA evidence located moral-domain information within a largely shared response-level architecture. Population prevalence established that shared architecture; complete configurations carried significant omnibus domain information, with Fairness and Authority providing the reliable diagonals; ablation narrowed the conditional evidence to one Loyalty identifier and four directed pair-separation effects across three AUs; and MI/PMI localised a broader marginal repertoire within response categories. These findings are cumulative rather than interchangeable: moral context selectively differentiated a response pattern organised primarily by apparent apology intensity.

As Table 4.8 shows, all three ablation-active signals occurred among the 32 unique mild/strong signed-PMI signals. They therefore represented 9.4% of the 32-signal union. Within AUs, the same three signals represented 10.0% of the 30-signal union. Neither signed-PMI gaze indicator had an ablation role.

Table 4.8

Signal-Level Overlap Between East Asian Ablation and Signed MI/PMI Results.

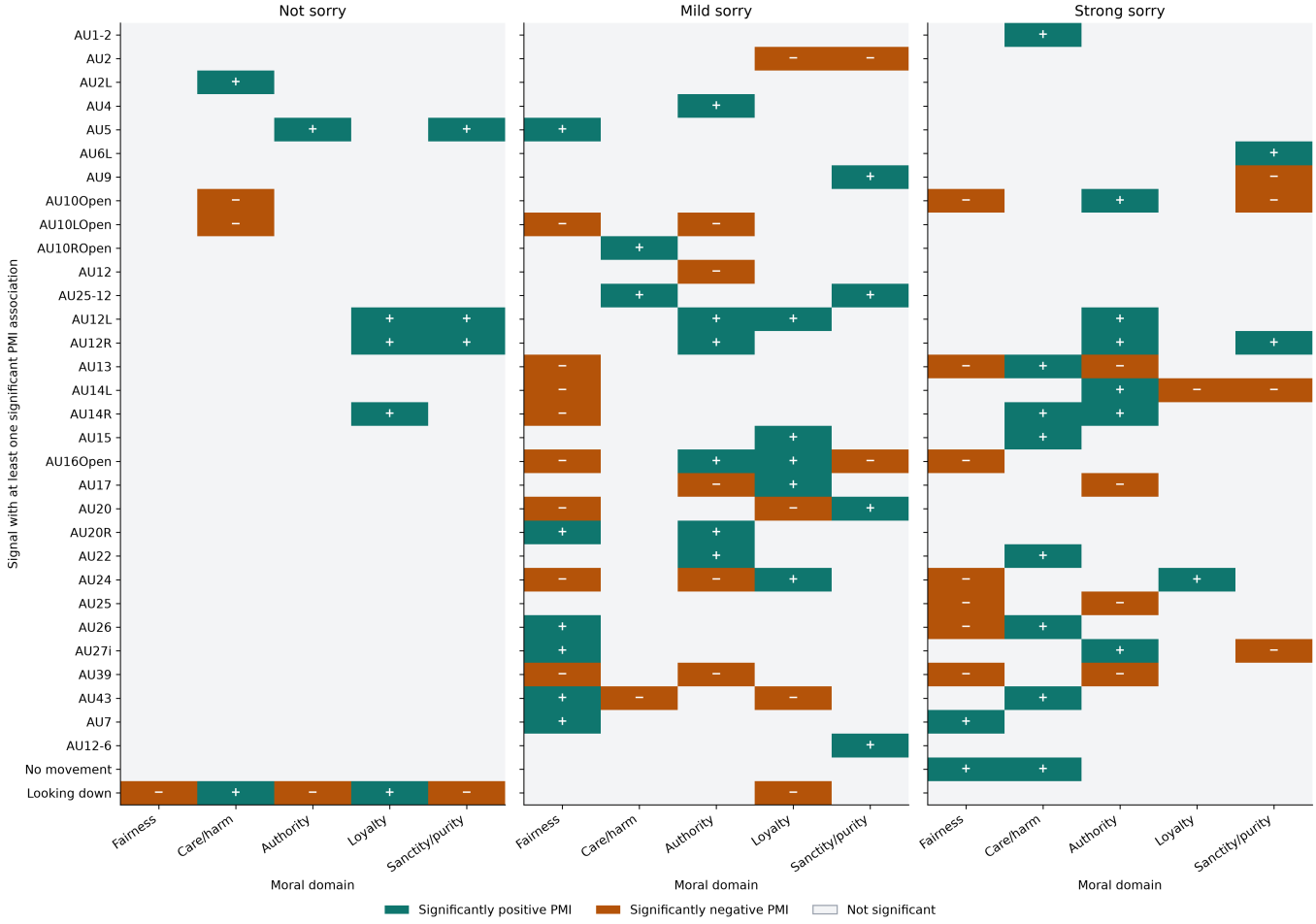
Channel	Ablation active	Signed PMI	Shared	Shared/ablation	Shared/PMI	Jaccard
Action Units	3	30	3	100.0%	10.0%	10.0%
Gaze	0	2	0	–	0.0%	0.0%
All signals	3	32	3	100.0%	9.4%	9.4%

Note. Signed PMI includes positive or negative associations in mild or strong sorry. The three shared signals were AU7, AU12R, and AU24.

Same-domain convergence was limited to the diagonal role. AU24 was the Loyalty class-specific identifier and was positively associated with Loyalty in both apology bins. Its convergence describes one coherent

Figure 4.9

Significant Positive and Negative PMI Associations Between Individual Signals and Moral Domains for East Asian Models in Each Response Bin.



Note. A (+) cell denotes overrepresentation of signal presence; a (−) cell denotes overrepresentation of signal absence and therefore underrepresentation of signal presence. Blank cells show no retained signed association.

component of the Loyalty configuration even though the complete Loyalty diagonal did not exceed the corrected cell threshold.

The other two signals require pairwise interpretation. AU7 supported separation of Care/harm, Authority, and Sanctity/purity from Fairness in the corresponding directed cells; it was positively associated with Fairness in both apology bins but had no signed PMI association at the other endpoints. AU12R supported Sanctity/purity-to-Authority separation and was positively associated with strong-sorry Sanctity/purity and with Authority in both apology bins. These two-endpoint patterns clarify the contrasts without turning AU7 or AU12R into independent identifiers of every endpoint domain.

No same-domain Jaccard was calculated from the four pairwise cells. The former 5.9% statistic depended on duplicated endpoint assignments and has therefore been withdrawn.

Moral-Domain Similarities and Differences. Fairness and Authority provided the clearest EA whole-vector separation, as shown by their significant classifier diagonals. Their conditional signal evidence was pairwise: AU7 supported three directed separations ending in Fairness, while AU12R supported Sanctity/purity-to-Authority separation. Neither signal was a class-specific diagonal identifier.

Loyalty had a non-significant .453 diagonal but contained the only class-specific ablation identifier, AU24, which also had positive Loyalty PMI in mild and strong sorry. Care/harm and Sanctity/purity remained overlapping at the classifier level, although the pairwise results identified selected boundaries involving their signals.

The combined EA evidence therefore supports a shared response-level architecture, organised mainly by apparent apology intensity, rather than five isolated moral-domain prototypes. Moral context nevertheless differentiated selected configurations: Fairness and Authority provided the clearest whole-vector separation, Loyalty supplied the only class-specific identifier, and a small AU set marked selected directed boundaries.

4.6 Discussion

4.6.1 Summary

In this study, I asked whether EA receivers used a shared visual-signal architecture across moral violations or different configurations for each moral domain. After validation, 80 participants and 240 response-category models were retained. Read collectively, the analyses show that apparent apology intensity supplied the dominant organisation: population prevalence progressed across not sorry, mild sorry, and strong sorry, while moral context selectively differentiated the configurations. The complete vectors carried significant omnibus domain information, with reliable Fairness and Authority diagonals; ablation isolated one Loyalty identifier and four directed pair-separation effects across three AUs; and MI/PMI localised a broader set of marginal associations.

The direct answer is therefore that EA receivers did not rely on five isolated apology prototypes. They shared an intensity-sensitive visual-signal pattern across domains, within which Fairness and Authority were most clearly separated and a small set of AUs marked selected domain boundaries. Response category defined the common architecture, while moral domain supplied selective differentiation.

4.6.2 Shared Response-Level Apology Pattern

The not-sorry–mild-sorry–strong-sorry progression extends Chapter 2's account of apology as a dynamic, multimodal configuration rather than one fixed set of facial expressions. Across the five EA moral

domains, not sorry was characterised by a compact AU profile led by AU43 and by upward gaze. Mild sorry broadened the AU and gaze repertoire, while strong sorry concentrated on AU4, AU14, AU12/AU25-12, AU20, AU12-6, and AU27i together with lateral gaze. The primary structure of EA receivers' judgements was whether, and how strongly, the face appeared sorry, with moral-domain information providing selective secondary differentiation.

Mild sorry was the least sharply bounded response category. Its mean validation accuracy of 48.8% was lower than the accuracies for not sorry and strong sorry, and it contained the largest signed-PMI repertoire. Its panel peaks and strict-majority counts were also less concentrated than those of strong sorry. The category therefore captured a heterogeneous region that overlapped with the adjacent responses; the evidence does not establish a uniformly intermediate state or a weak version of one consistent strong-sorry configuration.

The strong-sorry prevalence profile should be interpreted as a coordinated visual-signal configuration. AU4, AU14, AU12/AU25-12, AU20, AU12-6, AU27i, and lateral gaze were frequent at the population level, but no individual component can be equated directly with a discrete emotion, sincerity, or remorse. The results identify what receivers treated as stronger apology signals within the generated display space.

This response-level organisation corresponded descriptively with Chapter 3 in the prominence of AU43 and upward gaze for not sorry and in the broader heterogeneity of mild sorry. The leading strong-sorry configurations differed: EA emphasised AU4, AU14, AU12/AU25-12, and lateral gaze, whereas WC was led by AU20-related signals and downward gaze. These separate within-group patterns motivate, but do not replace, the direct comparison in Chapter 5.

Not-sorry models provide the baseline for this response pattern. Their compact AU and upward-gaze pattern helps identify the transition from the absence of apology to weaker and stronger reparative displays. However, not-sorry models were excluded from the formal domain classifier and ablation analysis. They therefore support the response-level progression but do not constitute formal evidence that moral domains were separable.

4.6.3 Selective Moral-Domain Differentiation

The classifier's balanced accuracy of 48.27% exceeded all 1,000 balanced accuracies reconstructed from the participant-label permutations ($p = .001$). The complete mild- and strong-sorry vectors therefore contained omnibus moral-domain information. Fairness-to-Fairness classification at .800 and Authority-to-Authority classification at .603 were the only diagonal cells exceeding their Bonferroni-corrected thresholds.

Fairness and Authority consequently provided the clearest EA domain-conditioned configurations. The ablation evidence for these domains was contrast-specific rather than class-specific: AU7 supported three directed separations ending in Fairness, while AU12R supported Sanctity/purity-to-Authority separation. Care/harm, Loyalty, and Sanctity/purity remained overlapping whole-vector classes. Loyalty nevertheless contained the only diagonal identifier, AU24, showing that a coherent component can occur within a configuration that is not globally separable at the corrected cell level.

EA and WC differed descriptively in their leading within-group cells, but separate significance results do not establish a receiver-group difference. Chapter 5 directly evaluates the WC–EA receiver-group contrasts using matched analyses.

4.6.4 Complementary Roles of Action Units and Gaze

AUs supplied all five significant EA ablation cells: one Loyalty diagonal identifier and four directed pair-separation effects. Gaze showed response-level prevalence and selected signed MI/PMI associations but no detectable conditional contribution to classifier cells under the present ablation settings.

All three ablation-active AUs occurred in the signed-PMI set, producing 9.4% signal-level overlap. Same-domain convergence applied only to Loyalty–AU24. AU7 and AU12R instead required two-endpoint pairwise interpretation. This distinction preserves the conditional question asked by ablation and prevents directed contrasts from being converted into independent domain markers.

Positive and negative PMI associations remained substantially broader than the ablation repertoire. Marginal enrichment or underrepresentation was therefore more common than a detectable contribution to multivariate classification, consistent with a shared repertoire that was selectively reweighted by response and moral context.

4.6.5 Theoretical and Methodological Implications

The results support a nested representational account in which reparative stance and apparent intensity provide the broadest organisation, while the damaged value or relationship is represented more selectively in complete configurations of present and absent AU and gaze components. This account explains how a stable response-level pattern can coexist with reliable Fairness and Authority classification and a small number of domain-specific signal contributions, without claiming a temporal order in receiver processing.

The profile also qualifies a simple opposition between *sad-sorry* and *smile-sorry*. Strong-sorry models contained AU12-related activity alongside AU4, AU14, AU20, AU27i, and lateral gaze. These labels can describe tendencies within the generated signal space, but the data do not show that one movement directly expresses sadness or happiness, nor that apologies divide into two discrete emotional types. Receiver judgements instead appear sensitive to configurations containing both smile-related and non-smile components.

The participant-level generative approach did not impose one fixed set of facial expressions for each domain. Population thresholding identified effects shared across receivers, classification and ablation tested domain information across complete configurations, and signed MI/PMI identified marginal enrichment and underrepresentation within response bins. Together, these analyses distinguished the shared response pattern from the smaller moral-domain differences.

The inference remains receiver-side. The models identify what EA receivers judged to look apologetic within the available generated AU and gaze space. They do not establish how offenders spontaneously produce apologies, whether a display is sincere, whether it elicits forgiveness, or whether it successfully repairs a relationship.

4.6.6 Receiver-Group Comparison Deferred to Chapter 5

Chapter 4 establishes the response-level and moral-domain pattern within the EA sample. Comparison with Chapter 3 shows that both samples displayed an intensity-sensitive response pattern and a heterogeneous mild-sorry repertoire, while differing descriptively in leading strong-sorry signals, the most separable domains, and the number of ablation effects. A within-EA analysis cannot determine whether those differences are cultural.

An effect reaching a corrected threshold in one receiver group but not another is not itself evidence

that the two effects differ. Chapter 5 therefore directly compares EA and WC receivers to test WC–EA receiver-group differences in signal prevalence, moral-domain separability, and the relative contributions of AUs and gaze. The EA–WC correspondences reported here remain descriptive unless supported by the direct tests in Chapter 5.

Chapter 4 is therefore the EA counterpart to the WC baseline established in Chapter 3. Its evidence for a shared response pattern with some moral-domain differences provides the within-EA basis for Chapter 5's formal receiver-group comparison.

4.6.7 Chapter Conclusion

EA receivers organised visual apology signals primarily by apparent apology intensity, not as five separate moral-domain prototypes. Moral context selectively differentiated this shared architecture: Fairness and Authority showed the clearest whole-vector separation, and AU7, AU12R, and AU24 supplied a small number of conditional contributions. AUs carried most moral-domain information, whereas gaze direction more consistently conveyed apparent apology intensity and contributed only selected marginal associations. Chapter 5 therefore directly tests whether WC and EA receivers share this organisation while differing in the configurations that realise it.

Chapter 5

Comparing Visual Apology Signals Across Western and East Asian Receiver Groups

5.1 Chapter Introduction

Chapters 3 and 4 examined visual apology signals within Western and East Asian receiver groups. This chapter brings the two datasets together to test their similarities and differences directly, within the same moral-domain contexts.

The moral-domain and receiver-centred rationale was introduced in those chapters. Cross-cultural research adds two reasons for a direct comparison. First, studies have reported cultural differences in moral identity, purity concerns and the moralisation of uncivilised conduct (Berniunas et al., 2022; Jia et al., 2019; Kim et al., 2012). Second, apology is evaluated through culturally organised expectations about face, responsibility and relationship repair (Gu, 1990; Y. Matsumoto, 1988). Studies have documented differences in apology strategies, their sensitivity to partner status, and their relationship-maintenance effects (Barnlund & Yoshioka, 1990; Komiya et al., 2020; Park & Guan, 2009).

These findings motivate asking whether a shared apparent-apology pattern is realised through different AU and gaze configurations across receiver groups. The comparison uses Western receivers judging Western faces and East Asian receivers judging East Asian faces. It tests whether complete configurations distinguish those sample-specific datasets, then examines the signal contributions and broader organisation underlying that distinction. The design preserves a common moral-domain framework while allowing the observed configurations to reflect the two groups' perceptual expectations.

5.2 The Current Study

Chapter 5 asked whether WC and EA receivers' apology-related configurations differed within the same moral-domain context, which signals contributed to those differences, and whether response category, receiver group or moral domain most strongly organised the signal space. Classification, ablation and signal-wise MI/PMI addressed the primary questions. Emotion-reference correspondence and Hamming geometry provided exploratory extensions. The comparison concerns the two receiver–face groupings in this design. Differences in classification accuracy among moral domains were descriptive.

5.3 Methods

5.3.1 Participants

Chapter 5 collected no new data. It reused the Chapter 3 and 4 datasets, reconstructing and revalidating participant-response models before comparison. The retained sample comprised 156 receivers (76 WC, 80 EA); EA participants 8, 9 and 26 were excluded from Fairness during validation. Recruitment and demographics are described in Chapters 3 and 4, Sections 3.3.1 and 4.3.1; realised model counts are reported in the present Results.

5.3.2 Procedure

The shared procedure is described in Chapter 3, Section 3.3.2. Chapter 5 analysed the resulting participant-by-response models: one binary AU and gaze vector per receiver and response category.

5.3.3 Stimuli

Stimulus generation followed Chapter 3, Section 3.3.3. Analyses used the same 47 binary signals (42 AUs and five gaze indicators), excluding AU temporal parameters and blushing. Face sets were receiver-matched: WC receivers viewed Western identities and EA receivers viewed East Asian identities. Matching followed evidence to avoid other-race effect (Dailey et al., 2010; Elfenbein & Ambady, 2002). Because receiver group and face-identity set varied together, the comparison could not isolate receiver culture from face identity.

5.4 Planned Data Analyses

5.4.1 Overview

The analyses used the population-thresholded participant-response models from Chapters 3 and 4. Input construction and internal validation followed Chapter 3, Section 3.4.2; descriptive summaries and primary tests were distinguished from exploratory extensions. The workflow was:

1. Stage 1: compare the separate WC and EA findings using matched accuracy and exploratory repertoire-breadth tests;
2. Stage 2: use primary classification and ablation to test direct receiver-group discrimination, alongside prevalence and matched-task analyses;
3. Stage 3: use primary MI/PMI to examine marginal signal associations and their variation across moral domains;
4. Stage 4: explore correspondence with literature-derived emotion and social-signal references; and
5. Stage 5: explore whole-profile geometry using Hamming distance and clustering.

5.4.2 Stage 1: Chapter 3 and Chapter 4 Response-First Comparison

Three aggregate comparisons summarised the earlier studies:

1. Five-domain balanced accuracy was compared over 50 design-matched D5 repetitions using the corrected repeated-holdout test of Nadeau and Bengio (2003), allowing for overlapping splits.
2. Separation-supporting and confusion-contributing ablation breadth used exact McNemar tests over 47 predictors.
3. Positive and negative MI/PMI discovery breadth used exact sign swaps over 141 signal–response opportunities, keeping each signal's three response values together.

Holm correction limited the chance of one or more false positives within each family. The D5 result retained its original two-comparison adjustment; separate families covered the two ablation tests and the two MI/PMI tests. Repertoire-breadth comparisons were exploratory feature-level analyses. Table 5.1 reports these comparisons.

5.4.3 Stage 2: Primary Receiver-Group Classification, Prevalence and Ablation

Classification (including matched comparisons), ablation and MI/PMI used MATLAB as described in Chapter 3. Python scripts using NumPy, SciPy and scikit-learn implemented the aggregate tests, prevalence analyses, literature-reference correspondence and Hamming analyses.

The inputs followed Chapters 3 and 4, with thresholds fitted within receiver-group-by-domain-by-response cells before classifier splitting. The analyses therefore evaluated the constructed representations under the validation limits in Chapter 3. Four linked tests addressed the direct comparison:

1. **Receiver-group classification.** A bagged-tree classifier with uniform priors distinguished WC from EA within each domain, using mild- and strong-sorry models and removing uniformly absent predictors. Fifty participant-level 80/20 holdouts kept each receiver's models together. Rows of the confusion matrix were normalised to 100%. Normalisation followed Chapter 3. Balanced accuracy averaged the two recalls. Inference used 1,000 participant-label permutations: Bonferroni correction covered four cells per domain, while plus-one balanced-accuracy values used Holm correction across five domains.

2. **Signal prevalence.** WC-minus-EA differences were summarised in percentage points and Cohen's *h*, as used in Chapter 3; positive values indicated greater WC prevalence. The 705 signal contrasts were descriptive. Panel tests summed squared studentised differences over 47 signals, using 5,000 participant-block label permutations within domain with thresholding repeated. Plus-one values used Holm correction across 15 response–domain panels.

3. **Matched classifier tasks.** Matched binary receiver-group (C2), binary pairwise-domain (D2) and joint five-domain (D5) tasks held participant selection, splits, rows per class and predictor dimensionality constant. The binary C2–D2 contrast used the corrected repeated-holdout test and retained the original two-comparison Holm adjustment. D5 provided descriptive context because its class count and decision problem differed. Appendix C reports the matched tasks and outcomes.

4. **Conditional ablation.** Each predictor was removed and the classifier refitted on the same 50 splits. Full-minus-ablated cell differences used 1,000 paired sign flips, with two-sided Bonferroni correction across four cells per predictor and no across-predictor correction. Identifiers supported correct classification; confusers impaired it. Off-diagonal roles concerned the WC–EA pair. Non-used predictors could remain in the thresholded inputs.

5.4.4 Stage 3: Primary MI/PMI Associations and Cross-Domain Variability

These analyses considered individual signals:

1. **Receiver-group MI/PMI.** Each of 47 signals was tested within response and domain, with one row per participant. MI used 1,000 label permutations and the signal's own 95th-percentile threshold; PMI localised states after the MI gate. These were candidate associations without correction across predictors. Positive and negative categories denoted overrepresented presence and absence. Signed associations were reported separately from thresholded prevalence.

2. **Cross-domain variability.** This analysis asked whether knowing the moral domain helped predict signal presence or absence. Within each of six receiver-group-by-response panels, raw MI was tested using 5,000 participant-block domain permutations with thresholding repeated. Plus-one values used Benjamini–Hochberg (BH) correction across 47 signals, controlling the expected proportion of false discoveries within that panel. Adjusted mutual information (AMI) provided a descriptive, chance-adjusted measure of association strength; prevalence range retained the frequency scale. The WC–EA mean-AMI contrast used 5,000 stratified participant bootstraps.

5.4.5 Stage 4: Exploratory Analyses: Emotion and Social-Signal Reference Correspondence

1. **Build the references.** Source representations (L3) encoded compatible AU and gaze evidence, keeping production and perception sources distinct. They supplied 17 named emotion or social-signal types (L2), grouped by source-defined construct or function, and three literature-defined valence categories (L1). The categories were determined before inspecting empirical correspondence; Table 5.12 records their literature rationale. Mappings retained source-supported alternatives where the literature did not resolve mouth opening, laterality or gaze direction.

2. **Compare profiles.** Pearson's r compared empirical prevalence with positive source support. A source used a common signal set across profiles within each branch, and within moral domain for M3, excluding a coordinate only when zero in that source and every profile. Correlations were averaged within source/evidence groups and then equally across groups for L2; L1 equally averaged eligible L2 scores. Source alternatives therefore contributed once. Correspondence described shared morphology rather than measured emotional states.

   Four empirical views retained the descriptive aliases: overall response profiles (M1); response profiles by receiver group, averaging domains (M2-1); response profiles by moral domain, averaging receiver groups (M2-2); and full condition profiles (M3). Relevant cells received equal weight, and these prevalence profiles were formed before the source correlations were calculated.

3. **Assess correspondence and contrasts.** L1/L2 scores and contrasts used participant-block bootstraps within the ten receiver-group-by-domain strata and restricted randomisation, reapplying prevalence cut-offs. Tests covered scores, response contrasts, receiver-group and domain differences, and the specified receiver-group-by-response contrasts. The 5,000 draws increased to 20,000 near decision thresholds. Scores and signed contrasts were tested two-sided against zero; domain omnibus tests used the upper tail. BH adjustment covered 28 families defined by empirical branch, literature level and scientific question. Retention required $q \leq .05$ and a 99% binomial interval for the Monte Carlo tail estimate that stayed on one side of the corresponding BH boundary. L3 remained an internal construction level.

5.4.6 Stage 5: Exploratory Analyses: Hamming Geometry and Clustering

1. **Construct the profiles.** Hamming distances compared 30 full-response and 20 apology-only condition profiles. A signal entered a profile when present in at least one thresholded participant model; globally absent predictors were removed. Average-linkage clustering and optimal leaf ordering displayed the geometry.

2. **Test label alignment.** Seven factor-alignment and mild–strong-change tests used 5,000 restricted participant-block permutations, plus-one values and one Holm family. Response labels were shuffled within participants, receiver-group labels within domain, and domain labels within receiver group. Thresholding and empty-predictor removal were repeated after relabelling.

3. **Assess cluster stability.** Cluster diagnostics considered $k = 2$ to 10 using 500 participant-block bootstraps with thresholding fixed. A cut required at least $\max\{2, \lceil .05n \rceil\}$ profiles in every cluster, where n was the number of condition profiles, and median adjusted Rand index of at least .75.

5.5 Results

The prespecified aggregate comparisons first synthesise the separate Chapter 3 and Chapter 4 findings. The direct analyses then examine receiver-group classification, literature-reference correspondence, and Hamming geometry.

5.5.1 Stage 1: Chapter 3 and Chapter 4 Within-Group Response-First Comparison

Both chapters showed the same ordinal response-validation pattern and located the broadest signal level differentiation in mild sorry. This common response structure is why the inferential comparison used whole-classifier and feature-repertoire summaries.

The prespecified aggregate comparisons are reported in Table 5.1. In the thresholded matched D5 design, EA exceeded WC by .039, but the difference was not significant after correction, $p_{\text{Holm}} = .628$. Separation-supporting ablation breadth also did not differ. The clearest aggregate contrast concerned confusion-contributing ablation breadth, which involved 15 WC predictors and no EA predictor, $p_{\text{Holm}} < .001$. Positive and negative MI/PMI discovery breadth did not differ. Thus, the separate chapters' descriptive differences did not translate into reliable aggregate differences in five-domain accuracy, separation-supporting ablation breadth, or signed MI/PMI breadth; the supported contrast was confined to confusion-contributing ablation roles.

Table 5.1

Compact WC–EA Comparisons of Moral-Domain Classification and Signal Repertoires.

Comparison	WC	EA	EA–WC	p	p_{Holm}
D5 balanced accuracy, thresholded	.431	.470	.039	.523	.628
Ablation separation-supporting breadth	5	3	-2	.688	.688
Ablation confusion-contributing breadth	15	0	-15	$< .001$	$< .001$
Positive MI/PMI discovery breadth	31	41	10	.264	.527
Negative MI/PMI discovery breadth	24	27	3	.773	.773

Note. D5 balanced accuracy used 50 design-matched repetitions and the Nadeau–Bengio corrected repeated-holdout paired test with $20/110 = 2/11$ as the test/train ratio; this pairing does not imply paired participants. Ablation breadth used exact McNemar tests over the fixed 47-predictor universe. MI/PMI breadth counted 141 signal–response opportunities while retaining each signal's three responses as one block in an exact sign-swap test. The reported D5 value retains its original two-comparison Holm adjustment. Holm correction was applied separately within the two ablation and two MI/PMI tests. The feature-level ablation and MI/PMI comparisons are exploratory repertoire analyses, not participant-level receiver-group inference.

Taken together, the response-first evidence establishes the common starting point for direct comparison: WC and EA shared an architecture organised by apparent apology intensity. The aggregate tests did not

support receiver-group differences in overall five-domain accuracy, separation-supporting ablation breadth, or signed MI/PMI breadth; the only supported contrast concerned confusion-contributing ablation breadth. The following analyses therefore ask whether this shared response structure is realised through different complete configurations within each moral domain.

5.5.2 Stage 2: Primary Direct WC–EA Receiver-Group Classification Results

As Table 5.2 and Figure 5.1 show, the complete apology-signal vectors differentiated the two sample-specific datasets within every moral domain. This contrast compares the WC and EA receiver groups defined above. All ten diagonal cells exceeded their within-domain Bonferroni-corrected permutation thresholds, whereas none of the off-diagonal cells did. The corrected diagonal thresholds ranged from .657 to .723 for WC and from .663 to .735 for EA. The upper-tail permutation values for the significant diagonal cells were $p_{\text{perm}} = .001$ in nine cases and $p_{\text{perm}} = .002$ for the WC Sanctity/purity cell. The complementary domain-level test led to the same conclusion: no label permutation produced a balanced accuracy as large as the observed value in any domain. Under the plus-one calculation, the resulting estimate was $1/1001 = .000999$, reported as $p_{\text{perm}} = .001$. Because all five domain-level values were at this Monte Carlo resolution floor, Holm correction yielded $5/1001 = .004995$, reported as $p_{\text{Holm}} = .005$ for every domain.

Table 5.2

Direct WC–EA Receiver-Group Classification Within Each Moral Domain.

Moral domain	WC → WC	WC → EA	EA → WC	EA → EA	BA	Holm *p*
Fairness	.900*	.100	.010	.990*	.9450	.005
Care/harm	.787*	.213	.127	.873*	.8300	.005
Authority	.830*	.170	.053	.947*	.8883	.005
Loyalty	.950*	.050	.063	.937*	.9433	.005
Sanctity/purity	.765*	.235	.043	.957*	.8608	.005

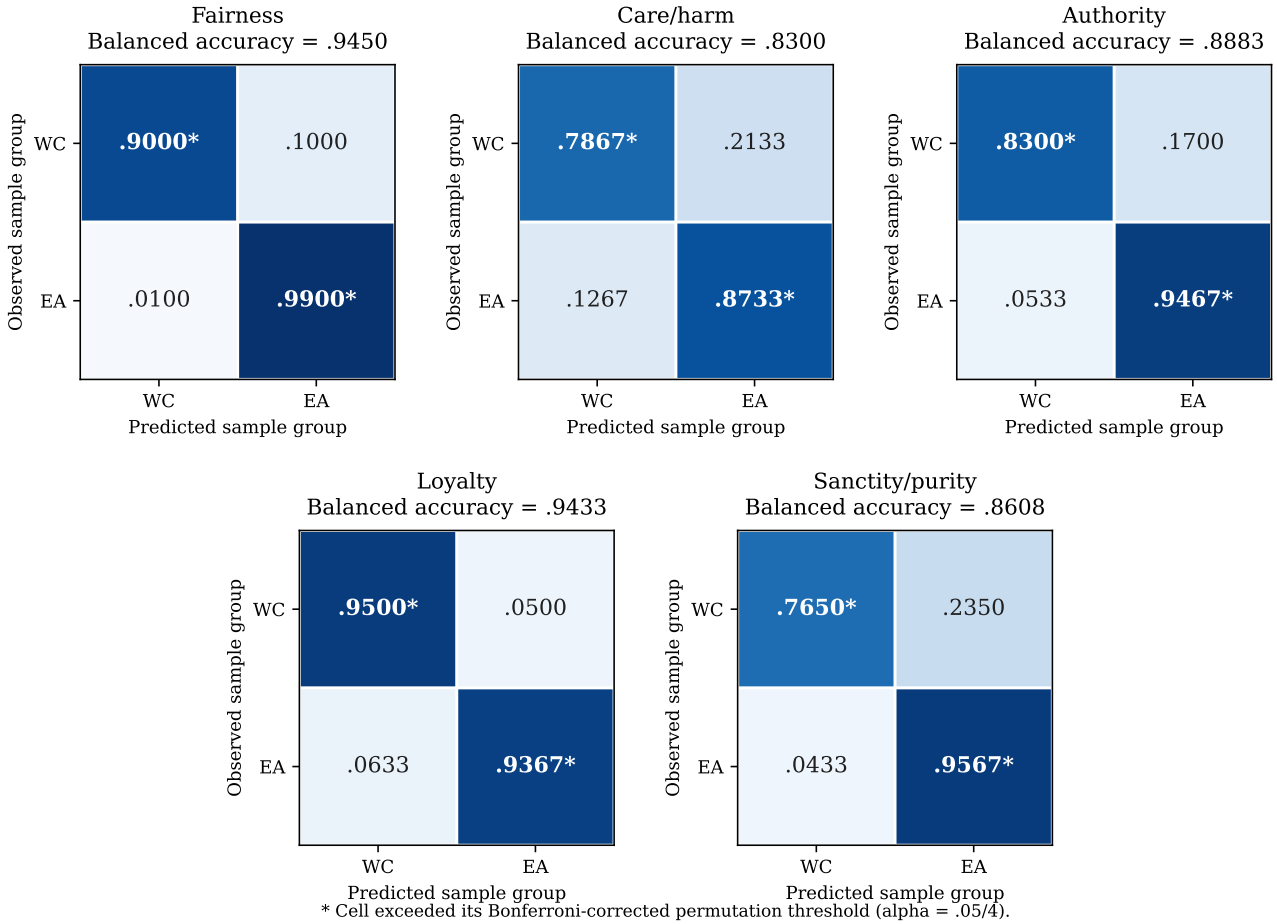
Note. Rows denote the observed receiver group and columns denote the predicted receiver group. Asterisks mark cells above their within-domain Bonferroni-corrected permutation thresholds; the final column gives the Holm-adjusted domain-level balanced-accuracy permutation value. Probabilities are row-normalised and averaged over 50 participant-grouped holdouts. Each cell was evaluated against 1,000 label permutations, each of which also contained 50 participant-grouped holdouts; correction was applied across the four cells within a domain. Balanced accuracy was reconstructed from the same permutation matrices and Holm corrected across the five domain-level tests. With 1,000 permutations and the plus-one calculation, the smallest attainable estimate is $1/1001 = .000999$; all five tests attained this value, so their equal Holm-adjusted value is $5/1001 = .004995$, reported as .005.

Fairness and Loyalty produced the highest balanced accuracies, at .9450 and .9433, respectively (Figure 5.1). Both Fairness recalls were high (.900 for WC and .990 for EA), whereas the Loyalty recalls were nearly equal (.950 and .937). Authority had a balanced accuracy of .8883, with recalls of .830 for WC and .947 for EA. Sanctity/purity had a balanced accuracy of .8608; its EA recall was .957, whereas its WC recall was .765. Care/harm had the lowest balanced accuracy, .8300, but both of its diagonal cells and its domain-level balanced accuracy remained significant. Thus, inferential evidence for sample-group differentiation was present in every domain, while the size and symmetry of the classification rates varied across domains. The balanced-accuracy tests compare each domain with its own label-permutation null; they do not test differences among the five domains.

Input Models and Prevalence Findings. The direct analyses used 156 validated participants and 468 participant-by-response rows. The classifiers retained only mild- and strong-sorry models, giving 312

Figure 5.1

Row-Normalised Direct Receiver-Group Classification Matrices for the Five Moral Domains.



Note. Asterisks mark cells above their within-domain Bonferroni-corrected permutation thresholds. Each row shows the predicted classifications within one true receiver group and sums to 1 (100%). The observed balanced accuracy is shown above each matrix; its domain-level permutation inference is reported in Table 5.2.

rows overall and 30–37 non-zero predictors across the five moral-domain classifiers. Table C1 reports the domain-specific analysis inputs.

Figure 5.2 pools the validated, population-thresholded models across moral domains to show the overall response profile of each sample group. The domain-resolved prevalence configurations are reported in Figures 3.1 and 4.1, avoiding their duplication in this chapter.

Direct Signal-Prevalence Contrasts. The 15 combinations of three response categories and five moral domains were then compared signal by signal. Figure 5.3 shows WC minus EA prevalence for all 47 predictors in the population-thresholded vectors. Positive values indicate that a signal occurred in a larger proportion of WC participant models and negative values indicate greater EA prevalence. The complete analysis contained 705 contrasts. Cohen's *h* was used to place differences near zero and one on a common proportion-difference scale; these effect sizes were descriptive and were not treated as 705 separate tests.

The largest absolute contrast in each response-by-domain panel is shown in Table 5.3. Not-sorry contrasts included greater WC prevalence of AU9 in Fairness and Authority, AU10Open in Care/harm, and AU25-12 in Sanctity/purity, whereas No movement was more prevalent in EA Loyalty. Mild sorry showed greater WC prevalence of AU20L in Fairness and Authority and AU12R in Care/harm, while AU24 in Loyalty and No movement in Sanctity/purity were more prevalent in EA. Strong sorry showed greater WC prevalence of AU13 in Fairness, AU12R in Care/harm, and AU20L/AU20R in Authority and Sanctity/purity; Looking left was more prevalent in EA Loyalty. Across the response-specific upper-quartile tables, after zero effects were excluded and boundary ties were retained, 17 contrasts were retained for not sorry, 41 for mild sorry, and 38 for strong sorry.

Table 5.3

Largest Absolute Population-Thresholded WC–EA Prevalence Contrast Within Each Response and Moral Domain.

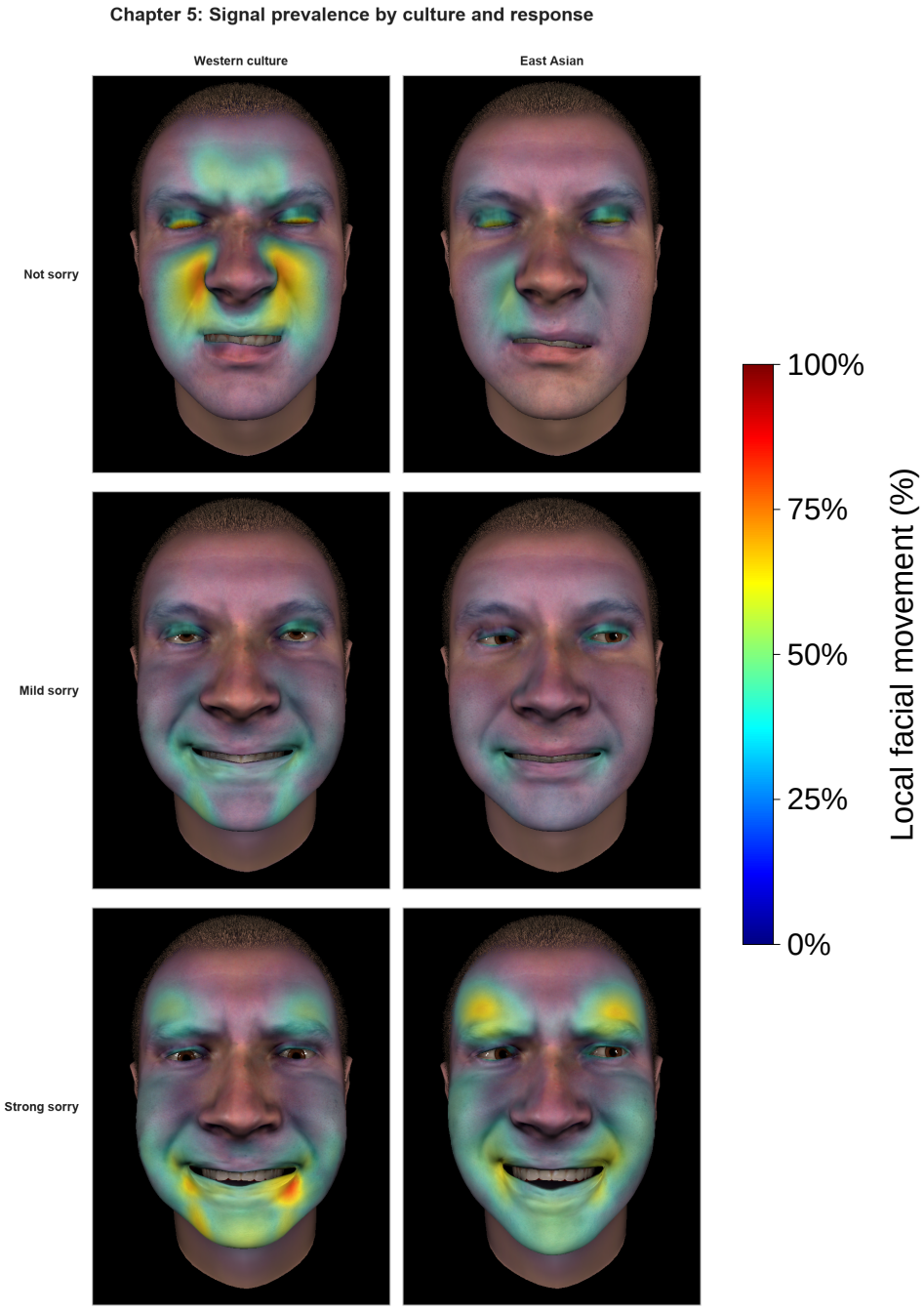
Response	Moral domain	Signal	WC prevalence	EA prevalence	WC–EA pp / Cohen *h*
Not sorry	Fairness	AU9	.800	.286	51.4 / 1.086
Not sorry	Care/harm	AU10Open	.467	.000	46.7 / 1.504
Not sorry	Authority	AU9	.929	.250	67.9 / 1.553
Not sorry	Loyalty	No movement	.000	.562	−56.2 / −1.696
Not sorry	Sanctity/purity	AU25-12	.500	.000	50.0 / 1.571
Mild sorry	Fairness	AU20L	.500	.000	50.0 / 1.571
Mild sorry	Care/harm	AU12R	.600	.000	60.0 / 1.772
Mild sorry	Authority	AU20L	.714	.000	71.4 / 2.014
Mild sorry	Loyalty	AU24	.000	.562	−56.2 / −1.696
Mild sorry	Sanctity/purity	No movement	.000	.625	−62.5 / −1.823
Strong sorry	Fairness	AU13	.800	.000	80.0 / 2.214
Strong sorry	Care/harm	AU12R	.533	.000	53.3 / 1.638
Strong sorry	Authority	AU20L	.714	.000	71.4 / 2.014
Strong sorry	Loyalty	Looking left	.000	.625	−62.5 / −1.823
Strong sorry	Sanctity/purity	AU20R	.714	.000	71.4 / 2.014

Note. pp denotes percentage points. Cohen's *h* is a descriptive effect size for the WC–EA difference in prevalence proportions; positive values indicate greater WC prevalence and negative values greater EA prevalence. Values describe the separately population-thresholded vectors. Individual effects localise the panel-level contrasts and were not assigned separate *p*-values.

The threshold-aware omnibus tests evaluated each 47-signal panel jointly (Table 5.4). Strong-sorry Fairness was the only panel that survived Holm correction, $T = 160.65$, $p = .0004$, $p_{\text{Holm}} = .0060$; the remaining thresholded panels had adjusted values from .0896 to .3359.

Figure 5.2

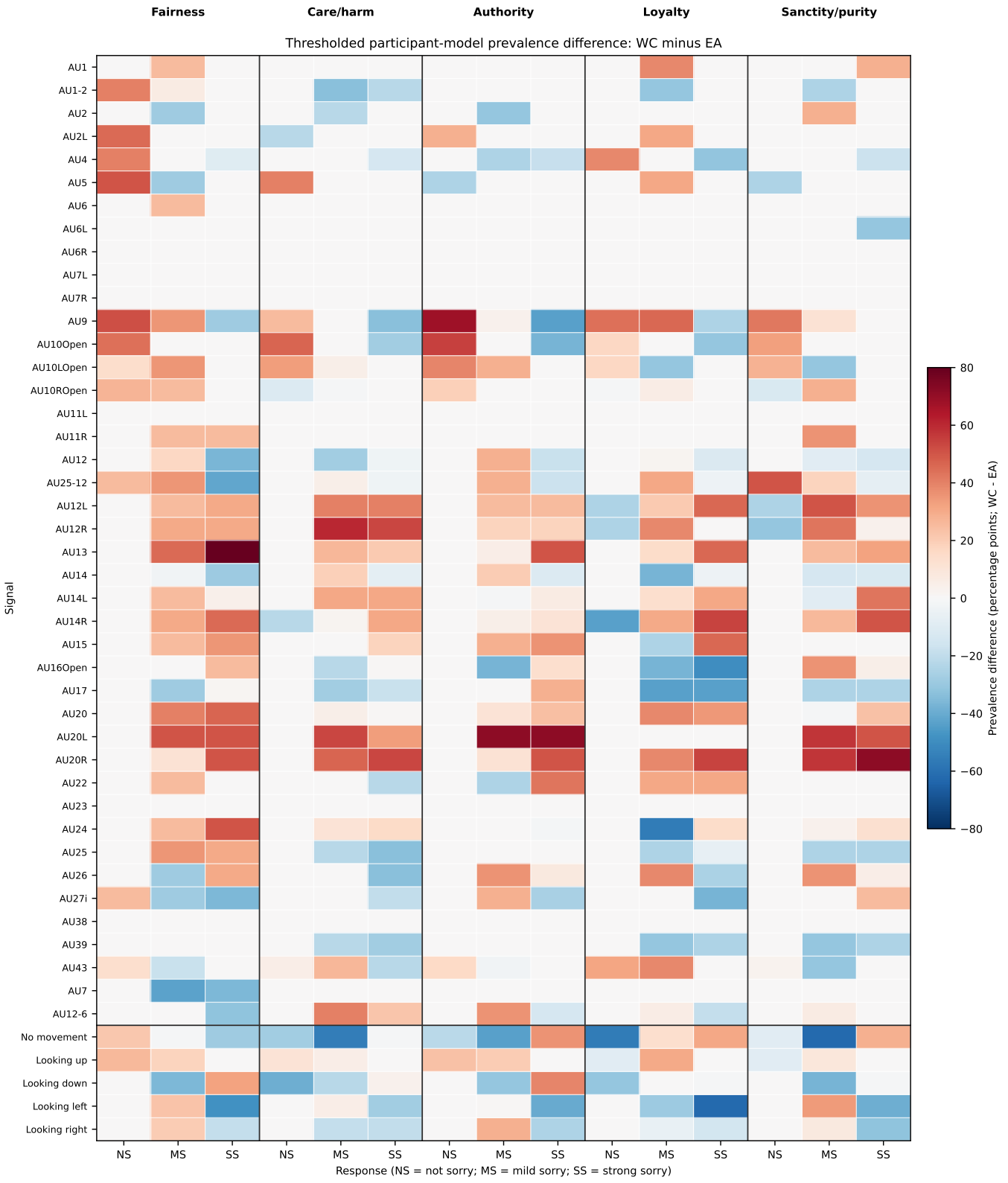
Pooled Population-Level Signal Prevalence by Response Category and Receiver Group.



Note. Rows show not sorry, mild sorry, and strong sorry; columns show WC and EA models. Colour highlights facial changes from AUs weighted by pooled prevalence. Only the most prevalent gaze state is displayed; ties are drawn with neutral-looking eyes. The bar shows the normalised local facial movement scale (0–100%); colours fade into the facial texture at lower values.

Figure 5.3

WC Minus EA Population-Thresholded Prevalence for All 47 Signals in Each Response-by-Domain Panel.



Note. Positive cells indicate greater WC prevalence and negative cells indicate greater EA prevalence. The display localises the panel-level permutation results; no signal-wise significance is assigned.

Table 5.4

Primary Thresholded Panel-Level Permutation Tests of WC–EA Signal-Prevalence Differences.

Response	Moral domain	T	p	Holm p
Not sorry	Fairness	66.18	.0110	.1430
Not sorry	Care/harm	47.24	.0536	.3215
Not sorry	Authority	42.53	.0298	.2683
Not sorry	Loyalty	50.77	.0392	.3135
Not sorry	Sanctity/purity	36.22	.0716	.3359
Mild sorry	Fairness	156.25	.1014	.3359
Mild sorry	Care/harm	117.40	.0672	.3359
Mild sorry	Authority	99.67	.1634	.3359
Mild sorry	Loyalty	140.97	.0244	.2683
Mild sorry	Sanctity/purity	133.33	.0194	.2328
Strong sorry	Fairness	160.65	.0004	.0060
Strong sorry	Care/harm	97.49	.0982	.3359
Strong sorry	Authority	104.96	.0266	.2683
Strong sorry	Loyalty	123.59	.0064	.0896
Strong sorry	Sanctity/purity	99.49	.0418	.3135

Note. T is the sum across 47 signals of squared WC–EA prevalence differences divided by their pooled binomial sampling variances; zero-variance signals contribute zero. Complete participants were shuffled between sample groups within moral domain, with all three response rows moving together and group sizes preserved. Population thresholding was repeated after every primary shuffle. The analysis used 5,000 permutations and a 15-test Holm correction.

Matched C2–D2–D5 Comparison. The design-matched analysis controlled participant selection, participant-level splits, rows per class, and predictor dimensionality across the classifier tasks. On the primary thresholded vectors, receiver-group classification exceeded pairwise moral-domain classification by .1299, 95% CI [.0794, .1804], $t(49) = 5.166$, $p_{\text{Holm}} < .001$. This result supports greater WC–EA receiver-group than pairwise-domain separability within the matched model vectors. The original two-comparison Holm adjustment is retained. Appendix C gives the matched task counts and corrected C2–D2 comparison.

Ablation Analysis. Here each classifier distinguished WC from EA within one moral domain. An ablation-active signal reliably changed at least one classifier cell when removed and the classifier refitted, with the other signals available. A diagonal identifier helped correct receiver-group classification, so removing it reduced correct classifications; a diagonal confuser showed the opposite pattern. A pairwise identifier instead helped separate the directed WC–EA pair: removing it increased that misclassification. The ablation result was sparse relative to the classification result. Across the five domain-specific classifiers, five unique AUs had a significant role and no gaze predictor did (Table 5.5; Figure 5.4). Fairness contained one active predictor, AU16Open, which acted as a WC diagonal confuser. Care/harm contained no ablation-active predictor. In Authority, AU15 and AU20L each supported both the WC and EA diagonals and both off-diagonal WC–EA contrasts. In Loyalty, AU20 supported the EA diagonal and the EA-to-WC contrast; AU20R showed the same pattern in Sanctity/purity. The off-diagonal effects indicate pairwise separation and are not independent identifiers for both sample groups.

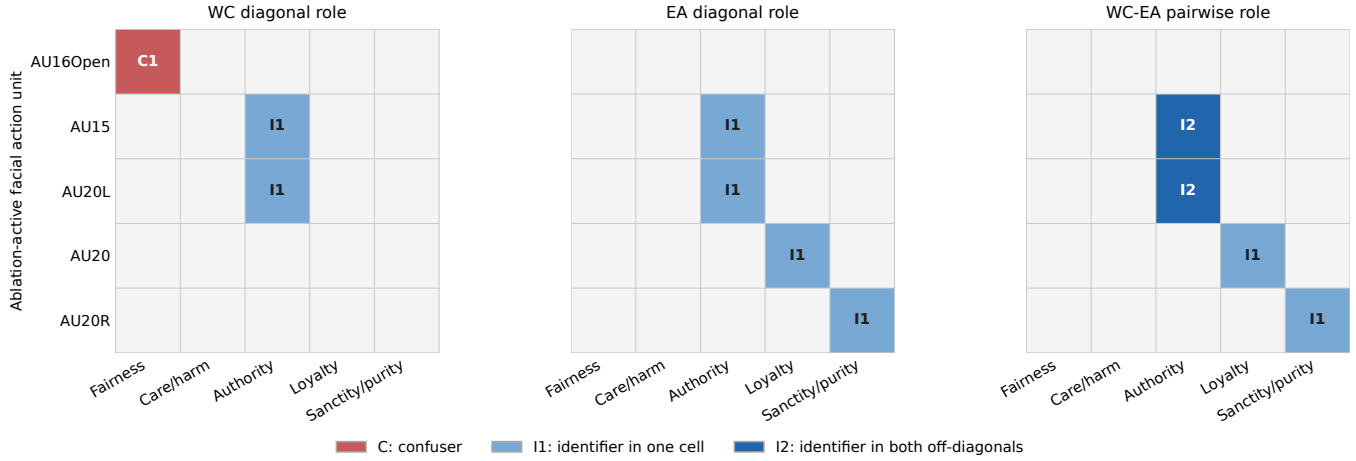
Table 5.5

Ablation-Active Signals in the Direct Receiver-Group Classifiers, Distinguishing Class-Specific Diagonal Roles From WC–EA Pairwise Off-Diagonal Roles.

Moral domain	Retained predictors	Globally non-used	Reliable ablation role(s)
Fairness	37	36	AU16Open: WC diagonal confuser
Care/harm	34	34	None
Authority	30	28	AU15 and AU20L: WC and EA diagonal identifiers; identifiers in both pairwise off-diagonals
Loyalty	35	34	AU20: EA diagonal identifier; EA-to-WC pairwise identifier
Sanctity/purity	34	33	AU20R: EA diagonal identifier; EA-to-WC pairwise identifier

Figure 5.4

Direct Receiver-Group Ablation Roles Separated by Classifier-Cell Type.



Note. The first two panels show diagonal roles: removing an identifier reduced correct classification of the labelled receiver group, whereas removing a confuser improved it. In the third panel, pairwise identifiers supported WC–EA separation: removing them increased the directed error. Blank cells indicate no reliable role. Care/harm contained no active signal, and no gaze predictor had an ablation role.

The direct analyses add a second level to the story: complete visual-signal configurations differentiated WC and EA within every moral domain, even though corrected prevalence and ablation evidence was concentrated in selected response–domain contexts and AUs. In the matched binary comparison, receiver-group classification exceeded pairwise moral-domain classification within these vectors, whereas the D5 contrast did not support a WC–EA difference in overall five-domain accuracy. The receiver-group difference is therefore robust at the configuration level but not a uniform signal-by-signal divergence; the next section localises the smaller set of candidate contributions.

5.5.3 Stage 3: Primary Ablation and MI/PMI Findings

Integrated Ablation-Role and MI/PMI Interpretation. Ablation and MI/PMI localised different parts of the same receiver-group classification result. The five ablation-active AUs were conditional con-

tributors: their removal changed a classifier cell while the remaining predictors stayed available. MI/PMI instead identified marginal over-representation of signal presence or absence within a response and domain. Four of the five direct ablation-active domain–signal roles also had a signed MI/PMI association in the same domain. Authority AU15 and AU20L, Loyalty AU20, and Sanctity/purity AU20R showed this convergence, whereas the Fairness confuser AU16Open did not. The overlap supports the relevance of selected signals without implying that the two methods estimated the same mechanism.

Figure 5.5 localises these conditional contributions while retaining their class-specific or pairwise classifier roles.

The compact Chapter 3–4 comparison supplied a broader repertoire-level qualification. Separation-supporting ablation breadth did not differ between the two within-group classifiers, but confusion-contributing roles were more widespread in WC than EA. In the direct WC–EA classifiers reported here, by contrast, only Fairness AU16Open was a confuser; the remaining active roles supported a class diagonal, pairwise separation, or both. The broad within-WC confusion repertoire therefore describes organisation of the WC moral-domain classifier, not a general tendency for WC signals to confuse every classification problem.

Table 5.6 resolves the five domain–signal roles into the 13 significant classifier cells. AU16Open was the only confuser: removing it increased correct WC classification in Fairness by .0325. In Authority, removing AU15 or AU20L reduced both diagonal probabilities and increased both cross-group errors. In Loyalty, removing AU20 reduced EA recall and increased EA-to-WC error by .0467; AU20R produced the same two-cell pattern in Sanctity/purity. No gaze predictor showed a detectable conditional contribution.

Table 5.6

Ablation-Active Signals and Their Significant Direct Receiver-Group Classification Cells.

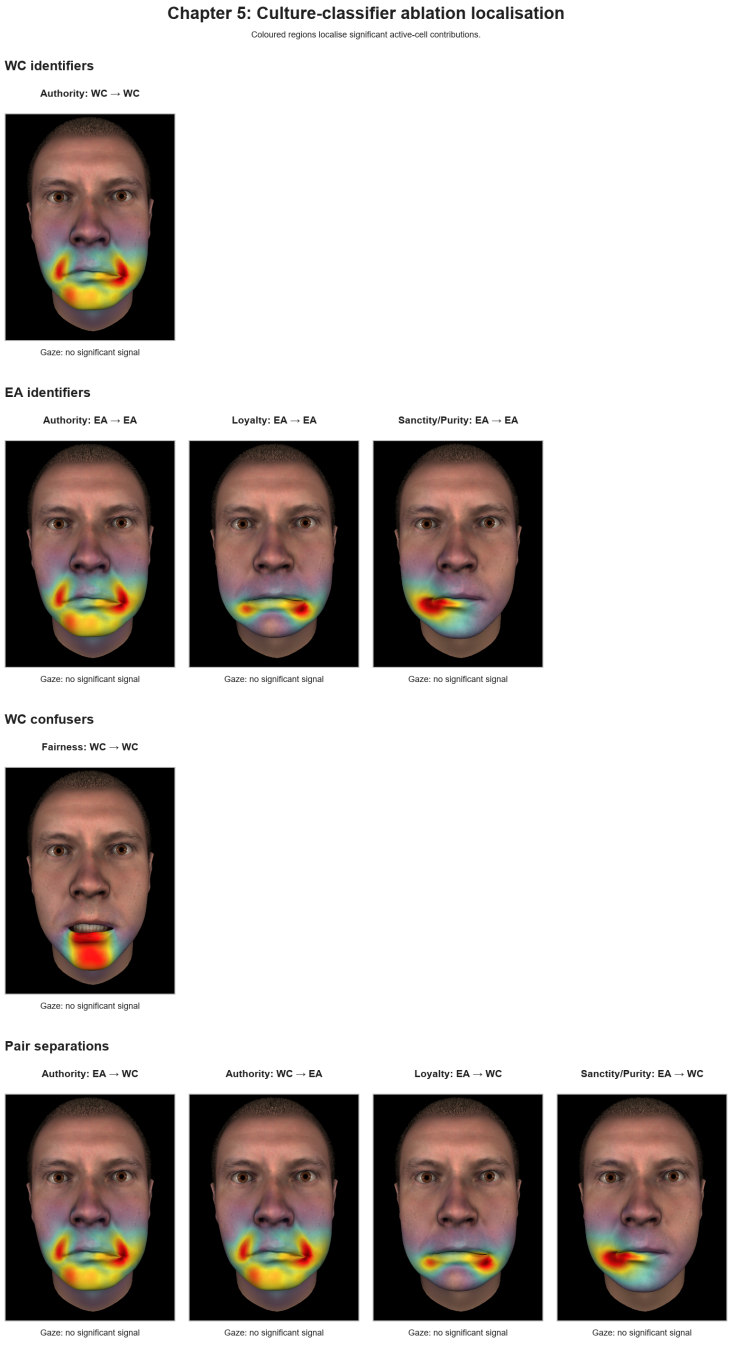
Domain	Signal	Observed	Predicted	Δp	Effect of removal	Role assignment
Fairness	AU16Open	WC	WC	−.0325	Increased correct WC classification	WC confuser
Authority	AU15	WC	WC	.0800	Reduced correct WC classification	WC identifier
Authority	AU15	WC	EA	−.0800	Increased WC-to-EA error	WC–EA pairwise identifier
Authority	AU15	EA	WC	−.0467	Increased EA-to-WC error	WC–EA pairwise identifier
Authority	AU15	EA	EA	.0467	Reduced correct EA classification	EA identifier
Authority	AU20L	WC	WC	.1350	Reduced correct WC classification	WC identifier
Authority	AU20L	WC	EA	−.1350	Increased WC-to-EA error	WC–EA pairwise identifier
Authority	AU20L	EA	WC	−.1600	Increased EA-to-WC error	WC–EA pairwise identifier
Authority	AU20L	EA	EA	.1600	Reduced correct EA classification	EA identifier
Loyalty	AU20	EA	WC	−.0467	Increased EA-to-WC error	WC–EA pairwise identifier
Loyalty	AU20	EA	EA	.0467	Reduced correct EA classification	EA identifier
Sanctity/purity	AU20R	EA	WC	−.0467	Increased EA-to-WC error	WC–EA pairwise identifier
Sanctity/purity	AU20R	EA	EA	.0467	Reduced correct EA classification	EA identifier

Note. Δp is the row-normalised full-model cell probability minus the probability after removing the signal, averaged across 50 matched holdouts. Positive values mean that removal reduced the displayed cell; negative values mean that removal increased it. On diagonals, removing an identifier reduced correct receiver-group classification; removing a confuser improved it. For an off-diagonal pairwise identifier, removal increased that directed WC–EA error. This is a pair-separation role, not an independent identifier of each receiver group.

Specific and Shared Signal Repertoires. The primary MI/PMI analysis tested the 47 predictors separately within each of the 15 moral-domain-by-response panels. The signed output contained 281 receiver-group-associated rows: 146 positive and 135 negative. Here, a positive association indicates that signal presence was disproportionately associated with the named sample group, whereas a negative association indicates that signal absence was disproportionately associated with that group. These signed associations

Figure 5.5

Facemap Localisation of Ablation-Active Signals in the Direct Receiver-Group Classifiers.



Note. Removing a diagonal identifier reduced correct WC or EA classification; removing the Fairness confuser improved it. Pair-separation roles concern the named WC–EA error. Care/harm contained no active signal. Colours locate the selected AUs, each displayed at maximum rendering amplitude.

were kept separate from population-thresholded presence, which records whether a signal occurred in at least one retained model within a panel.

WC contained 97 positive and 49 negative associations, while EA contained 49 positive and 86 negative associations. Mild sorry had the broadest signed repertoire in both groups, with 33 unique WC and 32 unique EA signals and 69 and 63 receiver-group-associated rows, respectively. Strong sorry involved 25 WC and 24 EA signals, and not sorry involved 13 signals in each group. The thresholded-presence totals followed a related but non-identical pattern: 44 signal–domain cells were present for not sorry in each group, compared with 110 WC and 97 EA cells for mild sorry and 111 WC and 98 EA cells for strong sorry (Table 5.7).

Table 5.7

Response-Level Direct MI/PMI Associations and Independently Derived Thresholded Presence.

Sample group	Response	Positive associations	Negative associations	Unique signed signals	Thresholded-present cells
WC	Not sorry	17	5	13	44
WC	Mild sorry	43	26	33	110
WC	Strong sorry	37	18	25	111
EA	Not sorry	5	16	13	44
EA	Mild sorry	26	37	32	97
EA	Strong sorry	18	33	24	98

Note. "Thresholded-present cells" counts non-zero signal–domain cells across the five domains for the named response (maximum 235). It is a prevalence-based quantity and is not inferred from the positive or negative association category.

WC–EA Receiver-Group Contrasts. The 281 receiver-group-associated rows represented 147 unique domain–response–signal relations. Of these relations, 134 (91.2%) formed reciprocal contrasts across the two groups: 85 paired WC-associated signal presence with EA-associated signal absence, and 49 paired WC-associated signal absence with EA-associated signal presence. The remaining 13 relations (8.8%) were one-sided, comprising 12 WC-presence associations and one EA-absence association. The two group-specific entries of a reciprocal relation therefore localised opposite states of the same signal contrast rather than two unrelated findings.

Figure 5.6 displays the 134 reciprocal relations once on their group-specific ON side, with the corresponding other-group OFF association stated in the column heading. These opposing state associations constitute valid WC–EA receiver-group differences in the present design. The 13 one-sided relations are retained separately in Figure C1 in Appendix C.

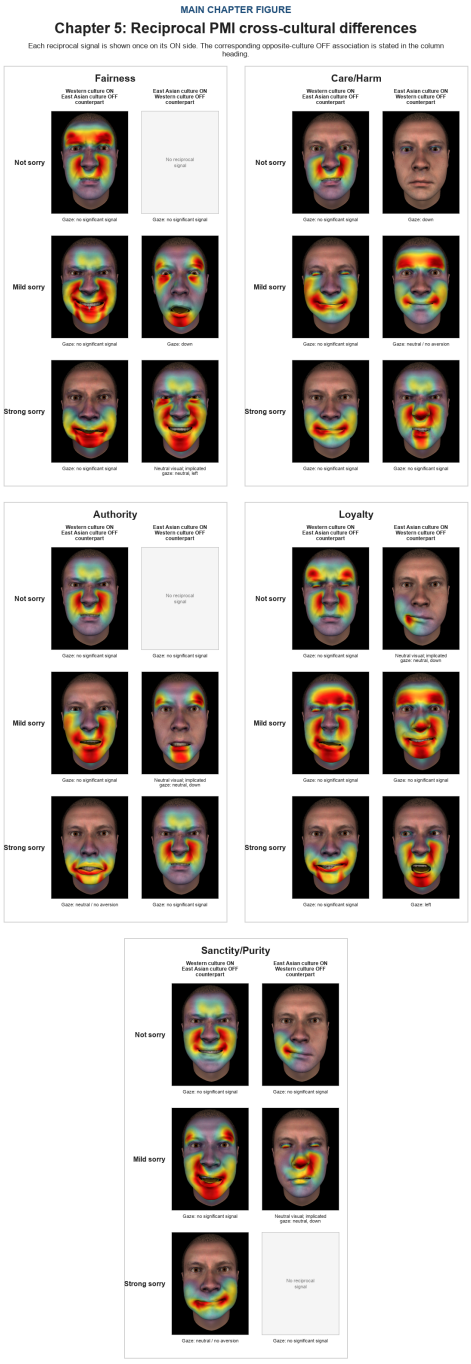
Table 5.8 reports the complete panel-level count check. The broadest Fairness panels occurred in mild and strong sorry, while Care/harm carried fewer signed rows. Every moral domain nevertheless contained both signed associations and thresholded-present signals in each response category.

The complete WC and EA membership tables report the signal membership. The accompanying signed-association heatmap displays the same pattern across all 47 candidate signals. Every listed entry passed both the signal-level MI gate and the relevant state-specific PMI threshold.

WC-Associated Repertoire. As Table 5.9 shows, WC associations were predominantly presence-based: 97 of 146 WC rows indicated overrepresented signal presence and 49 indicated overrepresented absence.

Figure 5.6

Facemap Localisation of Reciprocal WC–EA Receiver-Group MI/PMI Differences.



Note. Each relation is shown for the group in which signal presence was overrepresented; the column heading names the group in which absence was overrepresented. Colours locate the selected AUs, each displayed at maximum rendering amplitude. Blank panels contain no retained reciprocal relation.

Table 5.8

Signed MI/PMI Associations and Independently Thresholded Signal Presence in All 15 Domain-by-Response Panels.

Moral domain	Response	WC	EA
Fairness	Not sorry	6 / 0 / 13	0 / 6 / 7
Fairness	Mild sorry	9 / 7 / 27	7 / 10 / 16
Fairness	Strong sorry	11 / 7 / 23	7 / 11 / 15
Care/harm	Not sorry	3 / 1 / 8	1 / 3 / 8
Care/harm	Mild sorry	6 / 3 / 18	3 / 6 / 21
Care/harm	Strong sorry	4 / 4 / 23	4 / 4 / 26
Authority	Not sorry	3 / 0 / 8	0 / 2 / 8
Authority	Mild sorry	8 / 4 / 21	4 / 4 / 19
Authority	Strong sorry	7 / 2 / 24	2 / 6 / 19
Loyalty	Not sorry	3 / 3 / 7	3 / 3 / 11
Loyalty	Mild sorry	11 / 7 / 22	7 / 8 / 20
Loyalty	Strong sorry	8 / 5 / 19	5 / 6 / 19
Sanctity/purity	Not sorry	2 / 1 / 8	1 / 2 / 10
Sanctity/purity	Mild sorry	9 / 5 / 22	5 / 9 / 21
Sanctity/purity	Strong sorry	7 / 0 / 22	0 / 6 / 19

Note. Each entry is positive associations / negative associations / thresholded-present cells.

EA-Associated Repertoire. As Table 5.10 shows, EA had the complementary balance: 49 of 135 rows indicated overrepresented signal presence and 86 indicated overrepresented absence; reciprocal rows represent the opposite signal state of the corresponding WC contrast.

Shared Thresholded Presence and Non-Detection. As Figure 5.7 shows, shared occurrence was evaluated independently of signed association. Among the 235 possible signal–domain cells in each response, both groups had non-zero thresholded prevalence in 32 not-sorry cells, 58 mild-sorry cells, and 74 strong-sorry cells. WC-only presence occurred in 12, 52, and 37 cells, respectively, and EA-only presence occurred in 12, 39, and 24. A cell that was present in both groups was therefore not necessarily statistically unassociated with receiver group, because prevalence could still differ; conversely, a blank MI/PMI cell indicated that the association criteria were not met and did not demonstrate equal prevalence or a shared signal. Shared presence and non-detection were retained as separate concepts throughout the interpretation.

Action Units. Action units accounted for 252 of the 281 signed rows and involved 34 unique AUs. WC contained 94 positive and 37 negative AU associations; EA contained 37 positive and 84 negative AU associations. AU20L and AU20R were the most recurrent paired relations: each occurred in eight WC-positive and eight EA-negative domain–response cells. AU9, AU12L, and AU12R also recurred across multiple panels. This AU result was substantially broader than the five-AU ablation set because MI/PMI evaluated marginal signal associations within response panels, whereas ablation evaluated conditional contributions to the multivariate receiver-group classifier fitted to the apology rows.

Gaze Direction. Gaze accounted for 29 signed rows involving three states: no movement, looking down, and looking left. WC had three positive and 12 negative gaze associations, while EA had 12 positive and two negative associations. Looking down formed five reciprocal WC-absence/EA-presence relations: Fairness mild sorry, Care/harm not sorry, Authority mild sorry, Loyalty not sorry, and Sanctity/purity mild

Table 5.9

Western Receiver-Group (WC) Signed MI/PMI Associations by Response and Moral Domain.

Response	Domain	Positive AUs	Positive gaze	Negative AUs	Negative gaze
Not sorry	Fairness	AU1-2, AU2L, AU4, AU5, AU9, AU10Open	–	–	–
Not sorry	Care/harm	AU5, AU10Open, AU10LOpen	–	–	Looking down
Not sorry	Authority	AU2L, AU9, AU10Open	–	–	–
Not sorry	Loyalty	AU4, AU9, AU43	–	AU14R	No movement, Looking down
Not sorry	Sanctity/ purity	AU9, AU25-12	–	AU12R	–
Mild sorry	Fairness	AU9, AU10LOpen, AU25-12, AU12R, AU13, AU14R, AU20, AU20L, AU25	–	AU2, AU5, AU17, AU26, AU27i, AU7	Looking down
Mild sorry	Care/harm	AU12L, AU12R, AU20L, AU20R, AU43, AU12-6	–	AU1-2, AU12	No movement
Mild sorry	Authority	AU10LOpen, AU12, AU25-12, AU15, AU20L, AU26, AU27i, AU12-6	–	AU2, AU16Open	No movement, Looking down
Mild sorry	Loyalty	AU1, AU2L, AU5, AU9, AU25-12, AU12R, AU20, AU20R, AU22, AU26, AU43	–	AU1-2, AU10LOpen, AU14, AU16Open, AU17, AU24, AU39	–
Mild sorry	Sanctity/ purity	AU2, AU10ROpen, AU11R, AU12L, AU12R, AU16Open, AU20L, AU20R, AU26	–	AU10LOpen, AU39, AU43	No movement, Looking down
Strong sorry	Fairness	AU12L, AU12R, AU13, AU14R, AU15, AU20, AU20L, AU20R, AU24, AU25, AU26	–	AU9, AU12, AU25-12, AU27i, AU7	No movement, Looking left
Strong sorry	Care/harm	AU12L, AU12R, AU20L, AU20R	–	AU9, AU10Open, AU25, AU39	–
Strong sorry	Authority	AU13, AU15, AU17, AU20L, AU20R, AU22	No movement	AU9, AU10Open	–
Strong sorry	Loyalty	AU12L, AU13, AU14L, AU14R, AU15, AU20R, AU22	No movement	AU10Open, AU16Open, AU17, AU27i	Looking left
Strong sorry	Sanctity/ purity	AU1, AU12L, AU14L, AU14R, AU20L, AU20R	No movement	–	–

Note. Positive denotes overrepresentation of signal presence. Negative denotes overrepresentation of signal absence and is not a numerically negative PMI value. A dash denotes no significant association in that direction.

Table 5.10

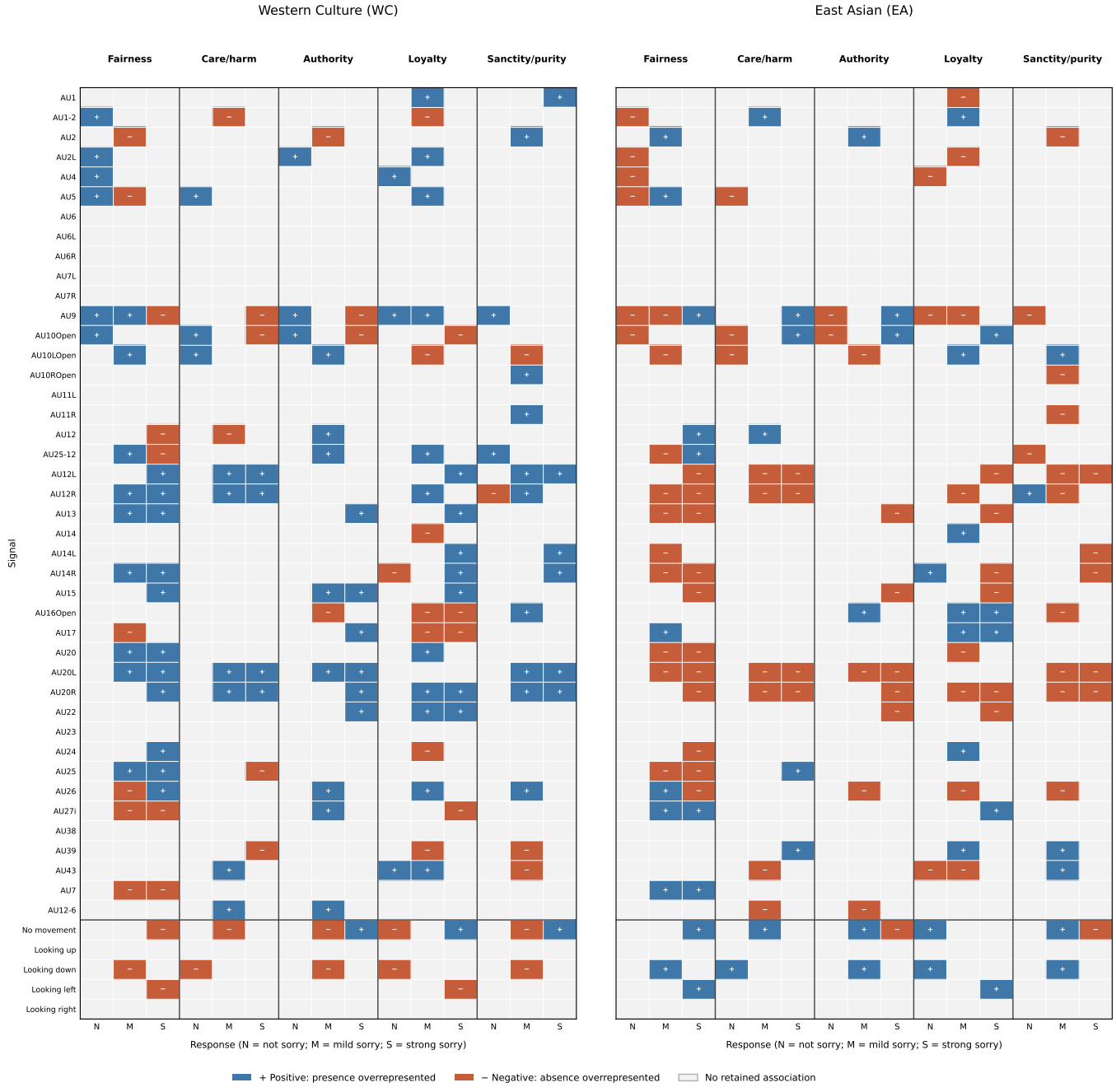
East Asian (EA) Signed MI/PMI Associations by Response and Moral Domain.

Response	Domain	Positive AUs	Positive gaze	Negative AUs	Negative gaze
Not sorry	Fairness	–	–	AU1-2, AU2L, AU4, AU5, AU9, AU10Open	–
Not sorry	Care/harm	–	Looking down	AU5, AU10Open, AU10LOpen	–
Not sorry	Authority	–	–	AU9, AU10Open	–
Not sorry	Loyalty	AU14R	No movement, Looking down	AU4, AU9, AU43	–
Not sorry	Sanctity/ purity	AU12R	–	AU9, AU25-12	–
Mild sorry	Fairness	AU2, AU5, AU17, AU26, AU27i, AU7	Looking down	AU9, AU10LOpen, AU25-12, AU12R, AU13, AU14L, AU14R, AU20, AU20L, AU25	–
Mild sorry	Care/harm	AU1-2, AU12	No movement	AU12L, AU12R, AU20L, AU20R, AU43, AU12-6	–
Mild sorry	Authority	AU2, AU16Open	No movement, Looking down	AU10LOpen, AU20L, AU26, AU12-6	–
Mild sorry	Loyalty	AU1-2, AU10LOpen, AU14, AU16Open, AU17, AU24, AU39	–	AU1, AU2L, AU9, AU12R, AU20, AU20R, AU26, AU43	–
Mild sorry	Sanctity/ purity	AU10LOpen, AU39, AU43	No movement, Looking down	AU2, AU10ROpen, AU11R, AU12L, AU12R, AU16Open, AU20L, AU20R, AU26	–
Strong sorry	Fairness	AU9, AU12, AU25-12, AU27i, AU7	No movement, Looking left	AU12L, AU12R, AU13, AU14R, AU15, AU20, AU20L, AU20R, AU24, AU25, AU26	–
Strong sorry	Care/harm	AU9, AU10Open, AU25, AU39	–	AU12L, AU12R, AU20L, AU20R	–
Strong sorry	Authority	AU9, AU10Open	–	AU13, AU15, AU20L, AU20R, AU22	No movement
Strong sorry	Loyalty	AU10Open, AU16Open, AU17, AU27i	Looking left	AU12L, AU13, AU14R, AU15, AU20R, AU22	–
Strong sorry	Sanctity/ purity	–	–	AU12L, AU14L, AU14R, AU20L, AU20R	No movement

Note. Positive denotes overrepresentation of signal presence. Negative denotes overrepresentation of signal absence and is not a numerically negative PMI value. A dash denotes no significant association in that direction.

Figure 5.7

Signed Receiver-Group-Associated Signal Presence and Absence Across All Moral-Domain and Response Panels.



Note. Blue positive cells denote overrepresented signal presence, orange negative cells denote overrepresented signal absence, and blank cells denote no retained signed association.

sorry. Looking left formed the same reciprocal relation in Fairness and Loyalty strong sorry. No movement formed five WC-absence/EA-presence relations, two reciprocal WC-presence/EA-absence relations in Authority and Sanctity/purity strong sorry, and one additional WC-presence relation in Loyalty strong sorry. Looking up and looking right had no signed association. Independently, thresholded gaze presence occurred in 9, 17, and 18 WC gaze-signal–domain cells and 12, 24, and 17 EA gaze-signal–domain cells for not, mild, and strong sorry, respectively.

Cross-Domain Variability. MI supplied the permutation tests, with BH correction across the 47 signals within each panel; AMI summarised chance-adjusted association strength descriptively. In the population-thresholded vectors, mild sorry showed the largest mean AMI in both groups: .0371 in WC and .0444 in EA (Table 5.11). The corresponding means were .0192 and .0139 for not sorry and .0181 and .0309 for strong sorry. However, no individual signal survived BH correction in any of the six thresholded receiver-group-by-response panels. The smallest adjusted value was .0846, for the EA mild-sorry panel.

Descriptively, the largest thresholded AMI values were carried by AU5 for WC not sorry (.1660), No movement for WC mild sorry (.1234), AU27i for WC strong sorry (.1495), Looking down for EA not sorry (.1213), AU7 for EA mild sorry (.1281), and AU26 for EA strong sorry (.1003). These descriptive leaders included both facial action and gaze, but none supported a signal-specific inferential claim after within-panel FDR control.

The stratified participant bootstrap did not support an overall EA–WC difference in mean AMI for any response. The EA-minus-WC differences were −.0053 for not sorry, .0073 for mild sorry, and .0129 for strong sorry; the corresponding descriptive 95% intervals were [−.0180, .0166], [−.0226, .0212], and [−.0228, .0248]. The corrected primary analysis therefore supplied neither signal-level evidence of cross-domain variability nor evidence that either receiver group was generally more domain-variable across the 47 signals.

Table 5.11

Receiver-Group-Specific Cross-Domain Variability Averaged Across 47 Signals.

Vectors	Response	WC mean AMI	EA mean AMI	EA–WC	95% bootstrap CI	FDR signals (WC/EA)
Thresholded	Not sorry	0.0192	0.0139	-0.0053	[-0.0180, 0.0166]	0/0
Thresholded	Mild sorry	0.0371	0.0444	0.0073	[-0.0226, 0.0212]	0/0
Thresholded	Strong sorry	0.0181	0.0309	0.0129	[-0.0228, 0.0248]	0/0

Note. Signal-level inference used 5,000 participant-block domain-label permutations, with the population cut-offs reapplied after every rearrangement. FDR counts use BH correction across 47 signals within each receiver-group-by-response panel. Confidence intervals are descriptive percentile intervals from 5,000 stratified participant-block bootstrap resamples of the EA–WC difference in mean AMI; they are not response-level hypothesis tests.

Read together, ablation and MI/PMI localise rather than redefine the configuration-level receiver-group difference. Five ablation-active AUs supplied sparse conditional contributions, whereas MI/PMI described a broader marginal repertoire of receiver-group-associated signal presence and absence; their convergence was confined to a small candidate set. The threshold-aware cross-domain analysis retained no signal after FDR correction and showed no overall EA–WC difference in mean AMI. The evidence therefore supports different configurations. The literature-reference analysis next asks how those configurations correspond descriptively with emotion and social-signal representations.

5.5.4 Stage 4: Exploratory Emotion and Social-Signal Reference Correspondence

The preceding analyses established the organisation of the 47 AU and gaze-direction signals. The hierarchical comparison measured correspondence between empirical AU/gaze retention profiles and literature-derived representations of emotion and social signals. It therefore compared positive signal-support patterns. Formal inference was restricted to valence categories (L1) and emotion or social-signal types (L2), yielding 3,540 rows in 28 separately corrected families for the thresholded primary representation. Table 5.12 groups the mapped literature first by valence (L1) and then by the 17 emotion or social-signal types (L2).

Table 5.12

Literature Source and Categorisation Map for Valence Categories (L1) and Emotion or Social-Signal Types (L2)

L2 type	Mapped sources	L3	Mapped AU and gaze support
L1: Positive			
Affiliation smile	J. D. Martin et al., 2021; Rychlowska et al., 2017	2	AU: AU12, AU14, AU24; gaze: None
Amused smile	Ambadar et al., 2009	3	AU: AU6, AU25-12, AU25, AU26, AU27i, AU12-6; gaze: None
Happiness/joy	Hyniewska et al., 2019; D. Matsumoto, Keltner, et al., 2008; Snoek et al., 2023; ten Brinke & Adams, 2015	6	AU: AU6, AU12, AU25-12, AU24, AU25, AU26, AU7, AU12-6; gaze: None
Reward smile	J. D. Martin et al., 2021; Rychlowska et al., 2017	2	AU: AU1-2, AU12, AU25-12, AU13, AU14, AU16Open, AU26; gaze: None
L1: Negative			
Anger	Hyniewska et al., 2019; D. Matsumoto, Keltner, et al., 2008; Snoek et al., 2023	5	AU: AU1, AU4, AU5, AU10Open, AU13, AU17, AU22, AU23, AU24, AU25, AU26, AU7; gaze: None
Contempt	Hyniewska et al., 2019	1	AU: AU1, AU4, AU6, AU10Open, AU23; gaze: None
Disgust	D. Matsumoto, Keltner, et al., 2008; Snoek et al., 2023	4	AU: AU9, AU10Open, AU15, AU16Open, AU17, AU25, AU26, AU7; gaze: None
Dominance smile	Rychlowska et al., 2017	2	AU: AU5, AU6, AU9, AU10Open, AU12L, AU12R; gaze: None
Embarrassment/appeasement	Ambadar et al., 2009	3	AU: AU6, AU25-12, AU25, AU26, AU27i, AU12-6; gaze: None
External-standard violation	Hyniewska et al., 2019	1	AU: AU12, AU17; gaze: None
Fear	Hyniewska et al., 2019; D. Matsumoto, Keltner, et al., 2008; Snoek et al., 2023	5	AU: AU1-2, AU4, AU5, AU16Open, AU17, AU20, AU25, AU26, AU27i, AU7; gaze: None
Guilt	Julle-Danière et al., 2020	1	AU: AU4, AU5, AU17, AU20; gaze: None
Internal-standard violation	Hyniewska et al., 2019	1	AU: AU4, AU5, AU6, AU12, AU24, AU26, AU27i; gaze: None
Sadness	Hyniewska et al., 2019; D. Matsumoto, Keltner, et al., 2008; Snoek et al., 2023; ten Brinke & Adams, 2015; Yamamoto et al., 2021	10	AU: AU1, AU2, AU4, AU5, AU6, AU11L, AU11R, AU15, AU17; gaze: Neutral, Down, Left, Right
Shame	Halmesvaara et al., 2020; Hyniewska et al., 2019; Keltner, 1995; Tracy et al., 2009	4	AU: AU2, AU5, AU20, AU25; gaze: Down
L1: Neutral/mixed			
Polite smile	Ambadar et al., 2009	3	AU: AU6, AU25-12, AU25, AU26, AU27i, AU12-6; gaze: None
Surprise	D. Matsumoto, Keltner, et al., 2008; Snoek et al., 2023	4	AU: AU1-2, AU5, AU25, AU26, AU27i; gaze: None

Note. The L3 column counts source representations. Evidence scope, grouping rationale, and mapping details are in Table C3.

Overall Response Correspondence (M1).

Valence Categories (L1). Positive-valence correspondence increased across both response transitions. Mild minus Not was $\Delta = .174$, 95% CI $[.106, .217]$, $p_{MC} = .00765$, $q = .01377$; Strong minus Mild was .192, $[.146, .252]$, $p_{MC} = .00260$, $q = .00585$; and Strong minus Not was .366, $[.319, .397]$, $p_{MC} = .00005$, $q = .00023$. Negative-valence correspondence decreased from Not to Mild and increased from Mild to Strong, whereas the full-range Strong-minus-Not contrast was not retained. No neutral/mixed response contrast was retained.

Emotion or Social-Signal Types (L2). The L2 contrasts separated a broad response effect from claims of a simple monotonic sequence. From Not to Mild, affiliation ($\Delta = .319$) and reward smile (.215) increased, whereas dominance smile ($-.439$), contempt, disgust, and anger decreased. From Mild to Strong, stable increases included internal-standard violation (.453), guilt (.404), fear, contempt, anger, sadness, shame, happiness/joy (.171), and reward smile (.159). Across the full Not-to-Strong range, affiliation (.517), reward (.374), internal-standard violation, happiness/joy, guilt, embarrassment/appeasement, polite and amused smiles, fear, and shame increased, while dominance smile ($-.496$), contempt, and disgust decreased. Contempt and reward smile had stable changes for all three planned response contrasts. Affiliation and dominance clearly separated apology from Not sorry but did not show a retained Mild-to-Strong change; happiness changed from Mild to Strong and across the full range, but not from Not to Mild.

Figure 5.8 shows the M1 L2 point scores. Stars identify stable score-family retention and question marks identify scores whose significance remains uncertain because of the limited number of randomisation draws; the separately corrected response contrasts are reported in Table 5.13.

Table 5.13

Selected Thresholded-Primary Correspondence Estimates and Planned Contrasts for the Overall Response Profiles (M1)

	Response score, r [95% CI], q			Response contrast, Δ [95% CI], q		
Reference	Not sorry	Mild sorry	Strong sorry	Mild − Not	Strong − Mild	Strong − Not
Positive valence (L1)	-.132 [-.143, -.111], .000	.042 [-.019, .083], .150	.234 [.197, .261], .000	.174 [.106, .217], .014	.192 [.146, .252], .006	.366 [.319, .397], .000
Affiliation smile (L2)	-.148 [-.170, -.131], .000	.171 [.058, .244], .001	.369 [.307, .419], .000	.319 [.209, .401], .004	.198 [.126, .304], .077	.517 [.456, .578], .000
Happiness/joy (L2)	-.119 [-.133, -.093], .000	-.010 [-.080, .041], .800	.162 [.114, .196], .000	.109 [.036, .161], .140	.171 [.112, .240], .019	.281 [.224, .318], .000
Reward smile (L2)	-.136 [-.148, -.107], .000	.078 [.001, .124], .030	.237 [.194, .276], .000	.215 [.124, .259], .004	.159 [.112, .235], .035	.374 [.313, .413], .000

Note. Score and response-contrast q values come from separate BH families within each literature level. Confidence intervals are unadjusted participant-block bootstrap intervals. Scores whose significance remains uncertain are not treated as stably retained even when their current $q \leq .05$. Thresholded results are primary.

Not Sorry in the Sad-Versus-Smiling Apology Debate. The thresholded Not-sorry profile was led by upward gaze (.769), AU43 (.731), AU9 (.526), AU10ROpen (.513), AU10Open (.462), and AU10LOpen (.404). Its highest L2 type was nevertheless dominance smile ($r = .231$), whereas affiliation ($r = -.148$) and reward ($r = -.136$) were negative. Mild and Strong sorry reversed that functional pattern: affiliation was .171 and .369, reward was .078 and .237, and dominance was negative. The results distinguish affiliation and reward from dominance, consistent with accounts in which smiles can reinforce, bond, appease, or negotiate hierarchy rather than express one positive state (J. Martin et al., 2017; J. D. Martin et al., 2021; Rychlowska et al., 2017). Culture, intensity, duration, symmetry, and timing can also change smile inter-

pretation, but those dimensions were not preserved by the static 47-signal representation (Fang et al., 2020; Hoque et al., 2011).

In summary, M1 provided stable evidence for response-linked changes at L1 and L2, led by increasing positive valence and reward correspondence, a non-monotonic negative-valence trajectory, and a functional separation between affiliation/reward and dominance. The evidence supports graded morphological correspondence and motivates direct testing of whether the response pattern differs between WC and EA.

Response Correspondence by Receiver Group (M2-1). M2-1 tested the WC–EA difference directly after giving the five moral domains equal weight within each response and sample group. At L1, EA minus WC was more negative for Not-sorry negative valence ($\Delta = -.074$, 95% CI $[-.128, -.049]$, $q = .00157$) and more positive for Strong-sorry negative valence (.104, $[.062, .143]$, $q = .00045$). Strong-sorry correspondence was also higher in EA for positive valence (.087, $[.043, .164]$, $q = .01920$) and neutral/mixed valence (.094, $[.040, .162]$, $q = .02441$).

The L1 differences-in-differences showed that the EA-minus-WC contrast changed with response rather than remaining a constant offset. Stable effects occurred for negative valence from Not to Mild, and for all three valence categories from Not to Strong. No L1 differences-in-differences for Strong minus Mild were retained. At L2, direct receiver-group contrasts included lower EA than WC correspondence for Not-sorry contempt, disgust, and dominance; lower EA dominance and higher EA guilt and shame correspondence for Mild sorry; and higher EA contempt, disgust, embarrassment/appeasement, external-standard violation, guilt, and internal-standard violation correspondence for Strong sorry. Receiver-group-by-response effects included the Not-to-Strong changes in contempt, disgust, dominance, guilt, and internal-standard violation, the Not-to-Mild change in guilt, and opposing Strong-minus-Mild changes in dominance and shame. Four Monte Carlo-boundary rows were withheld. Figure 5.9 shows the corresponding L2 profiles.

Taken together, M2-1 provides direct evidence that literature correspondence was calibrated differently between the WC and EA groups as the response category changed, especially relative to Not sorry.

Response Correspondence by Moral Domain (M2-2). M2-2 gave WC and EA equal weight within each response-by-domain profile. Some within-domain response contrasts were stable. Positive-valence Strong-minus-Not contrasts were retained in all five domains. Negative-valence Strong-minus-Mild contrasts were retained in Care/harm, Authority, and Sanctity/purity. The retained L2 response contrasts comprised an Authority reward-smile Strong-minus-Not increase, an Authority internal-standard-violation Strong-minus-Mild increase, and a Sanctity/purity dominance-smile Strong-minus-Not decrease. These tests establish changes within named domain contexts.

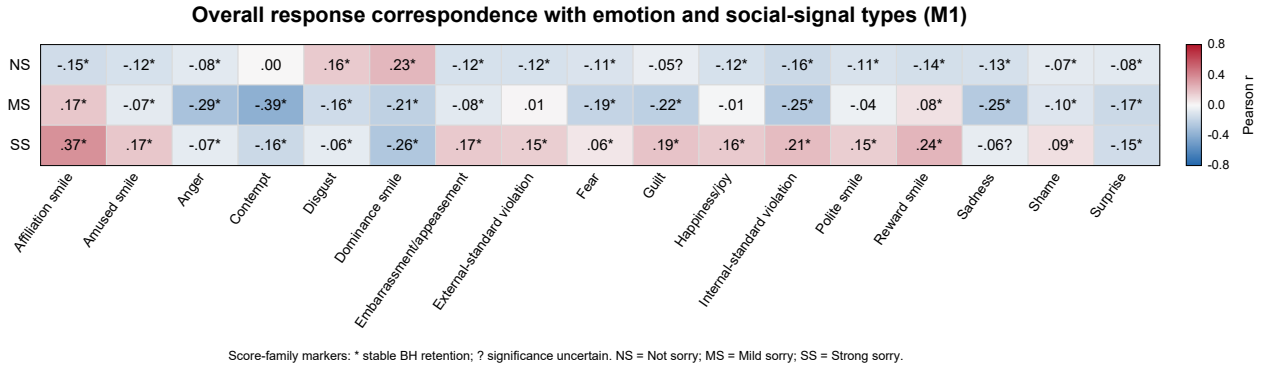
The direct domain tests supplied no reliable evidence of such differences. No thresholded L1 or L2 five-domain omnibus test or domain-pair contrast was stably retained. The smallest omnibus q values were .726 at L1 and .856 at L2; the corresponding pairwise minima were .934 and .964. Figure 5.10 shows the response-by-domain L2 profiles to which these separately corrected tests apply.

In summary, M2-2 recovered selected response changes within domain-labelled contexts but no stable difference among the five domains in L1 valence or L2 emotion/social-signal correspondence.

Full Condition Correspondence (M3). M3 retained the 30 fully conditional response-by-sample-group-by-domain profiles and formally tested response contrasts, WC–EA contrasts, and receiver-group-by-response differences within each domain at L1 and L2. No response contrast, receiver-group contrast, or receiver-group-by-response difference met the retention criterion. The smallest thresholded L2 q values

Figure 5.8

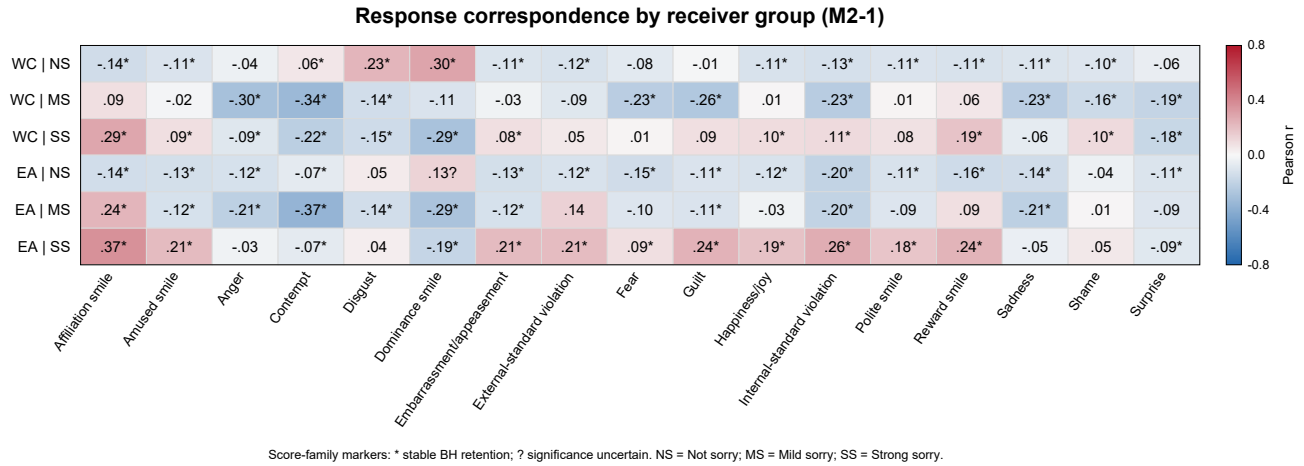
Thresholded-Primary Correspondence of the Overall Response Profiles (M1) With 17 Literature-Derived Emotion and Social-Signal Types.



Note. Stars mark stable score-family retention (BH $q \leq .05$ and a stable Monte Carlo decision); question marks identify scores whose significance remains uncertain. Response contrasts belong to separate M1–L2 families and are not encoded by the cell markers.

Figure 5.9

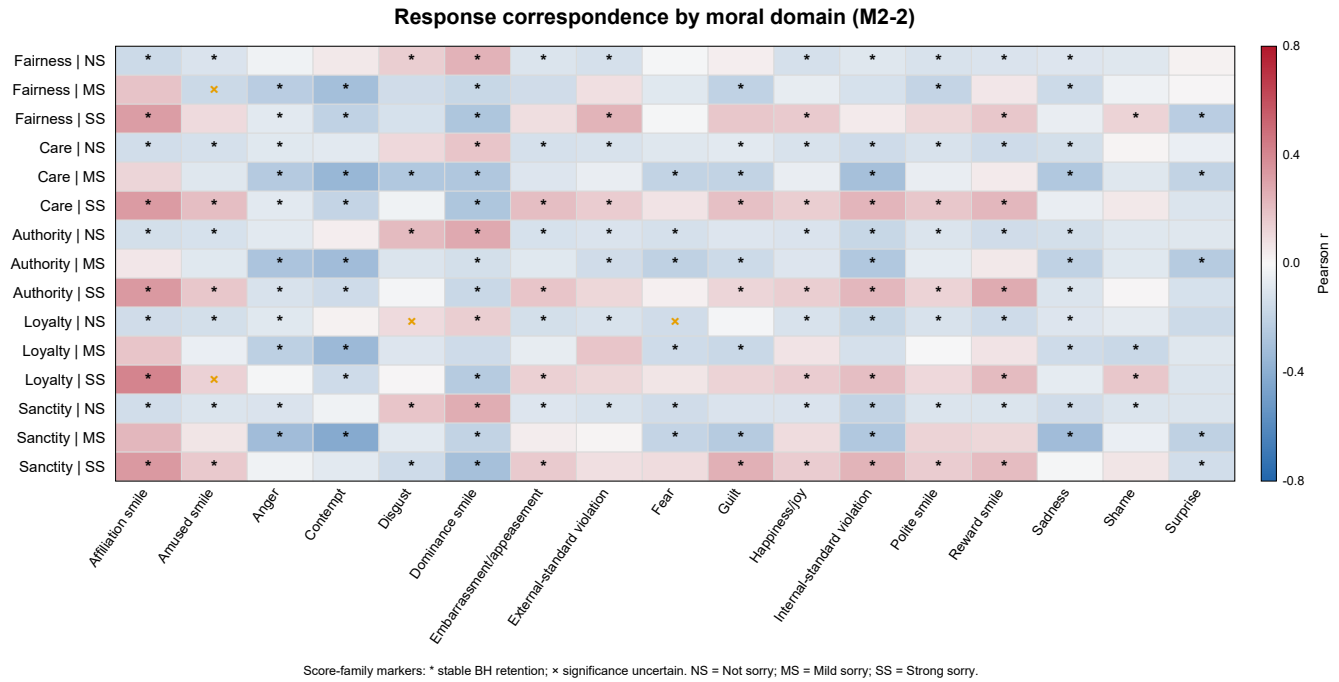
Thresholded-Primary Correspondence of Response Profiles by Receiver Group (M2-1) With Emotion or Social-Signal Types (L2), After Equal Averaging Over Moral Domains.



Note. Stars mark stable score-family retention; question marks identify scores whose significance remains uncertain. Direct WC–EA and receiver-group-by-response contrasts use separate correction families.

Figure 5.10

Thresholded-Primary Correspondence of Response Profiles by Moral Domain (M2-2) With Emotion or Social-Signal Types (L2), After Equal Averaging Over Receiver Groups.



Note. Stars mark stable score-family retention; crosses identify scores whose significance remains uncertain. Response, five-domain omnibus and domain-pair contrasts use separate correction families.

were .121 for receiver group and .097 for receiver group by response. Two L1 contrast candidates remained Monte Carlo-boundary-sensitive and were withheld. Figure 5.11 shows the fully conditional L2 profiles.

Taken together, M3 supplies detailed descriptive localisation but no stable evidence that the WC–EA correspondence difference or its response trajectory varies within a particular moral domain. The result does not establish equivalence: the fully conditional cells are smaller and the correction families are broad.

Taken together, the correspondence analyses add functional context to the same response-led, receiver-group-calibrated architecture. Stable evidence was concentrated at M1 response level and M2-1 response-by-group level: reward, affiliation, and dominance were functionally distinct, and WC–EA correspondence changed with response category. No stable L1/L2 difference among moral domains or fully conditional M3 contrast survived correction. These coefficients describe similarities between empirical AU/gaze profiles and literature-derived representations. The Hamming analyses therefore return to the geometry of the empirical profiles.

5.5.5 Stage 5: Exploratory Hamming Geometry and Clustering

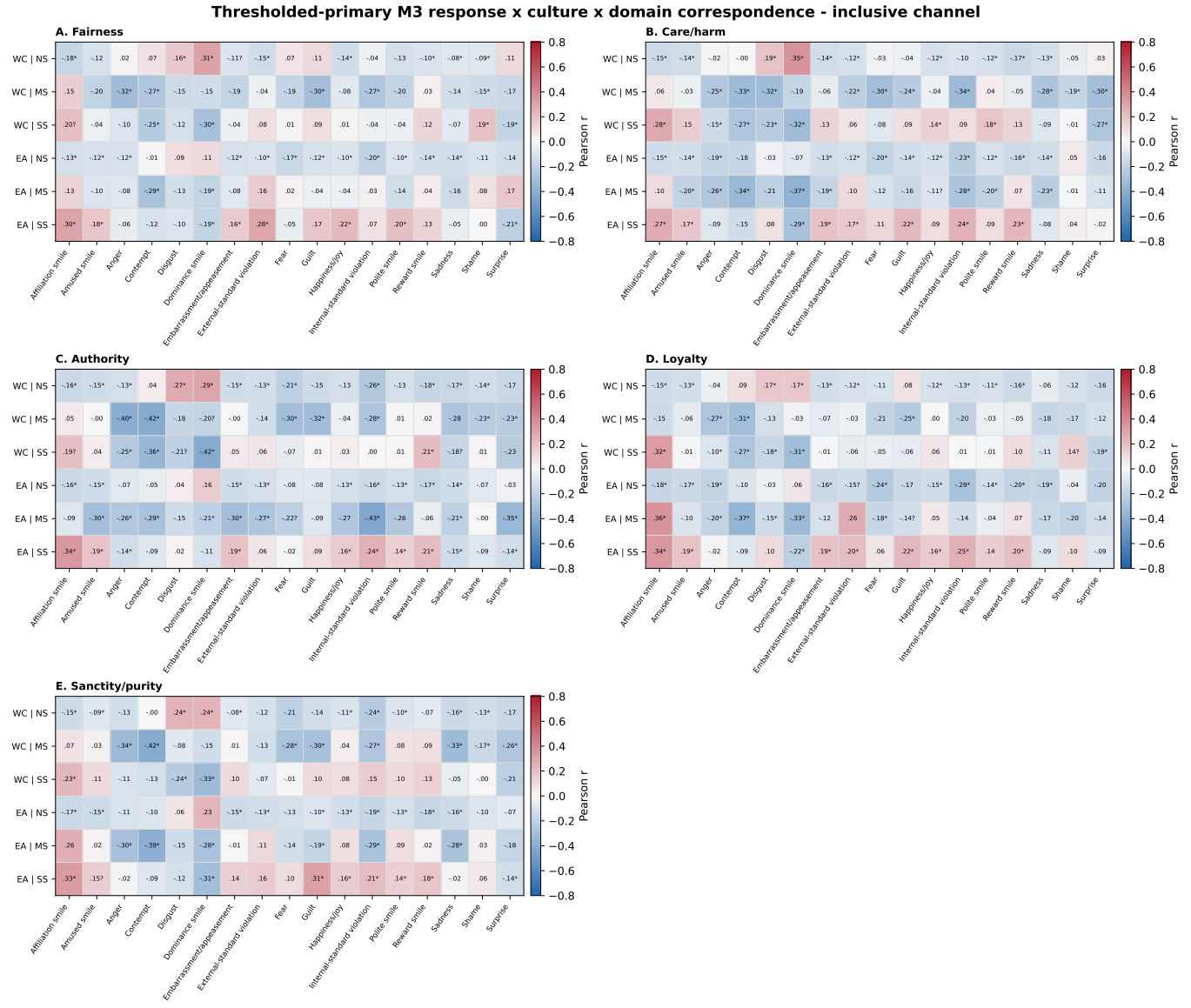
The Hamming analyses constructed condition profiles from the population-thresholded participant response models. Six predictors were globally zero—AU6R, AU7L, AU7R, AU11L, AU23, and AU38—and were removed before distances were calculated, leaving 41 signals. The broad overview retained all three responses in 30 condition profiles; the apology-focused view retained mild and strong sorry in 20 condition profiles. Within every heatmap, the dendrogram leaf order was applied identically to both axes.

Brief Full-Response Overview. Twenty-nine of the 30 condition profiles were unique; the only exact tie joined WC not-sorry Care/harm with EA not-sorry Authority. Across the 15 WC–EA pairs matched on response and moral domain, profile Hamming distance averaged .2862 (Table 5.14). The matched mean was .1171 for not sorry, .4439 for mild sorry, and .2976 for strong sorry. Thus, the mild-sorry profiles showed the greatest aggregate cross-sample separation, while not sorry was most similar.

As Figure 5.12 shows, the full condition-profile dendrogram retained one admissible geometric summary at $k = 2$: silhouette .4785, fusion-height gap .1192, cluster sizes 19 and 11, and median stability ARI 1.000 across 500 participant-block resamples. The 11-profile cluster contained all ten not-sorry profiles plus EA Fairness mild sorry; the 19-profile cluster contained all ten strong-sorry profiles and the remaining nine mild-sorry profiles. This response-led split was a stable description of the aggregated profiles.

Figure 5.11

Thresholded-Primary Correspondence of Full Condition Profiles (M3) With Emotion or Social-Signal Types (L2).



Note. Each domain facet shows WC and EA response profiles. Stars mark stable score-family retention; question marks identify scores whose significance remains uncertain. Response, receiver-group and receiver-group-by-response contrasts use separate correction families.

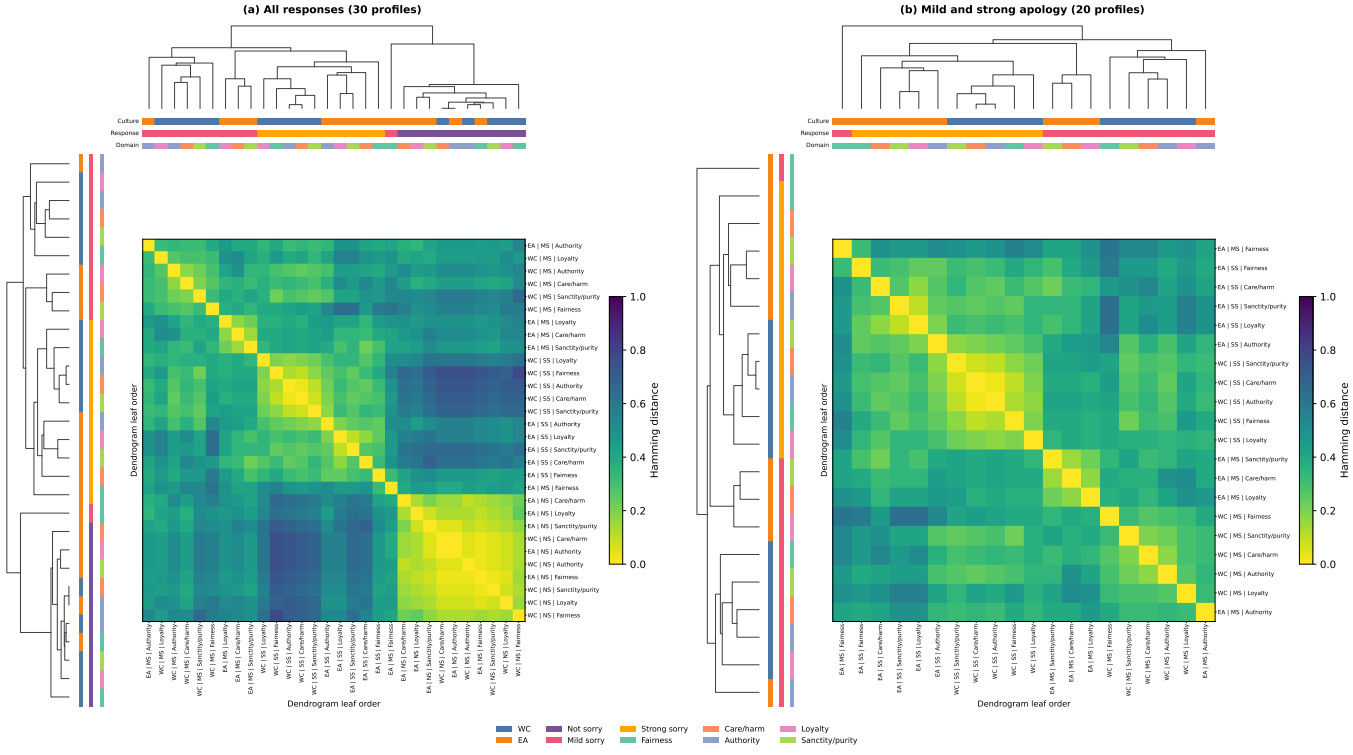
Table 5.14

Hamming Distance Between Matched WC and EA Binary Profiles Within Each Response and Moral Domain.

Response	Moral domain	Hamming distance
Not sorry	Fairness	.1463
Not sorry	Care/harm	.1463
Not sorry	Authority	.0488
Not sorry	Loyalty	.1463
Not sorry	Sanctity/purity	.0976
Mild sorry	Fairness	.6098
Mild sorry	Care/harm	.3659
Mild sorry	Authority	.3415
Mild sorry	Loyalty	.4878
Mild sorry	Sanctity/purity	.4146
Strong sorry	Fairness	.3902
Strong sorry	Care/harm	.2683
Strong sorry	Authority	.2195
Strong sorry	Loyalty	.3415
Strong sorry	Sanctity/purity	.2683

Figure 5.12

Hamming-Distance Heatmaps and Average-Linkage Dendrograms for (a) the 30 Full-Response Receiver-Group-by-Response-by-Domain Profiles and (b) the 20 Mild- and Strong-Sorry Profiles.



Note. Receiver group, response, and moral domain are shown by the three annotation strips. Both matrix axes follow the corresponding dendrogram leaf order. The apology-only panel is the focused view of this exploratory analysis.

Apology-Only Focus. Removing the not-sorry profiles eliminated the stable two-group summary visible in the full-response overview (Figure 5.12). No apology-only cut from $k = 2$ to $k = 10$ met both admission criteria. At $k = 2$, the clusters contained 19 profiles and one singleton, despite a median resampling ARI of 1.000. The highest silhouette occurred at $k = 4$, at .2862, but the cluster sizes of 10, 6, 3, and 1 again included a singleton. The apology profiles therefore formed stable local branches without supporting a discrete global cluster count. The annotation strips showed response- and receiver-group-contiguous sections, but moral domains remained interspersed.

Table 5.15

Participant-Level Threshold-Aware Permutation Tests of Hamming Alignment.

View	Contrast	Within	Between	Difference	Raw p	Holm p
All responses	Response alignment	.2192	.2993	.0801	.0002	.0014
All responses	Receiver-group alignment	.2666	.2789	.0122	.0002	.0014
All responses	Moral-domain alignment	.2709	.2732	.0023	.7361	1.0000
Apology only	Mild/strong alignment	.2725	.3135	.0410	.0002	.0014
Apology only	Receiver-group alignment	.2800	.3061	.0261	.0002	.0014
Apology only	Moral-domain alignment	.2890	.2941	.0051	.9062	1.0000
Apology only	EA–WC mild–strong change	–	–	.0007	.9700	1.0000

Note. Alignment is mean between-label minus within-label Hamming distance. Each null used 5,000 participant-restricted permutations with thresholding and global-zero predictor removal repeated after relabelling. Alignment tests used the upper tail; the EA–WC change test was two-sided. Holm correction covered all seven tests.

Restricted Permutation Inference. The restricted permutation tests showed that the geometry aligned reliably with response and receiver group, but not with moral domain (Table 5.15). Across all responses, mean distance was .2192 within response labels and .2993 between them, a difference of .0801, $p_{\text{Holm}} = .0014$. The corresponding receiver-group difference was smaller, .0122, but also exceeded its threshold-aware null, $p_{\text{Holm}} = .0014$. Moral-domain alignment was .0023, $p_{\text{Holm}} = 1.0000$. In the apology-only analysis, the mild–strong alignment difference was .0410 and the receiver-group-alignment difference was .0261; both had $p_{\text{Holm}} = .0014$. Moral-domain alignment remained unsupported, difference .0051, $p_{\text{Holm}} = 1.0000$. Thus, removal of not sorry reduced but did not eliminate response organisation and made the relative contribution of receiver group more visible, while the five moral domains did not show overall alignment in these restricted tests.

The mean within-participant mild-to-strong Hamming change was almost identical across sample groups: the EA-minus-WC difference was .0007, $p_{\text{Holm}} = 1.0000$. Domain-specific estimates varied in direction but every 95% bootstrap interval included zero (Table 5.16). The largest positive estimate was Care/harm, .041, and the largest negative estimate was Loyalty, $-.038$. These intervals localise the overall comparison and were not treated as five additional tests.

Table 5.16

Domain-Specific Receiver-Group Differences in Within-Participant Mild-to-Strong Hamming Change.

Domain	n_{WC}	n_{EA}	WC mean	EA mean	EA–WC [95% CI]
Fairness	20	14	.235	.202	−.033 [−.099, .029]
Care/harm	15	18	.208	.249	.041 [−.030, .114]
Authority	14	16	.225	.244	.019 [−.033, .070]
Loyalty	13	16	.251	.213	−.038 [−.102, .025]
Sanctity/purity	14	16	.197	.207	.010 [−.042, .062]

Note. Confidence intervals are percentile intervals from 5,000 stratified participant bootstrap samples. No domain-specific hypothesis tests were added.

Cluster Diagnostics and Stability. Only the 30-profile full-response view supported an admissible cut; no apology-only profile cut met both the minimum-size and stability criteria. These condition-profile diagnostics evaluated $k = 2 \ldots 10$ with 500 participant-block resamples.

Taken together, the formal Hamming results show a nested geometry rather than discrete apology types. Across the four threshold-aware permutation schemes, 36–47 predictors remained active after rethresholding and global-zero removal, with medians of 39–42. Response led the full-profile geometry and remained aligned within apology; receiver group also differentiated the apology profiles, whereas moral domain did not align with the overall geometry. No apology-only condition-profile cut was stable and sufficiently sized to name as a discrete cluster solution. These conclusions remain conditional on the population-supported representation, Hamming distance, average linkage, and preserved sample sizes.

5.6 Discussion

5.6.1 Moral-Domain Organisation Across the WC and EA Receiver Groups

Both WC and EA groups contained recoverable moral-domain information, but the evidence did not show that this information was stronger overall in one group than in the other. In the separate analyses reported in Chapters 3 and 4, the five-domain classifiers achieved balanced accuracies of .4492 in WC and .4827 in EA, and both exceeded their respective permutation nulls. The thresholded matched comparison reproduced the same numerical ordering while holding participant counts, predictor columns, and participant splits constant, but the EA-minus-WC difference of .039 was not reliable, $p_{\text{Holm}} = .628$. The descriptive difference between the original chapter-level accuracies therefore does not establish that moral domains were generally more distinguishable in EA. The two groups instead showed similar overall domain-classification strength alongside different domain-specific configurations.

The small number of corrected domain-specific diagonal cells does not contradict the significant omnibus balanced accuracies. The omnibus tests aggregate information across the complete confusion pattern, whereas each diagonal cell must separately exceed a stricter corrected threshold. Accordingly, the WC Loyalty result and the EA Fairness and Authority results identify where recall was clearest, not the only domains contributing to the multivariate classifiers. Their different leading domains are most cautiously interpreted as differences in which complete configuration supplied the strongest contrast within each receiver group; they do not establish one-to-one domain-specific expressions or different moral meanings for individual signals.

The direct receiver-group classifiers showed that the complete mild- and strong-apology vectors nevertheless differentiated the WC and EA groups within every moral domain. Fairness and Loyalty produced the strongest complete-vector separation, followed by Authority and Sanctity/purity, while Care/harm produced the weakest but still reliable separation. Because receiver group and face-identity set varied together, these results establish differentiation between the coupled datasets within each domain, not an effect of receiver culture independent of face identity.

The prevalence results demonstrated that strong complete-vector classification did not require a corrected panel-wide difference in every response and domain. Only strong-sorry Fairness survived correction in the 15-panel thresholded primary family. The strongest primary prevalence evidence was therefore concentrated in one response-by-domain context rather than distributed uniformly across the chapter. The 705 signal-level prevalence effects show where the panel contrasts were located, but they are descriptive components of the omnibus panel tests rather than 705 separately tested discoveries.

Ablation and MI/PMI refined this conclusion by addressing different properties of the same configurations. Direct ablation isolated five AUs and no gaze predictor. AU16Open acted as a WC confuser in Fairness; AU15 and AU20L supported both receiver-group diagonals and their pairwise separation in Authority; AU20 supported the EA diagonal and pairwise separation in Loyalty; and AU20R had the corresponding role in Sanctity/purity. Care/harm had no reliable ablation-active signal despite its significant complete-vector classifier. MI/PMI identified a much broader marginal repertoire across 34 AUs and three gaze states. Four of the five ablation-active domain–signal roles also had a same-domain MI/PMI association, but this convergence remained selective because ablation measured conditional classifier contributions whereas MI/PMI measured marginal association within a response and domain.

The compact comparison of the Chapter 3 and Chapter 4 repertoires similarly separated overall organisation from exploratory localisation. Separation-supporting ablation breadth did not differ reliably between WC and EA, whereas confusion-contributing roles were more widespread in the WC classifier. Positive MI/PMI discovery breadth was numerically greater in EA and negative breadth was similar, but neither comparison was reliable. These feature-repertoire tests used predictors or signal blocks rather than participants as their matched units. They therefore describe exploratory differences between the two repertoires rather than participant-level evidence that one receiver group used a broader signal repertoire.

The corrected cross-domain variability analysis further limited signal-specific conclusions. Mild sorry had the largest descriptive mean AMI in both group-specific thresholded representations, but no signal survived within-panel BH correction in any of the six panels. The descriptive EA-minus-WC mean-AMI differences changed direction across responses, and every participant-bootstrap interval included zero. The data therefore did not support a general difference in domain variability between the two groups.

At the level of the complete Hamming geometry, moral domain was not the dominant organising axis. Participant-level domain alignment was unsupported both when all three responses were included and when the analysis was restricted to mild and strong apology. Moral domains were interspersed in the reordered condition-profile matrices, and no apology-only condition-profile cut met both the size and stability criteria. Domain information was therefore detectable through classification, prevalence contrasts, conditional ablation roles, and marginal associations without producing five globally separated regions of the binary signal space.

The domain-specific pattern illustrates why these analyses should not be reduced to one rank order. Fairness combined the highest receiver-group classification accuracy with the only corrected thresholded prevalence panel, but its sole ablation-active signal was a WC confuser. Loyalty was almost equally classifiable and contained an EA-linked AU20 role without corrected primary panel-level prevalence evidence. Authority contained the densest direct ablation result despite intermediate classifier accuracy. Sanctity/purity remained reliably classifiable and contained an AU20R role without a corrected prevalence panel. Care/harm remained classifiable from the complete vector and contained signal-wise MI/PMI associations despite having neither a corrected prevalence panel nor an ablation-active predictor. Moral domains therefore modified particular configurations within a broader response- and receiver-group-aligned geometry rather than forming five isolated global clusters.

5.6.2 Apology-Response Organisation Across the WC and EA Receiver Groups

Response supplied the most consistent shared organisation across the two groups. The separate validation analyses produced the same ordinal pattern: not sorry was most accurately distinguished, strong sorry was intermediate, and mild sorry was least accurately distinguished. The prevalence summaries were consistent with this ordering. Not sorry had a concentrated repertoire led in both groups by AU43 and upward gaze, whereas mild sorry had a broader, less concentrated configuration. Strong sorry was more coherent than mild sorry but showed clearer descriptive differences between the groups, including predominantly downward gaze in WC and lateral gaze in EA. The shared response hierarchy therefore concerned the concentration and separability of the configurations, while the AUs and gaze directions occupying those configurations could differ.

The literature-reference results can also be compared with Chapter 2. Chapter 2's rank-two NNMF

separated a smile-related lower-face configuration from a configuration dominated by lip stretching, mouth opening, and jaw movement, while its gaze analysis associated upward gaze with not-sorry or weaker-apology interpretations and downward gaze with stronger sorry judgements. The present hierarchical analysis is consistent with that mixed-configuration account and distinguishes different smile-related configurations. Reward- and affiliation-smile correspondence increased across apparent apology intensity, whereas dominance-smile correspondence was positive for Not sorry and negative for both apology profiles. The empirical profiles also contained gaze and non-smile movements. The two chapters point to the same broad interpretation. Chapter 2 derived two exploratory AU factors and analysed gaze separately within a Western receiver sample, whereas the present analysis compared 47-signal feature-retention profiles across the WC and EA groups with heterogeneous constructed literature references. Together, the analyses support a mixed, context-sensitive repertoire of candidate apology signals.

The full-response Hamming geometry made the not-sorry distinction especially visible. Its only admissible condition-profile cut separated all ten not-sorry profiles, together with one EA mild-sorry profile, from every strong-sorry profile and the remaining mild-sorry profiles. Cross-group matched-profile distances were also lowest for not sorry and highest for mild sorry. This profile-level split did not mean that the data reduced to a simple Not-sorry-versus-sorry response dichotomy. At the participant level, response alignment remained reliable after not sorry was removed: mild- and strong-sorry rows were more similar within response than between responses. The response organisation therefore weakened but persisted within apology.

Receiver-group differentiation also persisted when the analysis was restricted to apology. The direct classifiers used only mild- and strong-sorry vectors and distinguished WC from EA in all five moral domains. In the thresholded matched task comparison, binary receiver-group classification exceeded binary moral-domain-pair classification. Because class number, observations per class, predictor columns, and participant splits were matched between C2 and D2, the difference was not explained by those design quantities. It remains a comparison within the present signal vectors rather than evidence that receiver group is universally easier to classify than moral domain. D5 provides a descriptive reference only because it has a different chance level and represents a joint five-class decision problem.

The formal Hamming tests placed response and receiver group within the same apology geometry without reducing them to the same effect. In the full-response restricted tests, both response alignment and receiver-group alignment exceeded their threshold-aware nulls, while domain alignment did not. In the apology-only restricted tests, mild-versus-strong alignment and receiver-group alignment again survived joint correction. Removing not sorry also removed the only admissible condition-profile partition, and no apology-only condition-profile cut met both the cluster-size and stability requirements. The apology space therefore contained reproducible response and receiver-group gradients but not a defensible small set of discrete global clusters.

WC and EA nevertheless showed almost identical total amounts of within-person mild-to-strong signal change. The overall EA-minus-WC Hamming difference was effectively zero, and all five domain-specific bootstrap intervals included zero. This comparison concerns the proportion of binary signal states that changed, not which signals changed. It is therefore compatible with the receiver-group classifiers, prevalence contrasts, and signed MI/PMI results: two groups can move a similar Hamming distance from mild to strong apology while taking that step through different AU or gaze configurations. The response evidence

consequently indicates a common degree structure together with receiver-group- and domain-dependent signal realisations.

5.6.3 Chapter-Level Empirical Synthesis

Across analyses, the WC and EA receiver groups shared a response-led architecture. Both showed concentrated not-sorry configurations, broader and less separable mild-sorry repertoires, more coherent strong-sorry configurations, and reliable mild–strong alignment after not sorry was removed. They also contained comparable overall moral-domain information: the matched comparison did not support a WC–EA difference in five-domain balanced accuracy.

That shared organisation was realised through different complete visual-signal configurations. WC and EA were classifiable within every moral domain, receiver-group alignment survived in both Hamming views, and matched binary receiver-group classification exceeded pairwise moral-domain classification. Prevalence, ablation, and MI/PMI localised candidate contributors, but their signal-level evidence was narrower than the complete-vector result. Moral domain likewise supplied local information in classifiers and selected signal patterns without overall alignment in the restricted tests or five global condition-profile clusters. The formal classifier, prevalence, and Hamming tests establish patterns within the present WC and EA groups; primary ablation and MI/PMI analyses locate candidate signal contributions under their stated correction limits, while emotion-reference and Hamming analyses remain exploratory.

The chapter's research question is therefore answered by a hierarchy: visual apology signals were organised similarly by apparent apology intensity in WC and EA receivers, but the complete configurations realising that shared structure differed between the groups in this coupled receiver-group/face-identity design, while moral context selectively differentiated local configurations rather than partitioning the apology space. Chapter 6 integrates this shared architecture, receiver-group calibration, and moral-domain differentiation to explain apology as a receiver-interpreted, contextually calibrated social signal.

Chapter 6

General Discussion

6.1 Overview

In this thesis, I investigated how receivers use dynamic visual signals to judge how sorry an offender appears after a transgression, and what those judgements reveal about moral organisation and apology as social communication. The studies did not assume that apology reproduces a canonical emotion expression. Instead, a data-driven generative approach recovered the facial movements, gaze directions, and, in Chapter 2, blushing states that receivers treated as evidence of apparent apology. The claims therefore concern receiver-interpreted visual evidence, not what offenders intended, felt, or achieved.

Across Chapters 2–5, the analyses converge on a three-level organisation. First, apparent apology intensity supplied the broadest structure, expressed in Chapters 3–5 through the model-derived Not-sorry, Mild-sorry, and Strong-sorry response categories. Receivers used mixed visual-signal configurations rather than one fixed set of facial expressions; gaze direction supplied the clearest graded cue in the initial Western study, whereas blushing showed no population-level pattern. Second, the two receiver groups whose stimuli were ethnically matched—Western receivers judging Western faces (WC) and East Asian receivers judging East Asian faces (EA)—shared this response-level organisation but differed reliably in the complete Mild- and Strong-sorry configurations through which it was expressed within each moral domain. Third, moral-domain information was recoverable within each group but remained selective and local: it did not divide the global signal geometry into five separated regions or produce stable apology-only clusters.

The Chapter 5 literature-correspondence analysis refined rather than replaced this hierarchy. Positive- and negative-valence correspondence followed different response trajectories, and functional-smile types did not behave as one undifferentiated category (Rychlowska et al., 2017). Selected domain-averaged trajectories differed between the WC and EA groups, but the analysis did not supply stable valence- or type-level separation among moral domains or reliable local group contrasts within domains. These correspondences therefore identify a graded repertoire of possible relational functions, not a dictionary linking each configuration to one expressed, perceived, or experienced emotion.

Taken together, the findings support a nested or graded account of moral-domain organisation and a receiver-centred account of apology as a multimodal, contextually calibrated social signal. The following sections develop these conclusions, define their empirical limits, and identify the next step: connecting controlled receiver judgements to natural production and consequential interaction.

6.2 Apology Signals and the Organisation of Moral Domains

The first theoretical question is where moral context enters the interpretation of an apology signal. Across Chapters 3–5, the five moral domains were neither absent from receivers' judgements nor expressed as five self-contained facial prototypes. Instead, the generated configurations were organised primarily by apology-response category and apparent apology intensity; the WC and EA groups differentiated the configurations within that broad organisation; and moral domain selectively modified particular configurations. This result connects pluralist and domain-general accounts of morality. It suggests that common processes of recognising and responding to wrongdoing can coexist with context-sensitive information about the kind

of norm or relationship at stake. The findings do not suggest that apologies reveal five natural moral kinds. Instead, they show how moral context can affect receivers' interpretation of possible social repair within a shared response pattern.

6.2.1 The Moral-Domain Debate: Shared and Domain-Specific Response Patterns

The disagreement between pluralist and domain-general theories is sometimes presented as a choice between five moralities and one. That contrast is useful as a starting point, but it can obscure the more qualified claims made on both sides. Moral Foundations Theory proposes that care, fairness, loyalty, authority, and sanctity address partly distinct and recurrent social problems. Its pluralism is pragmatic and revisable: foundations are intended to identify useful regularities in moral judgement, not to establish five mutually isolated or immutable natural kinds (Graham et al., 2013). The theory consequently allows different foundations to share psychological processes and to be elaborated through cultural learning. Its relevant prediction for the present thesis is not that every domain must possess a unique facial display. It is that the normative context of a violation may contribute information that cannot be reduced completely to a single undifferentiated judgement of wrongdoing.

Domain-general approaches locate the coherence of morality at a broader level. Mind-perception and dyadic accounts emphasise a perceived moral agent, a vulnerable or suffering patient, a norm violation, negative evaluation, and some representation of harm (Gray et al., 2012; Schein & Gray, 2018). On this view, apparently different transgressions can be organised through a common moral template even when their surface contents differ. The strongest version of such an account would lead one to expect that, once the seriousness of wrongdoing and the offender's relation to a victim were represented, little additional organisation would be gained by dividing apology signals into five domains. Yet domain-general theory does not imply that moral content is literally absent or that all contexts must elicit identical judgements. It instead asks whether the shared structure explains more than the proposed domain boundaries do.

The present studies approach this debate from an unusual direction. Most tests of moral-domain structure examine evaluations of transgressions, foundation questionnaires, or verbal descriptions. Here, the question is whether receivers' selected visual configurations retain information about the moral context in which an offender is judged to look more or less sorry. This is a deliberately narrower question than whether the mind contains one or five moral systems. It concerns the organisation of a receiver-side perceptual code: whether a common apology-response structure exhausts the information in that code, or whether moral-domain context also conditions the configurations that receivers use. Standardised moral-foundation vignettes make this comparison tractable (Clifford et al., 2015), but the assigned domain remains a design category rather than independent evidence that receivers themselves represent five discrete kinds.

The clearest evidence for common organisation was the repeated response pattern. In both WC and EA groups, the model-derived Not-sorry, Mild-sorry, and Strong-sorry configurations were organised primarily by response category, with Mild sorry forming the least sharply bounded category. This structure appeared across the five vignette contexts. It also persisted when the comparison was restricted to Mild- and Strong-sorry profiles, indicating that the broad organisation was not merely a contrast between the Not-sorry comparator and apology. A domain-general reading is therefore indispensable: receivers were solving a common judgement problem across all contexts, distinguishing configurations that afforded weaker or stronger inferences of apparent apology. Nothing in this shared response structure requires five independent

apology mechanisms.

Moral context nevertheless contributed reliable information within that pattern. In the population-thresholded representations, the five-domain classifiers were above their restricted participant-label permutation nulls in both groups, with balanced accuracies of .4492 in the WC group and .4827 in the EA group. These values are well below what would be expected if every domain possessed a perfectly distinct prototype, but substantially above the five-class chance level. Moreover, corrected cell-wise differentiation was selective rather than uniform: Loyalty was clearest in the WC analysis, whereas Fairness and Authority were clearest in the EA analysis. The Chapter 5 direct comparison did not support a reliable difference in the overall strength of five-domain separation. Thus, domain information was recoverable in each group without forming a uniform five-part code or being generally stronger in one group.

The localisation analyses clarify what that recoverability means. Domain information was distributed across combinations whose contribution depended on response level, the comparison being made, and the surrounding signals. Conditional ablation and marginal association analyses overlapped only partly, because they asked different questions of the same modelled profiles. A signal that helped distinguish one domain from another in a multivariate classifier did not thereby become an invariant marker of either domain; nor did a marginal association identify an independent domain mechanism. The configuration, rather than an action unit or gaze state, is therefore the appropriate unit of analysis. Moral context changed the diagnostic value of signals within a configuration, but the data did not yield a fixed dictionary in which one facial movement denoted care, another fairness, and so forth.

The shared response pattern and the conditional domain information do not fit either theoretical extreme. Complete invariance would predict that domain labels add no recoverable information once apparent apology intensity is represented; the classifiers contradict that expectation. Conversely, five isolated apology prototypes would predict broad and stable separation among domains; the selective corrected results and the later global geometry do not support that expectation. The pattern is better characterised as *nested or graded pluralism*. It is nested because domain-sensitive configurations occur within a more general organisation of apparent apology response. It is graded because the distinctiveness of domain information varies across contexts and groups rather than appearing as an all-or-none boundary.

Nested pluralism is consistent with the pragmatic form of Moral Foundations Theory, which permits shared processes and culturally elaborated expression (Graham, 2015; Graham et al., 2013). It is also compatible with the proposition that perceived agency, harm, responsibility, and relational disruption supply a common moral template (Schein & Gray, 2018). The evidence therefore does not decisively choose one complete theory over the other. Instead, it specifies the level at which each captures the data. Domain-general organisation is clearest in the broad response pattern, whereas plural organisation appears in the conditional information carried by local configurations. A theory that treats these levels as mutually exclusive would miss part of the empirical structure.

This interpretation also responds to the methodological exchange surrounding moral-domain effects. Scenario severity, unusualness, and perceived harm can create apparent domain differences, but acknowledging these influences does not demonstrate that all residual differentiation is illusory (Graham, 2015; Gray & Keeney, 2015). The present evidence is informative because domain recoverability appeared in both groups and across several analytic views, yet it cannot eliminate the scenario problem: one adapted vignette represented each domain. Domain was therefore inseparable from that vignette's particular ac-

tors, relationships, severity, familiarity, and plausibility. The findings warrant the conclusion that the five designed contexts selectively tuned apology judgements; they do not establish that an abstract foundation, independent of scenario content, caused each difference. Replication with multiple matched vignettes per domain is required before stronger foundation-level claims can be made.

The resulting answer to the moral-domain debate is consequently qualified but substantive. Receivers did not apply one invariant visual apology code regardless of what had been violated. Nor did they sort apologies into five autonomous facial kinds. They interpreted moral-context information within a shared structure of apparent apology intensity. The next question is whether this organisation appeared similarly across the WC and EA groups or whether cultural calibration altered its form.

6.2.2 Cross-Cultural Moral Domains: Shared Organisation and Different Signal Configurations

The cross-cultural evidence reinforces the distinction between common organisation and identical signal configurations. Cultural-development accounts of morality do not require a choice between universal moral content and unrestricted cultural difference. Moral sensitivities may be partly prepared by recurrent social problems while their meanings, virtues, practices, and relative importance are elaborated within particular communities (Graham et al., 2013; Haidt & Joseph, 2004). What is widespread at the level of a social problem need not be expressed through the same judgement, display, or communicative convention. For apology signals, this means that receivers in different settings may confront a common task—assessing whether an offender recognises a breach and is oriented towards repair—while learning different cues and configurations as evidence relevant to that task.

Cross-cultural moral research supplies strong reasons not to treat one population's category boundaries as universal. In some tasks, Western samples have connected immorality more strongly with harmfulness, whereas Chinese samples have more readily moralised conduct described as uncivilised or uncultured (Berniunas et al., 2022; Buchtel et al., 2015). These are sample- and task-specific differences, not fixed properties of "West" and "East", but they show that the extension of the moral category can be culturally calibrated. Likewise, face loss, relational context, and autonomy- or community-oriented construals alter responses to transgression and forgiveness across the cultural comparisons examined by Hui and Bond and by Laham and colleagues (Hui & Bond, 2009; Laham et al., 2010). The relevant implication is that a violation's relational meaning can shape what counts as an appropriate reparative signal.

Measurement research offers a parallel caution. Five-domain instruments can recover recognisable moral dimensions across populations, while their factor structure, item loadings, and associations with other constructs vary (Atari et al., 2023; Dogruyol et al., 2019; Graham et al., 2011). A shared broad pattern therefore does not guarantee an equivalent psychological meaning. This point maps closely onto the thesis findings: moral-domain information was available in both WC and EA groups, but the configurations that carried it need not have been identical. The comparison should accordingly focus on which organisational properties were shared and which multivariate signal configurations differed.

At the broadest domain level, the two groups were similar. Each five-domain classifier performed above its own restricted null, and the Chapter 5 direct comparison found no reliable difference in overall domain separability. The numerical difference between their balanced accuracies could not support the claim that domains were more strongly organised in the EA group. This absence of a reliable difference is not proof of equivalence, but it removes a tempting and unsupported rank-order interpretation. Both groups retained

selective moral-context information within a response-led pattern.

At the level of complete configurations, however, the two groups were reliably distinguishable. Classifiers using the population-thresholded Mild- and Strong-sorry vectors separated the WC and EA groups within all five domains, with balanced accuracies from .830 to .945 and all five corrected permutation tests retained. In matched binary tasks, group classification also exceeded pairwise domain classification for the population-thresholded representation. These comparisons showed that, within this signal space and this analytic design, the complete configurations carried more readily recoverable information about the two groups than about a matched pair of moral domains.

The contrast between whole-configuration and signal-level evidence is theoretically important. Reliable group differentiation did not translate into a list of facial actions or gaze directions that uniformly marked one group. Only one of the fifteen primary response-by-domain prevalence panels survived family-wise correction; direct ablation isolated a small number of facial actions and no gaze predictor; and no individual signal survived the corrected cross-domain variability panels. Broader association analyses identified candidate signal-wise associations without joint correction across signals and answered a different question. The two groups can therefore be distinguished through a distributed pattern even when no single component is consistently diagnostic. The receiver-group difference is more plausibly represented as a change in configuration and cue weighting than as one culturally exclusive movement.

This evidence supports shared organisation together with different signal configurations. The shared component is that domain information was recoverable in both groups and that the direct comparison did not support a difference in its overall strength. The differentiated component is the multivariate route through which apparent apology intensity and moral context were represented in each group. Such a pattern fits accounts in which recurrent moral and relational problems provide common constraints while social learning calibrates the perceptual code used to interpret them. Cultural practices could influence attention, display rules, expectations about deference or affiliation, and the meaning attached to gaze or facial movement. The present studies did not isolate those mechanisms, so they remain explanatory possibilities rather than tested causes.

The term “culture” must itself be used cautiously in describing the empirical contrast. Receiver group varied together with the corresponding generated face-identity set. The estimand is therefore a comparison between the WC and EA groups, not a purified effect of receiver culture. Face morphology, identity sampling, receiver experience, culturally patterned expectations, or interactions among these factors may have contributed to classification. Nor can two sampled groups establish a general West–East division or show that individuals within either group use a homogeneous code. The appropriate cultural interpretation is that the signal configurations were calibrated across the two groups, with the source of that calibration left open.

This qualification does not make the cross-group result theoretically incidental. It indicates where future research must intervene: receiver background and face identity should be crossed factorially, cultural orientation should be measured at the individual level, and multiple populations should be sampled within each broad region. Within the present design, the critical conclusion remains that common moral-domain organisation did not require identical visual configurations. The Chapter 5 correspondence analysis added a complementary result at a different level: after averaging the five domains, selected valence and emotion/social-signal trajectories differed between the groups, but those aggregate differences did not resolve into stable local group or group-by-response contrasts within individual domains. This pattern is evidence of receiver-

group calibration of correspondence, not evidence that one sample was more emotional than the other.

6.2.3 Moral Domains and Moral Emotions: Contextual Correspondence Without One-to-One Labels

Moral emotions offer one possible mechanism through which domain context could tune apology judgements. Emotions such as guilt, shame, embarrassment, anger, contempt, disgust, and compassion connect appraisals of a transgression to action tendencies and relationship regulation. They can orient a person towards avoidance, self-protection, norm enforcement, responsibility, or constructive repair. Importantly, these categories are not interchangeable. Guilt is more consistently associated with evaluating a particular behaviour and taking reparative action, whereas global shame can direct attention towards a devalued self and withdrawal (Tangney et al., 2007). These functional distinctions make moral emotions relevant to apology, but they do not imply that an observed facial configuration transparently displays the named internal state.

Nor does the literature support a simple allocation of one emotion to each of the five foundations. Landmann and Hess found partial specificity: care violations were more clearly related to compassion and purity violations to disgust, whereas fairness, loyalty, and authority were less uniquely differentiated (Landmann & Hess, 2018). Moral emotions also vary with the relational model and cultural context through which a violation is understood (Sunar et al., 2021). More broadly, a disposition to apologise is associated with care, compassion, and positive moral-emotional orientation, but this is a general dispositional bridge rather than a domain-specific nonverbal code (Howell et al., 2011). The literature therefore motivates a contextual analysis while warning against five emotion labels mapped onto five facial prototypes.

Differences between moral contexts were detectable in the complete visual configurations, but they did not correspond reliably to a distinct emotion or social-signal family for each domain. Chapter 5 compared the retained profiles with literature-derived configurations at two levels: valence (L1) and emotion or social-signal type (L2). The comparisons used overall response profiles (M1), response profiles by receiver group (M2-1), response profiles by moral domain (M2-2), and full condition profiles (M3).

The moral-domain result was notably restrained. In the population-thresholded primary analysis, no M2-2 L1 or L2 domain omnibus and no pairwise domain contrast was stably retained after within-family false-discovery-rate control and the Monte Carlo stability gate. At M3, no receiver-group or receiver-group-by-response contrast was stably retained within any individual domain; nor did any within-cell response contrast clear that gate. The results indicated that the present data did not support a reliable allocation of different valence or emotion/social-signal correspondences to the five domain contexts.

Domain classification and emotion correspondence addressed different questions. The classifiers asked whether complete 47-signal configurations carried enough distributed information to recover a domain label within each group. The correspondence tests instead asked whether those profiles differed in their balanced similarity to selected literature-defined valences and signal types. Distributed domain information can be recoverable through contrasts spread across signals without producing a reliable shift towards one named emotion family, one social function, or one pair of domains.

The aggregate group result provides a separate qualification. Once the five domains were weighted equally at M2-1, the response trajectory differed between the groups for negative valence and for several L2 types. Retained patterns included opposite Not-sorry and Strong-sorry group contrasts for contempt and disgust, higher guilt correspondence in the EA group for Mild and Strong sorry, higher internal-standard

correspondence in the EA group for Strong sorry, and higher shame correspondence in the EA group for Mild sorry. Dominance also showed group-dependent response change, with lower EA correspondence for Not and Mild sorry. These are differences in the resemblance of the WC and EA groups to the same literature-derived encodings.

The theoretical implication is that moral emotions may contribute to apology through relational functions rather than through one-to-one facial labelling. Depending on the violation, receivers may attend to configurations compatible with acknowledging a norm, accepting lowered status, signalling non-hostility, expressing self-evaluation, or maintaining affiliation. Guilt-related cues could be relevant when responsibility and corrective action are salient; shame- or embarrassment-related cues could matter when face, status, or public exposure is central; affiliation and reward-related smile morphology could indicate approach or reconciliation rather than pleasure at the offence. These functions are supported as possibilities by the independent literature, not demonstrated causal functions of the generated profiles.

Taken together, the moral-emotion evidence places a firmer boundary around contextual correspondence. The literature makes moral emotions plausible mechanisms, but Chapter 5 did not find stable L1 or L2 separation among the five domains or reliable local group interactions within domains. Moral context remained legible in the complete multivariate configurations, while the available emotion and social-signal references did not provide a one-to-one explanation of that information. This distinction preserves the moral-domain result without treating exploratory source resemblance as confirmatory evidence.

6.2.4 Moral-Domain Synthesis

The moral-domain findings operate at three levels. Apparent apology intensity supplied the broadest and most reproducible organisation across the five contexts. The WC and EA groups expressed that shared response pattern through reliably distinguishable multivariate configurations and differed in selected domain-averaged literature-correspondence trajectories. Moral-domain information was also recoverable within each group, but it neither formed five global regions nor yielded stable domain-to-domain differences in valence or emotion/social-signal-type correspondence.

This combination is best described as *nested or graded pluralism*. The account is pluralist because the kind of violation changed the configuration through which possible repair was read; nested because that differentiation operated within a common response problem; and graded because its strength varied across contexts and groups. Above-null domain classification shows that distributed configurations carried domain information, whereas the absence of global domain alignment shows that this information remained subordinate to broader response and receiver-group gradients. Neither result converts the five domains into autonomous facial kinds or makes moral context irrelevant. The interpretation is therefore compatible with pragmatic pluralism (Graham et al., 2013) while granting domain-general accounts an important role at the level of agency, harm, responsibility, negative evaluation, and repair (Gray et al., 2012; Schein & Gray, 2018).

The evidence remains bounded by five adapted scenarios, two receiver groups coupled to different face-identity sets, generated visual signals, and a specific analytic design. One scenario per domain prevents foundation from being separated cleanly from vignette content; emotion correspondence does not identify felt or recognised emotion; and neither classification nor clustering establishes psychological modules, universality, sincerity, forgiveness, causal social function, or evolutionary origin. Within those limits, the

studies show that the type of wrong conditions what receivers treat as looking apologetic even when the broader response pattern is shared.

The moral-domain answer is therefore hierarchical: receivers interpreted apology-related configurations through a common response pattern whose form differed between groups and was selectively tuned by moral context. This conclusion provides the basis for considering what the same evidence contributes to the apology literature.

6.3 Contributions to the Apology Literature

The findings also revise how the appearance of apology should be conceptualised. Much previous work has asked whether an apologetic person should look sad or remorseful, whether smiling makes an apology appear insincere, and whether particular emotions are necessary for repair. Those questions identify important contrasts, but they can imply that apology has a single facial prototype whose meaning transfers unchanged between situations and cultures. Across the present thesis, receivers instead treated apology as a graded judgement supported by configurations of facial movements and gaze direction. This organisation varied with apparent apology intensity, moral context, and receiver group, and combined smile-related movements with non-smile facial tension and gaze-direction signals. The central contribution is therefore not the identification of one expression that constitutes apology. It is an account of how receivers use mixed visual evidence to distinguish weaker from stronger apparent apology under uncertainty.

This receiver-centred account also requires several distinctions that are sometimes compressed in apology research. A face that looks sorry is not necessarily produced by an offender who feels remorse; a movement associated with a smile does not necessarily communicate happiness; and a display judged appropriate does not necessarily restore trust or elicit forgiveness. The studies measured receiver judgements of controlled generated signals, while the Chapter 5 emotion analysis compared retained-feature profiles with literature-derived positive-support encodings. The resulting correspondences can clarify the possible social information available to a receiver, but they do not identify an internal emotional state, communicative intention, sincerity, or repair outcome. With those boundaries in place, the findings contribute to three connected questions in the apology literature, followed by their synthesis.

6.3.1 Smiling Versus Sad-Looking Apologies

A prominent expectation in the apology literature is that an appropriate apology should display negative affect. Reverse-correlation research has recovered apologetic facial representations characterised by sadness, regret, remorse, and submissiveness, and later work found that generated apologetic faces were judged to be more remorseful and sad than their corresponding base faces (George et al., 2023, 2025). Role-play research has likewise operationalised apology-relevant behaviour through sadness-related facial action and gaze (Yamamoto et al., 2021). This expectation is socially intelligible. Following wrongdoing, a sad or remorse-like appearance can indicate that the offender recognises that something undesirable has occurred and is not pleased by the victim's loss. Conversely, happiness-like behaviour that is conspicuous or norm-incongruent can make a public apology appear less sincere and produce less favourable evaluations (ten Brinke & Adams, 2015). On this evidence, sadness and the absence of pleasure are important resources through which an offender may appear to acknowledge the seriousness of a transgression.

The difficulty is that the category *smile* covers movements with markedly different forms and functions. Smiles can communicate reward and enjoyment, but they can also communicate affiliation, politeness, em-

barrassment, nervousness, deference, or appeasement. Their interpretation depends on morphology, timing, co-occurring movements, and the interactional context in which they occur (Ambadar et al., 2009; J. Martin et al., 2017). This functional heterogeneity matters directly after wrongdoing. A pleasure or dominance smile can contradict the normative work of apology by implying indifference or satisfaction. A restrained affiliative or embarrassed smile may instead soften the encounter, reduce interpersonal threat, or indicate a wish to remain engaged. Blushing research illustrates the point: an embarrassed expression can contain non-Duchenne smile-like movement while still improving evaluation after a mishap or transgression, particularly when combined with visible redness (Dijk et al., 2009). Cross-cultural reports of smiling as one possible nonverbal apology strategy provide a further reason not to assign all lip-corner pulling the same meaning (Park & Guan, 2009). The theoretically relevant contrast is thus not simply whether a smile is present, but which smile-like configuration occurs, when it occurs, and what relational stance it affords a receiver.

Chapter 2 supplies the thesis's principal empirical answer to this debate. Its data-driven models did not converge on either a pure sad face or a pure smiling face. The exploratory factorisation separated a lower-face configuration containing AU14 and AU12-family activity from another configuration dominated by lip stretching, mouth opening, jaw movement, and some AU4- and AU12-containing activity. The first was compatible with smile-related affiliation, whereas the second was more compatible with strain or negative affect. These components coexisted within the wider set of receiver-selected signals rather than defining two mutually exclusive apology types. Gaze also graded the judgement: lateral gaze was most associated with lower-intensity sorry responses, whereas downward gaze became more associated with stronger sorry judgements. Blushing appeared in some individual models but did not form a population-level pattern. The result was therefore a multimodal and configurational repertoire, not a categorical choice between positive and negative expression.

This pattern changes the interpretation of AU12-family activity. Lip-corner pulling contributed to some faces that receivers judged to look sorry, but it cannot be translated directly into happiness. In the generated displays, its meaning was conditioned by dimpling, lip stretch, mouth opening, brow movement, gaze, and knowledge that a transgression had occurred. It could consequently support an affiliative or appeasing reading even while other movements conveyed tension or discomfort. The same qualification applies to downward gaze: it is compatible with shame, submission, reduced dominance, or withdrawal, but Chapter 2 did not ask receivers to identify any of those states. What the study establishes is that these movements formed perceptual evidence for apparent apology. It does not establish why an offender produced them or what the receiver would subsequently do.

The Chapter 5 correspondence analysis reinforces this multifunctional interpretation without replacing the Chapter 2 result. Across response-level profiles, positive-valence correspondence increased at both response transitions, but the component types followed different trajectories. Reward increased from Not to Mild, Mild to Strong, and Not to Strong sorry; affiliation distinguished both apology profiles from Not sorry without a reliable Mild-to-Strong difference; and happiness did not increase at the first transition. The detailed constellation and paired contrasts are considered in Section 6.3.3; here, their relevance is narrower. They allow the smile-related component recovered in Chapter 2 to be situated within a broader graded repertoire in which affiliation, reward, self-evaluation, appeasement, and non-smile signals can overlap. At the same time, the correspondence analysis did not manipulate smile timing or distinguish every pro-

posed smile morphology. It therefore cannot decide whether a particular generated movement was read as affiliative, polite, amused, embarrassed, or rewarding on a given trial.

The smiling-versus-sad-looking debate is consequently better resolved at the level of social function and configuration. Sadness- and strain-compatible cues can support the recognition that harm or norm violation matters. Smile-like cues can support affiliation, politeness, embarrassment, or an attempt to reduce antagonism. Either can fail when it conflicts with the offence, verbal content, timing, or receiver's expectations. The thesis does not show that smiling apologies are generally perceived as more sincere or more effective, nor that sadness is dispensable. It shows that receiver judgements did not converge on a unitary sad-looking prototype and that smile-related movement was not uniformly disqualifying. Direct experiments that cross smile morphology and timing with gaze, verbal content, transgression type, and downstream evaluations will be needed to establish when these candidate functions help or hinder repair.

6.3.2 Cross-Cultural Apologies

Apology is a widely recognisable response to wrongdoing, but its social meaning is culturally situated. Comparative research shows variation in when an apology is expected, how readily it implies personal blame, how it responds to the status of the other person, and which strategies accompany it. Japanese and American participants, for example, have differed in the use of compensation and explanation and in the extent to which apology is adjusted to partner status; other work indicates that Americans may connect apology more closely with individual responsibility than Japanese participants do (Barnlund & Yoshioka, 1990; Maddux et al., 2011). Comparisons involving the United States, China, and Korea likewise show that desire, obligation, and intention to apologise depend on the focal situation rather than following a single national rank order (Guan et al., 2009). Perspective taking can facilitate apology acceptance in more than one cultural setting while operating through different mediating processes (Takaku et al., 2001). These findings make culture relevant not because each culture possesses one fixed apology code, but because the act is embedded in locally learned understandings of responsibility, face, status, and relationship repair.

Socio-ecological evidence strengthens this contextual interpretation. Apology and compensation need not have equal repair value across settings: their usefulness can depend on residential mobility, expectations of continuing relationships, and whether words or material redress provide the more credible indication of future conduct (Komiya et al., 2020). Culture should therefore not be treated merely as a label attached to different surface forms. It can calibrate what relational problem an apology is expected to solve and which form makes that solution plausible. This argument also cautions against transferring findings from verbal strategy directly to facial behaviour. A culture in which verbal apologies are frequent does not necessarily require more of any particular AU or gaze direction; the mapping between verbal practice and visual interpretation is an empirical question.

Adjacent nonverbal research suggests how such calibration may occur. Facial-expression contexts show both cross-cultural commonality and variation, while display rules affect when and how strongly an expression is shown (Cowen et al., 2021; D. Matsumoto, Yoo, et al., 2008). Visual attention also differs: British and Japanese observers do not necessarily allocate attention to the eyes, mouth, and gaze in the same way (Senju et al., 2013). Studies of verbal and nonverbal apology strategies similarly indicate that smiling and other visible behaviours can be used differently across cultural and relational contexts (Park & Guan, 2009). Shared expressive resources can therefore coexist with differences in display, attention, and pragmatic in-

terpretation. This avoids two equally restrictive positions: that the same visual cue must mean the same thing everywhere, or that cultural groups lack any common communicative organisation.

The cross-chapter findings support precisely this combination of common organisation and differentiated realisation. In both WC and EA groups, receiver judgements were organised along a comparable progression of apparent apology intensity. The Chapter 5 comparison nevertheless showed that the complete Mild- and Strong-sorry configurations reliably distinguished the groups within every moral domain, and receiver-group alignment remained when the geometry was restricted to apology profiles. Thus, the groups were not using identical AU and gaze solutions. The difference was clearest at the level of complete configurations rather than as a universal list of isolated culture markers, which is consistent with culturally calibrated codes operating through relations among signals.

A shared progression of apparent apology intensity was compatible with different AU and gaze configurations in the two groups. The within-person Mild-to-Strong comparison showed nearly identical observed mean proportions of binary signal states changing: the EA-minus-WC estimate was .0007 (Holm-adjusted $p = 1.0000$), and every domain-specific interval included zero. Complete configurations classification nevertheless remained strongly group-classifiable. The amount of observed individual change was therefore similar even though the signal space as a whole differ.

The domain-marginal M2-1 analysis provides a complementary direct comparison. Negative-valence correspondence was lower in the EA than the WC group for Not sorry but higher for Strong sorry, and the associated response differences between groups were retained. At L2, the retained group calibration included the same Not-versus-Strong reversal for contempt and disgust, higher guilt correspondence in the EA group for Mild and Strong sorry, higher internal-standard correspondence there for Strong sorry, and higher shame correspondence there for Mild sorry. Dominance also showed retained group differences in the Strong-minus-Mild and Strong-minus-Not trajectories, with lower EA correspondence for Not and Mild sorry. These findings concern how the WC and EA groups resemble a common literature library after the five domains have been weighted equally.

The cultural comparison therefore concerns the WC and EA groups. Receiver group covaried with the corresponding face-identity set, so the direct comparison cannot attribute every difference to receivers' cultural learning independently of face morphology, identity sampling, or receiver-by-face interaction. Nor can two samples represent all Western and East Asian populations. Within that estimand, however, the evidence advances cross-cultural apology research beyond a comparison of verbal frequencies. It shows that observers can share a graded communicative problem—distinguishing weaker from stronger apparent apology—while relying on different visual configurations within the WC and EA groups. Future work should cross receiver populations with common and varied face identities, measure individual cultural orientations and display-rule beliefs, and build culture-matched emotion references before assigning the source of that calibration.

6.3.3 Apologies and Emotions

Emotion is central to apology because an apology asks a receiver to revise an interpretation of the offender. Visible remorse, guilt, shame, sadness, or embarrassment may imply that the offender recognises the wrong, evaluates it negatively, and is oriented towards changed conduct. Yet several distinct phenomena sit between a transgression and that inference: the emotion an offender feels, the expression the offender pro-

duces, the label a receiver applies to it, the configuration encoded in an empirical literature, and the degree to which a receiver judges a generated face to look sorry. These phenomena can covary without being identical. The present thesis directly models the last of them. Its emotion analysis adds a structured comparison with literature-derived configurations, not a measurement of felt, displayed, or recognised emotion.

The distinction is also required by apology theory. Guilt, shame, remorse, and embarrassment are related moral emotions, but they imply different objects of evaluation and behavioural tendencies. Guilt is commonly linked to evaluation of a specific action, responsibility, and reparation, whereas global shame can centre the self, status, or exposure and can motivate concealment as well as corrective conduct (Tangney et al., 2011; Tangney et al., 2007). The offender and receiver can also approach the same apology from different affective positions: offenders' willingness to apologise has been linked to guilt, while victims' desire for an apology has been linked to anger (Leunissen et al., 2013). A receiver's judgement that a face looks sorry should therefore not be treated as a direct reading of the emotion that caused the apology. It may instead be an inference that the visible configuration fits one or more culturally available models of responsibility, non-hostility, self-evaluation, or repair.

Embarrassment makes the functional overlap particularly visible. Its characteristic display has been interpreted as appeasement, and features such as gaze aversion, head movement, smile-like action, and facial redness can alter social evaluation (Dijk et al., 2009; Keltner, 1995; Thorstenson et al., 2020). Embodied displays of remorse can increase perceived remorse, empathy, satisfaction, or favourable appraisal without necessarily increasing forgiveness (Hornsey et al., 2020). Likewise, victim anger and willingness to reject an apology remain sensitive to offence severity and offender responsibility even when the apology appears sincere and remorseful (Bennett & Earwaker, 1994). These results separate the information conveyed by an apparent emotional display from the downstream success of repair. A cue may make responsibility or deference more legible while still being insufficient to restore trust, reduce anger, or reconcile the relationship.

The Chapter 5 response-level analysis supplies the strongest evidence for a mixed repertoire. In the population-thresholded primary representation, positive-valence correspondence increased from Not to Mild and again from Mild to Strong sorry. Negative-valence correspondence did not form the inverse of that trajectory: it decreased at the first transition and increased at the second. Thus, stronger apparent apology did not simply move the profiles from negative to positive affect. It changed the balance among literature-defined signal families. The individual scores described in Chapter 5 show that affiliation was prominent in the apology profiles and that Strong sorry also overlapped with reward, guilt, standard-violation, embarrassment/appeasement, and happiness-related references. Those correspondences do not mean that the Strong-sorry profile expressed several emotions or that the Not-sorry comparator was itself an emotional state.

The L2 response contrasts make the functional distinction sharper. Contempt and reward smile changed reliably across all three planned response contrasts. Reward increased at each response transition, whereas contempt fell from Not to Mild and rose from Mild to Strong. Affiliation increased from Not to Mild and from Not to Strong, but the Mild-to-Strong contrast was not retained. Dominance moved away from its Not-sorry correspondence for both apology profiles, yet Mild and Strong did not differ reliably from one another. Happiness likewise failed to follow a monotonic three-step pattern: the Strong profile differed from Mild and Not, but Mild did not differ reliably from Not. These conclusions use the retention criterion in Chapter 5: an effect was retained only when its within-family q value was at most .05 and its Monte Carlo

decision was stable.

Taken together, the response scores and contrasts indicate a mixed emotion and social-signal repertoire. Functional-smile research independently distinguishes reward, affiliation, and dominance configurations and shows that dominance smiles communicate low positive feeling and social connection despite belonging to the broad smile category (J. Martin et al., 2017; Rychlowska et al., 2017). The Chapter 5 results follow that functional distinction without reducing the response continuum to it: reward tracked all three contrasts, affiliation marked apology relative to Not sorry without differentiating Mild from Strong sorry, dominance separated the apology profiles from Not sorry in the opposite direction, and happiness was non-monotonic. Guilt-, standard-, and appeasement-related correspondence further broadens the repertoire.

The response-by-receiver-group analysis (M2-1) qualifies this response-level account by showing that some trajectories were calibrated differently in the WC and EA groups. These direct receiver-group tests support the aggregate patterns described above. Source-specific encodings document the published morphology contributing to each L2 summary.

This interpretation also clarifies the role of the Not-sorry category. It is a model-derived response-category comparator retained to organise the full response hierarchy. Its negative correspondence with many reference types means that its retained positive-signal profile was dissimilar to those constructed positive-support encodings. Dominance smile was the highest numerical L2 match for Not sorry ($r = .231$) and met the score-retention criterion. Both apology profiles had lower dominance-smile correspondence than Not sorry. The conclusion concerns graded resemblance across profiles.

The contribution to the emotion literature on apology is therefore a disciplined separation of morphology, receiver judgement, and psychological state. Apparent apology intensity reorganised a mixed repertoire rather than adding a single emotion in equal increments: positive valence rose at both transitions, negative valence changed direction, reward graded all three response contrasts, and affiliation, dominance, and happiness followed distinct trajectories. That repertoire can inform hypotheses about norm acknowledgement, lowered threat, responsibility, relationship value, and willingness to repair. Testing those hypotheses will require direct emotion ratings, spontaneous production, culture-matched reference models, and behavioural outcomes. The current evidence establishes which receiver-modelled profiles overlap with encoded signal support; it does not establish that an offender felt remorse, that a receiver recognised a particular emotion, that the resulting apology was perceived as sincere, or that it was effective.

6.3.4 Apology-Literature Synthesis

Across the thesis, visual apology is best understood as a graded, multimodal, and context-sensitive configuration interpreted by receivers. The shared response pattern is clearest in the validation results. Both groups showed the same separability order: Not sorry was most accurately distinguished, Strong sorry was intermediate, and Mild sorry was least accurately distinguished. Mean accuracies for Not sorry, Strong sorry, and Mild sorry, respectively, were 97.8%, 86.2%, and 63.1% in WC and 92.7%, 79.5%, and 48.8% in EA. This is an ordering of model performance, not an assertion that Not sorry is a psychologically stronger state than apology. It indicates that Mild sorry was the least sharply bounded receiver-modelled category, whereas the Not-sorry comparator and Strong-sorry profiles were more concentrated.

Three refinements to apology research follow. First, the appearance of apology cannot be reduced to a choice between smiling and sadness because smiles are not functionally uniform. Reward correspondence

graded every planned response contrast, affiliation separated apology from Not sorry without differentiating Mild from Strong, and dominance followed an opposing apology-versus-Not pattern; happiness itself was not monotonic. Politeness, embarrassment, and appeasement add further functions whose interpretation depends on the wider configuration. Second, the WC and EA groups shared a graded response problem while calibrating both its complete visual solution and selected domain-averaged literature-correspondence trajectories differently. This is group-specific resemblance, not evidence that one group was more emotional. Third, apology-related emotion cues form a mixed repertoire in which valence and type operate at different levels: positive valence rose across both transitions, negative valence changed direction, and guilt, standard violation, embarrassment/appeasement, and other references contributed selectively. No finding requires one canonical apology emotion or a small set of discrete facial apology types.

The resulting account is deliberately narrower than a theory of apology success. The studies show how controlled facial-action and gaze configurations support receivers' judgements that someone looks more or less sorry. They do not show what apologising offenders spontaneously produce, whether the display is sincere, whether responsibility has genuinely been accepted, or whether the victim forgives, restores trust, or resumes the relationship. Those distinctions explain why the same display can be apology-relevant without being sufficient for repair. They also motivate the broader social-signal account developed in the next section: the meaning of an apology display lies not in a single facial category, but in the relational inferences that a configuration makes available to a receiver.

6.4 Apology as a Social Signal: Social Communication, Function, and Evolutionary Origins

The moral-domain and apology-literature findings converge on a receiver-centred problem: how observers infer an offender's orientation towards a breach, a relationship, and possible repair from configurations of behaviour. Social-signal theory provides the framework for integrating those findings. It locates apology meaning in the inferences that perceivers draw from configurations of behaviour, and it allows those configurations to serve several relational functions without assuming that any one movement transparently reveals an inner state. Evolutionary and social-functional perspectives can then help explain why cues concerning affiliation, appeasement, status and norm alignment are plausible resources for social repair. Importantly, however, those perspectives provide an interpretation of the empirical pattern rather than an additional empirical result: the present studies did not test the evolutionary origins, natural production or causal functions of apology displays.

6.4.1 Receiver-Interpreted Social Communication

Social signals convey information about social interactions, relationships, attitudes and emotions. In the broad psychological account adopted here, this information need not have been intentionally encoded by a sender. Poggi and D'Errico distinguish intentionally communicative behaviour from behaviour that is informative because an observer assigns social meaning to it (Poggi & D'Errico, 2010, 2011). A deliberate statement of remorse can therefore be communicative, whereas averted gaze, facial tension or hesitation might be informative even when it was not produced with the purpose of influencing another person. The distinction is particularly important for apology. Following a transgression, receivers must interpret not only what an offender says, but what the offender appears to understand about the wrong, the relationship and the obligations that follow from it. An apology can consequently be treated as a proposal to revise the receiver's beliefs under uncertainty rather than as a formula whose meaning is fixed by its words or

constituent movements.

Several accounts specify the kinds of belief that such a proposal might address. In signalling models, an apology can alter expectations about an offender's trustworthiness or future conduct after a damaging outcome (B. Ho, 2012). Relationship-value accounts propose that an apology can communicate that the damaged relationship continues to matter to the offender (Forster et al., 2021). Responsibility exchange theory instead emphasises the public reassignment of responsibility: apologising accepts a greater share of blame for a negative outcome, although doing so may also incur an image cost (Chaudhry & Loewenstein, 2019). These are related but distinct mechanisms. A receiver might regard an offender as accepting responsibility without believing that the relationship is highly valued, or might recognise a conciliatory orientation without yet restoring trust. The present findings do not adjudicate among these mechanisms. They show that receivers treated visual configurations as graded evidence of how sorry an offender looked, thereby identifying a perceptual basis upon which inferences about responsibility, relationship value or future conduct could subsequently be made.

This receiver-centred emphasis clarifies what is meant by a perceptual "code" in this thesis. The term refers to systematic relations between generated visual configurations and receiver judgements, not to a fixed vocabulary in which each action unit has one stable social meaning. Chapter 2 identified the facial-action and gaze configurations that observers selected when constructing weaker and stronger appearances of apology. Chapters 3 and 4 examined how related configurations were organised across moral-domain contexts separately in the WC and EA receiver groups, and Chapter 5 compared that organisation directly while relating the retained profiles to literature-derived emotion and social-signal references. Across these chapters, the dependent evidence concerns what receivers selected or judged. It does not show which configurations offenders spontaneously produce, whether those configurations were intended as apologies, or what the offenders would have felt.

The model-derived response categories should be interpreted within the same boundary. A face classified as relatively strong apology evidence is not thereby a sincere offender, and the Not-sorry comparator is not a measured emotional state or a naturally occurring absence of apology. Rather, the graded structure describes how observers differentiated configurations within the experimental task. Its theoretical value lies in showing that apology-related meaning can be organised on the receiver side before a verbal formula, an attributed intention or a repair outcome is introduced. This is consistent with a social-communication account in which meaning emerges through inference from available evidence, but it leaves open whether the same evidence would be produced or interpreted in a natural interaction.

The distinction between informative and intentionally communicative behaviour also prevents two common inferential shortcuts. First, a cue can help a receiver judge apology without having been strategically designed or deliberately displayed by an offender. Secondly, a deliberately produced display can fail to convey the intended message if the receiver interprets it differently. Sender intention and receiver uptake are therefore analytically separable. The present studies isolate the latter side of this relation: they estimate which visual configurations receivers recruit as apology evidence under specified moral and cultural contexts. This makes the findings relevant to social communication while avoiding the stronger claim that they reveal an offender's apology-production programme.

6.4.2 Multimodal and Multifunctional Signalling

Receiver inference usually operates over constellations of cues rather than isolated movements. Social Signal Processing treats facial behaviour, gaze, gesture, posture and vocal behaviour as interdependent sources of information about interaction (Vinciarelli et al., 2009). Similarly, the face is a dynamic communication system in which movement, morphology, timing and complexion can jointly inform judgements about emotion, intention, identity, status and relational stance (Jack & Schyns, 2015). Even when an analysis focuses on the face, its constituent movements need not carry invariant meanings independently of their timing or co-occurrence. The empirical emphasis on configurations is therefore not only a modelling choice; it corresponds to the way social information can be distributed across multiple simultaneous and sequential cues.

The findings across Chapters 2–5 fit this configurational account. The initial discovery study did not yield one fixed set of facial expressions. Smile-related lower-face movement occurred alongside facial tension or opening, and gaze changed the strength with which some configurations were interpreted. The later studies likewise found that complete response profiles distinguished the experimental conditions more consistently than any short, universally diagnostic inventory of individual signals. Signal-level analyses helped localise selected differences, but they did not justify assigning one action unit or gaze direction a stable apology meaning across response levels, moral domains and WC and EA groups. Thus, the relevant unit of interpretation is the configuration as encountered by a receiver, not a decontextualised facial component.

The emotion-linkage findings reinforce this point by showing that stronger apology profiles corresponded with a mixed set of literature-derived references. Affiliation and reward-related smile morphology appeared alongside references concerning guilt, standards, embarrassment and appeasement. This pattern makes the familiar opposition between a smiling and a sad-looking apology too coarse. Smiles themselves can differ in morphology, dynamics and perceived social meaning (Ambadar et al., 2009). In particular, reverse-correlation work distinguishes reward and affiliation smiles from an asymmetric dominance smile associated with superiority, low positive feeling, and low social connection (Rychlowska et al., 2017); social-functional theory treats these patterns as addressing reinforcement, bonding, and hierarchy negotiation rather than as interchangeable variants of happiness (J. Martin et al., 2017). Perception studies further show that expresser culture and smile intensity alter ratings of positivity, negativity, authenticity, and politeness, while spontaneous polite and amused smiles differ in duration and temporal symmetry (Fang et al., 2020; Hoque et al., 2011). Chapter 5 followed the source-defined functional distinction that its static AU/gaze representation could encode: reward correspondence increased across all three response contrasts; affiliation and dominance distinguished both apology profiles from Not sorry without separating Mild from Strong; and happiness was not monotonic. Embarrassment displays can have appeasing consequences (Keltner, 1995), and emotion displays more generally can coordinate relationships rather than merely externalise private feeling (Keltner & Haidt, 1999). Accordingly, smile-related movement in an apology configuration might contribute to affiliation, politeness, nervousness, appeasement or positive re-engagement, whereas a superficially similar movement in another configuration might be read as amusement, pleasure at the offence or dominance. Its meaning depends on the rest of the signal, the transgression and the receiver's expectations.

The same reasoning applies to cues commonly regarded as difficult to control. Blushing can provide unbidden information about sensitivity to social evaluation and shared standards (Castelfranchi & Poggi,

1990). More conspicuous embodied displays, such as crying or kneeling, can increase attributed remorse in public-apology settings (Hornsey et al., 2020). Such evidence explains why receivers may attend to visible bodily conduct when words alone are ambiguous. It does not establish that embodied conduct is inherently truthful. People can misread spontaneous behaviour, deliberately manage apparently spontaneous displays, or apply a learned stereotype to a cue that has a different cause. Moreover, an increase in perceived remorse need not produce forgiveness or restore a relationship. Embodiment, controllability, attributed sincerity and effectiveness are therefore separate dimensions rather than interchangeable indicators of a genuine apology.

This multifunctional interpretation also resolves an apparent tension in the empirical results. Positive-valence correspondence can rise at both response transitions without implying that receivers think the offender feels happy about the transgression; happiness itself did not show the same monotonic trajectory. The relevant positive information may concern orientation towards renewed interaction rather than pleasure. The separation among reward, affiliation, and dominance strengthens that interpretation: reward graded the response continuum, whereas affiliation and dominance distinguished apology from Not sorry without differentiating its Mild and Strong forms. At the same time, guilt-, shame- or standard-related cues may indicate recognition of wrongdoing without being either necessary or sufficient for a strong apology judgement. What appears to be a heterogeneous emotional repertoire can thus be coherent at the level of relational work: different elements may contribute to reducing antagonism, accepting a lowered position, acknowledging a norm, inviting re-engagement or showing readiness to repair. These functions remain theory-guided interpretations of correspondence with reference configurations; the analyses did not recognise emotions or measure their experience.

The thesis nevertheless samples only part of a fully multimodal apology. Chapter 2 used controlled dynamic visual stimuli, while the later profile analyses concentrated on retained facial-action and gaze states. They did not include speech, prosody, posture, gesture, interpersonal distance, touch, turn-taking or the receiver's reciprocal response. Nor did the reduced profiles preserve all temporal qualities that can distinguish otherwise similar smiles and expressions. The present configurational account should therefore be understood as a visual foundation for a broader multimodal model. It establishes that facial action and gaze already support graded, multifunctional receiver judgements, while leaving open how verbal, vocal, bodily and temporal information would reinforce, qualify or reverse those judgements in interaction.

6.4.3 Shared Social Functions and Cultural Calibration

Social-signal theory also offers a way to interpret the combination of shared organisation and cultural differentiation. Human groups repeatedly confront similar interactional problems after wrongdoing: whether the offender recognises the breach, remains antagonistic, accepts responsibility, values continued association and is likely to change future conduct. Comparable problems do not require identical signal forms. Cultural conventions, display rules and learned attention can calibrate which behaviours are appropriate, conspicuous and persuasive in a particular setting. Research on emotional expression accordingly points to both cross-cultural regularities and substantial variation in the categories, configurations and meanings that people perceive (Cowen et al., 2021). Display rules shape when and how emotion-related behaviour is shown (D. Matsumoto, Yoo, et al., 2008), while gaze is attended to and interpreted through culturally patterned expectations (Senju et al., 2013). A common social task can therefore be realised through more than one perceptual code.

The cross-chapter findings are compatible with this account of shared function and calibrated form. The WC and EA groups exhibited a common graded organisation of apology response, and the analyses detected no reliable between-group difference in the overall amount of signal-state change between weaker and stronger profiles. At the same time, their complete signal configurations were reliably distinguishable, and the domain-marginal emotion analysis found selected group differences in valence- and type-level response trajectories. The combination matters: similarity in the structure or degree of apology judgement did not entail identity in the facial-action and gaze states through which that judgement was realised or in their resemblance to a common literature library. Conversely, differentiated configurations and correspondence trajectories did not eliminate the shared organisation of apparent apology intensity. Social-signal theory accommodates these observations by treating the communicative problem and its conventional solution as related but analytically distinct.

This conclusion must remain tied to the empirical estimand described above. The observed difference is between the WC and EA groups rather than an isolated effect of receiver culture. It may reflect learned display expectations, attention, experience with faces, stimulus identity or interactions among these factors. The direct M2-1 tests established aggregate differences in resemblance to the common reference library, including robust group calibration for negative valence and selected contempt, disgust, guilt, standard, shame, and dominance trajectories. Because the reference models were not culture matched, these differences do not reveal which mechanism produced the calibration or whether the named category had an equivalent meaning in the two populations. They particularly do not support the claim that one group was more emotional than the other.

Moral context adds a second, empirically distinct question. Moral Foundations Theory proposes that concerns involving care, fairness, loyalty, authority and purity can acquire partly distinct social meanings through an interaction between prepared sensitivities and cultural development (Graham et al., 2013; Haidt & Joseph, 2004). In the present studies, moral-domain information was recoverable from complete configurations within the broader response structure, yet the domains did not form five globally separate apology types. Moreover, the Chapter 5 correspondence analysis found no stable L1 or L2 domain omnibus or domain-pair contrast and no stable local group or group-by-response contrast within domains. From a social-communication perspective, different violations could still alter relations and obligations—harm may foreground welfare, unfairness responsibility and balance, disloyalty or disrespect group commitment or status, and purity local concepts of contamination or propriety—but the present emotion-reference library did not identify a distinct signal function for each. These remain hypotheses about relational content, not direct labels for the observed facial profiles.

Culture and moral context should therefore not be treated as independent catalogues of fixed expressions. They jointly shape the inference problem faced by the receiver: what kind of breach occurred, what response the relationship requires, and which visible conduct counts as adequate evidence in that setting. The empirical pattern suggests a nested organisation in which a broadly shared pattern of apparent apology intensity is realised through configurations that differed across the groups, while domain information was carried by distributed contrasts rather than by reliably different literature-defined emotion profiles. This is more precise than claiming either one universal apology configuration or one receiver-group-specific configuration for each offence. It also explains why configuration-level group differences can coexist with the absence of a simple set of receiver-group-specific action units. Calibration can affect relations among cues,

their relative diagnosticity and the contexts in which they are combined, rather than assigning exclusive ownership of a movement to one receiver group.

The phrase "shared function, calibrated form" describes a synthesis that still needs to be tested; it does not demonstrate equivalent functions. The studies measured whether configurations looked more or less sorry, not whether receivers separately inferred threat reduction, norm acknowledgement, affiliation or willingness to repair. Similar ratings could arise through partially different inferences, and culturally distinguishable configurations could ultimately serve different relational purposes. Direct tests would need to manipulate candidate configurations across crossed receiver and face-identity groups, measure the proposed inferences separately, and examine whether those inferences mediate downstream responses. The current evidence establishes the organisation of receiver judgements and makes functional calibration a plausible account of that organisation.

6.4.4 Evolutionary and Functional Foundations

An evolutionary perspective can broaden this account, provided that evolutionary origin is kept separate from current receiver interpretation. Facial expressions have been analysed as adaptations that make socially relevant intentions and motivations visible (Schmidt & Cohn, 2001). Social-functional theories similarly propose that emotions coordinate action at individual, dyadic, group and cultural levels, including affiliation, hierarchy, conflict and relationship regulation (Keltner & Haidt, 1999). These approaches suggest why human perceivers may be especially responsive to facial and bodily information after a transgression: recurrent problems of conflict, exploitation, status and reintegration create value in detecting whether another person remains threatening or is oriented towards renewed cooperation. They do not imply that a contemporary apology, or any specific experimental configuration, is itself an evolved unit.

Indeed, biological capacity and culturally articulated moral practice need not be competing explanations. Ayala distinguishes the evolved capacities that make moral behaviour possible from the culturally developed rules through which morality is expressed (Ayala, 2010). The same distinction can be applied to apology. Contemporary apologies are institutionally and linguistically elaborated practices, but they may recruit older perceptual sensitivities to gaze, submission, distress, approach and affiliation. The social meaning of a constituent movement can then be reshaped by conventions concerning politeness, hierarchy and appropriate remorse. On this view, evolutionary history may help explain the availability of certain communicative resources, whereas cultural learning and local moral norms help explain how those resources are assembled and interpreted.

Embarrassment and appeasement provide a concrete illustration. Embarrassment displays can reduce negative evaluation and signal an orientation towards prevailing social standards (Keltner, 1995); blushing has likewise been theorised as counter-voluntary evidence of concern about others' judgements (Castelfranchi & Poggi, 1990). Affiliative or appeasing displays could, in principle, lower perceived threat and make reintegration less risky after wrongdoing. The present correspondence of stronger apology profiles with affiliation, embarrassment/appeasement, guilt and standard-related references is compatible with this functional possibility. It does not show that the selected configurations actually evolved for appeasement, that an offender displaying them would feel embarrassed, or that a receiver would forgive. The function is supplied by the independent literature; the thesis contributes receiver-side evidence that the relevant visual resources overlap with what observers use to judge apparent apology intensity.

Costly-signalling theory addresses a related problem of credibility. Where offenders benefit from appearing conciliatory regardless of their true intentions, a cheap statement can be imitated by both sincere and insincere actors. In general signalling theory, costs or constraints can stabilise credibility when they make deceptive production comparatively disadvantageous (McAndrew, 2019; Zahavi, 1975). Apology research applies this reasoning by showing that reparative acts carrying meaningful costs can influence receivers' judgements of sincerity or conciliatory intention (Ohtsubo & Watanabe, 2009; Ohtsubo et al., 2018), including in group-apology contexts (Ohtsubo et al., 2020). This literature helps explain why receivers might search beyond the words "I am sorry" for evidence of commitment, sacrifice or reduced exploitation risk.

It would nevertheless be incorrect to convert the present facial profiles into costly signals. No production cost, constraint or differential ability to fake a configuration was measured. A facial movement that looks effortful is not necessarily costly in the technical sense, and a less controllable cue is not automatically honest. Even embodied remorse can alter perceived remorse without producing forgiveness (Hornsey et al., 2020). The thesis therefore cannot show that affiliation, gaze aversion, tension or embarrassment-related morphology guarantees sincerity. At most, its findings identify visual evidence that receivers treat as apology-relevant and that could participate in credibility judgements when combined with information about timing, context, voluntary control and material reparation.

The evolutionary interpretation must therefore remain conditional. A broadly shared response pattern is compatible with common constraints on social coordination after transgression. Cultural and moral-domain tuning are compatible with the developmental and conventional calibration of signal forms. A mixed repertoire of affiliative, appeasing and norm-related correspondences is compatible with the reuse of social behaviours that regulate approach, threat, status and belonging. None of these compatibilities demonstrates ancestral selection pressure, universality or adaptive function. Current usefulness also cannot, by itself, establish historical origin: a behaviour may serve a contemporary communicative role without having been selected for that role, and a convention may recruit older movements while giving them a new local meaning.

Establishing function more directly would require evidence from both sides of the interaction. Naturalistic and experimental production studies could test when offenders generate candidate configurations, whether their occurrence depends on felt emotion, intention, relationship value or anticipated cost, and how readily the displays can be controlled. Receiver studies could independently manipulate morphology, timing, gaze, verbal content and reparative cost, then measure attributed responsibility, threat, affiliation, perceived sincerity and expected future conduct as distinct outcomes. Finally, interactional studies would need to determine whether these inferences change trust, forgiveness, reconciliation or reoffending. Until such evidence is available, evolutionary and functional perspectives are best used to generate mechanisms that the current receiver-centred pattern makes plausible, not to redescribe those mechanisms as findings.

Taken together, the social-signal account represents visual apology as interpreted by receivers, configurational, multifunctional and contextually calibrated. It links the moral-domain and apology contributions through a shared problem of social repair whose visual form differed across the two sampled groups and moral contexts. Evolutionary and social-functional literatures make affiliation, appeasement, credibility and norm alignment plausible dimensions of repair. The evidential boundary remains clear: the thesis maps a receiver-side perceptual organisation of generated signals, not offenders' intentions or emotions, signal honesty, evolutionary adaptation, or successful relationship repair. These boundaries motivate the limitations

considered next.

6.5 Limitations and Future Directions

The main limitations concern the scenarios, receiver judgements, sampled groups and faces, available signals, and analysis choices. These constrain conclusions about moral domains, culture, emotion, natural apology production and successful repair.

6.5.1 Moral Domain and Scenario Are Not Separately Identified

Each of the five moral domains in Chapters 3–5 was represented by one adapted transgression vignette. Moral domain was therefore confounded with the content of its scenario. A configuration classified as domain-related could instead reflect severity, perceived harm, weirdness, disgust, intentionality, familiarity, plausibility, relationship structure, victim relevance, or another feature that differed among the vignettes. This concern is particularly important in the Graham–Gray debate because scenario sampling and unmatched weirdness or severity can create apparent domain specificity or conceal it (Graham, 2015; Gray & Keeney, 2015). The use of standardised moral-foundation vignettes improves conceptual control (Clifford et al., 2015), but adaptation into victim-centred apology contexts means that validation of the original items cannot automatically validate every property of the adapted scenarios.

Intentionality warrants particular attention. The scenario-selection criteria constrained social status and several contextual features, but intentionality was neither independently manipulated nor explicitly matched across the adapted vignettes. Previous research shows that intentionality shapes victims' desire for an apology, offenders' willingness to apologise, and the relationship between apology and forgiveness (Leunissen et al., 2013; Struthers et al., 2008). These findings motivate the possibility that receivers expect different visual apology signals after accidental and deliberate wrongdoing; they do not establish a difference in facial signals. Though some scenario properties may be substantive features of different kinds of wrongdoing, as Graham's discussion of severity and typicality illustrates (Graham, 2015), this does not diminish the relevance of intentionality or establish that it explains the present domain differences. The current design cannot separate contributions from perceived intentionality of the transgression, moral content, and their combination.

Severity and interpersonal relevance require similar caution. Adapting the scenarios placed participants in the affected person's position, but this did not establish that the violations were equally severe or equally direct in their interpersonal consequences. The purity scenario referred to the participant's own rubbish and discarded underwear and could therefore be interpreted as a personal intrusion; nevertheless, its interpersonal meaning may differ from that of anticipated bodily harm, cheating, or direct interruption. Such differences could contribute to the recovered configurations alongside moral content. They do not establish that purity violations are inherently less interpersonal.

Relationship type was constrained only partly. Scenario screening sought a direct offender–victim relation and approximately equal, otherwise unspecified social status, but equal status does not imply equal familiarity or closeness. Leaving the relationship unspecified can also allow receivers to infer different relationships; the loyalty vignette explicitly described a fellow employee in the participant's company. Previous vignette research found that describing the same accidental transgression as involving a friend or a stranger changed reported apology obligations and intentions (Guan et al., 2009). This supports treating relationship context as relevant, without demonstrating a corresponding difference in facial signals. The present design

therefore cannot determine whether the same configurations guide apparent-apology judgements when the offender is a close friend or a stranger.

The domain results should consequently be interpreted as differences among the five domain-labelled transgression contexts used here. Above-null classification establishes that their apology configurations were distinguishable within the present representations; it does not establish that the irreducible source of that difference was a moral foundation. Likewise, the absence of global five-domain alignment does not demonstrate that domains are psychologically unreal. Both conclusions depend on how moral content was instantiated.

Future studies should sample multiple independently constructed scenarios within each domain and treat scenarios as sampled stimuli in a hierarchical design. They should measure perceived severity, harm, disgust, weirdness, intentionality, familiarity, social distance, relationship type, and domain membership rather than relying solely on the assigned category. A cross-classified model could then estimate what generalises across scenarios within a proposed domain and what is specific to one event. Including ambiguous and cross-domain violations would also test whether apology signals track categorical foundations, continuous appraisals, relational models, or combinations of these properties. In particular, accidental and deliberate versions of several comparable transgressions should be crossed with moral domain, with perceived intentionality checked independently. Checks should distinguish intending the action from intending its harmful or norm-violating consequences. Presenting the same generated facial configurations in both versions would test whether describing a transgression as accidental or deliberate changes apparent-apology judgements, while separate production studies could establish whether offenders also display different signals. Severity and interpersonal relevance should be checked for the adapted versions and, where feasible, matched or varied independently. Relationship closeness should likewise be manipulated within otherwise comparable transgressions, for example by describing the offender as a close friend or a stranger.

6.5.2 Response Categories and Receiver-Side Construct Scope

The studies measured how sorry a generated face appeared, a narrower question than whether a complete apology had occurred. An apology can also involve acknowledgement, responsibility, explanation, repair, sincerity, and interaction with a recipient. Although *apology* and *sorry* are sometimes used interchangeably for readability, they should not be treated as equivalent constructs.

Not sorry, Mild sorry and Strong sorry were derived from each receiver's ratings; they were not three kinds of apology presented experimentally. Not sorry is neither a no-apology condition nor an observed emotional state. The less clearly bounded Mild category may reflect flexible intermediate signals, greater variation in moderate judgements, or difficulty modelling those judgements.

Future work should examine the original graded ratings when possibly and test whether intermediate judgements form one broad pattern or several distinct patterns, rather than assuming fixed apology types. It should also compare apparent sorriness with judgements of sincerity, remorse, responsibility, willingness to repair and expected future conduct.

The studies identify signals receivers use to judge apparent apology; they do not show what offenders naturally produce, intend to communicate, or feel. An offender may look sorry without feeling remorse or changing subsequent behaviour, while a sincere offender may not display the signals a receiver expects.

Apparent apology intensity is not a measure of perceived sincerity, forgiveness, trust or reconciliation.

Physical displays can change perceived remorse without necessarily increasing forgiveness (Hornsey et al., 2020). Even an apology described as sincere and remorseful may be accepted under social pressure while anger and rejection vary with responsibility and offence severity (Bennett & Earwaker, 1994). The identified signals are candidates for tests of repair, rather than evidence that repair occurred.

Future work should record natural apologies and test reconstructed configurations alongside verbal apology components. Separate measures of perceived emotion, responsibility and sincerity, together with forgiveness, trust and later behaviour, would distinguish what offenders produce, what receivers infer, and whether the interaction improves.

6.5.3 Cultural, Identity, and Demographic Scope

The WC and EA groups judged different face-identity sets, so the comparison cannot isolate receiver culture from face identity. Group differences may reflect learned expectations, experience with particular faces, differences between the faces, or interactions between receivers and faces. The findings therefore describe these two sampled groups rather than a pure causal effect of culture.

The labels Western and East Asian conceal variation within societies, including language, migration history, religion, social position and local ecology. Such variation can be as informative as average national differences (Atari et al., 2023; Komiya et al., 2020). Two samples cannot establish a universal East–West distinction.

Future studies should ask receivers from several populations to judge common and varied face identities, so that receiver and face effects can be distinguished. They should check whether "sorry" judgements function comparably across languages and samples, and examine cultural experience and demographic variation directly. Gender, status and offence severity can affect accounts and their evaluation (Gonzales et al., 1990); the present studies could not isolate these interactions.

6.5.4 Signal-Space and Interactional Coverage

The generative method offered unusually broad and controlled sampling of dynamic facial behaviour, but it still captured only part of an apology. Chapter 2 varied 42 AUs, five gaze states, and blushing. The later moral-domain analyses retained binary support for 42 AUs and five gaze states. This reduction made population-level comparison tractable, but it compressed temporal parameters, movement intensity, and coordination, and it omitted blushing from the later 47-signal profiles. Head orientation, tears, posture, gesture, interpersonal distance, touch, vocal prosody, speech content, turn timing, and the receiver's response were also outside those profiles. Given the limited number of trials, interactions between AUs were not systematically tested.

Non-facial cues can therefore contribute substantive information rather than merely accompany a facial expression. In medical-error disclosures, keeping the verbal content constant while changing a bundle of nonverbal involvement cues: including body orientation, touch, vocal qualities, and interaction management—changed patients' evaluations of the disclosure (Hannawa et al., 2016). Because these cues varied together, this evidence does not isolate the contribution of any one channel. In public-apology experiments, physical displays such as kneeling and crying increased perceived remorse, while their effects on forgiveness were negligible (Hornsey et al., 2020).

These omissions matter because social signals are configurational and sequential. The same smile related movement may be affiliative, embarrassed, nervous, dominant, or amused depending on its morphol-

ogy, intensity, onset, duration, symmetry, accompanying gaze, and interactional context (Ambadar et al., 2009; Fang et al., 2020; Hoque et al., 2011; J. Martin et al., 2017; Rychlowska et al., 2017). The analysis of static representation also cannot preserve the dynamic parameters through which functional smile models were constructed. Downward gaze may communicate submission or shame in one setting but disengagement or deference in another. A verbal acknowledgement of responsibility can stabilise one interpretation of an ambiguous face, while an excuse or delayed response may destabilise it. Binary feature support cannot represent these temporal and cross-modal contingencies. The present studies consequently cannot establish the relative contribution of facial signals within a complete apology, whether those signals add information beyond speech and bodily behaviour, or whether their meaning changes when channels agree or conflict.

The findings nevertheless provide a stepping stone for such research by identifying empirically grounded candidate visual configurations. Instead of choosing facial displays solely from theoretical prototypes or varying every possible facial combination alongside additional channels. Future experiments could begin with these candidates and cross them with selected verbal, vocal, and bodily cues. This would make targeted tests of cross-channel interaction more tractable. The candidates are not a validated minimal or exhaustive inventory, however, and their contribution must be re-established in the richer setting. Dynamic reconstructions could vary smile morphology and timing, gaze trajectory, head movement, posture, prosody, and verbal acknowledgement of responsibility while holding the transgression constant. Information-theoretic or hierarchical models could test whether channels combine additively, redundantly, synergistically, or in a temporally ordered manner. Independent receiver judgements and interactive trust or cooperation tasks should assess apparent apology, perceived sincerity, and behavioural repair separately, including whether an interpretation changes after the receiver responds and whether repair is sustained by the offender's subsequent behaviour.

6.5.5 An Integrated Programme for Future Research

Future research should connect natural production, receiver interpretation and the consequences of apology. Candidate configurations should be tested for their effects on perceived norm recognition, sincerity, forgiveness, trust and cooperation. Following these effects through ongoing interactions would test whether the offender's later behaviour supports the initial display.

Studies should include multiple scenarios per moral domain, receiver populations and face identities, alongside dynamic multimodal signals and behavioural outcomes. Separating what is shared from what varies across these contexts would test the thesis's central proposal: apology addresses common problems of social repair through forms that vary with moral, cultural and interactional context.

6.6 Conclusion

This thesis answers the moral-domain question with a hierarchical account. Apparent apology intensity supplied the broadest shared organisation, represented in Chapters 3–5 by the model-derived response categories; the WC and EA groups differed in the complete visual-signal configurations through which that organisation was expressed; and moral context selectively tuned distributed configurations without partitioning the global signal space into five regions or producing one emotion-related profile per domain. The findings are therefore most consistent with nested or graded pluralism: common social-repair organisation coexists with local, context-sensitive differentiation.

The apology question has a similarly configurational answer. Receivers treated dynamic visual signals

as graded evidence about how an offender appeared to orient towards the norm, the relationship, and possible repair, while literature correspondence showed resemblance to a mixture of potential social functions rather than one-to-one emotional labels. These findings concern receiver interpretation, not the offender's felt emotion, sincerity, or success in repairing the relationship. The next step is to test whether offenders naturally produce these configurations and whether they causally shape consequential interaction. Within the scope of this thesis, looking sorry is neither one emotion nor one fixed set of facial expressions; it is a receiver-interpreted, multimodal, and contextually calibrated social signal.

References

Aijmer, K. (2019). 'Ooh whoops I'm sorry!' teenagers' use of English apology expressions. *Journal of Pragmatics*, *142*, 258–269. https://doi.org/10.1016/j.pragma.2018.10.017

Allan, A., Allan, M. M., Kaminer, D., & Stein, D. J. (2006). Exploration of the association between apology and forgiveness amongst victims of human rights violations. *Behavioral Sciences & the Law*, *24*(1), 87–102. https://doi.org/10.1002/bsl.689

Ambadar, Z., Cohn, J. F., & Reed, L. I. (2009). All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior*, *33*, 17–34. https://doi.org/10.1007/s10919-008-0059-5

Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S. T., & Dehghani, M. (2023). Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, *125*, 1157–1188. https://doi.org/10.1037/pspp0000470

Ayala, F. J. (2010). The difference of being human: Morality. *Proceedings of the National Academy of Sciences*, *107*(Supplement 2), 9015–9022. https://doi.org/10.1073/pnas.0914616107

Barnlund, D. C., & Yoshioka, M. (1990). Apologies: Japanese and American styles. *International Journal of Intercultural Relations*, *14*(2), 193–206. https://doi.org/10.1016/0147-1767(90)90005-H

Beal, B. (2020). What are the irreducible basic elements of morality? a critique of the debate over monism and pluralism in moral psychology. *Perspectives on Psychological Science*, *15*(2), 273–290. https://doi.org/10.1177/1745691619867106

Bennett, M., & Earwaker, D. (1994). Victims' responses to apologies: The effects of offender responsibility and offense severity. *The Journal of Social Psychology*, *134*(4), 457–464. https://doi.org/10.1080/00224545.1994.9712196

Berniunas, R., Silius, V., & Dranseika, V. (2022). Moralization east and west: Moralizing different transgressions among Chinese, Americans, and Lithuanians. *Asian Journal of Social Psychology*, *25*, 185–197. https://doi.org/10.1111/ajsp.12479

Beyens, U., Yu, H., Han, T., Zhang, L., & Zhou, X. (2015). The strength of a remorseful heart: Psychological and neural basis of how apology emolliates reactive aggression and promotes forgiveness. *Frontiers in Psychology*, *6*, 1611. https://doi.org/10.3389/fpsyg.2015.01611

Bhushan, B., Basu, S., & Dutta, S. (2020). Revisiting guilt, shame, and remorse. *Psychological Studies*, *65*(3), 247–260. https://doi.org/10.1007/s12646-020-00561-z

Bippus, A. M., & Young, S. L. (2020). How to say "I'm sorry:" ideal apology elements for common interpersonal transgressions. *Western Journal of Communication*, *84*(1), 43–57. https://doi.org/10.1080/10570314.2019.1610787

Blatz, C. W., & Philpot, C. (2010). On the outcomes of intergroup apologies: A review. *Social and Personality Psychology Compass*, *4*(11), 995–1007. https://doi.org/10.1111/j.1751-9004.2010.00318.x

Blum-Kulka, S., & Olshtain, E. (1984). Requests and apologies: A cross-cultural study of speech act realization patterns (CCSARP). *Applied Linguistics*, *5*(3), 196–213. https://doi.org/10.1093/applin/5.3.196

Bonensteffen, F., Zebel, S., & Giebels, E. (2020). Sincerity is in the eye of the beholder: Using eye tracking to understand how victims interpret an offender's apology in a simulation of victim–offender mediation. *Frontiers in Psychology*, *11*, 835. https://doi.org/10.3389/fpsyg.2020.00835

Bosche, E. P., Ludwig, J. M., & Schumann, K. (2026). You're not so bad after all: Apologies and guilt promote perceptions of moral restoration after wrongdoing. *Journal of Experimental Social Psychology*, *124*, 104879. https://doi.org/10.1016/j.jesp.2026.104879

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press. https://doi.org/10.1017/CBO9780511813085

Buchtel, E. E., Guan, Y., Peng, Q., Su, Y., Sang, B., Chen, S. X., & Bond, M. H. (2015). Immorality east and west: Are immoral behaviors especially harmful, or especially uncivilized? *Personality and Social Psychology Bulletin*, *41*(10), 1382–1394. https://doi.org/10.1177/0146167215595606

Burgoon, J. K. (2015). Expectancy violations theory. In C. R. Berger, M. E. Roloff, S. R. Wilson, J. P. Dillard, J. P. Caughlin, & D. H. Solomon (Eds.), *The international encyclopedia of interpersonal communication* (pp. 1–9). Wiley. https://doi.org/10.1002/9781118540190.wbeic102

Castelfranchi, C., & Poggi, I. (1990). Blushing as a discourse: Was Darwin wrong? In R. Crozier (Ed.), *Shyness and embarrassment: Perspectives from social psychology* (pp. 230–251). Cambridge University Press.

Chang, W.-L. M. (2008). Australian and Chinese perceptions of (im)politeness in an intercultural apology. *Griffith Working Papers in Pragmatics and Intercultural Communication*, *1*(2), 59–74.

Chang, W.-L. M., & Haugh, M. (2011). Evaluations of im/politeness of an intercultural apology. *Intercultural Pragmatics*, *8*(3), 411–442. https://doi.org/10.1515/iprg.2011.019

Chaudhry, S. J., & Loewenstein, G. (2019). Thanking, apologizing, bragging, and blaming: Responsibility exchange theory and the currency of communication. *Psychological Review*, *126*(3), 313–344. https://doi.org/10.1037/rev0000139

Chiu, C.-y., Dweck, C. S., Tong, J. Y.-y., & Fu, J. H.-y. (1997). Implicit theories and conceptions of morality. *Journal of Personality and Social Psychology*, *73*(5), 923–940. https://doi.org/10.1037/0022-3514.73.5.923

Cho, M., Bae, J., Cho, E., & Bae, S. H. K. (2026). Polite or playful? how service robot type and apology style influence negative word-of-mouth after service failures. *Computers in Human Behavior*, *181*, 108995. https://doi.org/10.1016/j.chb.2026.108995

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods*, *47*, 1178–1198. https://doi.org/10.3758/s13428-014-0551-2

Cohen, A. D., Olshtain, E., & Rosenstein, D. S. (1986). Advanced EFL apologies: What remains to be learned? *International Journal of the Sociology of Language*, *1986*(62), 51–74. https://doi.org/10.1515/ijsl.1986.62.51

Cohen, J. R. (1999). Advising clients to apologize. *Southern California Law Review*, *72*(4), 1009–1070.

Costa, P. J. C., & Moreira, P. A. S. (2024). The dimensionality of the moral foundations: Contributions from the moral foundations sacredness scale in four societies. *Journal of Personality Assessment*, *106*(3), 361–371. https://doi.org/10.1080/00223891.2023.2245895

Cowen, A. S., Keltner, D., Schroff, F., Jou, B., Adam, H., & Prasad, G. (2021). Sixteen facial expressions occur in similar contexts worldwide. *Nature*, *589*(7841), 251–257. https://doi.org/10.1038/s41586-020-3037-7

Dailey, M. N., Joyce, C., Lyons, M. J., Kamachi, M., Ishi, H., Gyoba, J., & Cottrell, G. W. (2010). Evidence and a computational explanation of cultural differences in facial expression recognition. *Emotion*, *10*(6), 874–893. https://doi.org/10.1037/a0020019

Darby, B. W., & Schlenker, B. R. (1982). Children's reactions to apologies. *Journal of Personality and Social Psychology*, *43*(4), 742–753. https://doi.org/10.1037/0022-3514.43.4.742

Darby, B. W., & Schlenker, B. R. (1989). Children's reactions to transgressions: Effects of the actor's apology, reputation and remorse. *British Journal of Social Psychology*, *28*(4), 353–364. https://doi.org/10.1111/j.2044-8309.1989.tb00879.x

De Cremer, D., & Schouten, B. C. (2008). When apologies for injustice matter. *European Psychologist*, *13*(4), 239–247. https://doi.org/10.1027/1016-9040.13.4.239

Deutschmann, M. (2003). *Apologising in British English*. Umeå University, Department of Modern Languages.

Dijk, C., de Jong, P. J., & Peters, M. L. (2009). The remedial value of blushing in the context of transgressions and mishaps. *Emotion*, *9*(2), 287–291. https://doi.org/10.1037/a0015081

Dogruyol, B., Alper, S., & Yilmaz, O. (2019). The five-factor model of the moral foundations theory is stable across WEIRD and non-WEIRD cultures. *Personality and Individual Differences*, *151*, 109547. https://doi.org/10.1016/j.paid.2019.109547

Eaton, J., Struthers, C. W., Shomrony, A., & Santelli, A. G. (2007). When apologies fail: The moderating effect of implicit and explicit self-esteem on apology and forgiveness. *Self and Identity*, *6*(2-3), 209–222. https://doi.org/10.1080/15298860601118819

Ebesu Hubbard, A. S., Hendrickson, B., Fehrenbach, K. S., & Sur, J. (2013). Effects of timing and sincerity of an apology on satisfaction and changes in negative feelings during conflicts. *Western Journal of Communication*, *77*(3), 305–322. https://doi.org/10.1080/10570314.2013.770160

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, *48*(4), 384–392. https://doi.org/10.1037/0003-066X.48.4.384

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press. Palo Alto, CA.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*(2), 203–235. https://doi.org/10.1037/0033-2909.128.2.203

Fang, X., Sauter, D. A., & van Kleef, G. A. (2020). Unmasking smiles: The influence of culture and intensity on interpretations of smiling expressions. *Journal of Cultural Cognitive Science*, *4*, 293–308. https://doi.org/10.1007/s41809-019-00053-1

Fehr, R., Gelfand, M. J., & Nag, M. (2010). The road to forgiveness: A meta-analytic synthesis of its situational and dispositional correlates. *Psychological Bulletin*, *136*(5), 894–914. https://doi.org/10.1037/a0019993

Fischbacher, U., & Utikal, V. (2013). On the acceptance of apologies. *Games and Economic Behavior*, *82*, 592–608. https://doi.org/10.1016/j.geb.2013.09.003

Forster, D. E., Billingsley, J., Burnette, J. L., Lieberman, D., Ohtsubo, Y., & McCullough, M. E. (2021). Experimental evidence that apologies promote forgiveness by communicating relationship value. *Scientific Reports*, *11*, 13107. https://doi.org/10.1038/s41598-021-92373-y

Frantz, C. M., & Bennigson, C. (2005). Better late than early: The influence of timing on apology effectiveness. *Journal of Experimental Social Psychology*, *41*(2), 201–207. https://doi.org/10.1016/j.jesp.2004.07.007

Freedman, G., Burgoon, E. M., Ferrell, J. D., Pennebaker, J. W., & Beer, J. S. (2017). When saying sorry may not help: The impact of apologies on social rejections. *Frontiers in Psychology*, *8*, 1375. https://doi.org/10.3389/fpsyg.2017.01375

Galdon, F., & Wang, S. J. (2020). From apology to compensation: A multi-level taxonomy of trust reparation for highly automated virtual assistants. In T. Ahram, W. Karwowski, S. Pickl, & R. Taiar (Eds.), *Human interaction and emerging technologies* (pp. 42–46, Vol. 1018). Springer. https://doi.org/10.1007/978-3-030-25629-6_7

García, C. (1989). Apologizing in English: Politeness strategies used by native and non-native speakers. *Multilingua*, *8*(1), 3–20. https://doi.org/10.1515/mult.1989.8.1.3

George, M., Guilfoyle, J. R., Steele, J. R., & Struthers, C. W. (2023). A sorry excuse for an apology: Examining people's mental representations of an apologetic face. *Journal of Nonverbal Behavior*, *47*(1), 57–81. https://doi.org/10.1007/s10919-022-00422-5

George, M., Guilfoyle, J. R., Struthers, C. W., & Steele, J. R. (2025). An investigation of how gender shapes the appearance and judgment of apologetic faces. *Group Processes & Intergroup Relations*, *28*(6), 1193–1212. https://doi.org/10.1177/13684302251322754

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Harvard University Press.

Giner-Sorolla, R., Castano, E., Espinosa, P., & Brown, R. (2008). Shame expressions reduce the recipient's insult from outgroup reparations. *Journal of Experimental Social Psychology*, *44*(3), 519–526. https://doi.org/10.1016/j.jesp.2007.08.003

Glikson, E., & Asscher, O. (2023). AI-mediated apology in a multilingual work context: Implications for perceived authenticity and willingness to forgive. *Computers in Human Behavior*, *140*, 107592. https://doi.org/10.1016/j.chb.2022.107592

Goffman, E. (1955). On face-work. *Psychiatry*, *18*(3), 213–231. https://doi.org/10.1080/00332747.1955.11023008

Goffman, E. (1971). *Relations in public: Microstudies of the public order*. Basic Books.

Gold, G. J., & Weiner, B. (2000). Remorse, confession, group identity, and expectancies about repeating a transgression. *Basic and Applied Social Psychology*, *22*(4), 291–300. https://doi.org/10.1207/s15324834basp2204_3

Gonzales, M. H., Pederson, J. H., Manning, D. J., & Wetter, D. W. (1990). Pardon my gaffe: Effects of sex, status, and consequence severity on accounts. *Journal of Personality and Social Psychology*, *58*(4), 610–621. https://doi.org/10.1037/0022-3514.58.4.610

Graham, J. (2015). Explaining away differences in moral judgment: Comment on Gray and Keeney (2015). *Social Psychological and Personality Science*, *6*(8), 869–873. https://doi.org/10.1177/1948550615592242

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (pp. 55–130, Vol. 47). Academic Press. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847

Gray, K., & Keeney, J. E. (2015). Impure or just weird? scenario sampling bias raises questions about the foundation of morality. *Social Psychological and Personality Science*, *6*(8), 859–868. https://doi.org/10.1177/1948550615592241

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387

Grunenberg, M., Esch, C., Toussaint, L. L., Maes, J., & Kals, E. (2026). Sincerity of apologies: Do it right or don't do it at all. *International Journal of Conflict Management*, *37*(1), 7–28. https://doi.org/10.1108/ijcma-03-2025-0078

Gu, Y. (1990). Politeness phenomena in modern Chinese. *Journal of Pragmatics*, *14*(2), 237–257. https://doi.org/10.1016/0378-2166(90)90082-O

Guan, X., Park, H. S., & Lee, H. E. (2009). Cross-cultural differences in apology. *International Journal of Intercultural Relations*, *33*(1), 32–45. https://doi.org/10.1016/j.ijintrel.2008.10.001

Guilfoyle, J. R., Struthers, C. W., van Monsjou, E., Shoikhedbrod, A., Eghbali, N., & Kermani, M. (2022). Sorry, not sorry: The effect of social power on transgressors' apology and nonapology. *Journal of Experimental Psychology: Applied*, *28*(4), 883–897. https://doi.org/10.1037/xap0000392

Guilfoyle, J. R., Struthers, C. W., van Monsjou, E., & Shoikhedbrod, A. (2019). Sorry is the hardest word to say: The role of self-control in apologizing. *Basic and Applied Social Psychology*, *41*(1), 72–90. https://doi.org/10.1080/01973533.2018.1553715

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). Oxford University Press.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, *133*(4), 55–66. https://doi.org/10.1162/0011526042365555

Halmesvaara, O., Harjunen, V. J., Aulbach, M. B., & Ravaja, N. (2020). How bodily expressions of emotion after norm violation influence perceivers' moral judgments and prevent social exclusion: A socio-functional approach to nonverbal shame display. *PLOS ONE*, *15*(4), e0232298. https://doi.org/10.1371/journal.pone.0232298

Hamilton, W. D. (1964). The genetical evolution of social behaviour. ii. *Journal of Theoretical Biology*, *7*(1), 17–52. https://doi.org/10.1016/0022-5193(64)90039-6

Hannawa, A. F. (2011). Shedding light on the dark side of doctor–patient interactions: Verbal and nonverbal messages physicians communicate during error disclosures. *Patient Education and Counseling*, *84*(3), 344–351. https://doi.org/10.1016/j.pec.2011.04.030

Hannawa, A. F., Shigemoto, Y., & Little, T. D. (2016). Medical errors: Disclosure styles, interpersonal forgiveness, and outcomes. *Social Science & Medicine*, *156*, 29–38. https://doi.org/10.1016/j.socscimed.2016.03.026

Hareli, S., & Eisikovits, Z. (2006). The role of communicating social emotions accompanying apologies in forgiveness. *Motivation and Emotion*, *30*(3), 189–197. https://doi.org/10.1007/s11031-006-9025-x

Harland, H., Dazeley, R., Nakisa, B., Cruz, F., & Vamplew, P. (2023). AI apology: Interactive multi-objective reinforcement learning for human-aligned AI. *Neural Computing and Applications*, *35*(23), 16917–16930. https://doi.org/10.1007/s00521-023-08586-x

Harland, H., Dazeley, R., Senaratne, H., Vamplew, P., Cruz, F., & Nakisa, B. (2025). AI apology: A critical review of apology in AI systems. *Artificial Intelligence Review*, *58*(12), 369. https://doi.org/10.1007/s10462-025-11305-8

Haugh, M., & Chang, W.-L. M. (2019). "the apology seemed (in)sincere": Variability in perceptions of (im)politeness. *Journal of Pragmatics*, *142*, 207–222. https://doi.org/10.1016/j.pragma.2018.11.022

Hess, U., Beaupré, M. G., & Cheung, N. (2002). Who to whom and why: Cultural differences and similarities in the function of smiles. In M. H. Abel (Ed.), *An empirical reflection on the smile* (pp. 187–216). Edwin Mellen Press.

Ho, B. (2012). Apologies as signals: With evidence from a trust game. *Management Science*, *58*(1), 141–158. https://doi.org/10.1287/mnsc.1110.1410

Ho, D. Y.-F., Fu, W., & Ng, S. M. (2004). Guilt, shame and embarrassment: Revelations of face and self. *Culture & Psychology*, *10*(1), 64–84. https://doi.org/10.1177/1354067x04044166

Holmes, J. (1990). Apologies in New Zealand English. *Language in Society*, *19*(2), 155–199. https://doi.org/10.1017/s0047404500014366

Hoque, M. E., Morency, L.-P., & Picard, R. W. (2011). Are you friendly or just polite? analysis of smiles in spontaneous face-to-face interactions. *Affective Computing and Intelligent Interaction*, *6974*, 135–144. https://doi.org/10.1007/978-3-642-24600-5_17

Hornsey, M. J., Wohl, M. J. A., Harris, E. A., Okimoto, T. G., Thai, M., & Wenzel, M. (2020). Embodied remorse: Physical displays of remorse increase positive responses to public apologies, but have negligible effects on forgiveness. *Journal of Personality and Social Psychology*, *119*(2), 367–389. https://doi.org/10.1037/pspi0000208

Howell, A. J., Dopko, R. L., Turowski, J. B., & Buro, K. (2011). The disposition to apologize. *Personality and Individual Differences*, *51*, 509–514. https://doi.org/10.1016/j.paid.2011.05.009

Howell, A. J., Turowski, J. B., & Buro, K. (2012). Guilt, empathy, and apology. *Personality and Individual Differences*, *53*(7), 917–922. https://doi.org/10.1016/j.paid.2012.06.021

Hui, V. K.-Y., & Bond, M. H. (2009). Target's face loss, motivations, and forgiveness following relational transgression: Comparing Chinese and US cultures. *Journal of Social and Personal Relationships*, *26*(2–3), 123–140. https://doi.org/10.1177/0265407508100312

Hyniewska, S., Sato, W., Kaiser, S., & Pelachaud, C. (2019). Naturalistic emotion decoding from facial action sets. *Frontiers in Psychology*, *9*, 2678. https://doi.org/10.3389/fpsyg.2018.02678

Ince, R. A. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J., & Schyns, P. G. (2017). A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human Brain Mapping*, *38*(3), 1541–1573. https://doi.org/10.1002/hbm.23471

Ince, R. A. A., Paton, A. T., Kay, J. W., & Schyns, P. G. (2021). Bayesian inference of population prevalence. *eLife*, *10*, e62461. https://doi.org/10.7554/eLife.62461

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*(2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241–7244. https://doi.org/10.1073/pnas.1200155109

Jack, R. E., & Schyns, P. G. (2015). The human face as a dynamic tool for social communication. *Current Biology*, *25*, R621–R634. https://doi.org/10.1016/j.cub.2015.05.052

Jeter, W. K., & Brannon, L. A. (2018). 'I'll make it up to you:' examining the effect of apologies on forgiveness. *The Journal of Positive Psychology*, *13*(6), 597–604. https://doi.org/10.1080/17439760.2017.1291854

Jia, F., Krettenauer, T., & Li, L. (2019). Moral identity in cultural context: Differences between Canadian and Chinese university students. *Journal of Moral Education*, *48*(2), 247–262. https://doi.org/10.1080/03057240.2018.1499504

Jucker, A. H. (2023). "he offered an apologetic smile": The politeness of apologetic gestures. In A. H. Jucker, I. Hübscher, & L. Brown (Eds.), *Multimodal im/politeness: Signed, spoken, written* (pp. 327–351). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.333.12juc

Julle-Danière, E., Whitehouse, J., Mielke, A., Vrij, A., Gustafsson, E., Micheletta, J., & Waller, B. M. (2020). Are there non-verbal signals of guilt? *PLOS ONE*, *15*(4), e0231756. https://doi.org/10.1371/journal.pone.0231756

Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, *68*(3), 441–454. https://doi.org/10.1037/0022-3514.68.3.441

Keltner, D., & Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition and Emotion*, *13*(5), 505–521. https://doi.org/10.1080/026999399379168

Kim, K. R., Kang, J.-S., & Yun, S. (2012). Moral intuitions and political orientation: Similarities and differences between South Korea and the United States. *Psychological Reports*, *111*(1), 173–185. https://doi.org/10.2466/17.09.21.pr0.111.4.173-185

Kirchhoff, J., Wagner, U., & Strack, M. (2012). Apologies: Words of magic? the role of verbal components, anger reduction, and offence severity. *Peace and Conflict: Journal of Peace Psychology*, *18*(2), 109–130. https://doi.org/10.1037/a0028092

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347–480). Rand McNally.

Kohlberg, L. (1971). Stages of moral development as a basis for moral education. In C. M. Beck, B. S. Crittenden, & E. V. Sullivan (Eds.), *Moral education* (pp. 23–92). University of Toronto Press. https://doi.org/10.3138/9781442656758-004

Komiya, A., Ozono, H., Watabe, M., Miyamoto, Y., Ohtsubo, Y., & Oishi, S. (2020). Socio-ecological hypothesis of reconciliation: Cultural, individual, and situational variations in willingness to accept apology or compensation. *Frontiers in Psychology*, *11*, 1761. https://doi.org/10.3389/fpsyg.2020.01761

Kraig, A. F., Barraza, J. A., Montgomery, W., & Zak, P. J. (2022). The neurophysiology of corporate apologies: Why do people believe insincere apologies? *International Journal of Business Communication*, *59*(4), 531–550. https://doi.org/10.1177/2329488419858391

Kubo, K., Okanoya, K., & Kawai, N. (2012). Apology isn't good enough: An apology suppresses an approach motivation but not the physiological and psychological anger. *PLoS ONE*, *7*(3), e33006. https://doi.org/10.1371/journal.pone.0033006

Laham, S. M., Chopra, S., Lalljee, M., & Parkinson, B. (2010). Emotional and behavioural reactions to moral transgressions: Cross-cultural and individual variations in India and Britain. *International Journal of Psychology*, *45*(1), 64–71. https://doi.org/10.1080/00207590902913434

Landmann, H., & Hess, U. (2018). Testing moral foundation theory: Are specific moral emotions elicited by specific moral transgressions? *Journal of Moral Education*, *47*(1), 34–47. https://doi.org/10.1080/03057240.2017.1350569

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791. https://doi.org/10.1038/44565

Leech, G. (2014). *The pragmatics of politeness*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195341386.001.0001

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

Leunissen, J. M., De Cremer, D., Reinders Folmer, C. P., & van Dijke, M. (2013). The apology mismatch: Asymmetries between victim's need for apologies and perpetrator's willingness to apologize. *Journal of Experimental Social Psychology*, *49*(3), 315–324. https://doi.org/10.1016/j.jesp.2012.12.005

Levi, D. L. (1997). The role of apology in mediation. *New York University Law Review*, *72*(5), 1165–1210.

Lewicki, R. J., Polin, B., & Lount, R. B., Jr. (2016). An exploration of the structure of effective apologies. *Negotiation and Conflict Management Research*, *9*(2), 177–196. https://doi.org/10.1111/ncmr.12073

Lim, J. S., Hong, N., & Schneider, E. (2025). How warm-versus competent-toned AI apologies affect trust and forgiveness through emotions and perceived sincerity. *Computers in Human Behavior*, *172*, 108761. https://doi.org/10.1016/j.chb.2025.108761

Ma, R., & Wang, W. (2021). Smile or pity? examine the impact of emoticon valence on customer satisfaction and purchase intention. *Journal of Business Research*, *134*, 443–456. https://doi.org/10.1016/j.jbusres.2021.05.057

Maddux, W. W., Kim, P. H., Okumura, T., & Brett, J. M. (2011). Cultural differences in the function and meaning of apologies. *International Negotiation*, *16*, 405–425. https://doi.org/10.1163/157180611X592932

Mapson, R. (2014). Polite appearances: How non-manual features convey politeness in British Sign Language. *Journal of Politeness Research*, *10*(2), 157–184. https://doi.org/10.1515/pr-2014-0008

Martin, J., Rychlowska, M., Wood, A., & Niedenthal, P. (2017). Smiles as multipurpose social signals. *Trends in Cognitive Sciences*, *21*(11), 864–877. https://doi.org/10.1016/j.tics.2017.08.007

Martin, J. D., Wood, A., Cox, W. T. L., Sievert, S., Nowak, R., Gilboa-Schechtman, E., Zhao, F., Witkower, Z., Langbehn, A. T., & Niedenthal, P. M. (2021). Evidence for distinct facial signals of reward,

affiliation, and dominance from both perception and production tasks. *Affective Science*, *2*, 14–30. https://doi.org/10.1007/s42761-020-00024-8

Matsumoto, D., Keltner, D., Shiota, M. N., O'Sullivan, M., & Frank, M. (2008). Facial expressions of emotion. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 211–234). Guilford Press.

Matsumoto, D., Yoo, S. H., Fontaine, J., Anguas-Wong, A. M., Arriola, M., Ataca, B., Bond, M. H., Boratav, H. B., Breugelmans, S. M., Cabecinhas, R., Chae, J., Chin, W. H., Comunian, A. L., Degere, D. N., Djunaidi, A., Fok, H. K., Friedlmeier, W., Ghosh, A., Glamcevski, M., … Grossi, E. (2008). Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism. *Journal of Cross-Cultural Psychology*, *39*(1), 55–74. https://doi.org/10.1177/0022022107311854

Matsumoto, Y. (1988). Reexamination of the universality of face: Politeness phenomena in Japanese. *Journal of Pragmatics*, *12*(4), 403–426. https://doi.org/10.1016/0378-2166(88)90003-3

McAndrew, F. T. (2019). Costly signaling theory. In T. K. Shackelford & V. A. Weekes-Shackelford (Eds.), *Encyclopedia of evolutionary psychological science*. Springer Nature. https://doi.org/10.1007/978-3-319-16999-6_3483-1

Meier, A. (1998). Apologies: What do we know? *International Journal of Applied Linguistics*, *8*(2), 215–231. https://doi.org/10.1111/j.1473-4192.1998.tb00130.x

Mu, F., & Bobocel, D. R. (2019). Why did i say sorry? apology motives and transgressor perceptions of reconciliation. *Journal of Organizational Behavior*, *40*(8), 912–930. https://doi.org/10.1002/job.2376

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*, 239–281. https://doi.org/10.1023/A:1024068626366

Newman, L. S., & Kraynak, L. R. (2013). The ambiguity of a transgression and the type of apology influence immediate reactions. *Social Behavior and Personality: an international journal*, *41*(1), 31–45. https://doi.org/10.2224/sbp.2013.41.1.31

Nunney, S. J., & Manstead, A. S. R. (2021). Improving the effectiveness of intergroup apologies: The role of apology content and moral emotions. *Peace and Conflict: Journal of Peace Psychology*, *27*(3), 436–448. https://doi.org/10.1037/pac0000571

Ogiermann, E. (2009). *On apologising in negative and positive politeness cultures*. John Benjamins Publishing Company. https://doi.org/10.1075/pbns.191

O'Hara, E. A., & Yarn, D. (2002). On apology and consilience. *Washington Law Review*, *77*(4), 1121–1192.

Ohbuchi, K.-i., Kameda, M., & Agarie, N. (1989). Apology as aggression control: Its role in mediating appraisal of and response to harm. *Journal of Personality and Social Psychology*, *56*(2), 219–227. https://doi.org/10.1037/0022-3514.56.2.219

Ohtsubo, Y., Matsunaga, M., Himichi, T., Suzuki, K., Shibata, E., Hori, R., Umemura, T., & Ohira, H. (2020). Costly group apology communicates a group's sincere intention. *Social Neuroscience*, *15*(2), 244–254. https://doi.org/10.1080/17470919.2019.1697745

Ohtsubo, Y., Matsunaga, M., Tanaka, H., Suzuki, K., Kobayashi, F., Shibata, E., Hori, R., Umemura, T., & Ohira, H. (2018). Costly apologies communicate conciliatory intention: An fMRI study on forgive-

ness in response to costly apologies. *Evolution and Human Behavior*, *39*, 249–256. https://doi.org/10.1016/j.evolhumbehav.2018.01.004

Ohtsubo, Y., & Watanabe, E. (2009). Do sincere apologies need to be costly? test of a costly signaling model of apology. *Evolution and Human Behavior*, *30*, 114–123. https://doi.org/10.1016/j.evolhumbehav.2008.09.004

Ohtsubo, Y., Yamaura, K., & Yagi, A. (2015). Development of Japanese measures of reconciliatory tendencies: The Japanese trait forgivingness scale and the Japanese proclivity to apologize measure. *Japanese Journal of Social Psychology*, *31*(2), 135–142. https://doi.org/10.14966/jssp.31.2_135

Okada, Y., Kimoto, M., Iio, T., Shimohara, K., & Shiomi, M. (2023). Two is better than one: Apologies from two robots are preferred. *PLOS ONE*, *18*(2), e0281604. https://doi.org/10.1371/journal.pone.0281604

Park, H. S., & Guan, X. (2009). Cross-cultural comparison of verbal and nonverbal strategies of apologizing. *Journal of International and Intercultural Communication*, *2*(1), 66–87. https://doi.org/10.1080/17513050802603471

Petrucci, C. J. (2002). Apology in the criminal justice setting: Evidence for including apology as an additional component in the legal system. *Behavioral Sciences & the Law*, *20*(4), 337–362. https://doi.org/10.1002/bsl.495

Poggi, I., & D'Errico, F. (2010). Cognitive modelling of human social signals. *Proceedings of SSPW 2010—Social Signal Processing Workshop*, 21–26. https://doi.org/10.1145/1878116.1878124

Poggi, I., & D'Errico, F. (2011). Social signals: A psychological perspective. In A. A. Salah & T. Gevers (Eds.), *Computer analysis of human behavior* (pp. 185–225). Springer-Verlag London. https://doi.org/10.1007/978-0-85729-994-9_8

Polin, B., Doyle, S. P., Kim, S., Lewicki, R. J., & Chawla, N. (2024). Sorry to ask but … how is apology effectiveness dependent on apology content and gender? *Journal of Applied Psychology*, *109*(3), 339–361. https://doi.org/10.1037/apl0001128

Rieger, C. L. (2017). "I want a real apology": A discursive pragmatics perspective on apologies. *Pragmatics*, *27*(4), 553–590. https://doi.org/10.1075/prag.27.4.04rie

Risen, J. L., & Gilovich, T. (2007). Target and observer differences in the acceptance of questionable apologies. *Journal of Personality and Social Psychology*, *92*(3), 418–433. https://doi.org/10.1037/0022-3514.92.3.418

Rizvi, S., Struthers, C. W., Shoikhedbrod, A., & Guilfoyle, J. R. (2022). Take a moment to apologize: How and why mindfulness affects apologies. *Journal of Experimental Psychology: Applied*, *28*(3), 661–675. https://doi.org/10.1037/xap0000387

Robbennolt, J. K. (2003). Apologies and legal settlement: An empirical examination. *Michigan Law Review*, *102*(3), 460–516. https://doi.org/10.2307/3595367

Robinson, J. D. (2004). The sequential organization of "explicit" apologies in naturally occurring English. *Research on Language & Social Interaction*, *37*(3), 291–330. https://doi.org/10.1207/s15327973rlsi3703_2

Roschk, H., & Kaiser, S. (2013). The nature of an apology: An experimental study on how to apologize after a service failure. *Marketing Letters*, *24*(3), 293–309. https://doi.org/10.1007/s11002-012-9218-x

Russell, P. S., Hepper, E. G., Fasoli, F., Hopkins-Doyle, A., & Boone, F. (2025). Examining the dynamic impact of emotional apologies on forgiveness. *Emotion*, *25*(5), 1337–1348. https://doi.org/10.1037/emo0001504

Rychlowska, M., Jack, R. E., Garrod, O. G. B., Schyns, P. G., Martin, J. D., & Niedenthal, P. M. (2017). Functional smiles: Tools for love, sympathy, and war. *Psychological Science*, *28*(9), 1259–1270. https://doi.org/10.1177/0956797617706082

Sandlin, J. K., & Gracyalny, M. L. (2018). Seeking sincerity, finding forgiveness: YouTube apologies as image repair. *Public Relations Review*, *44*(3), 393–406. https://doi.org/10.1016/j.pubrev.2018.04.007

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Scher, S. J., & Darley, J. M. (1997). How effective are the things people say to apologize? effects of the realization of the apology speech act. *Journal of Psycholinguistic Research*, *26*(1), 127–140. https://doi.org/10.1023/a:1025068306386

Schlenker, B. R., & Darby, B. W. (1981). The use of apologies in social predicaments. *Social Psychology Quarterly*, *44*(3), 271–278. https://doi.org/10.2307/3033840

Schmidt, K. L., & Cohn, J. F. (2001). Human facial expressions as adaptations: Evolutionary questions in facial expression research. *Yearbook of Physical Anthropology*, *44*, 3–24. https://doi.org/10.1002/ajpa.20001

Schmitt, M., Gollwitzer, M., Förster, N., & Montada, L. (2004). Effects of objective and subjective account components on forgiving. *The Journal of Social Psychology*, *144*(5), 465–486. https://doi.org/10.3200/socp.144.5.465-486

Schumann, K., Ritchie, E. G., & Forest, A. (2023). The social consequences of frequent versus infrequent apologizing. *Personality and Social Psychology Bulletin*, *49*(3), 331–343. https://doi.org/10.1177/01461672211065286

Schumann, K., & Ross, M. (2010). Why women apologize more than men. *Psychological Science*, *21*(11), 1649–1655. https://doi.org/10.1177/0956797610384150

Senju, A., Vernetti, A., Kikuchi, Y., Akechi, H., Hasegawa, T., & Johnson, M. H. (2013). Cultural background modulates how we look at other persons' gaze. *International Journal of Behavioral Development*, *37*(2), 131–136. https://doi.org/10.1177/0165025412465360

Shao, W., Moffett, J. W., Quach, S., Surachartkumtonkun, J., Thaichon, P., Weaven, S. K., & Palmatier, R. W. (2022). Toward a theory of corporate apology: Mechanisms, contingencies, and strategies. *European Journal of Marketing*, *56*(12), 3418–3452. https://doi.org/10.1108/ejm-02-2021-0069

Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In A. M. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119–169). Routledge.

Slocum, D., Allan, A., & Allan, M. M. (2011). An emerging theory of apology. *Australian Journal of Psychology*, *63*(2), 83–92. https://doi.org/10.1111/j.1742-9536.2011.00013.x

Smith, C. E., Chen, D., & Harris, P. L. (2010). When the happy victimizer says sorry: Children's understanding of apology and emotion. *British Journal of Developmental Psychology*, *28*(4), 727–746. https://doi.org/10.1348/026151009x475343

Snoek, L., Jack, R. E., Schyns, P. G., Garrod, O. G. B., Mittenbühler, M., Chen, C., Oosterwijk, S., & Scholte, H. S. (2023). Testing, explaining, and exploring models of facial expressions of emotions. *Science Advances*, *9*, eabq8421. https://doi.org/10.1126/sciadv.abq8421

Struthers, C. W., Eaton, J., Santelli, A. G., Uchiyama, M., & Shirvani, N. (2008). The effects of attributions of intent and apology on forgiveness: When saying sorry may not help the story. *Journal of Experimental Social Psychology*, *44*(4), 983–992. https://doi.org/10.1016/j.jesp.2008.02.006

Struthers, C. W., Santelli, A. G., Khoury, C., Pang, M., Young, R. E., Kashefi, Y., Marjanovic, Z., Phills, C., Nash, K., & Vasquez, N. A. (2014). The role of victim embarrassment in explaining why apologies affect reported (but not actual) forgiveness. *Journal of Language and Social Psychology*, *33*(5), 517–525. https://doi.org/10.1177/0261927x14520983

Sunar, D., Cesur, S., Piyale, Z. E., Tepe, B., Biten, A. F., Hill, C. T., & Koc, Y. (2021). People respond with different moral emotions to violations in different relational models: A cross-cultural comparison. *Emotion*, *21*(4), 693–706. https://doi.org/10.1037/emo0000736

Syme, K. L., & Hagen, E. H. (2019). When saying "sorry" isn't enough: Is some suicidal behavior a costly signal of apology? *Human Nature*, *30*(1), 117–141. https://doi.org/10.1007/s12110-018-9333-3

Takaku, S., Weiner, B., & Ohbuchi, K.-i. (2001). A cross-cultural examination of the effects of apology and perspective taking on forgiveness. *Journal of Language and Social Psychology*, *20*(1–2), 144–166. https://doi.org/10.1177/0261927X01020001007

Tangney, J. P., Stuewig, J., & Hafez, L. (2011). Shame, guilt, and remorse: Implications for offender populations. *The Journal of Forensic Psychiatry & Psychology*, *22*(5), 706–723. https://doi.org/10.1080/14789949.2011.617541

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*, 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145

ten Brinke, L., & Adams, G. S. (2015). Saving face? when emotion displays during public apologies mitigate damage to organizational performance. *Organizational Behavior and Human Decision Processes*, *130*, 1–12. https://doi.org/10.1016/j.obhdp.2015.05.003

Thorstenson, C. A., Pazda, A. D., & Lichtenfeld, S. (2020). Facial blushing influences perceived embarrassment and related social functional evaluations. *Cognition and Emotion*, *34*(3), 413–426. https://doi.org/10.1080/02699931.2019.1634004

Tomlinson, E. C., Dineen, B. R., & Lewicki, R. J. (2004). The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise. *Journal of Management*, *30*(2), 165–187. https://doi.org/10.1016/j.jm.2003.01.003

Tracy, J. L., Robins, R. W., & Schriber, R. A. (2009). Development of a FACS-verified set of basic and self-conscious emotion expressions. *Emotion*, *9*(4), 554–559. https://doi.org/10.1037/a0015766

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57. https://doi.org/10.1086/406755

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge University Press.

Vinciarelli, A., Salamin, H., & Pantic, M. (2009). Social signal processing: Understanding social interactions through nonverbal behavior analysis. *IEEE Computer Vision and Pattern Recognition Workshops*, 42–49. https://doi.org/10.1109/CVPR.2009.5204290

Wagatsuma, H., & Rosett, A. (1986). The implications of apology: Law and culture in Japan and the United States. *Law & Society Review*, *20*(4), 461–498. https://doi.org/10.2307/3053463

Wang, Y., & McGlone, M. S. (2020). Did i offend you or did it? agency assignment in interpersonal apology. *Discourse Processes*, *57*(1), 66–80. https://doi.org/10.1080/0163853x.2019.1576471

Wenzel, M., Anvari, F., de Vel-Palumbo, M., & Bury, S. M. (2017). Collective apology, hope, and forgiveness. *Journal of Experimental Social Psychology*, *72*, 75–87. https://doi.org/10.1016/j.jesp.2017.05.003

Wooten, D. B. (2009). Say the right thing: Apologies, reputability, and punishment. *Journal of Consumer Psychology*, *19*(2), 225–235. https://doi.org/10.1016/j.jcps.2009.02.017

Xie, C., Zhu, J., Xie, Y., & Liang, C. (2025). Apologizing with a smile or crying face? exploring the impact of emoji types on customer forgiveness within chatbots service recovery. *Electronic Commerce Research and Applications*, *70*, 101488. https://doi.org/10.1016/j.elerap.2025.101488

Yamamoto, K., Kimura, M., & Osaka, M. (2021). Sorry, not sorry: Effects of different types of apologies and self-monitoring on non-verbal behaviors. *Frontiers in Psychology*, *12*, 689615. https://doi.org/10.3389/fpsyg.2021.689615

Yang, K., & Qian, S. (2024). Your smiling face is impolite to me: A study of the smiling face emoji in Chinese computer-mediated communication. *Social Science Computer Review*, *42*(4), 947–960. https://doi.org/10.1177/08944393231219481

Yu, H., Garrod, O. G. B., & Schyns, P. G. (2012). Perception-driven facial expression synthesis. *Computers & Graphics*, *36*(3), 152–162. https://doi.org/10.1016/j.cag.2011.12.002

Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of Theoretical Biology*, *53*(1), 205–214. https://doi.org/10.1016/0022-5193(75)90111-3

Zeelenberg, M., & Breugelmans, S. M. (2008). The role of interpersonal harm in distinguishing regret from guilt. *Emotion*, *8*(5), 589–596. https://doi.org/10.1037/a0012894

Zhan, J., Garrod, O. G. B., van Rijsbergen, N., & Schyns, P. G. (2019). Modelling face memory reveals task-generalizable representations. *Nature Human Behaviour*, *3*(8), 817–826. https://doi.org/10.1038/s41562-019-0625-3

Zhang, H., Yu, F., Xu, L., & Xuan, Z. (2023). Monism and pluralism in morality: Origins, connotations and debates. *Advances in Psychological Science*, *31*(2), 301–314. https://doi.org/10.3724/SP.J.1042.2023.00301

Appendix A

Supplementary Details for Chapter 3

This appendix reports the sample counts, screening outcomes, and supplementary OFF-state PMI localisation supporting Chapter 3.

A.1 Validated Participants and Models

Table A1

Validated Western Receiver-Group Participant and Model Counts by Moral Domain.

Domain	*N*	Female	Male	Non-binary	Models	Threshold
Fairness	20	10	10	0	60	5
Care/harm	15	9	6	0	45	4
Authority	14	10	4	0	42	4
Loyalty	13	10	3	0	39	4
Sanctity/purity	14	10	4	0	42	4
Total	76	49	27	0	228	–

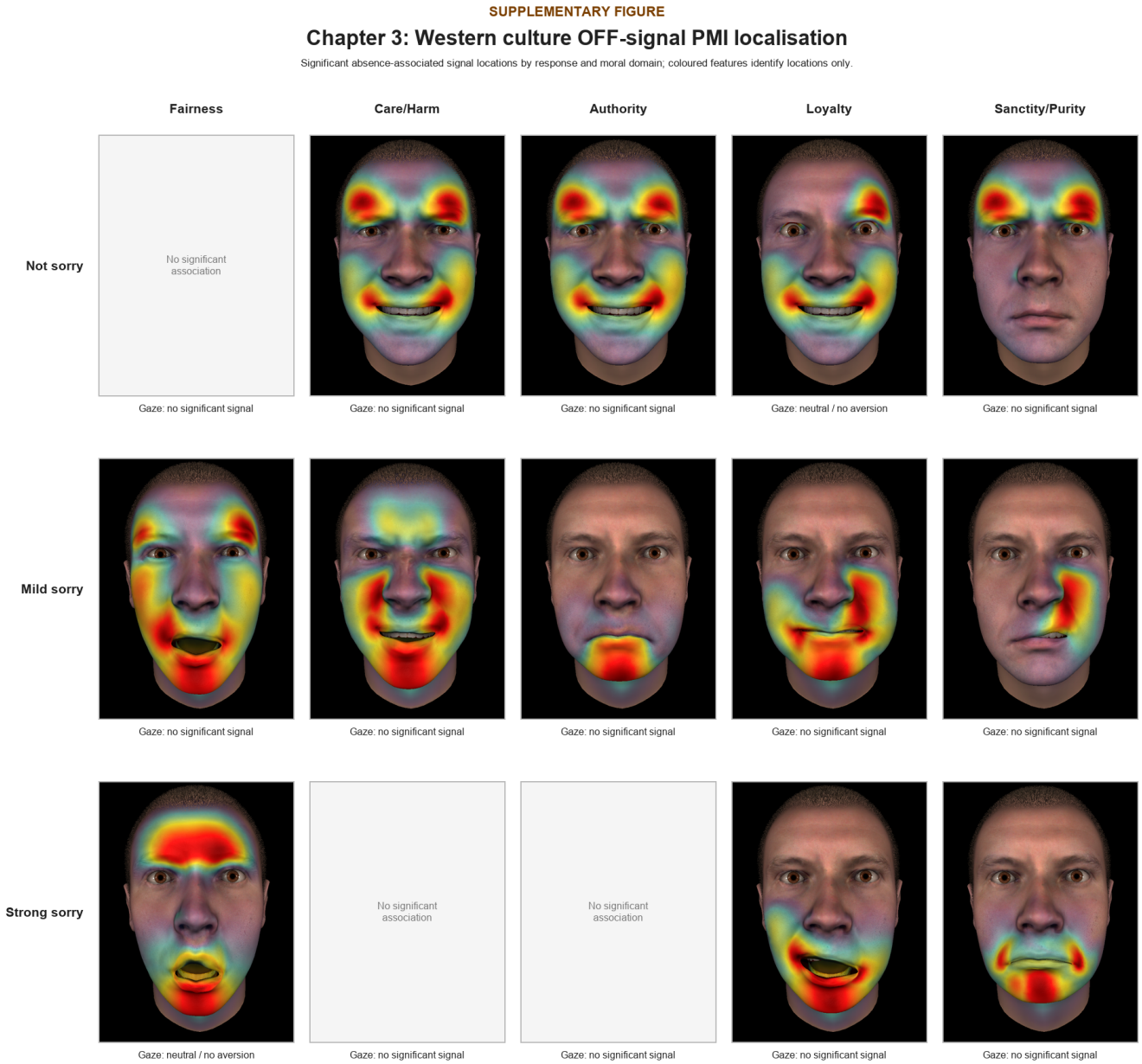
A.2 Observed Screening Outcomes

Of 182,400 trial records, 3,716 were "don't know" responses and 178,684 were usable. Participant-specific rebinning gave 29.1% not-sorry, 40.5% mild-sorry, and 30.4% strong-sorry responses. All 76 WC participants entering validation were retained, providing 228 response-bin models. Mean validation accuracy across domains was 97.8% for not sorry, 63.1% for mild sorry, and 86.2% for strong sorry.

A.3 Supplementary PMI Facemap Localisation

Figure A1

Facemap Localisation of Significant WC OFF-State PMI Associations by Response Category and Moral Domain.



Note. Coloured facial regions identify the locations of signals whose absence was overrepresented; blank panels contain no retained OFF-state association. Each selected AU is illustrated at maximum rendering amplitude to show its location; ON/OFF labels identify the associated signal state.

Appendix B

Supplementary Details for Chapter 4

This appendix reports the sample counts, screening outcomes, and supplementary OFF-state PMI localisation supporting Chapter 4.

B.1 Validated Participants and Models

Table B1

Validated East Asian Participant and Model Counts by Moral Domain.

Domain	*N*	Female	Male	Non-binary	Models	Threshold
Fairness	14	11	3	0	42	4
Care/harm	18	10	8	0	54	4
Authority	16	10	6	0	48	4
Loyalty	16	11	5	0	48	4
Sanctity/purity	16	10	6	0	48	4
Total	80	52	28	0	240	–

B.2 Observed Screening Outcomes

Of 199,200 trial records, 4,277 were "don't know" responses and 194,923 were usable. Rebinning produced 66,951 not-sorry responses (34.4%), 77,427 mild-sorry responses (39.7%), and 50,545 strong-sorry responses (25.9%). The validation procedure excluded participants 8, 9, and 26 from Fairness, leaving 80 of 83 EA participants and 240 response-bin models. Mean validation accuracy across domains was 92.7% for not sorry, 48.8% for mild sorry, and 79.5% for strong sorry. Table B2 reports the domain-level diagnostics.

Table B2

East Asian Model-Validation Performance by Moral Domain.

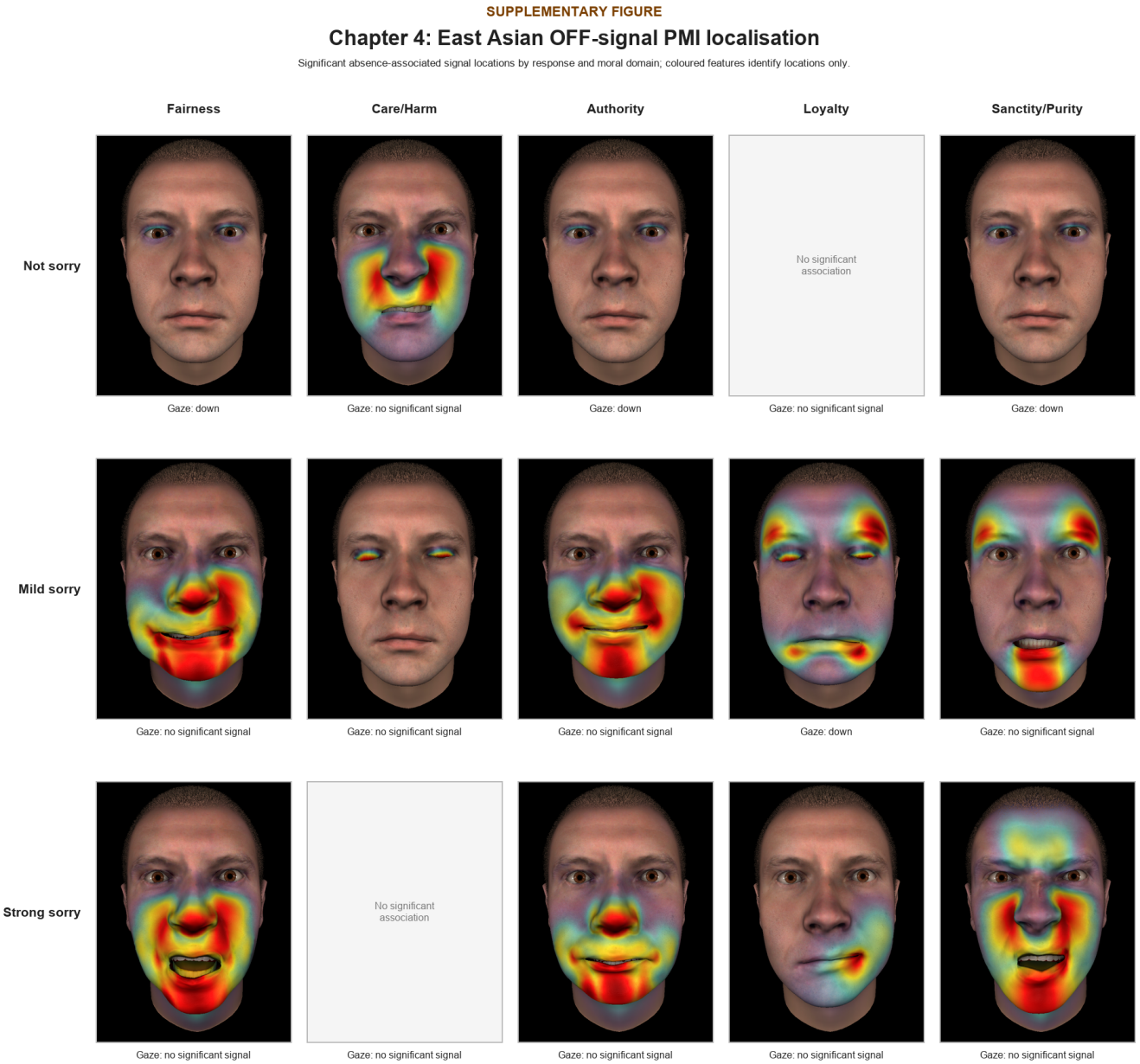
Moral domain	Retained/original *N*	Loss (%)	*SD*	Not sorry	Mild sorry	Strong sorry
Fairness	14/17	28.33	16.31	1.000	.375	.775
Care/harm	18/18	26.67	11.05	.967	.400	.833
Authority	16/16	28.89	8.38	.833	.533	.767
Loyalty	16/16	20.00	10.57	.867	.700	.833
Sanctity/purity	16/16	27.78	13.25	.967	.433	.767

Note. Loss and its standard deviation are expressed as percentages. Response-bin columns report diagonal validation accuracy.

B.3 Supplementary PMI Facemap Localisation

Figure B1

Facemap Localisation of Significant East Asian OFF-State PMI Associations by Response Category and Moral Domain.



Note. Coloured facial regions identify the locations of signals whose absence was overrepresented; blank panels contain no retained OFF-state association. Each selected AU is illustrated at maximum rendering amplitude to show its location; ON/OFF labels identify the associated signal state.

Appendix C

Supplementary Details for Chapter 5

This appendix reports the direct-analysis input counts, supplementary one-sided PMI localisation, design-matched classifier comparison, and literature reference mapping supporting Chapter 5.

C.1 Direct-Analysis Inputs

Table C1

Realised Input to the Direct WC–EA Receiver-Group Analyses.

Moral domain	WC			EA			Retained predictors
	Participants	Validated response-model rows	Direct-classifier rows	Participants	Validated response-model rows	Direct-classifier rows	
Fairness	20	60	40	14	42	28	37
Care/harm	15	45	30	18	54	36	34
Authority	14	42	28	16	48	32	30
Loyalty	13	39	26	16	48	32	35
Sanctity/purity	14	42	28	16	48	32	34
Total	76	228	152	80	240	160	–

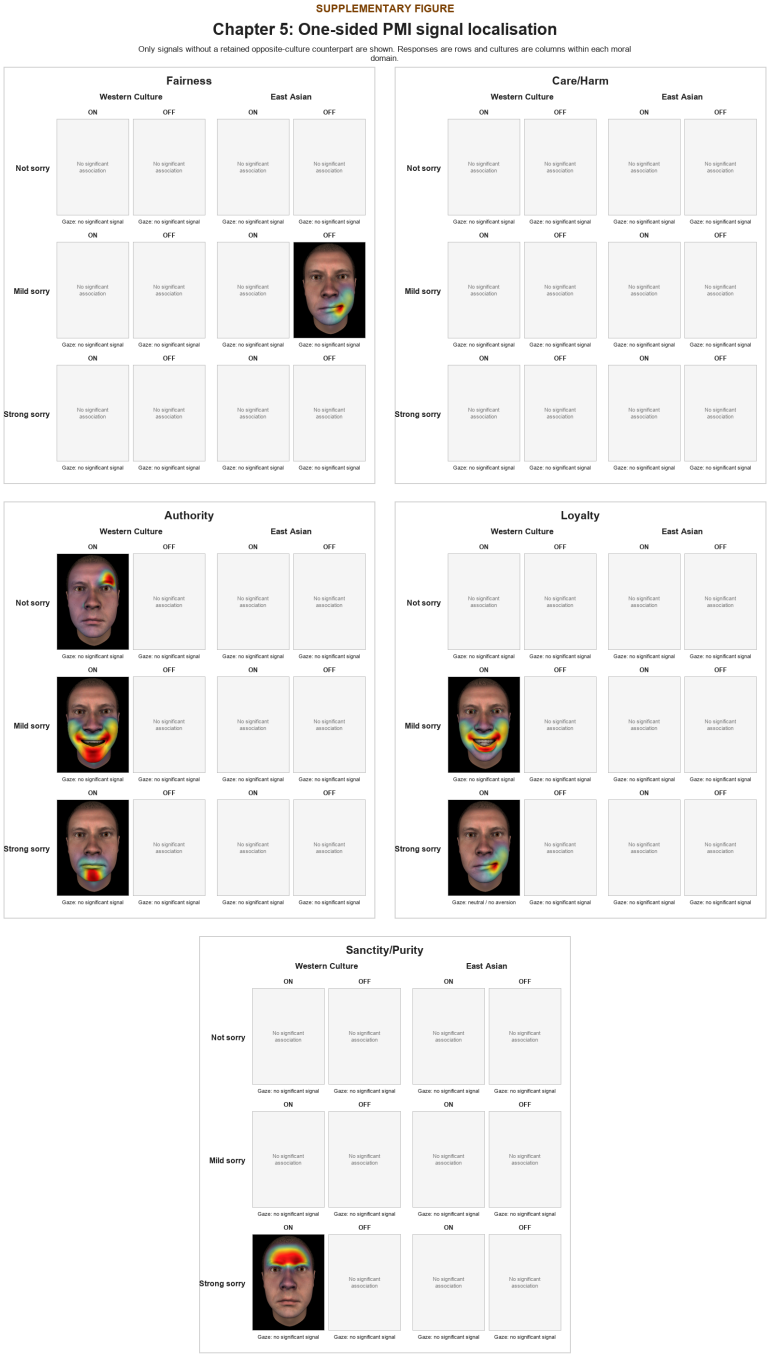
Note. Each participant contributed three validated response-model rows (not sorry, mild sorry, and strong sorry) and two rows to the direct receiver-group classification task (mild and strong sorry).

C.2 Supplementary One-Sided PMI Facemap Localisation

Figure C1 shows the 13 domain–response–signal relations without a retained opposite-group counterpart.

Figure C1

Facemap Localisation of One-Sided WC–EA Receiver-Group MI/PMI Associations.



Note. Rows show response categories; moral-domain blocks contain WC and EA ON- and OFF-state panels. Coloured regions identify retained one-sided associations, while blank panels contain no association without an opposite-group counterpart. Each selected AU is illustrated at maximum rendering amplitude to show its location; ON/OFF labels identify the associated signal state.

C.3 Design-Matched C2–D2–D5 Analysis

Thirteen participants per receiver-group-by-domain cell were selected for each of 50 matched participant-level splits, with 11 training and two test participants; both apology-model rows from each participant stayed together. C2 distinguished receiver groups within domains, D2 distinguished domain pairs within each group, and D5 distinguished all five domains jointly. The same participants and 47 predictors supplied 250 C2, 1,000 D2, and 100 D5 fits. The corrected C2–D2 comparison supported greater receiver-group than pairwise-domain separability under this matched design (Table C2). Comparisons of D5 with the binary tasks remained descriptive: the raw D2–D5 difference of .2886 was close to the .30 chance-level difference, and separate pairwise and joint classifiers address different questions. The repeated splits reused participants and did not constitute independent samples.

Table C2

Design-Matched Classifier Performance and Corrected C2–D2 Inference for Population-Thresholded Vectors.

Vector representation	Task	Classes	Tasks/rep.	Rows/class	Train/test	Chance BA	Mean BA	Chance-corrected
Population thresholded	C2	2	5	26	44/8	.50	.8690	.7380
Population thresholded	D2	2	20	26	44/8	.50	.7391	.4783
Population thresholded	D5	5	2	26	110/20	.20	.4505	.3131

Vector representation	C2	D2	C2–D2	Corrected 95% CI	$t(49)$	Holm p
Population thresholded	.8690	.7391	.1299	[.0794, .1804]	5.166	$< .001$

Note. All fits used 47 predictors and the original 50-cycle bagged-tree classifier. Chance-corrected balanced accuracy is $(\mathrm{BA} - 1/K)/(1 - 1/K)$, where K is the number of classes. C2 and D2 were averaged across five and 20 classifiers per repetition before subtraction. Corrected repeated-holdout inference used the Nadeau–Bengio variance multiplier $1/50 + 40/220$ and 49 degrees of freedom (Nadeau & Bengio, 2003). The original two-comparison Holm adjustment was retained (exact $p = 8.738 \times 10^{-6}$); D5 was not included in this test.

C.4 Literature Reference Mapping

Table C3

Evidence Scope, Categorisation, and Signal Mapping for the Chapter 5 Literature References

L2 type and sources	Evidence scope	Grouping and valence	Signal mapping
L1: Positive			
Affiliation smile (J. D. Martin et al., 2021; Rychlowska et al., 2017)	Perception and production; direct sources.	Bond-maintenance function; positive relational intent.	—
Amused smile (Ambadar et al., 2009)	Perception; direct source.	Perceived amusement; pleasant emotion.	Unspecified mouth opening maps to AU25, AU26 or AU27i and the corresponding project compounds.
Happiness/joy (Hyniewska et al., 2019; D. Matsumoto, Keltner, et al., 2008; Snoek et al., 2023; ten Brinke & Adams, 2015)	Perception, production, and configurations or unspecified process; direct and secondary sources.	Happiness and joy across sources; positive emotion.	—
Reward smile (J. D. Martin et al., 2021; Rychlowska et al., 2017)	Perception and production; direct sources.	Reward and reinforcement function; positive feeling and social reinforcement.	—
L1: Negative			
Anger (Hyniewska et al., 2019; D. Matsumoto, Keltner, et al., 2008; Snoek et al., 2023)	Perception, and configurations or unspecified process; direct and secondary sources.	Anger across sources; negative emotion.	—
Contempt (Hyniewska et al., 2019)	Perception; direct source.	The source's contempt category; negative emotion.	—
Disgust (D. Matsumoto, Keltner, et al., 2008; Snoek et al., 2023)	Configurations or unspecified process; direct and secondary sources.	Disgust across sources; negative emotion.	—
Dominance smile (Rychlowska et al., 2017)	Perception; direct source.	Status and superiority function; negative relational valence, reflecting low positive feeling and social connection.	Source-supported left- and right-AU12 alternatives.
Embarrassment/ appeasement (Ambadar et al., 2009)	Perception; direct source.	The source's pooled embarrassed/nervous smile category, linked to appeasement; negative self-conscious emotion.	Unspecified mouth opening maps to AU25, AU26 or AU27i and the corresponding project compounds.

Continued on next page

Table C3 continued

L2 type and sources	Evidence scope	Grouping and valence	Signal mapping
External-standard violation (Hyniewska et al., 2019)	Perception; direct source.	Decoding a violation of an external standard; negative norm-violation category.	—
Fear (Hyniewska et al., 2019; D. Matsumoto, Keltner, et al., 2008; Snoek et al., 2023)	Perception, and configurations or unspecified process; direct and secondary sources.	Fear across sources; negative emotion.	—
Guilt (Julle-Danière et al., 2020)	Perception and production; direct source.	Produced and perceived evidence for guilt; negative self-conscious emotion.	—
Internal-standard violation (Hyniewska et al., 2019)	Perception; direct source.	Decoding a violation of an internal standard; negative norm-violation category.	—
Sadness (Hyniewska et al., 2019; D. Matsumoto, Keltner, et al., 2008; Snoek et al., 2023; ten Brinke & Adams, 2015; Yamamoto et al., 2021)	Perception, production, and configurations or unspecified process; direct and secondary sources.	Sadness across sources; negative emotion.	Yamamoto's direct gaze maps to Neutral; unspecified gaze aversion maps to Down, Left or Right.
Shame (Halmesvaara et al., 2020; Hyniewska et al., 2019; Keltner, 1995; Tracy et al., 2009)	Mixed perception and production evidence; direct sources.	Shame across sources, with each source configuration retained separately; negative self-conscious emotion.	—
L1: Neutral/mixed			
Polite smile (Ambadar et al., 2009)	Perception; direct source.	Polite social intention; neutral/mixed because the construct has no fixed positive or negative valence.	Unspecified mouth opening maps to AU25, AU26 or AU27i and the corresponding project compounds.
Surprise (D. Matsumoto, Keltner, et al., 2008; Snoek et al., 2023)	Configurations or unspecified process; direct and secondary sources.	Surprise across sources; neutral/mixed because the construct has no fixed positive or negative valence.	—

Note. Reference profiles were not separated by culture. Secondary sources reproduce or specify configurations from earlier work. A dash indicates no additional mapping detail.

Appendix D

Illustrative Videos of Facial Dynamics

This appendix indexes 30 illustrative videos: five moral-domain contexts, three apparent-apology response categories, and two receiver groups. WC denotes Western receivers judging Western faces; EA denotes East Asian receivers judging East Asian faces. Not-sorry, Mild-sorry and Strong-sorry are model-derived response categories.

Video collection: Facial dynamics videos. The collection uses the video IDs in Table D1.

Table D1

Video Index by Receiver Group, Moral-Domain Context and Response Category.

Group	Moral-domain context	Not-sorry	Mild-sorry	Strong-sorry
WC	Fairness	V01	V02	V03
WC	Care/harm	V04	V05	V06
WC	Authority	V07	V08	V09
WC	Loyalty	V10	V11	V12
WC	Sanctity/purity	V13	V14	V15
EA	Fairness	V16	V17	V18
EA	Care/harm	V19	V20	V21
EA	Authority	V22	V23	V24
EA	Loyalty	V25	V26	V27
EA	Sanctity/purity	V28	V29	V30

The clips illustrate configurations estimated from receiver judgements.

All videos use the default face shown in the facemaps. For each retained AU, peak amplitude equals the proportion of participant models containing that AU within the corresponding receiver group, moral domain and response category. The other five temporal parameters are averaged across models containing that AU within the same condition. Each video shows the most prevalent gaze direction; ties use a neutral gaze.

Glossary

This glossary explains the main terms used to describe the study inputs, analysis procedures, statistical measures, tests and decision rules. Related terms are grouped by function.

Stimuli, Responses and Model Inputs

The facial signals, recorded judgements, units of observation and representations used by the analyses.

Term	Definition
Facial signals	Action units and gaze indicators identify the facial movements and eye directions represented in the analyses.
Action unit (AU)	An anatomically defined visible facial movement, named within the Facial Action Coding System. An AU identifies what the face does, such as raising a lip corner, without by itself specifying an emotion or social meaning (Ekman & Friesen, 1978). Illustrative: AU12 denotes lip-corner pulling. The same movement can occur with other movements in displays interpreted differently across situations. Chapters 3–5 analyse 42 AU indicators alongside five gaze indicators. Selecting an AU for an animation differs from retaining it in a participant's response model.
Gaze direction	The direction in which the eyes are oriented, such as looking up, down, left or right. Five yes/no variables represent the four directions and no gaze movement in stimuli and response models. In Chapters 3–5 prevalence facemaps, the most prevalent gaze state is shown; tied maxima are displayed with neutral-looking eyes. Neutral-looking eyes caused by a display tie do not establish a retained no-movement signal.
Participant-by-response model	A summary of the facial and gaze signals associated with one response category for one participant. Each position in its binary vector records whether the corresponding signal was retained for that person's response. Illustrative: one receiver contributes separate not-sorry, mild-sorry and strong-sorry models, each summarising many trials rather than representing one face video. In Chapters 3–5, each model contains 42 AU and five gaze positions. Keeping all models from one participant together preserves their shared origin during splitting and resampling.

Term	Definition
Apparent apology intensity	How strongly an observer judges that another person looks sorry after a transgression. It measures the impression created by the display, rather than directly measuring the displayed person's feelings, sincerity or intention to repair the relationship. Illustrative: a face rated very sorry creates a strong impression of sorriness, even though the animation has no feelings to report. Participants rated generated faces. Chapters 3–5 rebinned ratings into not sorry, mild sorry and strong sorry; these are response categories rather than presented experimental conditions.

Analysis Procedures

Methods that turn the inputs into models, predictions or comparisons. They describe what we do with the data.

Term	Definition
Constructing models	
Methods for constructing representations of the facial signals associated with responses.	
Reverse correlation	A method for discovering which stimulus features are associated with people's judgements. It varies features across many stimuli, then works backwards from the judgements to identify the patterns linked to a response (George et al., 2023). Illustrative: a movement occurring unusually often in faces judged strongly sorry becomes part of the estimated strong-sorry pattern. Randomly generated facial signals and each receiver's apparent-sorriness judgements are used to construct response-specific models.
Prediction and model evaluation	
Methods for predicting categories, checking prediction on held-out cases, and comparing tasks or the contribution of inputs.	
Bagged-tree classifier	A model that combines many decision trees, each trained on a resampled version of the data. Each tree follows branching rules to predict a category; combining their predictions reduces reliance on any single tree. Illustrative: several trees use mouth and gaze signals differently; their combined prediction assigns a moral domain. The thesis uses complete signal patterns to predict moral domain or WC–EA receiver group.

Term	Definition
Model evaluation and classifier comparisons	
Prediction checks and matched task designs are explained together. The corrected repeated-holdout procedure is a statistical test of performance differences, kept beside the comparison it evaluates. The Statistical Tests group explains the role of statistical testing.	
Model validation	Checking how accurately a fitted classifier labels cases held out from its training. Repeated holdout repeats this check with different training and test groups, keeping each person's observations together.
	The thesis first checks whether models distinguish apparent-sorriness categories, then tests prediction of moral domains or receiver groups. Here, validation refers to these two different classification checks.
	Screening and thresholds set before later test splits limit the independence of these checks. They do not establish psychological validity.
Matched C2–D2–D5 comparison	A comparison designed to make classification tasks as alike as possible before comparing their accuracies. Matching the samples, training/test assignments and input signals helps separate task differences from differences in the analysis setup.
	C2 predicts WC versus EA within a domain; D2 predicts one domain versus another within a receiver group; D5 predicts among all five domains. The main matched test compares C2 with D2.
	D5 supplies descriptive context here because its five-class task has a different chance level.
Corrected repeated-holdout procedure	An approximate test comparing classifier performance across repeated training/test splits. Because the splits reuse participants, the performance differences are related; the correction increases estimated uncertainty to account for that overlap (Nadeau & Bengio, 2003).
	Chapter 5 uses the Nadeau–Bengio test for the matched C2–D2 contrast and the EA–WC D5 contrast. Holm separately adjusts for multiple comparisons.
Ablation and signal contribution	
Ablation assesses a signal with other inputs available; the named contrast explains how that differs from considering a signal alone.	
Ablation analysis	An analysis that removes one input and fits the model again, then checks what changes. This shows what difference that input makes when the other inputs remain available.
	Illustrative: if removing a mouth signal reduces correct classification, that signal helped the full classifier.
	Each AU or gaze signal is removed in turn, classifiers are refitted, and prediction changes are compared on the same test splits.
Marginal association and conditional contribution	Marginal association asks whether one signal, considered by itself, is related to a category. Conditional contribution asks what difference that signal makes to prediction when the other signals are already available to the model.
	Illustrative: two signals may both track a domain, but removing one barely changes prediction because the other carries similar information.
	MI/PMI assesses each signal separately; ablation removes it, refits the classifier and measures changes in prediction.

Term	Definition
Summarising and comparing patterns	
Methods for extracting recurring patterns or comparing observed profiles with reference descriptions. The literature-reference framework labels stay with that specific analysis.	
Non-negative matrix factorisation	A way to summarise a table of numbers that are zero or greater using a few recurring patterns. Each original column is approximated by adding those patterns in different amounts; subtraction is not allowed (Lee & Seung, 1999).
	Illustrative: one pattern combines mouth movements and another brow movements; a person's model can contain different amounts of both.
	The exploratory Chapter 2 analysis summarises 42-AU models with two retained movement patterns.
Literature-reference correspondence and its framework	
The retained framework terms are explained with the literature-reference comparison they describe.	
Literature-reference correspondence	A comparison between observed signal patterns and patterns described in earlier research. A similarity score shows how closely the signals in the new data match the signals represented in those earlier records.
	Illustrative: compare the AUs in an apology model with the AUs in a published smile description.
	Chapter 5 compares its facial-signal profiles with literature-based valence and emotion or social-signal records, as an exploratory analysis.
	Similarity does not show which emotion anyone actually felt.
Empirical branch	A version of the observed signal profiles defined by which experimental factors are kept separate and which are averaged. The branches make it possible to ask about overall responses, receiver groups, domains or complete conditions.
	M1: overall response profiles, averaging groups and domains equally. M2-1: response profiles by receiver group, averaging domains. M2-2: response profiles by moral domain, averaging groups. M3: full response-by-group-by-domain condition profiles. These contain 3, 6, 15 and 30 profiles respectively.
	These branches reuse the same participant models; they are different summaries rather than independent samples.
Literature level	A degree of grouping in the reference library: individual studies supply facial-signal descriptions, related descriptions form named emotion or social-signal types, and those types form broader valence categories. Each level summarises the one below it.
	L3 holds individual source representations. L2 groups them into 17 emotion or social-signal types. L1 groups these into Positive, Negative and Neutral/mixed. Balanced summaries prevent extra source alternatives from gaining weight.
	L3 supplies internal inputs; reported inference is at L1/L2. Resemblance does not establish felt or recognised emotion.

Term	Definition
Scientific question	Within the literature-reference correspondence analysis, the particular comparison used to group related tests. It specifies whether scores are compared with zero, between sorry levels, between receiver groups, across moral domains, or through group-by-response interactions. A group-by-response interaction asks whether the change between two response categories differs between EA and WC. For Strong minus Mild, this is the EA change minus the WC change. These question types define separate correction families within each empirical branch and literature level. Domain comparisons separate an overall five-domain test from the ten individual domain-pair contrasts.

Statistical Measures

Numbers calculated from data to describe an amount, association, difference or similarity. They describe what we quantify. A measure can also supply the statistic used in a test.

Term	Definition
Mutual information and its variants MI summarises an overall association; PMI describes particular state/category combinations; AMI adjusts a summary for chance agreement.	
Mutual information	A measure of how much knowing one thing helps predict another. It is zero when they are unrelated and larger when knowing one narrows down what to expect about the other (Ince et al., 2017). Illustrative: knowing whether a face looks away may help predict which sorry category a receiver chooses. MI links signal states to judgements during model construction, and later links individual signals to moral domains or receiver groups.
Pointwise mutual information	A score showing whether a particular signal state and category occur together more or less often than expected from their separate frequencies. Positive values mean more often; negative values mean less often. Illustrative: a mouth movement paired with strong-sorry judgements more often than expected has positive PMI for that pair. After MI identifies an association, PMI locates the particular signal states and categories contributing to it. Later 'negative associations' mean signal absence occurs unusually often, not simply a negative PMI number.

Term	Definition
Adjusted mutual information	A measure of association between two sets of category labels after accounting for the association expected by chance. Zero marks the chance baseline; one marks perfect agreement, and negative values can occur below that baseline. Illustrative: a small positive value means the two label sets share slightly more information than expected by chance. Chapter 5 uses AMI descriptively for binary signal state versus moral domain. Raw MI, participant-block permutations and BH correction provide the signal-level inference.
Prevalence and differences in prevalence	Prevalence describes how widely a signal recurs. Cohen's *h* describes the size and direction of a difference between two proportions.
Signal prevalence	The proportion of a specified set of models that contains a signal: the number containing it divided by the total number considered. It describes how widely that signal recurs across those models. Illustrative: if an AU occurs in 6 of 10 participant models for one response and domain, its prevalence is 60%. Chapters 3–5 summarise retained participant models within the relevant response, receiver group and domain. The denominator is models, rather than animation trials.
Cohen's *h*	An effect-size measure comparing two proportions after transforming their scale. A larger absolute value means a greater difference; a positive or negative sign shows which proportion is larger. It is not a percentage-point difference. Illustrative: 60% versus 40% gives $h \approx 0.403$. Reversing the comparison gives -0.403, with the same magnitude. Chapter 5 subtracts EA from WC, so positive *h* means higher WC prevalence. It describes effect size; panel-level tests supply the inferential decisions.
Hamming distance	A measure of how many positions differ between two equal-length patterns. For yes/no signal lists, it counts disagreements; dividing by the list length gives the proportion of signals that differ. Illustrative: with 1 meaning present and 0 absent, 1010 and 1001 differ in two of four positions: 0.5. Chapter 5 compares binary facial-signal profiles to explore how patterns differ across sorry levels, receiver groups and moral domains.

Statistical Tests

Procedures that assess how compatible an observed result is with a specified null hypothesis, such as no association or no group difference. A measure supplies the number being tested; the test supplies its reference calculation. For example, MI measures an association, permutation testing assesses it against a null reference, and BH adjusts across a specified family of tests.

Term	Definition
Permutation and paired rearrangement tests	Permutation testing provides a general testing approach based on allowed rearrangements under a specified null hypothesis. The panel-level prevalence test and the paired sign-flip and sign-swap tests below are named applications, each with its own comparison and rearrangement rules.
Permutation test	A test asking whether an observed difference or relationship is stronger than chance rearrangements would produce. It shuffles relevant labels, recalculates the result and compares the observed value with those shuffled results. Illustrative: shuffle moral-domain labels, refit the classifier and see how often its accuracy equals or exceeds the original. The thesis shuffles labels in ways suited to each question, keeping related observations from each participant together.
Panel-level prevalence test	A test comparing how often signals appear in two groups' models within one condition. Differences across signals are combined into one score; shuffled group labels show how large that score could be without a group difference. Thesis example: mild-sorry/Fairness is one panel; strong-sorry/Fairness is another. Three sorry levels $\times$ five moral domains give 15 panels. Each compares WC and EA across 47 signals, using the proportion of people's models containing each signal.
Paired sign-flip test	A test asking whether paired differences consistently favour one side. It flips each difference from plus to minus or vice versa, then compares the observed average with the averages produced by these flips. Illustrative: a full-minus-ablated cell difference of +0.04 becomes -0.04 when its sign is flipped. The thesis pairs full and signal-removed classifiers on the same data splits, testing changes in individual confusion-matrix cells.
Exact sign-swap test	A paired test comparing the number of findings in two sets. It judges the observed count difference against every allowed exchange of the set labels; using every swap makes the comparison exact. Illustrative: exchange one signal's WC and EA labels together across its not-sorry, mild-sorry and strong-sorry results. The thesis compares how many positive or negative MI/PMI associations occur in WC versus EA, keeping each signal's three responses together. This exploratory comparison uses signal blocks, not individual participants.
Exact McNemar test	A test comparing two yes/no assessments of the same items. It asks whether 'yes only in the first' is more common than 'yes only in the second', considering all possible allocations of those disagreements. Illustrative: six signals qualify only in WC, two only in EA; McNemar compares those disagreements. The thesis asks whether more of the same 47 signals support separation, or contribute to confusion, in WC than EA classifiers. This exploratory comparison concerns signals, not paired participants.

Term	Definition
Factorial ANOVA	A test comparing average values across several kinds of grouping, called factors. It checks each factor's overall pattern and whether that pattern changes with another factor; the latter is an interaction. Illustrative: downward versus upward gaze may differ more for strong-sorry than mild-sorry judgements; that is an interaction. Chapter 2 compares gaze-PMI values across gaze direction, sorry intensity and transgression. A separate participant-aware check then reassesses the conclusions.

Uncertainty, Corrections and Decision Rules

Methods that estimate precision, account for dependence or multiple testing, or set criteria for retaining findings or signals. Uncertainty estimation, multiple-testing correction and signal-selection thresholds have distinct roles.

Term	Definition
Uncertainty estimation	
Bootstrap resampling estimates how much a quantity varies; participant-clustered inference allows repeated measurements from each person to be related. These address uncertainty. Permutation tests instead construct a reference under a specified null hypothesis.	
Participant-block bootstrap	Repeatedly resampling whole people from the existing dataset to see how much an estimate varies. A person can be selected more than once, and all their records stay together. Illustrative: a resample may contain Alice twice and omit Ben, while carrying all of Alice's response models. Chapter 5 keeps the original receiver-group and moral-domain groupings when calculating uncertainty intervals for its summaries.
Participant-clustered inference	A way to calculate statistical uncertainty when each person contributes several measurements. It allows those measurements to be related, instead of treating them as observations from different people. Illustrative: five gaze values from one person are allowed to be related when judging how precise a group difference is. Chapter 2 uses this as a later check on its gaze-PMI results, allowing each person's repeated values to be related.

Term	Definition
Signal-selection thresholds	Rules determining which facial signals are retained for later summaries, at the stated stage of the analysis.
Bayesian population-prevalence thresholding	A method that updates an estimate of how widespread a detectable effect is using observed counts and a probability model. It accounts for chance detections before setting a rule for which signals to retain (Ince et al., 2021). Illustrative: a signal appearing in only one person's model is removed if the required count is higher. Within each response, moral domain and receiver group, the thesis sets the minimum number of participant models required before a signal is kept.
Multiple-testing correction	Adjustments that limit false-positive findings across a specified group of tests.
Multiple-testing correction: targets and procedures	FWER and FDR identify error targets; Bonferroni, Holm and BH are adjustment procedures; adjusted p/q values are their reported outputs.
Multiple-testing correction	Running many tests increases the opportunities to mistake random variation for a real association. A multiple-testing correction changes the decision rules to limit those mistaken findings across a specified group of tests. Illustrative example: a pattern that looks convincing in one test may no longer qualify after accounting for 20 tests. The thesis uses Bonferroni, Holm and BH adjustments for different groups of tests, each with its own stated scope. They do not create one correction covering the entire thesis.
Familywise error rate	The probability that a group of tests produces at least one false-positive conclusion: calling a difference or association real when it is absent. Even one such mistake counts as a familywise error. Illustrative example: a 5% familywise limit means at most a 5% chance of any mistaken finding across that family. Bonferroni and Holm limit the chance of getting even one false-positive result within each group of tests. It is not the probability that a particular finding is wrong.
False discovery rate	False discovery rate is the average fraction of declared findings that would be false positives across repeated studies. It concerns the proportion of mistakes among the findings, rather than the probability that any mistake occurs. Illustrative example: if two of ten declared findings were false, that study's false-discovery proportion would be 20%. In Chapter 5, BH controls this rate separately for each group of signal tests and each group of comparisons. A 5% target does not make every finding 95% likely to be true.

Term	Definition
Bonferroni correction	Bonferroni requires a smaller *p*-value when more tests are judged together. It divides the chosen overall cutoff by the number of tests, limiting the chance of at least one false-positive result in the family. Illustrative example: for five tests and a .05 overall cutoff, each *p*-value must be at most .01. It corrects 25 confusion-matrix cells in domain classifiers and four in receiver-group classifiers, with ablation corrected separately within each predictor.
Holm correction	Holm checks a family's *p*-values from smallest to largest. It begins with a strict cutoff, relaxes the cutoff after each successful step, and stops at the first failure, controlling the chance of any false-positive finding. Chapter 5 applies Holm to separate groups, including five classification tests, 15 prevalence panels and seven Hamming tests. Each adjustment applies only to its named family.
Benjamini–Hochberg (BH) correction	BH sorts *p*-values and uses cutoffs that increase with their rank. It keeps findings through the last qualifying value, aiming to limit the average fraction of false positives among significant findings over repeated studies. Chapter 5 applies BH within each 47-signal variability panel and within the specified literature-correspondence families. This controls a false-discovery fraction, rather than guaranteeing no mistakes.
Adjusted *p*-value and BH-adjusted value (*q*)	A *p*-value is the probability, under the test's no-effect assumption, of a result at least as extreme as the one observed. An adjusted value also accounts for the other tests in its family. Illustrative example: a raw *p*-value of .01 becomes .05 after Bonferroni adjustment for five tests. The thesis labels within-family BH-adjusted *p*-values *q* and reports them separately from the raw permutation *p*-values. Neither *p* nor *q* gives the probability that a finding is false.

Facial Displays and Rendering

How facial movements are visualised and how the figure's numerical colour key should be read.

Term	Definition
Facemap	A computer-rendered face showing where a set of facial movements acts. Colours highlight local movement, while the supplied AU values determine which movements appear and how they are rendered. Illustrative: rendering lip-corner pulling highlights movement around the mouth. The image locates the movement; its caption supplies the meaning of its values. Facemaps display prevalence-weighted AUs, NNMF patterns, or selected signals. Selected-signal localisation shows each AU at maximum rendering amplitude, rather than encoding prevalence or effect size.